Big data techniques to discover kidney problems at early stages: a prospective study

Omar García-González[1], Ivan E. Villalon-Turrubiates[1] and Pilar Pozos-Parra[2]
[1]Instituto Tecnológico y de Estudios Superiores de Occidente (ITESO)
Periférico Sur Manuel Gómez Morín 8585, 45604 Tlaquepaque, Jalisco, México
[2]Universidad Autónoma de Baja California
Calzada Universidad 14418, 22424 Tijuana, B.C, México
ng724433@iteso.mx, villalon@iteso.mx, maria.pozos@uabc.edu.mx

Abstract. Chronic Kidney Disease is a decrease in the kidney function, which can eventually derive in the cessation of the total function. It affects around 10 to 15% of the adults globally. This number is expected to grow as the diabetes disease is growing, and kidney disease is one of the consequences of diabetes. Some computational tools and techniques related to big data, machine learning, clustering, signal and image processing, and data mining among others, promises huge benefits for medical research through technologies that can provide a better categorization of the information, which will derive on an easier way to analyze data and convert it into valuable information for decision-making. This paper presents an analysis of the state of the art, and previous advances of the use of technology for discovering the presence of kidney diseases. This will lead to alternative and novel ways to detect the disease on its initial stages, with the aim of supporting the medical decision-making process.

Keywords: engineering in medicine, big data applications, machine learning

1 Introduction

As referred by Mayer [1], the purpose of the analysis of data is no longer simply answering existing questions but generating new hypotheses. Using the power of well-organized abundant information, guides us to the creation of knowledge in a faster ratio than we ever imagined. Some none invasive techniques are being used to predict kidney malfunction.

There are three key areas, where Big Data differentiates from any existing conventional analyses of a given data sample:

1. Data is being captured in a more comprehensive way.
2. Inclusion of new techniques like machine learning, which according to Murphy [2] is the set of methods that can automatically help us by detecting patterns in data, which we can use to predict future data, or to perform decision making under uncertainty.
3. Creation of new hypotheses.

This takes us to the conclusion that if we take advantage of the benefits of Big Data, we can accelerate the research in any field; medicine is not an exception.

2 The big data paradigm

2.1 Historical Review

The growth rate in the volume of data, popularly known as the "information explosion" [3] was first used according to the Oxford English Dictionary in 1941. Since then, multiple major milestones in the history of the big volumes of data have been reached:

1. On 1944, Fremont Rider [4] (Wesleyan University Librarian) estimated that the American university libraries, were doubling the size of their amount of information every sixteen years.
2. On April 1980, Tjomsland [5] said that the large amounts of data are being retained, because the users involved have no way of identifying obsolete data, besides the fact that the consequences associated for storing obsolete data, are less than the ones associated to discard potentially useful data.
3. On October 1998, Coffman [6] concluded that the growth rate of traffic on the public Internet was about 100% per year; data traffic overtook voice traffic on the U.S. by 2002.
4. On March 2007, Gantz [7] estimated that the information added year by year to the digital universe, will increase more than six-fold to 988 Exabyte's, doubling every 18 months. (On a follow up release of the same study, that forecast was surpassed reaching 1227 Exabyte's in 2010, and getting to 2837 Exabyte's in 2012).
5. On February 2011, Hilbert [8] estimated that in the year of 1986, 99.2% of the storage capacity was analog, while in 2007 94% was digital (it was on 2002, that digital information surpassed the non-digital for the first time).

2.2 Big Data Definition

Despite the fact that there is not a simple and widely accepted definition of the term Big Data, multiple authors and data engineer experts uses three V's to approach a definition (volume, velocity and variety).

Volume refers to the amount of data, variety to the different types and sources of data, and velocity to the speed needed to process it. The increase of the volume of data, as well as the variety of data sources and formats is exponentially growing year by year, while the users are demanding an increase in the velocity to access and/or process the data.

We define the big data management as the set of techniques and tools necessary to deal with the increasing amount of information, which has the purpose of organizing and categorizing the information in a better way, so that we can extract valuable information from huge data-bases, at the maximum possible speed.

Besides the challenges that the volume, variety and velocity causes, big data is becoming extremely valuable in the development of new knowledge for multiple business and industries.

3 Engineering for chronical kidney disease analysis

Multiple studies use technology to generate faster advances on a research. In the particular case of Chronical Kidney Disease (CKD), there are several examples where the analysis of data, has raised the interest of both medical and Big Data researches.

3.1 Medical Studies

A recent study analyses the early stage of kidney complication caused by Diabetes Mellitus (DM) through the analysis of the iris image of the patient [9]. Iridology is a technique based on the shape and structure inside the iris, which can picture the body system. According to this technique, anything that happens in the body is reflected as a sign in the eye-iris.

In his analysis, Prayitno [9] used 47 participants, from which 31 were preliminary diagnosed with DM, while 16 had no prior indication of DM or any kidney damage. They all had their eye images captured and analyzed, and also a blood test was taken from them. From the 47 patients, 36 of them (76%) were showing a broken tissue, related to the kidney location on the iris. From the 31 participants that were preliminary diagnosed with DM, a 100% were reflecting broken tissue. This concludes that all patients with DM, showed any type of complication with their kidney, and this was reflected in the image of their iris.

According to a recent study [10] there are strong links between depression and anxiety and CKD. Despite the fact that anxiety and depression in advanced stages sound logically related, studies have demonstrated higher prevalence rates of depression in patients with CKD than other chronic diseases.

Depression is an emotional state, which is characterized by causing somatic and cognitive symptoms like a constant feeling of sadness, sleeplessness, and in many occasions the loss of appetite and sexual desire, and the lack of interest in common activities [10].

Anxiety is an emotional state, which causes a person to feel intense fear, uncertainty, and dread from anticipation of any given threating situation. When anxiety becomes a disorder (in difference to brief anxiety states), remains at least 6 months, and can get worse if the patient is not under treatment [10].

According to Hedayati [11] and Cukor [12], patients with CKD have a ratio of depression five times more compared with general population. The range is between 20% and 30% of

patients with CKD, affected with depression. To prove this numbers, this was illustrated in the analysis of 249 studies conducted by Palmer et al. [13]. The patients treated with dialysis the rate of depression was 22.8% (this by using clinical interviews). But, when the technique used for the measurement was self-rated questionnaires, the occurrences increased to 39.3%. According to a study made on 2007 [14], the prevalence rate of anxiety is estimated to be between 12% and 52%; however, there is a limited number of studies, therefore the exact rate is uncertain.

As we can see, the exact rates of anxiety and depression is not well defined, but independently if whether the reality resides in the upper or lower boundaries of the rates, the levels are alarming, and there is a high correlation between CKD and these emotional states.

3.2 Algorithm Studies

On a different study, multiple factors for kidney dialysis such as creatinine, sodium and urea play an important role in deciding the survival prediction of the patients [15]. Clustering the information is important to identify the influence of kidney dialysis parameters. Using a simple K-means algorithm can help to determine the interaction between these parameters and patient survival.

Table 1 shows the range of parameters that are used for prioritization (if the patient falls into the high priority level, a kidney transplantation is needed), categorized in high, medium and low, based on the clinician feedback that was gathered from the analysis of 230 datasets taken from the Global Hospital Chennai [15].

Data mining is the process of extracting hidden information from a huge dataset. When choosing the appropriate data mining algorithms, in combination with applying a correct procedure on dialysis data set, will derive in a survival prediction of patients with CKD. On this analysis, the author used K-means. The study did not consider other parameters like chloride and bicarbonate levels, those will be analyzed in a future research.

On a different research, it was learned that using a support vector machine can help a doctor to detect if a patient is showing a chronic condition or not, with an accuracy of 98.35% [16]. This technique is divided in two phases: the classification modeling (which is in charge of finding rules and model in the classification of kidney disease) and the system development (which uses the input data and applies machine-learning techniques to give a result to the doctors).

The use of a selector named OneR attribute [17] which helps to extract action rules depending on the stages of the CKD condition, helps to prevent the advance of a chronical renal disease to beyond stages. On table 2, we can see the 5 stages and the description of each, depending on the level of GFR (glomerular filtration rate), which is the best way to measure the level of kidney function, to determine the stage of a kidney disease. A Naïve classifier uses the probability based on the Bayes theorem, with strong independence assumptions between the features. The results after applying this method, is the reduction of 80% of the attributes in the dataset, while improving the accuracy by 12.5%.

Table 1. Range of Parameters for prioritization decision making.

Sl.No	Parameters	Low Priority	Medium Priority	High Priority
1.	Creatinine	0.6 – 2.0	2.1 – 5.5	>5.5
2.	Urea	10 – 40	41 – 60	>60
3.	Sodium	135 – 150	NA	>150
4.	Potassium	3.5 – 5.3	NA	>5.3
5.	Chloride	95 – 106	NA	>106
6.	Bicarbonates	18 - 23	NA	>23

Table 2. Five stages of CKD, based on glomerular filtration rate

Stage	Description	GFR (ml/min/1.73m²)
1	Slightly diminished function	≥90
2	Kidney damage and Mild reduction in GFR	60-89
3	Moderate reduction in GFR	30-59
4	Severe reduction in GFR	15-29
5	Established kidney failure	<15

The system extracted the action rules, for any given chronical disease stage, so that the specific treatment can be taken. To avoid the CKD to advance to the next stage.

Over the last decades, the use of multiple data mining techniques to investigate diseases have become essential for the health care industry, and therefore its use have exponentially increased.

As we can see in Fig. 1, classification, which is the approach that assigns objects into groups that share common characteristics [18], has been the preferred method during the last 15 years to investigate breast cancer, heart and kidney disease.

After analyzing the behavior of two different classification techniques: Artificial Neural Network (ANN) and Naïve Bayes to predict and diagnose CKD [18], the conclusion is that Bayes is the most accurate classifier with 100% of accuracy, against a 72.73% of accuracy when using ANN.

3.3 The importance of Machine Learning

Besides the fact that machine learning and big data are closely related to each other, it is important to understand about machine learning, because the algorithms commonly used by it, determine how we are going to interpret and process big data.

Machine learning can be defined as the application of artificial intelligence, with the objective of providing a system the ability to automatically learn and improve from experience, without the need of being explicitly programmed [19].

Almost 70 years ago, Alan Turing on his Computing Machinery and Intelligence paper [20] asked the question: can machines think?

On his paper, Turing refuted multiple objections opposed to his opinion. As highlights, two of them can be addressed:

6. Theological objection: It argues that thinking is a unique function of man given by god through the soul. Hence animals and machines cannot think. Turing refuted this argument stating that machines would not usurp God's power more than humans do so when procreating children.

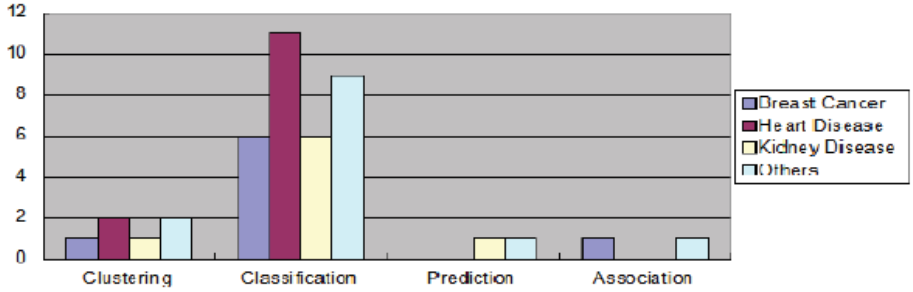

Fig. 1. Data mining techniques used the last 15 years for disease detection

7. Mathematical objection: It argues that multiple logical mathematical results show that there are limitations to the power of discrete-state machines.

Even do Turing acknowledged that there are limits to the power of machines, his said that it does not proof that the limits also apply to human intellect.

We cannot tell with 100% of certainty that a computer will be able to think, but no one has been able to prove it wrong.

A couple of popular examples of machine learning which are currently used:

8. Fraud detection on credit cards: the machine learning code that is embedded to the banks systems is capable to learn and become familiar with your spending patterns, so that when something unusual is detected on your account, an alert is triggered.
9. Customer service and support: nowadays it is possible that you call a company asking for support, and you end up talking with a computer without even noticing it. This because a computer can follow scripts to reply most of the questions that a regular customer might have, plus new technology that can perfectly simulate a real-life person's voice.

If we manage to properly use machine learning algorithms for analyzing and processing our data, we will have more and stronger tools to discover trends on the multiple variables that we want to consider for detecting CKD on its initial stages.

4 Future work

Besides making regular invasive techniques to determine the presence of CKD on its multiple stages, there are studies that look to correlate the kidney condition with other variables. We currently find efforts in iridology, emotional states (like anxiety and depression), data mining algorithms (like K-means, vector machines and Naïve Bayes).

We will also deeply use other computational tools and techniques related to big data, machine learning, and signal and image processing.

This research will focus in the discovery of the most adequate techniques and variables for analyzing and diagnose CKD, making special emphasis on the initial stages of the disease. To achieve it, it is important to include the analysis of psychological variables and their impact on the development of the disease. We will use real data of current patients, from a public health system database.

We will seek to find different ways of detecting CKD on initial stages, to develop alternative techniques that would help to prevent a fast deterioration of the kidneys, and start treatment as early as possible to increase the life expectancy, together with a high quality of life on patients affected with the condition.

It is important to highlight the fact that we will have support from urologists and other medical specialists during our investigation.

5 Conclusions

The increase of the amount of information (especially in digital sources) developed the need to create more optimal ways to store, process and analyze data. Big Data promises to be a support for the scientific world (including medicine).

Different methods are being used to show correlations of the deficiency of the kidney, with alternative methods for its detection. Studies suggests that the CKD is strongly related to other conditions like anxiety and depression.

Apart from invasive classical methods, iridology and mathematical methods applied to given sets of data from patients with CKD, such as K-means, Naïve Bayes and support vector machine, helps with the classification and categorization of data for a better analysis, and therefore become a tool that support doctors on decision making when deciding the best treatment for patients with CKD.

Machine learning techniques and algorithms will play an important role when deciding how to process and analyze the collected information.

The use of alternative techniques for the detection of malfunctioning of the kidneys and the diagnosis of CKD on its initial stages, is not intended to replace the traditional methods, but to provide the doctors with additional and valuable information so they can take a more informed decision when deciding the best treatment for their patients.

Acknowledgment

The authors would like to thank the **Instituto Tecnológico y de Estudios Superiores de Occidente (ITESO)** of Mexico for the resources provided for this research, and to the Mexican National Council for Science and Technology (Consejo Nacional de Ciencia y Tecnología CONACYT) for its support thru the scholarship number 498569 assigned to the main author.

References

1. V. Mayer-Schönberger and E. Inglesson, “Big data and medicine: a big deal?,” in NCBI Review Symp. Stanford, CA, USA, Jan. 2017, pp. 418-429.
2. K. Murphy, “Machine learning, a probabilistic perspective,” in The MIT Press, Massachusetts, US, Apr. 2014, pp. 62-63.
3. Press, G. (2013). A Very Short History Of Big Data. [online] Forbes. Available at: https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#6237ca465a18 [Accessed 12 Oct. 2018].
4. Rider, A. (1944). The scholar and the future of the research library. A problem and its solution. [Advocating the use of micro-cards.]. New York City: Hadham Press.

5. Tjomsland, I. (1980). The gap between MSS Products and User Requirements. [online] IEEE Computer Society. Available at: http://www.gbv.de/dms/tib-ub-hannover/017462509.pdf [Accessed 12 Oct. 2018].
6. Coffman, K. and Odlyzko, A. (1998). The size and growth rate of the Internet. [online] Dtc.umn.edu. Available at: http://www.dtc.umn.edu/~odlyzko/doc/internet.size.pdf [Accessed 12 Oct. 2018].
7. Gantz, J. and Reinsel, D. (2007). The Expanding Digital Universe: A Forecast Of Worldwide Information Growth Through 2010. [online] ECM Connection. Available at: https://www.ecmconnection.com/doc/the-expanding-digital-universe-mdash-a-foreca-0001 [Accessed 1 Sep. 2018].
8. Hilbert, M. and Lopez, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. [online] MartinHilbert.net. Available at: http://www.martinhilbert.net/WorldInfoCapacity.html/ [Accessed 1 Sep. 2018].
9. A. Prayitno, "Early detection study of kidney organ complication caused by diabetes mellitus using iris image color constancy," in ICTS International Conference on Information, Surabaya, Indonesia, Apr. 2017, pp. 146-149.
10. Z. Goh and K. Griva, "Anxiety and depression in patients with end-stage renal disease: impact and management challenges – a narrative review," International Journal of Nephrology and Renovascular Disease., vol. 11, pp. 93-102, Nov. 2017.
11. S. Hedayati and F. Finkelstein, "Epidemiology, diagnosis, and management of depression in patients with CKD," in American Journal of Kidney Diseases, vol. 54, pp. 741-752, July.2009.
12. D. Cukor, J. Coplan, et al. "Depression and anxiety in urban hemodialysis patients," in Clinical Journal of American Society of Nephrology, vol. 2, pp. 484-490, May.2007.
13. S. Palmer, M. Vecchio, et al. "Prevalence of depression in chronic kidney disease: systematic review and meta-analysis of observational studies," in Elsevier Kidney International, vol. 84, pp. 179-191, July.2013.
14. F. Murtagh, J. Hall and I. Higginson, "The prevalence of symptoms in end-stage renal disease: a systematic review," ACKD Advances in Chronic Kidney Disease, vol. 14, pp. 82-99, Jan. 2007.
15. B. V Ravindra and N. Siraam, "Discovery of significant parameters in kidney dialysis data sets by K-means algorithm," in IEEE International Conference on Circuits, Communication, Control and Computing, Bangalore, India, Nov. 2014, pp. 452-454.
16. M. Ahmad, V. Tundjungsari, D. Widianti, P. Amalia and U. Azizah, "Diagnostic decision support system of chronic kidney disease using support vector machine," in IEEE Second International Conference on Informatics and Computing, Jayapura, Indonesia, Nov. 2017, pp. 1-4.
17. U. Dulhare and M. Ayesha, "Extraction of action rules for chronic kidney disease using Naïve Bayes classifier," in IEEE International Conference on Computational Intelligence and Computing Research, Chennai, India, May. 2017, pp. 1-4.
18. V. Kunwar, K.Chandel, et al. "Chronic kisney disease analysis using data mining classification techniques," in IEEE International Conference –Cloud system and Big Data engineering, Noida, India, Jan. 2016, pp. 300-305.
19. Expert System. (2019). What is machine learning? A definition [Online]. Available: https://www.expertsystem.com/machine-learning-definition/
20. A.Turing, "Computing Machinery and Intelligence," Mind, vol. 59,pp. 433-460, Oct. 1950.