Proposal for a Master Thesis

Type of thesis / Line of study:	Master Thesis - Computer Science (6 months)
Title of the thesis:	Pattern-based Prediction of Life-Threatening Diseases in Intensive Care
Candidate:	Sayed Issa, Joud
Matriculation number:	387721
Advisors(s):	Dr. Ralf-Detlef Kutsche/ Anne Schwerk, Ph.D.
Reviewer(s):	Prof. Dr. Volker Markl
Planned period:	03.2019 – 08.2019

1. Introduction / Scientific Background / Related Work

There is a steady increase in data from all disciplines in life and science, making data omnipresent and thus significantly increasing data size ('volume') and the need for according processing pipelines. The term 'Big Data' is used for very large and potentially very complex datasets, which are difficult to process using standard database system tools [1]. In healthcare, an increasing amount of such high-volume data is produced, for example a single human genome requires to store 200 gigabytes of raw information [2], and a single functional magnetic resonance imaging file contains about 300 gigabytes [2]. Aside from the size, healthcare data is very varied, including not only structured but also unstructured and semi-structured data sources, such as demographics, diagnoses, images, procedures, and medications, lab results, clinical notes, and patient-generated data, such as information from body sensors [13].

Although most healthcare data resided in inaccessible silos, there are some Big Data collections of digital health records (EHRs) of hospitals open available, such as the Medical Information Mart for Intensive Care III (MIMIC-III) dataset [3]. As described in [3], it is a large, freely-available database comprised of de-identified health-related data associated with over 40,000 patients who stayed in critical care units at Beth Israel Deaconess Medical Center between 2001 and 2012. It consists of 26 tables that record every piece of data for the patient's admission to the hospital [4]. Data available in the MIMIC-III database ranges from time-stamped, nurse-verified physiological measurements made at the bedside to free-text interpretations of imaging studies provided by the radiology department [4].

The integration of (sometimes huge) data from the above-mentioned heterogeneous sources is crucial for improving and refining therapies and diagnoses. Yet, clinical environments lack the human and methodological ressources, including efficient and reliable software tools, for integrating and analyzing such Big healthcare Data [5] to uncover unknown patterns and leverage predictive analytics.

From the methodology viewpoint, the notion of 'patterns' is a key asset for applying knowledge from previous experiences to a new problem, which is well-known in software development and programming [7], but also applicable to other disciplines. Many researchers have attempted pattern recognition methods to discover hidden patterns in time series data from varying domains, i.e. social media, and in healthcare [8]. Pattern recognition is a computationally expensive process and most of the techniques are tailored to a small segment of data and when faced with exponential data volumes that are continuously generated from multiple systems they do not scale [8]. However, current dimension reduction approaches, are not able to maintain the temporal nature

of the underlying data. The ability to process huge volumes of data and at the same time understand relationships in time-oriented data could be key to discovery of hidden patterns in data [9].

Predictive Analytics is the branch of the advanced analytics which is used to make predictions about unknown future events [10]. Predictive analytics applies many techniques from data mining, statistics, modelling, machine learning, and artificial intelligence to investigate current findings to make predictions about future. Machine learning techniques have become progressively popular in conducting predictive analytics due to their outstanding performance in manage large scale datasets with uniform characteristics and noisy data. Observational studies show that machine learning is appropriate to build predictive models by extracting patterns from large datasets [10].

2. Problem Statement / Goals of the Thesis

Data mining in healthcare today is increased because the health sector is rich with information. Healthcare organizations generate and collect large volumes of information on a daily basis. Use of information technology enables automation of data mining and knowledge that help bring some interesting patterns which means eliminating manual tasks and easy data extraction directly from electronic records, electronic transfer system that will secure medical records, save lives and reduce the cost of medical services as well as enabling early detection of infectious diseases on the basis of advanced data collection [11]. Pattern discovery can enable healthcare organizations to anticipate trends in the patient's medical condition and behavior. The raw data from healthcare organizations are voluminous and heterogeneous needs to be stored in organized form and their integration allows to unite medical information system. Predictive analysis in health provides possibilities for analyzing different patterns. These patterns can be used by healthcare practitioners to make forecasts and set treatments for patients in healthcare organizations [11]. Despite the advances in healthcare mining, the characteristics of these data – complexity, volume, high dimensionality, etc. – still demand more efficient and effective techniques [12].

In order to enable the successful and efficient integration of such distributed, heterogeneous and very large data sources in the given setting of heart failure and its origin in Acute Kidney Injuries (AKI), the thesis proposed will apply complex data preprocessing techniques and ‘Big Data Analytics’ to gain knowledge from a big clinical data set derived from the intensive care unit (ICU) setting: the MIMIC III database.

The following steps must be taken in general in order to achieve this goal:

1. Analyze all electronic medical record (EMR) data including all patient’s diagnosis information and all lab values and demographic information.
2. Enable suitable preprocessing of heterogeneous data sources.
3. Discover patterns of the analyzed data to enable subtyping of patients.
4. Apply predictive modeling on historical EMR to verify patterns.
5. Establish an end-to-end processing pipeline that allows a high-level prediction in the ICU setting.

3. Thesis Approach / Plan of Implementation

- We will use an open public database, the MIMIC-III database [3], which contains deidentified, clinical data of patients admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts, allowing international researchers to use and replicate those data under a data use agreement.
- MIMIC-III includes different data sources and formats, including different data streams (EEG, ECG), demographic information, clinical notes, diagnoses, billing information, and laboratory test results – allowing for complex data modeling [4].

- The first step during data preprocessing is to select the proper cohort for the proposed study, followed by the extraction of the related records based on the patient's ID.
- The next preprocessing step involves cleaning the extracted data (e.g. inconsistent record units, multiple values for a variable or handling range variables).
- After preprocessing the raw data, an unsupervised machine learning (ML) method will be applied to all selected patients in the database to identify relevant features.
- Automatic feature selection methods (model-based selection or iterative selection, wrapper method) will then be applied.
- Different clustering methods (e.g. K-means and hierarchical clustering) will be applied to uncover data patterns.
- Then, the selected features will be used in a supervised model for survival/outcome prediction. This is done by comparing logistic regression (LR), support vector machine (SVM) and decision tree (DT) models.
- To evaluate the performance of our model we will calculate the testing error metric and then apply a cross-validation process by using another data set.
- The program Python will be used for ML models and R for simultaneous visualizations.

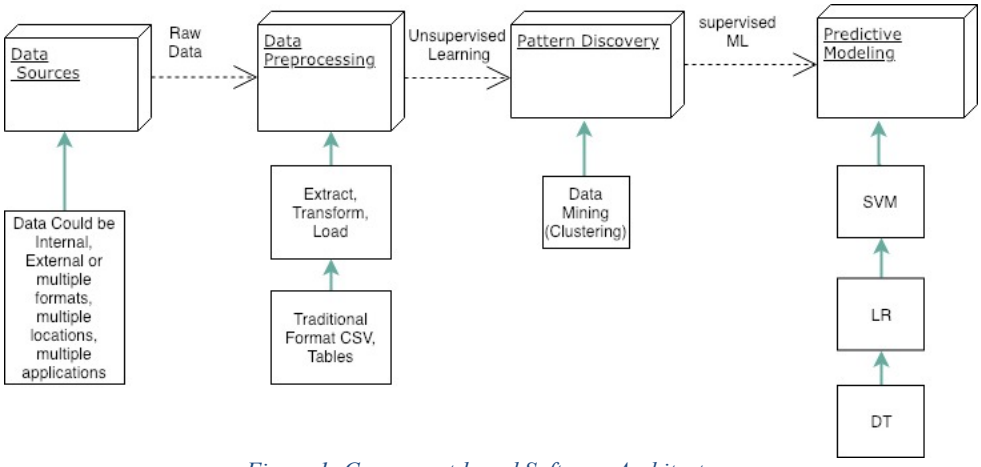

Figure 1: Component-based Software Architecture

4. Bibliography

[1] E. Wilder-James, "An introduction to the big data landscape.," O'Reilly, 11 January 2012. [Online]. Available: https://www.oreilly.com/ideas/what-is-big-data. [Accessed August 2018].

[2] J. Son, "Big Data Analytics in Healthcare," Udacity, [Online]. Available: https://eu.udacity.com/course/big-data-analytics-in-healthcare--ud758. [Accessed October 2018].

[3] T. J. Pollard and A. E. W. Johnson, "The MIMIC-III Clinical Database," 2016. [Online]. Available: http://dx.doi.org/10.13026/C2XW26. [Accessed 2018].

[4] A. E. Johnson, T. J. Pollard, . L. Shen, . L.-w. H. Lehman, . M. Feng, M. Ghassemi, . B. Moody, P. Szolovits, . L. A. Celi and R. G. Mark, "MIMIC-III, a freely accessible critical care database," Scientific Data, vol. 3, 2016.

[5] F. a. J. Y. a. Z. H. a. D. Y. a. L. H. a. M. S. a. W. Y. a. D. Q. a. S. H. a. W. Y. Jiang, "Artificial intelligence in healthcare: past, present and future," Stroke and Vascular Neurolog, vol. 2, no. 4, pp. 230--243, 2017.

[6] A. Bahga and V. . K. Madisetti, "Healthcare Data Integration and Informatics in the Cloud," IEEE, Vols. 48, no. 2, pp. 50-57, 2015.

[7] E. Gamma, R. Helm, R. Johnson and J. Vli, Design Patterns: Elements of Reusable Object-Oriented Software, Addison-Wesley Professional, 1994, p. 417.

[8] C. M. C. Inibhunu, "Machine learning model for temporal pattern recognition"," IEEE EMBS International, pp. 1-4, 2016.

[9] C. Inibhunu and C. M. Am, "State Based Hidden Markov Models for Temporal Pattern Discovery in Critical Care," in 2018 IEEE Life Sciences Conference (LSC), Montreal, QC, 2018.

[10] B. N. a. V. Ilango, "Predictive analytics in health care using machine learning tools and technique," in 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, 2017.

[11] A. Awad, M. Bader-El-Den, J. McNicholas and J. Briggsa, Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach, vol. 108, 2017, pp. 185-195.

[12] A. Silva and C. Antunes, "Finding Multi-dimensional Patterns in Healthcare," in Machine Learning and Data Mining in Pattern Recognition, Springer International Publishing, 2014, pp. 361--375.

[13] V. Afshar, "Cisco: Enterprises Are Leading The Internet of Things Innovation," [Online]. Available: https://www.huffingtonpost.com/entry/cisco-enterprises-are-leading-the-internet-of-things_us_59a41fcee4b0a62d0987b0c6?guccounter=1. [Accessed 21 August 2018].

[14] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," Bio Med Central - Health Information Science and Systems, 2014.

[15] A. Jain, "Watson Health:The 5vs of big data," IBM, 17 September 2016. [Online]. Available: https://www.ibm.com/blogs/watson-health/the-5-vs-of-big-data/. [Accessed December 2018].