Review of Prediction of Disease Trends using Big Data Analytics

Diellza Nagavci, Mentor Hamiti, Besnik Selimi
Faculty of Computer Science and Technologies
South East European University (SEEU)
Tetovo, Macedonia

***Abstract*—Big Data technologies promise to have a transformative impact in healthcare, public health, and medical research, among other application areas. Several intelligent machine learning techniques were designed and used to provide big data predictive analytics solutions for different illness. Nevertheless, there is no published research for prediction of allergy and respiratory system diseases. However, the impact of research and the finding of different cases is conducive to progress and further development of this. One of the goals of this paper is to devise a systematic mapping study, to explore and analyze existing research about disease prediction in healthcare information. According to the realized investigation of published research from 2012 up to today, we are focusing our research on studies that have been published around big data analytics. With this high number of secondary studies, it is important to conduct a review and provide an overview of the research situation and current developments in this area.**

Keywords*—Big data; algorithms; data analytics; healthcare; disease prediction; data mining

I. Introduction

Advancements in Information Technology have proven monumental in improving the quality of live throughout many dimensions. Medical Sciences is no exclusion to this relentless evolution. Internet, robots/ AI, and telemedicine have been very important in science when it is about adopting the medical science and profession with this trend; yet it is often argued that when it comes to critical thinking, no AI can beat the instincts of an experienced Doctor [1]. Insofar as 'preparing for the future' is concerned, Data Analytics have proven a highly reliable source of information in a plethora of sciences, and since 'all data is equal' for the AI, prognosis of health conditions and eventual epidemics using Big Data is particularly attainable, and immensely important.

Based on this premise, we are focusing our research on finding a suitable data mining algorithm for using Big Data to predict diseases. We believe that analyzing big data sets will lead to finding causes that lead to respiratory disease and allergies to populations in the future.

In this paper, we first introduce the methodology used, and the research questions defined. Secondly, we give a classification scheme of the fields of interest, big data, data mining, data sets, and optimization. Afterwards, we provide answers to five research questions and two are proposed as future aim research. On the discussion part, the time series of papers relating to health information field of interest have been included. As a future research we have proposed big data analytics and data mining learning algorithms.

II. Methodology

The main goal of this mapping study is to define each step of research fields answer the research questions based on analyzed articles [2]. The idea was to collect a series of publications in the field of interest, to determine the coverage of the research field. For categorization of reports and different results, we needed the structure and for that we had taken a systematic approach. It shows results by using the visual summary and a map. We have used different research questions that must be defined to obtain these objectives in a systematic manner. The main purpose of a structured mapping study is to present an overview of a certain research area as well as to identify research gaps. A systematic literature review is another kind of secondary study that answers specific research questions by identifying, analyzing and interpreting relevant evidence. The process begins with the definition of research questions, from which we can arrive to a research scope. The next step is to conduct the actual search by retrieving all papers that may be remotely related to the field. Then comes the screening of the papers, with the objective to filter all the relevant papers. The classification scheme is based on key wording by reading the abstracts. In the end is to extract the data and to show results.

A. Research Questions and Search Strategy

- What is the principle of interest discussed in the papers?
- What type of framework is used for Big Data?
- How publications have evolved over time? What the research and publication trends are?
- Which methods are used previously?
- Which algorithms are used for processing Big Data?

Most of the explored research publications are extracted from digital libraries as IEEE-Xplore, ACM, and some of articles are from Springer and IJIRCCE. The search strings in Table I are used to search in digital libraries.

Large number of articles appeared on different search strings. Have been selected just the ones that we saw reasonable to include as more appropriate and help achieve our goal. Most of papers have been published in recent years. In the Fig. 1 are shown the number of published papers by

year. The papers that have been published last year are from the first half of 2017. From the selected papers, further analysis is conducted and in this study papers related to Map-Reduce and Big Data, health information, data mining and prediction are included. As a result, after removing duplicates and irrelevant papers, only 119 articles have been filtered.

TABLE I. NUMBER OF PAPERS BY MAIN FIELD OF INTEREST

No.	Search String	No. of papers
SS1	((("Abstract": Map Reduce) OR "Abstract": big data) AND "Abstract": health information)	217
SS2	((("Map Reduce") OR " big data ") AND health information) AND data mining	133
SS3	(((("map reduce framework ") OR " big data ") AND health information) AND data mining) AND prediction	69

III. CLASSIFICATION SCHEME

The classification scheme is presented in three columns where we include the main fields of interests related to the research (Fig. 1). The Big Data analytics are the main fields on which we will focus to propose a more suitable algorithm in the future. The field of interests are defined in the first column to come up with framework types. And the third column is for different big data processing algorithms. Based on the analysis from the collected papers, we have found the research gap in the 'big data analytics', in which further contribution is expected from the research community. The third column shows different algorithms for processing big data is random walks that needs to be fulfilled in order the future goals be verified. More details are presented in research questions in the results part. Additionally, we classified papers according to the field of interest.

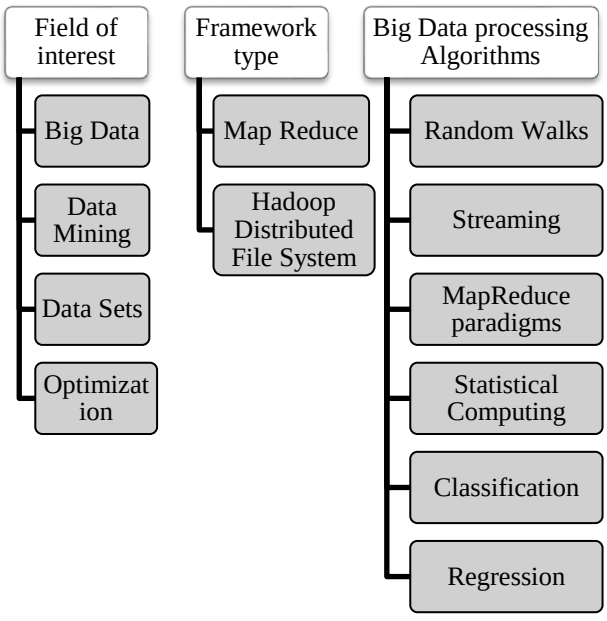

Fig. 1. Classification Schema.

IV. RESULTS

Each article was classified into the categories of each facet in order to answer the six research questions. The results of the systematic mapping study are presented as follows.

A. RQ1: What are the Principle of Interest Discussed in the Papers?

This question deals with the main field of interest that is investigated in each of the papers. We are interested in Big Data, but since we used several search strings, we got several results. In order to answer this question, we created the 'Field of Interest' classification for the papers.

From Table II, we can observe that about 58% of the papers have as main focus in the big data that refers to data sets and flows large enough that pose significant challenges when using commonly available tools and infrastructures for collecting, managing and processing the data within a tolerable amount of time.

The second most mentioned area is data mining with almost 21%. This category of papers includes tasks actions like as data extraction techniques and driving manual tools needed to adapt to new technologies to overcome time constraints.

TABLE II. NUMBER OF PAPERS BY MAIN FIELD OF INTEREST

Field of interest	Number of papers	Percentage
Big Data	69	58%
Data mining	25	21%
Data sets	15	13%
Optimization	10	8%

B. RQ2: What Type of Framework is used for Big Data?

A lot of proposals that we have searched during the research have been focused in MapReduce approach [3]. MapReduce paradigm has been used to implement Classification techniques. The data that is processed and disseminated in a cloud computing infrastructure is very convenient and very effective to accelerate the process of knowledge generation.

Here are two types of frameworks:

1) Hadoop–provides its own file system called HDFS (Hadoop Distributed File System). To find the solution of the data text in a Hadoop shows us that all the data are performing parallel operation in different clusters.

Hadoop also will keep the multiple copies of data in case of hardware failure [4].

2) The second framework is A MapReduce that consists of two functions: map and Reduce. These two functions take a set of important pairs/value data and generate a set of output key/value pairs when a Map Reduce job is given to the cluster. The job is divided in two pieces into map tasks and reduce tasks, where each Map task will process one block of input data. A Hadoop cluster takes slave nodes to execute Map and

reduce task. The slave node it is capable to except number of map and Reduce tasks and execute simultaneously. A slave node sends a signal to master node in a given period of time. To accept the signal it will request the master node to slave. The Map function waits for worker node that shows input key and value pairs outside of the block [5].

From the investigation of our papers, we found out the results in Table III, according to which the mass of the papers use Map Reduce framework (45%). The rest of the papers deal with Hadoop, about 55%.

TABLE III. NUMBER OF PAPERS BY FRAMEWORK TYPE

Framework Type	Number of papers	Percentage
Map Reduce	54	45%
Hadoop	65	55%

C. RQ3. How Publications have Evolved Over Time? What the Research and Publication Trends are?

While studying the year of publication for each paper, we notice that the time ranges between 2013 and 2017. The majority of the papers (31.93%) have been published in 2017 In fact if we look at graph in Fig. 2, we notice that the lot of papers increases from year to year.

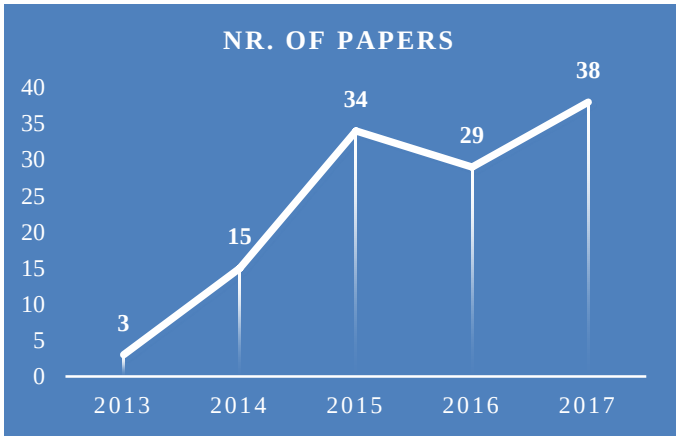

Fig. 2. Number of Papers per Year.

D. RQ4: Which Methods are used Previously to Process a Big Data Analysis?

In recent years, big data analytics is so important to heath care. Big data analytics is a very broad area that deals with the collection, storage and analysis of immense data sets to trace the unknown patterns and other key information. The data that are very important it can help us recognize the data that are integral component to the future business decisions. The second research question is concerned with the different methods that have been used in big data analytics.

Omar El-Gayar and PremTimsina [6] have proposed a model on Evidence Based Medicine and Big Data Analytics. The main purpose of the system is to improve the cost across the applications of business intelligence big data analytics.

Samir El-Masri et al. [7] in this paper the authors have designed a model for clinical Decision Support System. The model is described as an Adaptive Evidence based Medicine. Using this model, the patient data from Electronic Health Record (EHR) was collected in a data warehouse. The Clinical Practice Guidelines (CPG's) were taken and CPG rules were generated using an automated converter. These rules were applied on the data obtained from the warehouse and the standardized data was stored in the knowledge base. The inference engine processed the questions from the physician and searched the knowledge base for the most applicable guideline. The model lacked in handling distributed nature of the process, that is, the CDSS that were geographically distributed could not interact with each other. The rules required to be standardized before processing, there increasing the time complexity of the system.

Sankaranarayanan. S and Pramananda Perumal. T. [8] have invented a model for Diabetic prognosis using Data Mining techniques. The two major data mining algorithms Apriori and FPGrowth were applied on Diabetes Mellitus dataset to generate association rules. Frequent item sets were mined after which rules were generated using support and confidence threshold values. This model was not generic to diagnose variety of diseases. Accuracy of the prediction was not guaranteed, and the model was not scalable to support voluminous health records.

Mohamed Abouzahra et al. [9] have implemented a model on integrating data from Electronic Health Record (HER) to improve Clinical Decision making for Inflammatory Bowel disease. This model accumulated fragmented data of a patient that can take from different EHR systems. Analytical techniques were applied to this data to identify useful patterns. These techniques were based on physicians' input, literature, and existing guidelines to identify possible relationships between different components of patients' data. Predictive methods were used to predict future outcome of the patient and facilitated the diagnosis of disease. The patients' privacy and information security issues were not treated properly.

E. RQ5: Which Algorithms are used for Processing Big Data?

One important part of our research is to find out which algorithms are used for processing big data analytics. J. Qiu et al. has presented different machine learning algorithms for big data processing [10]. The first one is representation learning or feature learning which deals with learning data representations that make the data analysis process easier. It is found that the performances of the machine learning algorithms are strongly influenced by the selection of data representation (or features) [11]. Feature selection (variable selection) techniques are used to find those features of data which are most relevant for use in model construction. Feature extraction techniques transform the high dimensional data into a low dimensional space. In space metric learning, the function of distance is constructed to calculate the distance between different points of a data set. Table IV represents a list of some of the algorithms that are used in different research papers. The authors mentioned about another hot learning technique called deep learning in their paper.

TABLE IV. DIFFERENT ALGORITHMS SUMMARY

Algorithms	Used
Random Walks Distributed Hash Tables, Bulk Synchronous Parallel (BSP)	Random walk is designed to address wide range of problems in mobile and sensor networks. For every machine to know that information resides is used the hash table. The BSP computer is compiled of a set of processors connected by a communication network
CART, Recursive Partition Trees	Decision tree algorithms
K- Nearest neighbor, Bayesian, SVM, ANN, K-means,	A survey was done on the various machine Algorithms for classification, prediction and modelling
MapReduce, Linear regression	The main objective was to improvise the accuracy of rainfall forecasting.

V. DISCUSSION

After analyzing the 119 papers, our focus is to explore and to design an algorithm that will analyze and make prediction from the data sets, in different platforms. The idea is that in modern big data research, the suitability of different algorithms is solely dependent on the data characteristics. Therefore, there is a need for further in-depth analysis to find the suitable supervised and unsupervised machine learning algorithms to derive meaningful facts and actionable insights from HIS data.

J.L. Berral-Garcia has presented a paper describing the frequently used machine learning algorithms for big data analytics [12]. Several algorithms are used for performing modeling, prediction and clustering tasks. Decision tree algorithms (like CART, Recursive Partition Trees or M5), K-Nearest neighbors algorithms, Bayesian algorithms (using Byes theorem), and Support vector machines (SVM), Artificial Neural Network, K-means, DBSCAN algorithms, etc are presented in this paper. Several execution frameworks - Map-Reduce Frameworks (Apache Hadoop and Spark) were also mentioned. The implementations of the previously discussed algorithms are made available to the public through different tools, platforms and libraries such as R-cran, Python Sci-Kit, Weka, MOA, Elastic Search, Kibana etc. M. U. Bokhari et al. presented a three layered architecture model for storing and analyzing big data [13]. The three layers are data gathering layer, data storing layer and data analysis & report generation layer. In order to gather and handle the huge volume of big data coming from high speed sources such as sensors or social media, a cluster of high speed nodes or severs are kept in the data gathering layer. The data storage layer is responsible for storing the big data. The Hadoop Distributed File System (HDFS) can be used for data storage [14]. Principal Component Analysis, Singular Value Decomposition and tensor-based approaches are useful for feature extraction. For feature selection, filter-based and wrapper-based methods are helpful. All these are dimensionality reduction techniques. The authors compared different techniques for performing data mining tasks. Logistic regression, cox regression, local regression techniques are simple to interpret, but are prone to outliers. The authors discussed about the useful platforms for big data analytics. Apache Hadoop, IBM Platform, Apache Spark Streaming, Tableau, and other visual analytics tools are highly impactful platforms for providing big data analytics solutions. Two real world case studies such as integrative /omics data for the improved understanding of cancer mechanisms, and the incorporation of genomic knowledge into the EHR system for improved patient diagnosis and care were done to discuss the usefulness of biomedical big data analytics for precision medicine. Multi-omic TCGA [15] data and EHR data were used to conduct this study. Since we wanted to find the gap of where and how we can design an algorithm that will analyze and predict the data sets, in different platform. We will consider it for future goals and analysis.

VI. FUTURE RESEARCH

Our main objective in the recent future will be to analyze the different approaches and concepts previously used, determining which algorithms could potentially be appropriate for finding causes of respiratory diseases and allergies. These algorithms will be applied to data health information about patients and check what kind of prediction can be derived, how accurate are the predictions of each of these algorithms. Based on this, we want to derive a method that could provide satisfactory results in term of predicting disease trends.

Although we base on previous research, it should be remarked that it is a first attempt to suggest a concrete and detailed algorithm that will be implemented in the Kosovo Health Information System.

Two research questions proposed as future research are:

- Where does health big data come from?
- What value will give this algorithm in HIS?
- How can healthcare systems benefit from big data analytics?

VII. CONCLUSION

Big Data presents a unique discipline, which should have the most important role in the latest technology developments. We presented the different algorithms and technologies that are used in predicting diseases.

Although the main objective of this research is to apply existing algorithms in the prediction of a different disease, it is important to support it by practical example, limited to the data that can be made available on Kosovo Health. The collection, filtering, normalizing and processing of the data itself is an important problem – hence our focus on data mining and big data processing techniques.

In addition, it will serve as a recommended model for research and for further development of this field of research.

REFERENCES

[1] Athmaja S., H. M. (2017). A SURVEY OF MACHINE LEARNING ALGORITHMS FOR BIG DATA ANALYTICS. International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS) .

[2] Hirak Kashyap, H. A. (2014). Big Data Analytics in Bioinformatics: A Machine Learning Perspective. IEEE.

[3] Isaac Triguero, D. P. (2014). A MapReduce solution for prototype reduction in big data classification. IEEE.

[4] Krishna, P. R. (2015). Big Data Search and Mining. Springer India . IEEE.

[5] Matthew Herland, T. M. (2014). A review of data mining using big data in health informatics. IEEE.

[6] Megan Sheeran, R. S. (2017). A framework for big data technology in health and healthcare. . IEEE.

[7] Mohamed Abouzahra, K. S. (2012). "Integrating Data from EHRs toEnhance Clinical Decision Making: The InflammatoryBowel Disease Case". Proceedings of 27th InternationalSymposium on Computer-Based Medical System.

[8] Nadiya Straton1, R. R. (2017). Big Social Data Analytics for Public Health: Comparative Methods Study and Performance Indicators of Health Care Content on Facebook. IEEE International Conference on Big Data (BIGDATA), 2772-2777.

[9] Omar El-Gayar, P. D. (2014). "Opportunities for Business Intelligence and Big Data Analytics In Evidence Based Medicine". IEEE Trans, pp .749-757. .

[10] PramanandaPerumal.T, S. a. (2014). "Diabetic prognosis using Data Mining methods andtechniques". Proceedings of ICICA, Coimbatore, India.

[11] Prof. Sharmishta Desai, S. R. (2016). VERY FAST DECISION TREE (VFDT) ALGORITHM ON HADOOP. IEEE.

[12] Samir El-Masri, S.-S. (2012). "An Adaptive Evidence Based Medicine System Based on a Clinical Decision Support System". Science Series Data Report.

[13] Sayali D. Kadam, P. D. (2016). Big Data Analytics- Recommendation System with Hadoop Framework . IEEE.

[14] Shim, K. (2012). MapReduce Algorithms for Big Data Analysis. Proceedings of the VLDB Endowment, , Vol. 5(12),.

[15] Y. Bengio, A. C. (2013). "Representation Learning: A Review and New Perspectives. IEEE.