2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM) | 978-1-6654-2318-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/IMCOM51814.2021.9377403

Big Data Analytics for Prediction Modelling in Healthcare Databases

Ritu Chauhan
Amity Institute of Biotechnology
Amity University
Noida, india
rchauhan@amity.edu

Eiad Yafi
Malaysian Institute of Information Technology
Universiti Kuala Lumpur
Kuala Lumpur, Malaysia
eiad@unikl.edu.my

***Abstract*—Big data in healthcare has manifested as well as benefited healthcare practioners and scientists around the globe to detect hidden patterns for future clinical decision making. The major complexity faced in real world application domain is the volume of Electronic Health Records (EHR) which has gathered due to high end IT based technology which has boomed in past century for early detection of disease. The traditional technology tools adopted were incapable to discover hidden patterns due to its computational requirements. So, Big data has its generosity need in healthcare intervene technology due to diverse nature of data and accelerated speed of data that needs to processed for bet- ter diagnostic interventions. This study has been conducted using predictive data analytics on big data for discovery of knowledge for future decision making. The study consists of information about 3,56,507 patients from 1982-2010. Data curation has been done by organizing under various categories including Age, Year (1982-2010), Incidence Counts (1982-2010, all age groups and both genders), and Mortality Counts (1982-2010, all age groups). The results represents invariable patterns which can be utilized for future predictive modelling.**

***Index Terms*—Big Data, Big Data Analytics, Data Mining, Healthcare Databases, Breast Cancer**

I. Introduction

Data Mining in healthcare is usually considered as technology-based process which involves finding of trends as well as the extraction of predictive information, which is usually hidden in large scale databases. Apart from predicting trends and behaviours from varied databases, it also aids in discovering previously unknown or undiscovered patterns from related databases. In past several data mining techniques are widely anticipated in varied application domains to detect hidden knowledge from large scale databases. However, most commonly used data mining techniques are: Decision Trees, Nearest Neighbour methods, Rule Induction, Genetic Algo- rithms, Applied Artificial Neural Networks and Regression. In current study of approach, we have utilized data mining in medical databases for detecting hidden patterns for prediction of disease in big databases. Thus, medical databases are becoming voluminous in nature due to high end IT based technology, which required automated technological advances to analyse the data for future medical diagnosis.

The focus of researchers and scientists around the globe is to design and implement a model for detecting the prognosis and diagnosis of disease among the big databases. In context to same cancer research has become the most prospective area of reached field due to its big data-intensive nature and complexity research assisted by speedy developments in acquisition of specific methods. Several studies in the past correlated data mining with cancer is studied to detect the hidden knowledge from big databases. For example, for keeping a tab on changes occurring in tumour cells while they progress towards being invasive from a normal character, next generation sequencing has been used to gather data for genomic and proteomic nature of cancer cells. Another method used in field of cancer research is Clustering to mine data for developing useful hypothesis and give direction to a particular research or a study undertaken for future decision making. Another data mining technique called Decision tree has lend its helping hand in healthcare field. It has been used in prediction of different skin diseases in children as well as in adults along with the aid of artificial neural networking.

According to Australian Institute of Health and Welfare, cancer is among the leading cause of death in Australia, but the most commonly diagnosed cancer in Australia in 2014 was breast cancer. It is estimated that it will be the most commonly diagnosed cancer in 2018 among both males and females [1-2]. However, the It is estimated that, one in eight women will develop breast cancer in their entire life. On an average, it is said that, eight women die from breast cancer daily in Australia. Currently, there are more than 65,000 people living in Australia with breast cancer today. In 2017, 17,586 women were projected to be diagnosed with breast cancer, although mortality is predicted to continuously decline. Australian women diagnosed with breast cancer have a 90% chance of surviving five years after diagnosis [3]. In context to same, the major drawback for early diagnosis of disease is the large complexity and volume of data where the clinical prediction becomes invariable very difficult. To overcome the current flaws of clinical implications big data mining are generalized applied in medical application domain to detect hidden patterns for knowledge discovery[16-20].

In current study of research big data analytics techniques are utilized to discover the hidden information from breast cancer databases for future diagnosis and prognosis of disease. The data consists of information about 3,56,507 patients from 1982-2010. Data curation has been done by organizing under

978-0-7381-0508-6/21/$31.00 ©2021 IEEE

various categories including Age, Year (1982-2010), Incidence Counts (1982-2010, all age groups and both genders), and Mortality Counts (1982-2010, all age groups and both genders). The overall paper is contributed in varied sections, first we discuss in general healthcare with varied data analytics approach applied in past. Section III discuss the methodology adopted for detecting patterns among the big databases. Results are contributed in section IV, last section has conclusion.

II. Big Data Analytics in Healthcare Databases

Big data in healthcare has manifested as well as benefited healthcare practioners and scientists around the globe to detect hidden patterns for future clinical decision making. The major complexity faced in real world application domain is the volume of Electronic Health Records (EHR) which has gathered due to high end IT based technology which has boomed in past century for early detection of disease. The traditional technology tools adopted were incapable to discover hidden patterns due to its computational requirements. So, Big data has its generosity need in healthcare intervene technology due to diverse nature of data and accelerated speed of data that needs to processed for better diagnostic interventions.

The condensed nature of big data is creating implicative opportunities among the big data scientist to assist and understand avid nature of data and discover associated patters which can benefit healthcare practioners for future prediction of disease. Thus, big data provides enormous opportunities in healthcare domain which can benefit patient care by deriving insights of data for better decision making. The big data scenario is applied in healthcare for better prediction outcomes which include provides better understanding of varied parameters which can be cause of prognosis of disease and provide best patient care. The cost effectiveness of disease and provide timely outcomes which can invariable reduce the cost and provision of better offers. The patient fingerprinting could be adopted to determine the patient at risk of developing certain disease by invariably detecting the behaviour and dietary habits and support new prevention techniques for early prognosis of disease. The major challenge is to provide low cost-effective treatment in healthcare which can support wide variety of patient needs. The big data analytics are widely anticipated in fraud detection techniques where the patient's claiming fraud claims from insurance organizations must be detected prior for better clinical care and measuring accuracy of claims synthesized. The new mobile computation technology is ardently utilized to use application based healthcare diagnostic tools to measure heart rate, steps taken on daily basis and if analytics provide a platform for future prediction, so this can entirely change the scenario of human health and wellbeing.

Big data analytics provides several benefits to healthcare application domain by detecting early prognosis of disease where the treatment of disease in early stage can overall reduce the cost of the disease. However, the big data is widely dealing with issues in varied application domains which includes:

- Data Heterogeneity: The data available from varied resources in different formats is among the challenging task which needs to be overlooked for better decision making. For example, different healthcare organization are utilizing the software's and hardware's with need of patient care, hence acquiring relative predictive outcomes becomes relative complex situation for better patterns detection. The technological advancement should have provisions of low cost where each technology is shared at similar platform to benefit patient care.
- Data Volume: The new adoptive IT based technology in healthcare which includes MRI, CT Scan, Sensor based technology, Imaging and others have vastly increased the size and volume of data. Such nature of data has evolved due to invasive computational health data. The time has lapsed where EHR were the primary resource of data available in single platform thus, data has revolutionized with focus to determine better prognosis and diagnosis of disease in prediction modelling.
- Data Privacy: The patient care data consists of histological, socio economic, clinical Implications, hospital data, Financial data, causes of specific disease and others. Storing such immense nature of data is also adding high cost healthcare cost. Big data with cloud-based technology are intervened utilized to deal with such nature of data. Cloud computing provides varies service which Infrastructure as Service for providing infrastructure which includes storage, speed and processing time, Software as service for providing software's virtually to deal with time and purchasing cost of software's. Data Analytics as service which is the new era technology to provide data analytics as its service to predict future modelling of data. Big data technology mainstream needs to enhance the data privacy so confidential data is secure.
- Ease of Use: The platform should provide overall flexibility and scalability for its complete usage among each application domain. Thus, it should benefit end users for prediction modelling in future implications.

III. Methodology

Big databases have grown outrageously in past decade, big data analytics are widely utilized technique for detecting hidden patterns from big data for knowledge discovery. In current approach of study breast cancer data is utilized for detecting patterns for future prediction modelling. The data consists of 3,56,507 patient data records which are collected from 1982-2010. Data curation has been done by organizing under various categories including Age, Year (1982-2010), Incidence Counts (1982-2010, all age groups and both genders), and Mortality Counts (1982-2010, all age groups and both genders). The population under study was 23 million till 2010. The data has been sourced from Australian Institute of Health and Welfare (AIHW) for Breast Cancer in Australia. The data comprises of people from different age groups, viz. 0-4 years, 5-9 years, 10-14 years, 60-64 years ranging up to 85+ years of age.

The study involves varied techniques to discover hidden patterns using big data to interpret predictive results. It involves finding of trends as well as the extraction of predictive

978-0-7381-0508-6/21/$31.00 ©2021 IEEE

information, which is usually hidden, in Big dataset. The Big data analytics results can be used for predictive analysis of trends and behaviours which will lead to knowledge discovery of undiscovered patterns[7]. Predictive data analytics through clustering is particularly relevant when the characteristics of the data are unknown and it is required to group data points and study their collective behaviour. It has been regarded as the first line of method for analysis to mine data for developing useful hypothesis, which would give direction to the particular research [9-10].

In current study of research, we have utilized Linear regression as a parametric model which is widely used in research fields to develop a relationship between two different types of variables which are also known as factors. The variable to be predicted is known as the dependent variable. The variables or factors which are used in predicting the value of the other dependent variable to be calculated, is called the independent variable. The method of linear regression analysis is most commonly utilized in research areas as it provides with a platform that helps in establishing existence of correlation between the variables in consideration. The Simple Linear Regression Model is represented by,

$$D(y) = A_0 + A_1X \quad (1)$$

Where, X, Y are the variable types involved in calculating regression. $(A_0 + A_1X)$ represents the formula being studied. When the equation successfully describes the relationship between these two variables, a regression model develops. The regression line generated is a straight line where, A_0 is the X intercept of the regression line and A_1 is the slope. $D(y)$ represents the mean or the expected value of Y for a given value of X.

The descriptive methodology is adopted to provide the insights of the characteristic features of the data, gives statistical summary for scale variables and data measures. The method generates details of statistics which are univariate as well as the graphs for scale variables for a whole data or a subset of the same data. Explore procedure is widely used in assessment of normality for scale variables (numeric) using special type of statistics called inferential statistics along with descriptive diagnostics plots. The method of logistic regression generates coefficients for a formula for prediction of logit transformations to detect the probability of presence of that particular characteristic we are interested in.

$$logit(q) = a_1X_1 + a_2X_2 + a_3X_3 + \ldots + a_kX_k \quad (2)$$

where, q represents the probability of presence of the characteristic we are interested in. Also, logit transformation described as the logged odds,

$$odds = \frac{q}{1-q} = \frac{\text{probability of the presence of a characteristic}}{\text{probability of the absence of a characteristic}} \quad (3)$$

and,

$$logit(q) = ln\left(\frac{q}{1-q}\right) \quad (4)$$

The method of logistic regression selects the parameters that help in maximizing the likelihood of observing those samples.

For this study, time series method included in forecasting which has been used is called Spectral Plots. This procedure identifies the periodic behaviour in a time series. These spectral plots analyse the variants of series inputs as a whole, into periodic components which are of different frequencies. For the further analysis of this study, Kaplan- Meier method for Survivability. This method, is regarded as a non-parametric method which is widely used in estimation of time related events usually when censored cases are present. With the aid of this method, a comparison can be generated between distribution by considering levels of a particular factor variable. It can also achieve separate analysis by levels of a type of a variable known as stratification variable.

IV. Results

According to the results analyzed, the count incidence cases were only 0 or 1 in number for ages upto 9 years, which then rose to 1,516 cases in young adults. The incidence alarmingly increased to 35,261 in age group of 50-54 years. As the higher age groups were approached, the cases dropped 13,24 in 85+ years of age. According to the results analyzed, each age group was analyzed with respect to incidence and mortality for each case studied. The Table 1, represents that in 25-29 years of age group 1,522 highest cases of incidence were found. As the age increased, there is significant increase in the number of cases. A substantial decline occurred in the incidence as the cases dropped 14,066 in the age group of 85+ years. The results signify 100% death rates in the age groups of 0-9, a decline in the rates were noted A 3% drop was seen from 19% to 16% up to 19 years of age as well as in the age group of 45-49 years. As the old age approached, more cases of death were registered, resulting in increase in the mortality rates. The increase in mortality from 22% to 63% in 85+ years of age group.

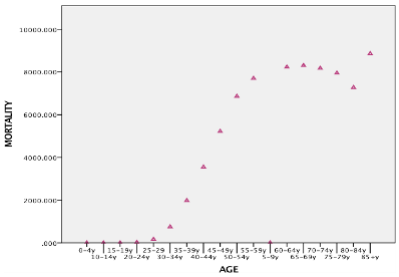

Fig. 1. Regression line representing mortality cases of breast cancer among women in different age groups 1

Fig 1 represents the total number of mortality cases using regression technique at age group the mortality rate is higher in age group of 80-84 years. In Fig 2 the incidence rate is calculated and to be higher in 50-54 years of age group. In table 2 a year wise trend is detected from 1982-2010, number of cases were found to be increasing subsequently each year.

978-0-7381-0508-6/21/$31.00 ©2021 IEEE

TABLE I
AGE-WISE INCIDENCE AND MORTALITY COUNTS ALONG WITH THE DEATH PERECENTAGE IN FEMALES

Age/Gender	Incidence	Mortality	Death Percentage
0-4y	0	0	0%
5-9y	1	1	100%
10-14y	1	0	0%
15-19y	15	2	13%
20-24y	249	20	8%
25-29y	1516	165	11%
30-34y	5247	739	14%
35-39y	11966	1947	16%
40-44y	22446	3491	16%
45-49y	31921	5121	16%
50-54y	35261	6724	19%
55-59y	35047	7473	21%
60-64y	35245	7961	23%
65-69y	32103	7973	25%
70-74y	26278	7826	30%
75-79y	21791	7640	35%
80-84y	16038	7039	44%
85+y	13924	8699	62%

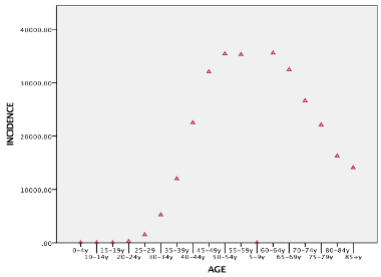

Fig. 2. Age-wise incidence of breast cancer in females in Australia.

The number of cases were found to be 5,303 in the year 1982. It jumped to about 10,061 in 1994. The case in 2010 reached an alarming number of 14,181. With the help of these results it was concluded that more females were being diagnosed and the cases only increased during the span of years studied. The case of mortality depicted the same picture. In 1982, 1987 cases of mortality were recorded. The number increased to 2,000 and reached 2,873 in 2010 showing increases in the cases of mortality.Even though the number of cases were seen increasing, the mortality rates were found to be decreasing. In the years 1982, 37% deaths due to breast cancer were recorded. These rates dropped to 28% in 1992. The rates, fortunately dropped to just 20% in 2010.

In Fig 3,4 a regression analysis is performed year wise for both incidence and mortality among the cases studied

The survivability rates were found to be 100% in younger age groups as no cases were found. The survivability rates varied in different age groups. The survival rates went down to 86.67% in the age group of 15-19 years of age, increased to 91.97% in the age group of 20-24 years. The rates, once again dropped to 80.9% in 50-54 years of age group. It further went down to 75% in 65-69 years of age. The age group 75- 59 years had 65% survival rate which dipped in the age group 80-84 years to 56%. The least survival rates were shown in

TABLE II
YEAR-WISE TREND IN FEMALES WITH BREAST CANCER RESIDING IN AUSTRALIA (1982-2010)

Year	Incidence	Mortality	Death Percentage	Year	Incidence	Mortality	Death Percentage
1982	5303	1987	37%	1997	10744	2604	24%
1983	5357	2040	38%	1998	10666	2541	24%
1984	5704	2166	38%	1999	11395	2512	22%
1985	6082	2196	36%	2000	11827	2521	21%
1986	6683	2165	32%	2001	12084	2594	21%
1987	6722	2293	34%	2002	11861	2681	23%
1988	7169	2361	33%	2003	12200	2710	22%
1989	7419	2449	33%	2004	12257	2624	22%
1990	8032	2422	30%	2005	12682	2710	21%
1991	8008	2526	32%	2006	12633	2624	21%
1992	8777	2429	28%	2007	13596	2722	20%
1993	9747	2611	27%	2008	13740	2746	20%
1994	10061	2669	27%	2009	14181	2785	20%
1995	9741	2635	27%	2010	14181	2837	20%
1996	10197	2620	26%				

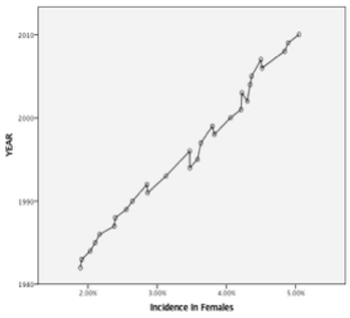

Fig. 3. Regression Year Wise Incidence in Females

women over 85 years of age.

TABLE III
SURVIVABILITY AGE GROUPS DIAGNOSED WITH BREAST CANCER

Age	Survivability	Age	Survivability
0-4y	100%	45-49y	84%
5-9y	0%	50-54y	81%
10-14y	100%	55-59y	79%
15-19y	87%	60-64y	77%
20-24y	92%	65-69y	75%
25-29y	89%	70-74y	70%
30-34y	86%	75-79y	65%
35-39y	84%	80-84y	56%
40-44y	84%	80+y	38%

Fig 5 represents the regression analysis of survivability rate at each age group, the survivability decrements as the age

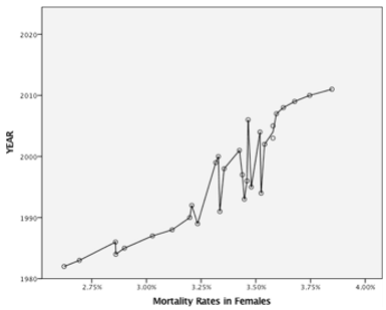

Fig. 4. Regression Year mortality in Females

978-0-7381-0508-6/21/$31.00 ©2021 IEEE

increases hence the early prognosis of disease can benefit patient care with medical implications.

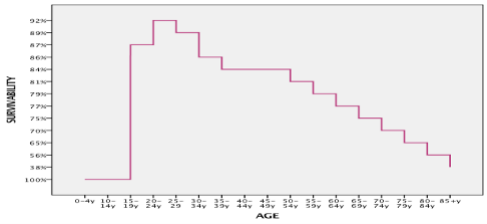

Fig. 5. Regression for survivability.

V. Conclusion

The aim of the study was to statistically estimate the incidence and mortality counts of breast cancer among females in Australia and to delve into trends in breast cancer among the same. With the help of this study, it was clear that breast cancer was the leading cause of cancer death among women in Australia. The most affected were the middle-aged group females. The maximum number of cases were counted in the year 2010. There was an increase in number of incidence counts from the year 1982 to 2010. This helped in reaching to a conclusion that there was more awareness about the disease, advancement in medical science helped to detect the cancer in time and elaborate chemotherapy and radiotherapy helped in treating the disease, thus reducing the overall mortality rates. Also, it was evident that cases of breast cancer incidence were maximum in the middle ages which is a cause of concern among the population. In the old ages, although the mortality rates decreased, the death count is still rising as the people in that particular age group are not able to bear the treatments. With the help of this study, innovative techniques can be developed to address every age group. Also, more awareness programs can be organized in order to bring the cases of breast cancer among men in light. It is hoped that timely diagnosis and treatment would reduce the number of incidences and mortality for a healthy and cancer-free life.

References

[1] DeSantis C, Ma J, Bryan L, Jemal A-CA Cancer J Clin-Breast cancer statistics, 2013, 2014 Jan-Feb;64(1):52-62, Epub 2013, Oct 1 https://www.ncbi.nlm.nih.gov/pubmed/21969133

[2] Asia Pac J Cin Oncol.2018 - Cancer in Australia: Actual inci- dence data from 1982 to 2013 and mortality data from 1982 to 2014 with projections to 2017, Feb;14(1):5-15, Epub 2017 Sep 20 https://www.ncbi.nlm.nih.gov/pubmed/28929586

[3] Rodney C. Richie, MD, FACP, FCCP; John O. Swanson, -Breast Cancer: A Review of the Literature, MD-JOURNAL OF IN- SURANCE MEDICINE, Journal of Insurance Medicine J Insur Med 2003;35:85–101, 2003 http://aaim.developmentwebsite.ca/journal- of-insurance-medicine/jim/2003/035-02-0085.pdf

[4] Xue Qin Yu, Roberta De Angelis, Qingwei Luo, Clare Kahn, Nehmat Houssami, and Dianne L O'Connell- A population based study of breast cancer prevalence in Australia: predicting the future health needs of women living with breast cancer, BMC Cancer, 14: 936, 2014 https://bmccancer.biomedcentral.com/articles/10.1186/1471-2407- 14-936

[5] Kriscia A Tapia, Gail Garvey, Mark Mc Entee, Mary Rickard, and Patrick Brennan, Asian Pac J Cancer Prev- Breast Cancer in Australian Indigenous Women: Incidence, Mortality and Risk Factors,18(4): 873–884, 2017 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5494235/

[6] Shreshtha Malvia, Sarangadhara Appalaraju, Bagadi, Uma S. Dubey, Sunita Saxena-Asia-Pacific Journal of Clinical Oncology- Epidemiology of Breast Cancer in Indian Women, Volume13, Issue4, August 2017 https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajco.12661

[7] Agarwal, R., & Dhar, V. (2014). Editorial-Big data, data science, and analytics: The opportunity and challenge for IS research. Information Systems Research, 25, 443–448

[8] Amarasingham, R., Patzer, R. E., Huesch, M., Nguyen, N. Q., & Xie, B. (2014). Implementing electronic health care predictive analytics: Considerations and challenges. Health Affairs, 33, 1148–1154

[9] Anderson, J. E., & Chang, D. C. (2015). Using electronic health records for surgical quality improvement in the era of big data. JAMA Surgery, 150, 24–29.

[10] Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. Health Affairs, 33, 1123–1131.

[11] Bengoa, R., Kawar, R., Key, P., Leatherman, S., Massoud, R., & Saturno, P. (2006). Quality of care: A process for making strategic choices in health systems. Geneva: World Health Organization. WHO press.

[12] Bonacina, S., Masseroli, M. & Pinciroli, F. (2005). Foreseeing promising bio-medical findings for effective applications of data mining. Biological and Medical Data Analysis, Springer.

[13] Caron, F., Vanthienen, J., Vanhaecht, K., van Limbergen, E., de Weerdt, J., & Baesens, B. (2014). Monitoring care processes in the gynecologic oncology department. Computers in Biology and Medicine, 44, 88–96.

[14] Bojesen, Stig E., Karen A. Pooley, Sharon E. Johnatty, Jonathan Beesley, Kyriaki Michailidou, Jonathan P. Tyrer, Stacey L. Edwards et al. "Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer." Nature genetics 45, no. 4 (2013): 371.

[15] Kitai, Toshiyuki, Takuya Inomoto, Mitsuharu Miwa, and Takahiro Shikayama. "Fluorescence navigation with indocyanine green for detecting sentinel lymph nodes in breast cancer." Breast cancer 12, no. 3 (2005): 211-215.

[16] Chauhan, R., Kaur, H. Big Data Application in medical domain. In Computational Intelligence for Big Data Analysis: Frontier Advances and Applications. Volume 19 of the series Adaptation, Learning, and Optimization pp 165-179.2015 Springer International Publishing Switzerland

[17] Chauhan, R., Kaur, H. "SPAM: An Effective and Efficient Spatial Algorithm for Mining Grid Data." Geo-Intelligence and Visualization through Big Data Trends. IGI Global, 2015. 245-263. Web. 9 Sep. 2015. doi:10.4018/978-1-4666-8465-2.ch010

[18] Chauhan, R., Kaur, H, Lechman, E., & Marszk, A. Big Data Analytics for ICT Monitoring and Development. Catalyzing Development through ICT Adoption. Springer International Publication 2017 pp25-36.

[19] Chauhan, R., Kaur, H. and Alam, A. M. Data Clustering Method for Discovering Clusters in Spatial Cancer Databases. International Journal of Computer Applications - Special Issue, Vol. 10 (6),2010 pp. 9-14

[20] X. Chauhan, R., Kaur, H. & Chang, V.Advancement and applicability of classifiers for variant exponential model to optimize the accuracy for deep learning, Journal Ambient Intelligent Human Computing, Springer 2017. https://doi.org/10.1007/s12652-017-0561-x.

978-0-7381-0508-6/21/$31.00 ©2021 IEEE