RESEARCH ARTICLE OPEN ACCESS

"Design and Analysis of Improvisation in Cancer detection technique in Patients using Machine learning"

Ravi Kumar Jangid, Mr. Prateek Singh
Centre for Cloud Infrastructure and Security Suresh GyanViharUniversity
Jaipur, Rajasthan, India
Centre for Cloud Infrastructure and Security Suresh GyanViharUniversity
Jaipur, Rajasthan, India

ABSTRACT:
Hospitalsupervision or healthcare administration is the field with reference to management, leadership and administration of hospitals, health care systems and hospital networks. Healthcare businessthesedaysgenerateshugeamountsofcomplicatedinformationconcerningwithpatients, medical devices, electronic patient records and sickness designation, hospital resources etc. The huge amounts of data's information and knowledge may be a key resource to be processed and analyzed for knowledge extraction that permits support for cost-savings and higher cognitive process. The new challenging and innovation part in healthcare Machine Learning is 'Big Data Analytics' revolution. The health care industry big data means electronic health data sets or flat file data which are unordered, so complex and very large that they suffer many problem and nearly impossible to manage with available tool or normal way and traditional hardware and software tools and techniques. For the health care data / information academic, there is very large amount of data is available to retrace and understanding the pattern and trends inside the data hence big data analytics has potential to improve healthcare services such as care, life and cost reduction.

Keyword: *Healthcare, Machine Learning, Information Processing, recognition, Hidden patterns, Association rule, Classification.*

Date of Submission: 13-05-2020 Date of Acceptance: 26-05-2020

I. INTRODUCTION:

The healthcare market in India is one of the largest and fastest growing industry in the world, it consume nearly 10 per cent of the GDP (Gross Domestic Product) most of the developed or developingnation, heal thcareindustrycontributeamaj or part for a country's economy. The Indian healthcare sector, one of the fastest growing industries, is expected to outstrip at a compound annual growth rate (CAGR) of 17 per cent during 2011-2020 to reach US$ 280 billion. It is expectedtorankamongst the to pthreehealthcaremarketsintermsofincremental growth by 2020. Indian Healthcare sector provides new and existing players with an only one and special opportunity to achieve and perform innovative research and profits. Healthcare in India also awarded as 'polio Free' country by World Health Organization(WHO). AccordingtoaresearchofMcKinsey&Companyinthe nextdecennary,consumerawarenessand demand for better service and facilities will increase and in India healthcare industry will become third largest service sectoremployer. Healthcare data or information is different in scope and very large in content and its segments is so extensive that routine / traditional analytical methods demonstration and little of the possible conclusions. Modern Machine Learning tool and techniques can be applied to this data / information to find out hidden / unknown phase of knowledge which may be of very importance to therapeutic, preventive and commercial aspects of healthcare. The new challenging and innovation part in healthcare Machine Learning is 'Big Data Analytics' revolution. The health care industry big data means electronic health data sets or flat file data which are unordered, so complex and very large that they suffer many problem and nearly impossible to manage with available tool or normal way and traditional hardware and software tools and techniques. For the health care data / information academic, there is very large amount of data is available to retrace and understanding the pattern and trends inside the data hence big data analytics has potential to improve healthcare services such as care, life and cost reduction.

1.1 Some Machine Learning Techniques

1.1.1 Logistic regression

Logistic regression is a special type of regression that is used to explain and predict a binary categorical variable, based on several independent variables that in turn it can be quantitative or qualitative. Due to their characteristics, logistic regression models allow two purposes: to. Quantify the importance of the relationship between each of the covariates or independent variables and the dependent variable, which also implies clarify the existence of interaction and confusion between covariates regarding the dependent variable (that is, knowing the “odds ratio” for each covariate). Classify individuals within the categories (present / absent) of the variable dependent, depending on the probability of belonging to one of them given the presence of certain covariates. There is no doubt that logistic regression is one of the statistical tools with the best ability to analyze data in clinical research and epidemiology, hence its Wide use. The primary objective that this technique solves is to model how it influences the probability of occurrence of an event, usually dichotomous, the presence or not of various factors and their value or level. It can also be used to estimate the probability of occurrence of each of the possibilities of an event with more than two categories.

1.1.2 Decision Trees

A decision tree is a structure in which each internal node denotes a test on one or more attributes, each branch represents an exit of the test and the Leaf nodes represent classes. The main feature of decision trees which are white box models in which you can directly see the frequency of appearance of each attribute. In addition, it allows the expert to know the attribute with greater classification power, that is, the one located in the root node (Mazo, Bedoya, 2010).

1.1.3 Vector Support Machines

The Vector Support Machines (of English Support Vector Machine) were developed in 1995 by Vladimir when he proposes a mathematical model for the resolution of classification and regression problems which they called MSV Model. They are based on the theory of statistical learning that allow solving problems of classification and regression efficiently (Cuba, 2010), (Dexheimera et al. (2007), (Rengifo, Juménez, 2010) The success of vector support machines lies in three key advantages: They have a solid mathematical foundation.

a) They are based on the concept of minimizing structural risk that is, minimizing the probability of a misclassification on new examples, particularly important when you have little training data.

b) Powerful tools and algorithms are available to find the solution so Fast and efficient Vector support machines, unlike neural networks, abstract the problem from an attribute space to a characteristic pattern space with larger dimension, so that they can be separated by a hyper plane.

Thus, through an appropriate nonlinear mapping function, which increases the dimension appropriately, is possible to separate samples that belong to two different categories by hyperplane (Maldonado, Weber, 2012)

1.1.4 Linearly Separable Classification:

Within the data sets that we cover with vector support machines, we find those that are linearly separable, and in them, minimize the function of Costs is very simple. The starting hypothesis is that the classes are linearly separable and therefore there are infinite hyper planes that separate the samples of a class, of the other. The points of space that fall within each of these hyper planes are those that satisfy the following expression (Reginfo, 2010).

Where w Y x Є R, being the dimension of the entrance space.

The resolution for this case would be to assume that there is a set of n separable data linearly(1 , 1), (2 , 2),…, (,) where 1 ∈ e 1 ∈ ^ {−1,1}. It will happen, according to the side in which they are with respect to the hyper plane the following:

There are cases of linearly separable data in which noise may exist due to errors in the measure of the data or by the presence of some atypical or extreme data. In sayings cases it is not convenient for the vector support machine to fully conform to the data. This point should not be considered to find the decision boundary since it could alter the desired results would lead us to incorrect classifications.

1.1.5 Linear Classification Not Separable

When the data is not linearly separable, there is the possibility of transforming the data to a larger space using a function, where you will find a hyper plane that can separate them. The decision frontier resulting in the input space will no longer be linear and will be given by another type of function that can be polynomial of degree superior to 1, Gaussian, sigmoid, among others, these functions are known as core or "kernel" functions. Samples once projected, can be used as a new training set, so it will look for a linear border in space, said border another in the starting space whose form will depend on the projection function used and the samples of the training set As you can see the data in the formulation of the machine Vector support only appear as product between samples , In this case, only the Kernel function is needed, such that Having this function you can apply the training algorithm of the Machine vector support without explicitly knowing what the transformation is or even the space The only necessary condition is that the kernel used is correctly definite.

1.2 **Machine Learning In Field Of HealthCare:**

Machine Learning: Machine Learning explores the hidden patterns or information from data warehouse. This knowledge 4i.e. information extracted from huge dataset is present in human understandable form. Therefore Machine Learning is also Known as KDD (Knowledge Discovery).Machine Learning tools and techniques helps to predict future trends and behavior of a business which helps in making proactive and knowledge driven decisions. Machine Learning techniques and tools can answer the complex business questions which were difficult and time consuming to resolve.The foundation of Machine Learning techniques are the result of a long process of analytical research and come up with a product development.

II. REVIEW OF LITERATURE

In 2012 ShwetaKharya [1] discussed on various Machine Learning approaches that are used for breast cancer diagnosis. This paper summarizes the various technical articles on breast cancer diagnosis which is one of the leading cancers for women in India. It is the third most cause of cancer death in women

In 2012 Patil, D.D.[2] presents an innovative wireless sensor network which helps to provide onlinehealthpredictionbymonitoringrealtimevitalbodysignals.Forimplementationandresult they applied clustering algorithm (Graph theoretic, K-means) on patient historical data. They use the comparative analysis on vital signals originate by the clustering algorithms which adds extra dimension to risk alert which helps doctor to diagnose moreaccurately

In 2012 Carel,Rafael and et.al [3] construct a predictive model for asthma drug utilization by applyingmethodsofKDDintimeseriesdatabasesto historicalasthmadrugdataset.Theyapplied clusteringanddecisiontreealgorithmonregionalpatients'database.Theresultsshows274asthma patients receives 9,319 prescriptions and classification shows that the corticosteroids medications is the main predictive factoring themodel.

In2012Reyes,A.J.Oandet.al[4]aimstodevelopanintegralsystemforanalysisofprimaryhealth care by using clustering technique(partitioned algorithms).The results provide complete analysis of clinical information of the patients. They develop the solution by using java 1.6 as language Eclipse 3.4 as Integrated and JBoss 4.2 asserver.

In 2011 Tan Zhongbing proposed a new clustering method based on gradient descent and genetic algorithm. And he concludes that gradient descent method can be used at the end of the genetic algorithm which is based on the clustering methods to get global optimal values and these values are better than K-means clustering methods.

In1999Goil,S.andet.al[5]addressthescalabilityanalysisofmultidimensionalsystemforOLAP, it also describe the integration of Machine Learningwith OLAP framework. It results with high dimensional data set on a distributed memory parallelmachine

In 2013 Balasubramanian, T.[6] explores Machine Learning techniques in health care, he discussed the risk factors associated with the high level of fluoride contains in water by using clustering algorithm and discover meaningful hidden patterns out of it which is helpful in decision making for socio-economic real world health hazard

In2013AdaandRajneetKaur[7]usedessentialtechniquetoperformmedicalimagemining,data processing, feature extraction, lung field segmentation, classification by using neural network techniques. They divide the digital X-ray chest films into two categories: normal and abnormal. TheymainlyfocusedonfeatureselectionandSVMswhichweretrainedwithdifferentparameters. They perform comparativeresult.

In2013KawsarAhmedandet.al[9]theycollected400cancerand non-cancerpatients’datafrom differentdiagnosticcentersandtheyappliedk-meansclusteringalgorithmtoidentifyrelevantand non-relevantdata,thentheyapplieddecisiontreeandanApriori Tidalgorithmtodiscoverfrequent patterns.Andfinallytheydeveloplungcancerpredictionsystemwhichisveryhelpfulindetection of a person’s predisposition for lungcancer

In 2013 V.Krishnaiah and et.al [10] they examine the use of classification based Machine Learning techniques such as Artificial Neural Network Rule based, Naïve Bayes and Decision tree on healthcare data. They work on generic lung cancer symptoms such as Wheezing, age, Pain in shoulder, arm, sex, Shortness of breath ,chest, it can predict the likelihood of patients getting a lungcancerdisease. Theyproposedamodelforearly detection and diagnosis of the disease which will help experts in saving the life ofpatients.In2008AbdelghaniB.andet.al[8]presentan analysis surveyofthepredictionofsurvivalrate of breast cancer patient using Machine Learning algorithm, they use the 'SEER Public-Use Data'. They mainly focused on the Naïve Bayes, the back-propagated neural network and the C4.5 decision tree algorithms. And they conclude that C4.5 algorithm has bunch better performance as compare to Naïve Bayes and back-propagated neuralnetwork.

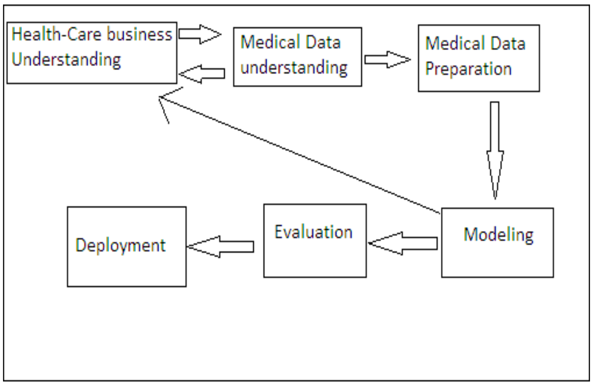

Health-Care business Understanding
Medical Data understanding
Medical Data Preparation
Modeling
Evaluation
Deployment

***Fig 1:** Basic concept of Machine Learning*

III. PROPOSED WORK

3.1 Steps involved in designing the research work

Initially Data is extracted, transform and load from the transaction database into Data-Warehouse as shown in fig2. After extracting data into electronic machine following tasks will be performed.

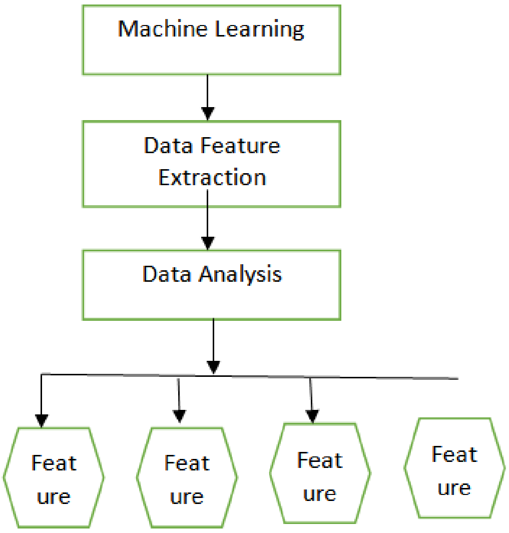


Fig 2: Proposed Algorithm

Health-Care Business Understanding:
Health-Care Business Understanding provide by study various type of research paper related to health care and by understanding health-care data.

Medical-Data Understanding:
This phase starts with collection of health care data from Laboratories, Operation Theater, Blood Bank, drug store, Therapy Modules etc., and also focuses on understanding of patient's data to discover knowledge out of it to generate HMIS repots.

Medical-Data Preparation:
This phase constructs final data set to feed into modeling tools and it is iterative process. Here various database artifacts such as attribute, table, records are selected as well as transformed and cleaned for modeling tools.

Modeling:
This phase apply various modeling techniques such as Naive Bays, Artificial neural network, decision tree, time series algorithm, clustering algorithm, sequence clustering algorithm etc. to generate optimal values.

Evaluation:
In this stage thorough evaluation and reviewing the model to check whether applied algorithms discovers proper hidden pattern. And this stage also checks for fast accessing on mined data.

Deployment:
In this phase most of the time customer carry out the deployment wizard with the help of analyst to generate HMIS reports. For generating useful knowledge out of data i.e. reports we can repeat Machine Learning process.

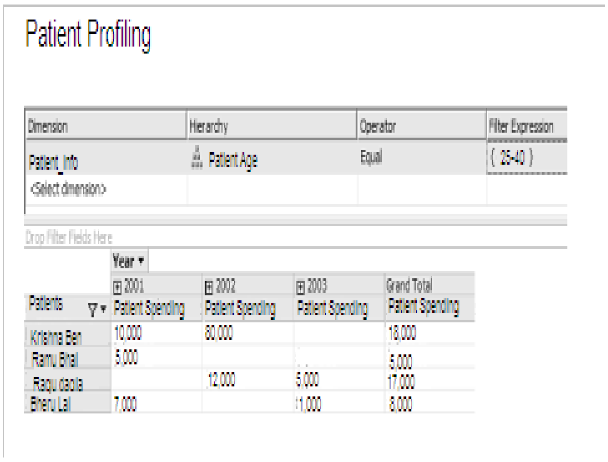
3.1 Details of experiment Carriedout:

The experiments carried out to achieve the above mentioned objectives are described in various stages. From data analysis to integration to reporting, following experiments are performed.

Stage1. Environmental setup:
The above mentioned hardware and software requirements are fulfilled and all the software's are downloaded. To establish the environment for experiment, below mentioned steps are followed. Windows 7 OS is installed on the system.
Visual studio is installed on the machine.
Installation of sql server 2008 r2:We have installed sql server 2008 r2 which include integration, analytical and reporting services,

Stage 2.Create OLAP cubes:
The data present in OLAP system is present in a multidimensional structure and it is created with the help of facts and dimension. Dimension have granularity of viewing data. Therefore, day □ monthyear is a TIME dimension hierarchy which specify various aggregation level. Other dimension

which we use is in/out patient data: Patient age, cancer diseases group, sex, diagnosis.



Patient Profiling

Dimension	Hierarchy	Operator	Filter Expression
Patient_Info	Patient Age	Equal	(25-40)
<Select dimension>			

Drop Filter Fields Here

Patients	Year 2001 Patient Spending	2002 Patient Spending	2003 Patient Spending	Grand Total Patient Spending
Krishna Ben	10,000	80,000		18,000
Ramu Bhai	5,000			5,000
Ragu daola		12,000	5,000	17,000
Bheru Lal	7,000		11,000	8,000

Fig 3: OLAP Cube creation

OLAP cube creation with the help of fact and dimensions, but it is difficult to find trends and patterns in large OLAP dimensions therefore we use Machine Learning techniques.

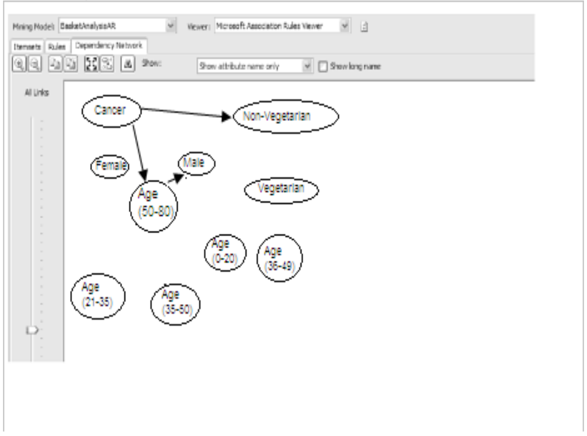
IV. RESULTS AND DISCUSSION

In the previous section, we have described the various experiments performed to achieve the objective. Now, in this section, we will discuss over the results obtained from the above experiments.

a) To determine cancer side and the morphology pattern among variouspatients.

b) To explores the Machine Learning applications & challenges in healthcare.

We choose Integration, analytical, reporting services. We are able to provide dependency network on cancer patients. In our research, we explore Machine Learning application & challenges in health care.



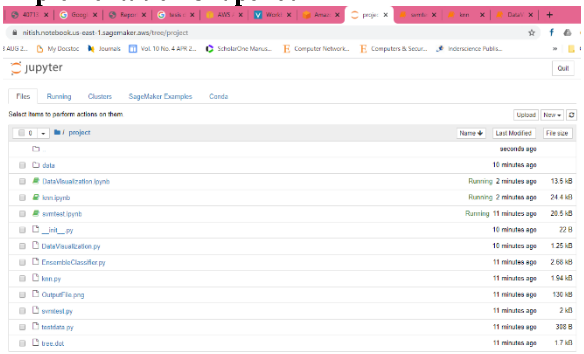
Mining Model: BasketAnalysisAR
Viewer: Microsoft Association Rules Viewer
Show: Show attribute name only
□ Show long name

Fig 4: Market Basket Analysis on patient data.

Market Basket Analysis:

It shows the dependency in attributes which are strongest correlated. Here we find very interesting patterns out of data. Initially we defined the attributes for patient's analysis such as age, gender, veg., non-veg. From the above analysis we come to know patient in the age group of 35-40, male and non-vegetarians are more in the cancer treatment category.Theaboveanalyticalgraphiscalculatedon8yearsofpatientdata.Itshowsveryinterestinggraphs overthecancerpatients.Mostofthecancerpatientspresentbetweenthe40-70yearsofagegroup and most of them are count of female patient is more than malepatients.

Implementation Snapshot



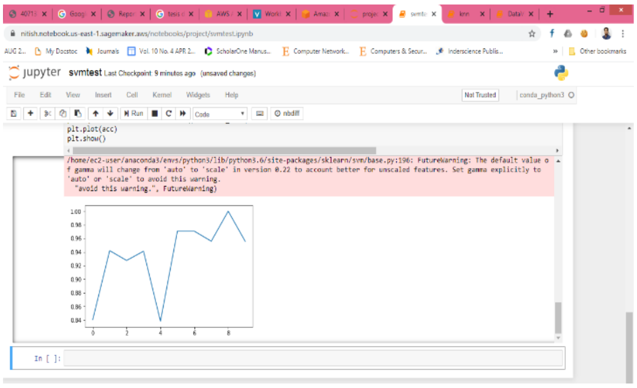
nitish.notebook.us-east-1.sagemaker.aws/tree/project

Select items to perform actions on them

/ project

Name		Last Modified	File size
..		seconds ago	
data		10 minutes ago	
DataVisualization.ipynb	Running	2 minutes ago	13.5 kB
knn.ipynb	Running	2 minutes ago	24.4 kB
svmtest.ipynb	Running	11 minutes ago	20.5 kB
__init__.py		10 minutes ago	22 B
DataVisualization.py		10 minutes ago	1.25 kB
EnsembleClassifier.py		11 minutes ago	2.68 kB
knn.py		11 minutes ago	1.94 kB
OutputFile.png		11 minutes ago	130 kB
svmtest.py		11 minutes ago	2 kB
testdata.py		11 minutes ago	308 B
tree.dot		11 minutes ago	1.7 kB

Fig 5: Machine Learning File Structure



nitish.notebook.us-east-1.sagemaker.aws/notebooks/project/svmtest.ipynb

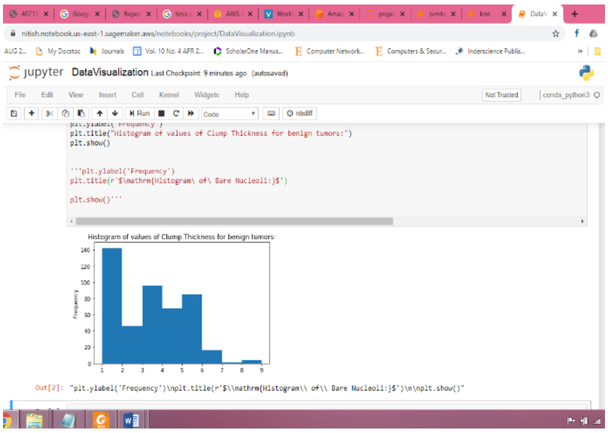
jupyter svmtest Last Checkpoint: 9 minutes ago (unsaved changes)

```
plt.plot(acc)
plt.show()
```

```
/home/ec2-user/anaconda3/envs/python3/lib/python3.6/site-packages/sklearn/svm/base.py:196: FutureWarning: The default value o
f gamma will change from 'auto' to 'scale' in version 0.22 to account better for unscaled features. Set gamma explicitly to
'auto' or 'scale' to avoid this warning.
  "avoid this warning.", FutureWarning)
```

```
In [ ]:
```

***Fig 6:** Decision Tree Graph Analysis*



nitish.notebook.us-east-1.sagemaker.aws/notebooks/project/DataVisualization.ipynb
jupyter DataVisualization Last Checkpoint: 9 minutes ago (autosaved)
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
conda_python3
Run
Code
nbdiff
Histogram of values of Clump Thickness for benign tumors:
Frequency
0
20
40
60
80
100
120
140
1
2
3
4
5
6
7
8
9

```
plt.ylabel('Frequency')
plt.title("Histogram of values of Clump Thickness for benign tumors:")
plt.show()


'''plt.ylabel('Frequency')
plt.title(r'$\mathrm{Histogram\ of\ Bare Nucleoli:}$')

plt.show()'''
```

```
Out[2]: "plt.ylabel('Frequency')\nplt.title(r'$\\mathrm{Histogram\\ of\\ Bare Nucleoli:}$')\n\nplt.show()"
```

Fig 7: SVM Accuracy Analysis

V. CONCLUSION

This research work provide association of OLAP and Machine Learning (OLAP Mining) for analysis. For future work, we have issues to address. First issue is to provide association of OLAP and predictiveanalyticsusingMachine Learningalgorithm.Secondistoprovideafacility fromwhichdoctor canquerytodatacubeonaspectsofbusinessproblemandtranslatethisproblemintoMDX(Multi dimension Expression) queriesautomatically.

REFERENCES

[1]. Shweta Kharya Discussed, "Using Machine Learning Techniques For Diagnosis And Prognosis Of Cancer Disease", International Journal Of Computer Science, Engineering And Information Technology (Ijcseit), 2(2), (2012)

[2]. Patil, D.D., "Dynamic Machine Learning Approach ToWmrhm", 7th Ieee Conference On Industrial Electronics And Applications (Iciea), Singapore, 978, (2012)

[3]. Carel, Rafael And Barak, Dotan, "Utilization Of Data-Mining Techniques For Evaluation Of Patterns Of Asthma Drugs Use By Ambulatory Patients In A Large Health Maintenance Organization", Seventh Ieee International Conference On Machine Learning Workshops, 2007. Icdm Workshops, Omaha, Ne, Usa, 169 – 174, (2007)

[4]. Reyes,A.J.O.;Garcia,A.O.;Mue,Y.L. "System For Processing And Analysis Of Information Using Clustering Technique", Latin America Transactions, Ieee (RevistaIeee America Latina) 12(2), 364 – 371, (2014).

[5]. Goil, S. And Choudhary, A., "A Parallel Scalable Infrastructure ForOlap And Machine Learning", International Symposium Proceedings On Database Engineering And Applications, Ideas '99, Montreal, Que, 178 – 186, (1999)

[6]. Balasubramanian T. , "An Analysis On The Impact Of Fluoride In Human Health (Dental) Using Clustering Machine Learning Technique", Ieee International Conference On Pattern Recognition, Informatics And Medical Engineering (Prime)", Salem, Tamilnadu, 370 – 375, (2012)

[7]. Ada AndRajneet Kaur, "A Study Of Detection Of Lung Cancer Using Machine Learning Classification Techniques", International Journal Of Advanced Research In Computer Science And Software Engineering, 3(3), (2013)

[8]. AbdelghaniBellaachia, ErhanGuven, "Predicting Breast Cancer Survivability Using Machine Learning Techniques", Department Of Computer Science The George Washington University Washington Dc Available At Eguven@Gwu.Edu

[9]. Kawsar Ahmed, Abdullah-Al-Emran, Tasnubajesmin, Roushney Fatima Mukti, Mdzamilurrahman, Farzana Ahmed, " Early Detection Of Lung Cancer Risk Using Machine Learning", Asian Pacific Journal Of Cancer Prevention, Vol.14, (2013)

[10]. V.Krishnaiah, Dr.G.Narsimha, Dr.N.Subhash Chandra "Diagnosis Of Lung Cancer Prediction System Using Machine Learning Classification Techniques", International Journal Of Computer Science And Information Technologies,Vol. 4 (1), 39 – 45 (2013)

[11]. Ruben D. And Canlas Jr. "Machine Learning In Healthcare: Current Applications And Issues", (2009)

[12]. Bellazzi, R. " Methods And Tools For Mining Multivariate Temporal Data In Clinical And Biomedical Applications" Ieee, 5629 – 5632, (2009)

[13]. P.Santhi, V.Muralibhaskaran, "Performance Of Clustering Algorithms In Healthcare Database", International Journal For Advances In Computer Science, 2(1), (2010)

[14]. Sung Ho Ha AndSeonghyeonjoo, "A Hybrid Machine Learning Method For The Medical Classification Of Chest Pain", World Academy Of Science, Engineering And Technology, 4, (2010)

[15]. Hemalatha M. And Megala S., "Mining Techniques In Health Care: A Survey Of Immunization", Little Lion Scientific R&D,

25(2), 63-70, (2012)

[16]. Al Iqbal, R. “Hybrid Clinical Decision Support System: An Automated Diagnostic System For Rural Bangladesh”, Ieee, Dhaka, 76 – 81, (2012) .

[17]. F. Xylogiannopoulos, “Developing An Efficient Health Clinical Application: Iiop Distributed Objects Framework”, International Conference On Advances In Social Networks Analysis And Mining, Istanbul Turkey, (2012)

Ravi Kumar Jangid, et. al. “Design and Analysis of Improvisation in Cancer detection technique in Patients using Machine learning.” *International Journal of Engineering Research and Applications (IJERA*), vol.10 (05), 2020, pp 42-48.