DATA VISUALIZATION AND DEPICTION INTERFACE TO DEVELOP CANCER EPIDEMIC DATASET

*[1]D. Roja Ramani, [2]G.Naveen Sundar,[3]N.Bhuvaneswary ,[4]D.Narmadha,[5]P.Nagaraj

ABSTRACT--Data *visualization is the vivid depiction of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide a versatile way to glimpse and understand trends, outliers, and patterns in data. In the world of Big Data, data visualization tools and technologies are vital part of analyzing the massive amounts of information and make data-driven decisions. As an outcome of the progressive technological trend, massive amounts of data from different subjects and areas are continuously generated and classified on a daily basis. In the past, the main problems were the preservation and publishing of data sets, but today, one of the main challenges is the presentation for more understanding of the data. The suitable visual representation of a given data set is the foundation for a precise and consistent interpretation, analysis and adoption of empirical conclusions related to the semantic meaning of information. In this paper, we present a synopsis of data visualization techniques and their practical application, starting from the acquisition of an unstructured set of publicly available data, their proper preprocessing and organization, and the visual representation for the end user. The nature of the data is related to the emergence, potential and development of various types of cancer diseases officially registered in different geographical regions. We present an interactive system that implements several visualization techniques.*

Keywords*--Cancer, data vislualization,bar plot, density,area,depiction*

I. INTRODUCTION

Visual analytics have been shown to be effective for data exploration [1], but the requirement of computational power is high for global comparisons of disease trends. Therefore, cloud computing is one of major enablers for explorations of data. This is a general shift of computer processing, storage, and software delivery away from the traditional desktop computers and local servers towards the cloud [2]. Cancer is one of the leading causes of morbidity and mortality worldwide. In 2017, 9.6 million people are estimated to have died from the various forms of cancer. Every sixth death in the world is due to cancer, making it the second leading cause of death – second only to cardiovascular diseases. Cancer is one of the most common causes of death, with nearly 7 million deaths each year worldwide. Right now 24.6 million people are living with cancer, and by 2020 it is projected that there will be 16 million new cancer cases and 10 million cancer deaths every year. In order to

[1] **Assistant Professor·Department of Information Technology, Sethu Institute of Technology, rosevsroja@gmail.com*

[2] *Assistant Professor,Department of Computer Science and Engineering , Karunya Institute of Technology and Sciences*

[3] *Assistant Professor,Department of Electronics and Communication Engineering , Kalasalingam Academy of Research and Education*

[4] *Assistant Professor,Department of Computer Science and Engineering , Karunya Institute of Technology and Science*

[5] *Assistant Professor,Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education*

Received: 21 Dec 2019 | Revised: 18 Jan 2020 | Accepted: 05 Feb 2020

address the rising health crisis, this WHA resolution gives special emphasis to the development and reinforcement of comprehensive national cancer control programmes that include prevention, early detection, improved treatment and palliative care, particularly for low- and middle-income countries [3]. The commonness of cancer varies by gender, age, ethnicity, geographical location, economic status, and so on. Generally, the cancer causes of death were common on breast, lung, liver, stomach, colon and rectum [4]. Although the age-standardized incidence rates on some cancer showed stable trends, but the prevalence of cancer has grown along with the ageing population. To better understand the progression of cancer, cancer registries were set up in different countries. The first population-based cancer registry was in Germany Hamburg in 1926 [4]. Cancer registries have been widely used in epidemiological research, so the World Health Organization (WHO) formed an International Agency for Research on Cancer (IARC) to collect cancer registry data across different countries. Descriptive studies use the registry database to examine differences in the incidence of cancer for different patient characteristics [5]. The data volume of the global cancer incidence is huge, so visual analytics can facilitate to broaden the data interpretation on disease distribution and trends.

Mainly, there are a number of techniques and approaches for visualizing data whose application depends on numerous factors and parameters, such as: the nature of the problem, the type of end-user, the purpose of visualization, the further applicability and scientific research goals, the structure and domain of data values, the size of the data and their precise semantic meaning.

The key part in this analysis is to use a visual analytics platform to distinguish cancer trends and patterns from WHO cancer registries. We wish to answer several major questions presented to us from cancer researchers, including: (i) What are the top-ranking cancers (ii) across different regions, (iii) Cases (iv) over the years, (v) across high- and mid-income regions, and (vi) across different age groups. The structure of this paper is organized as follows: related work in the academic field, data structure of cancer registry from WHO, data categorization for subgroup comparison, visual analytics environment, application scenario, and conclusion.

II. RELATED WORK

Data visualization is important to enhance the understanding of the overall dataset. It is widely applied to different academic areas. Fan et al. used different color regions to show the spatial distribution between weather temperature and light intensity, and developed a color coding scheme and showed on a map [8]. MuCusker et al. applied motion charts to demonstrate the association between tax payments and smoking prevalence. Visualization can initiatively show that when the taxes go up, the prevalence of smoking goes down [9]. Torres et al. designed a matrix of scatter plots to help researchers explore data patterns and formulate research hypotheses within the National Health and Nutrition Examination Survey [10]. Yung et al. developed an interactive platform to visualize multiple sets of proteomic data in huge data volumes [11]. Researchers can quickly judge the quality of selected proteomic features. One of the recent publications by Blevins presented an interactive data visualization application for human immunodeficiency virus (HIV) cohorts [12]. The platform was used to present the longitudinal plots, bubble plots and choropleth maps. Data visualization can demonstrate disease trends and distribution. The above platforms for data visualization were mainly conducted in the local computers. In the era of big data, the data volume will be a technical challenge to most of the existing platforms.

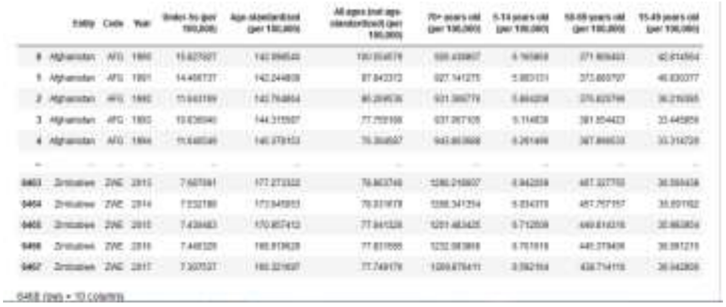
Received: 21 Dec 2019 | Revised: 18 Jan 2020 | Accepted: 05 Feb 2020

Therefore, visual analytics on the cloud will be an upcoming trend of application, and this study is a demonstration of visual analytics on the cloud platforms with global cancer incidence data using IBM Watson Analytics [13].

III. DATA STRUCTURE OF CANCER REGISTRY

The International Agency for Research on Cancer (IARC) is a specialized cancer agency of the WHO in order to promote international collaboration in cancer research. It has an important role in describing the burden of the global cancer through co-operation with cancer registries worldwide, monitoring geographical variations and assessing trends over time. IARC has developed a cancer database, called Cancer Incidence in Five Continents Time Trends (CI5 plus), providing access to detailed information on the incidence of cancer recorded by regional or national registries. Our World in Data is open access.

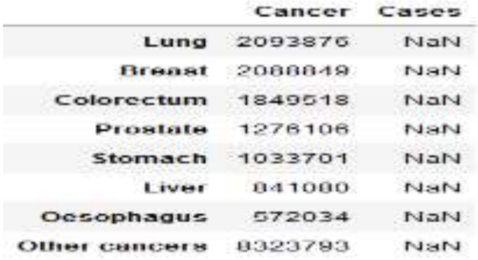
The database contains annual data of the population size, source of cancer registry. All histological data were available 1990 to 2017. All data were reported by Country, Year and by age at Under-5s (per 100,000), Age-standardized (per 100,000), All ages (not age-standardized) (per 100,000), 70+ years old (per 100,000), 5-14 years old (per 100,000), 50-69 years old (per 100,000), 15-49 years old (per 100,000). We selected the regions which reported 28 years of cancer incidence between 1990 and 2017.



	Entity	Code	Year	Under-5s (per 100,000)	Age-standardized (per 100,000)	All ages (not age-standardized) (per 100,000)	70+ years old (per 100,000)	5-14 years old (per 100,000)	50-69 years old (per 100,000)	15-49 years old (per 100,000)
0	Afghanistan	AFG	1990	15.627027	142.896542	100.054678	935.433807	6.765868	371.909403	42.614564
1	Afghanistan	AFG	1991	14.406737	142.244808	87.842312	927.141275	5.883131	373.800797	46.830077
2	Afghanistan	AFG	1992	13.643189	143.764864	80.294536	931.389779	5.864208	375.825798	36.216565
3	Afghanistan	AFG	1993	13.626940	144.315507	77.755108	937.267105	5.114838	381.854423	33.445859
4	Afghanistan	AFG	1994	11.648349	146.378153	76.364597	943.863868	6.261466	387.866533	33.314728
...	...	...	...	...	...	...	...	...	...	...
6463	Zimbabwe	ZWE	2013	7.607081	177.273322	76.863748	1280.216807	4.842318	487.327793	36.386436
6464	Zimbabwe	ZWE	2014	7.532180	173.945953	76.331978	1288.341254	5.034378	487.757157	35.697162
6465	Zimbabwe	ZWE	2015	7.438483	170.857412	77.041328	1251.463425	5.712658	449.614316	35.663854
6466	Zimbabwe	ZWE	2016	7.448329	168.819628	77.811565	1232.083868	5.761818	445.378408	36.391219
6467	Zimbabwe	ZWE	2017	7.307537	166.321607	77.749179	1209.876411	3.782164	438.714116	36.342808

6468 rows × 10 columns

Figure1 – Cancer deaths by Age



	Cancer	Cases
Lung	2093876	NaN
Breast	2088849	NaN
Colorectum	1849518	NaN
Prostate	1276106	NaN
Stomach	1033701	NaN
Liver	841080	NaN
Oesophagus	572034	NaN
Other cancers	8323793	NaN

Figure2 – Affected by types of cancer

IV. PROPOSED WORK

Data Visualization is the presentation of data in graphical format. It helps people understand the significance of data by summarizing and presenting a huge amount of data in a simple and easy-to-understand format and helps communicate information clearly and effectively.

In this paper, we use two datasets which are freely available. The cases of different types of cancer and Cancer death rates by age group, World, 2017 dataset. Figure1 & 2 shows that cancer death by age and affected by types of cancer.

4.1.1. Area

Figure3 shows that an area affected by the cancer death by age. An area chart or area graph displays graphically quantitative data. It is based on the line chart. The area between axis and line are commonly emphasized with colors, textures and hatchings. Commonly one compares two or more quantities with an area chart. Here plotted an area chart of Cancer deaths by Age.

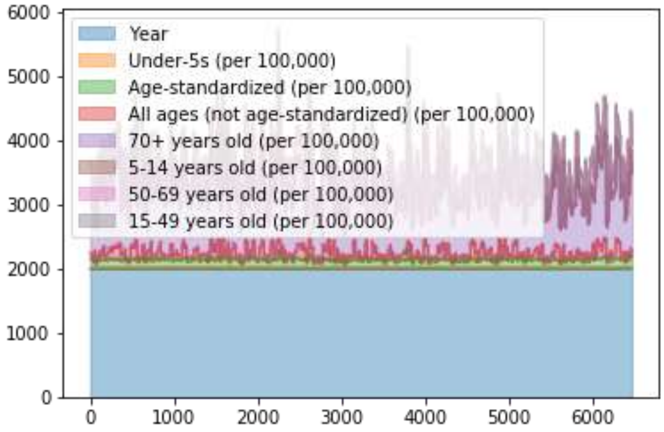


Figure3 – An area chart of Cancer deaths by Age

4.1.2. Bar plots

Figure4 shows that cancer deaths by age represented by bar chart. A bar chart or bar graph is a chart or graph that presents categorical data of Cancer deaths by Age with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a line graph.

Received: 21 Dec 2019 | Revised: 18 Jan 2020 | Accepted: 05 Feb 2020

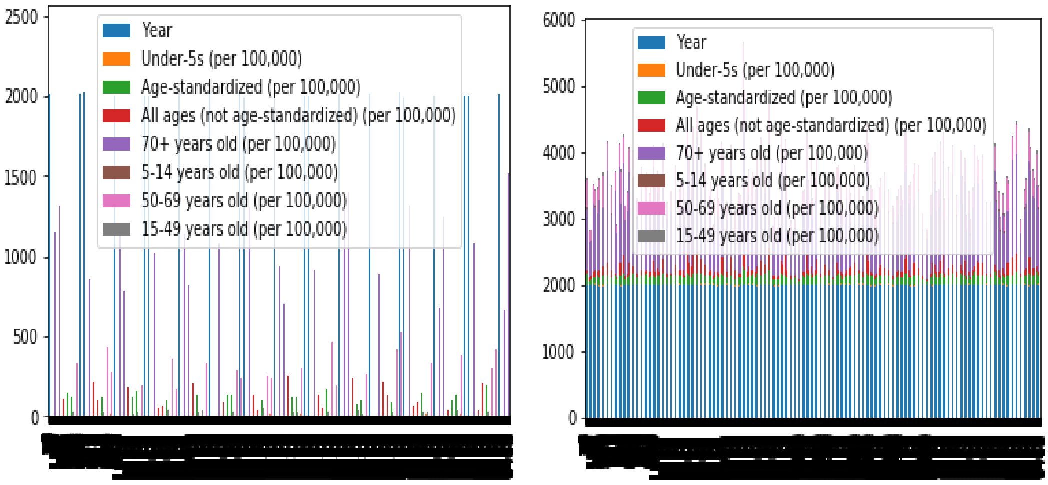


Figure4 – A bar chart of Cancer deaths by Age

4.1.3. Histograms

Figure5 histogram of cancer deaths by age and types of cancer cases, A plot that lets you discover, and show, the underlying frequency distribution (shape) of a set of continuous data of cancer death rate by age.

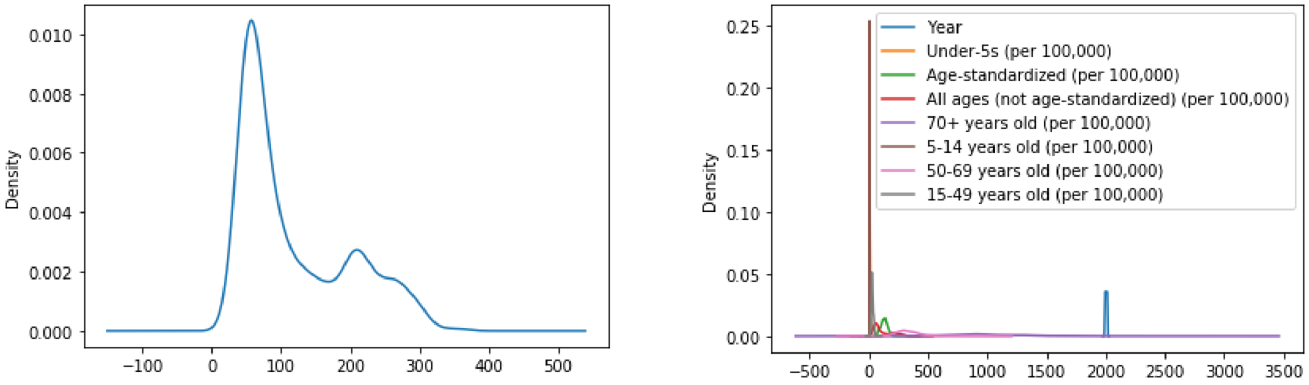
Figure5 – Histogram of Cancer deaths by Age and types of cancer cases

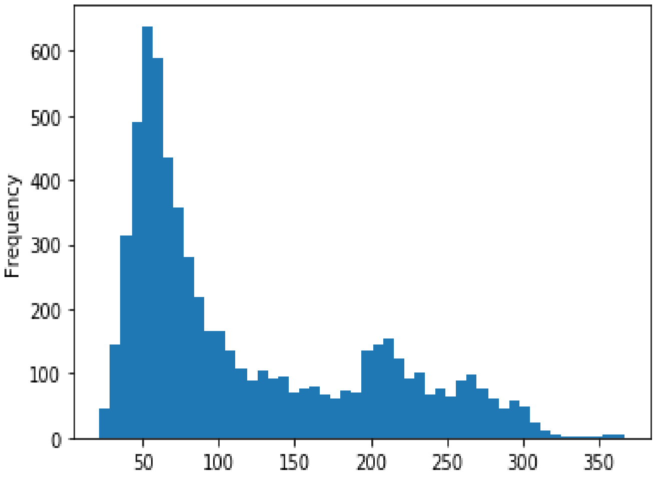
4.1.4. Density

To create a smooth curve given set of cancer deaths by age in figure6. This can be useful to visualize the "shape" of some data, as a kind of continuous replacement for the discrete histogram. It can also be used to generate points that look like they came from a certain dataset – this behavior can power simple simulations, where simulated objects are modeled off of real data.

Received: 21 Dec 2019 | Revised: 18 Jan 2020 | Accepted: 05 Feb 2020

Figure6 – Density of Cancer deaths by Age





4.1.5. Heat map

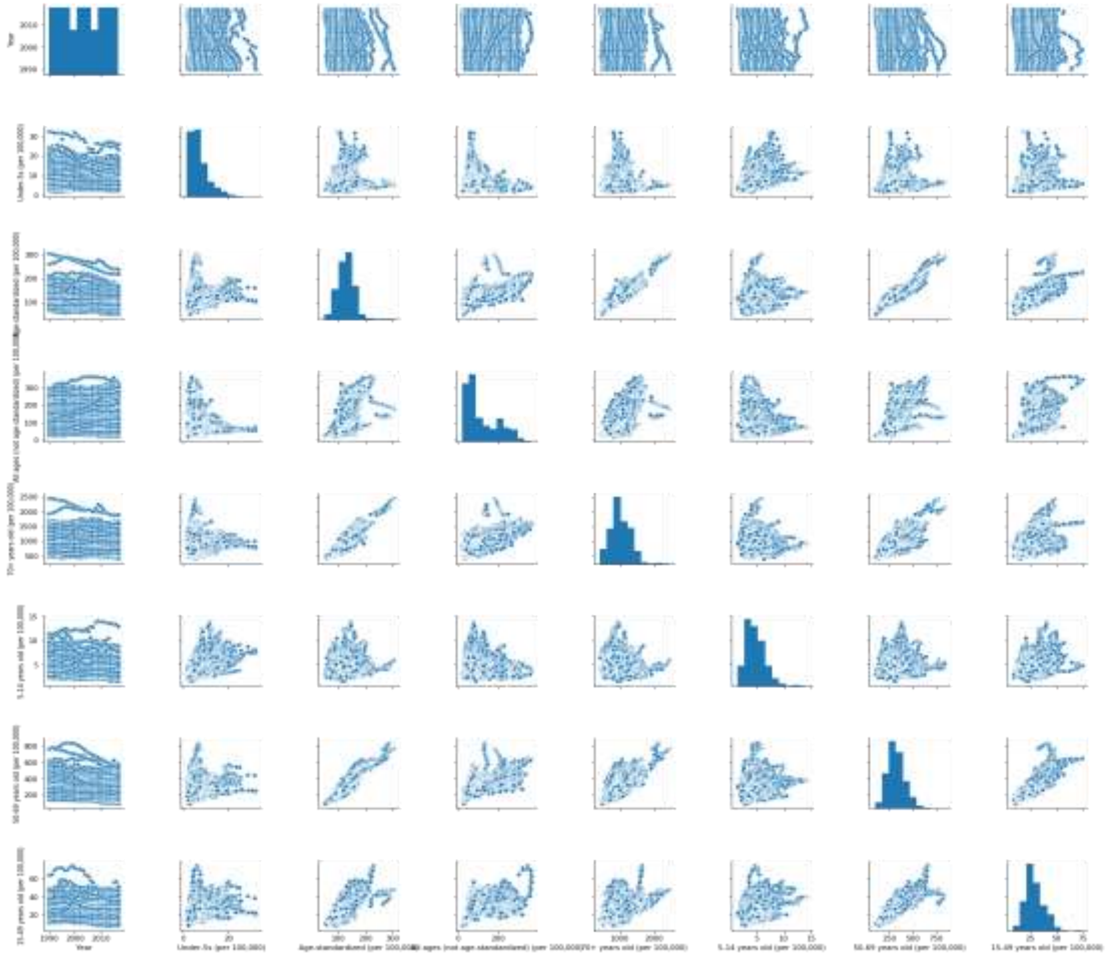
Figure7 shows that heat map for the cancer deaths by age, A Heat map is a graphical representation of data where the individual values contained in a matrix are represented as colors. Heat maps are perfect for exploring the correlation of features in a dataset. Get the correlation of the features inside a dataset.

4.1.6. Pair plot

Lastly, Seaborns pair plot and Pandas scatter_matrix, which enable you to plot a grid of pair wise relationships in a dataset.

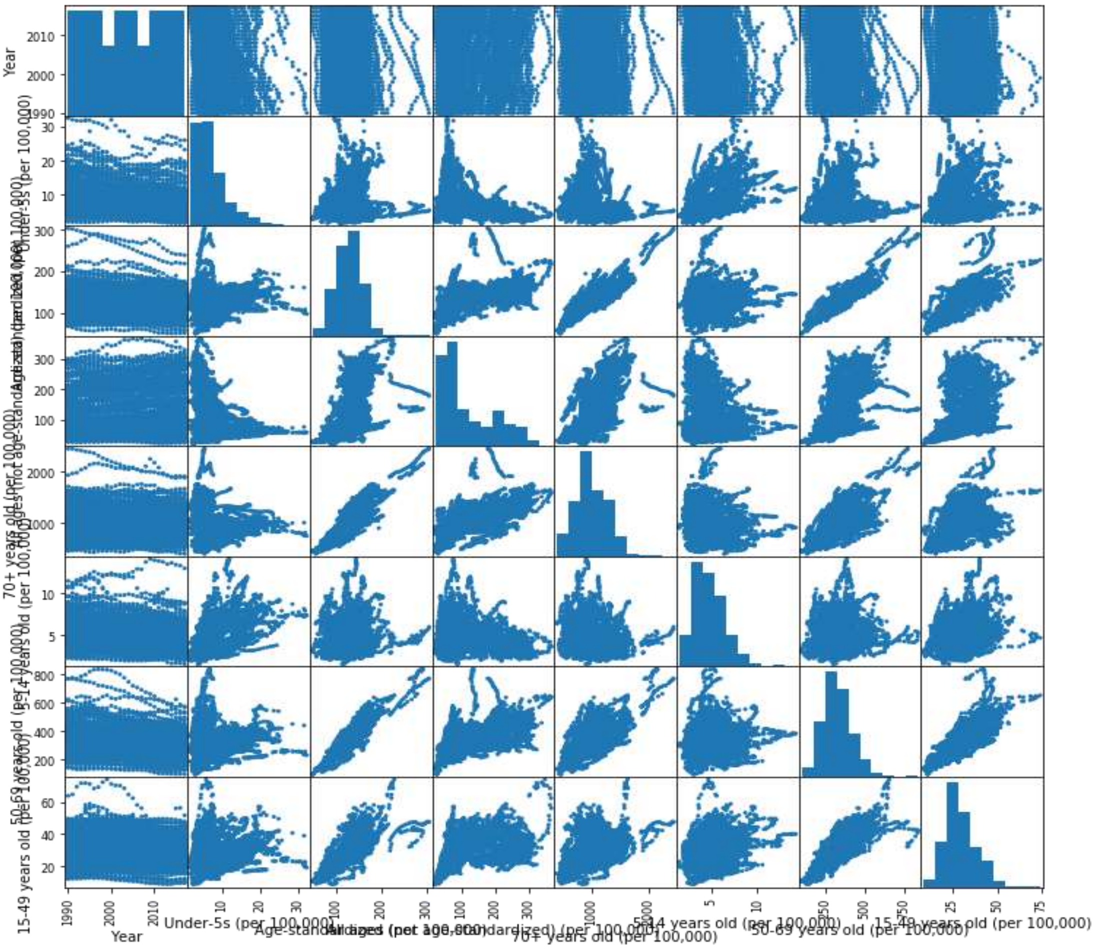
Received: 21 Dec 2019 | Revised: 18 Jan 2020 | Accepted: 05 Feb 2020

Figure8 shows that Pair plot map of Cancer deaths by Age techniques are always plotting two features with each other. The diagonal of the graph is filled with histograms and the other plots are scatter plots



(a)

Figure7 – Heat map of Cancer deaths by Age



(b)

Figure8 – Pair plot of Cancer deaths by Age

V. CONCLUSION

The software platform accentuate the semantic value and consequence of the data that are graphically presented, which enables a meticulous and efficient assessment of the process of development of diseases of this type. In this way, additional research can be made directly or indirectly related to other dynamic that may be solution in the progressive growth of registered cases of diseases of different cancer types. In this context, it can be accomplished that the precise implementation of data visualization techniques can speed up and make things easier. The process of interpretation of data and their significance, especially when it comes to huge data sets with a number of parameters and domains of values. The ease of this paper can provide as a basis for further related scientific research where graphic interpretation and interdependence of larger data sets is of enormous value. The acquired relational database can be integrated within already existing academic and scientific collections of data of a similar nature. Accordingly, the developed software platform can be used as an application module for integration with already existing visualization tools, but also independently, as a software component that can provide visual representation and interactive modules for review and systematic analysis to a predefined set of data structures.

In this work, we have described data visualization and to explore the open-sourced data on global cancer trends. The system makes the most of an interactive interface to exhibit the data distribution and trends. Future work can be extensive to the data projection on the particular type of cancer incidence.

REFERENCES

1. D.A.Keim, "Information Visualization and Visual Data Mining", IEEE Transactions on visualization and computer graphics, 2002, 8(1):1-8.
2. E. Saranya, A. Sunitha. "Identifying data integrity in Cloud Storage" International Journal of Computer Science, , Mar 2012, 9: 1694-14.
3. B. Stewart and C.P. Wild (eds.), International Agency for Research on Cancer, World Health Organization, (2014) World Cancer Report 2014 [Online], Available from: http://www.iarc.fr/en/publications/books/wcr/wcrorder.php [Accessed: 10th June 2016].
4. G. Wagner, History of cancer registration. In: O.M. Jensen, D.M. Parkin, R. MacLennan, et al, eds. Cancer registration: principles and methods. IARC Scientific Publications no 95. Lyon: International Agency for Research on Cancer, 1991: 3–6.
5. D.M.Parkin, The evolution of the population-based cancer registry, Nat Rev Cancer, 2006; 6: 603–12. [6] A.M. MacEachren, C.A. Brewer and L.W. Pickle, "Visualizing Georeferenced Data: Representing Reliability of Health Statistics", Environment and Planning, 1998, 30:1547-61.
6. F. Fan and E.S. Biagioni, "An approach to data visualization and interpretation for sensor networks", Proceeding of the 37th HICSS, 2004.
7. J.P. McCusker, D.L. McGuinness, J. Lee, C. Thomas, P. Courtney, Z. Tatalovich, N. Contractor, G. Morgan and A. Shaikh, "Towards next generation health data exploration: a data cube-based investigation into population statistics fortobacco", InSystem Sciences (HICSS), 2013 46th Hawaii International Conference on, 2013, 2725-2732 IEEE.
8. S.O. Torres, H. Eicher-Miller, C. Boushey, D. Ebert and R. Maciejewski, "Applied visual analytics for exploring the national health and nutrition examination survey", Proceeding of the 45th HICSS, 2012. [10] L.S. Yung, C. Yang and M. Dakna, "SyncPro: A synchronized visualization tool for differential analysis of proteomics data sets", InBioinformatics and Biomedicine Workshops (BIBMW), 2010 IEEE International Conference, 2010, 95-100 IEEE.
9. M. Blevins, F.H. Wehbe, P.F. Rebeiro PF, C.C McGowan and B.E. Shepherd, "Interactive Data Visualization for HIV Cohorts: Leveraging Data Exchange Standards to Share and Reuse Research Tools", PloS one, 2016, 11:e0151201.
10. C.R. Baquet, J.W. Horm, T. Gibbs and P. Greenwald, "Socioeconomic factors and cancer incidence", J Natl Cancer Inst, 1991, 83:551-57. [13] O.B. Ahmad, C. Boschi-Pinto, A.D. Lopez, C.J. Murray, R. Lozano and M. Inoue. Age standardization of rates: A new WHO standard. World Health Organization 2001. GPE Discussion Paper Series: No.31. Available from: http://www.who.int/healthinfo/paper31.pdf, accessed [6 Jun 2016]