Asian Journal of Information Technology 15 (14): 2337-2342, 2016
ISSN: 1682-3915
© Medwell Journals, 2016

A Novel Method for the Identification of Child Blood Cancer Using Data Mining Techniques

[1]M. Sangeetha, [2]N. K. Karthikeyan and [3]P. Tamijeselvy
[1]Department of IT, Sri Krishna College of Technology, Kovaipudur, 641 042 Coimbatore,
[2]Department of IT, Sri Krishna College of Engineering and Technology,
Kuniamuthur, 641 008 Coimbatore,
[3]Department of CSE, Sri Krishna College of Technology,
Kovaipudur, 641 042 Coimbatore, Tamil Nadu, India

Abstract: Generally the anatomy of human body is that the cells grow newly and after some period of time they became old and die and the new cells forms. This process continues till the breathe ends. But instead of this normal process when the body cells grows and spreads into surrounding tissues then it results in cancer. Sometimes these extra cells form tumours. Difference between normal cell and cancer cell is that the normal cells stops its growth at a certain period of time whereas cancer cells fail to stop the growth, i.e., they grows abnormally. Cancer cells will spoil the entire regular functioning of the body. Generally cancer centers treat patients upto age 12 under child cancer. Child cancer may be due to problem in DNA or due to deficiency in vitamin C or due to pesticides present in fruits or vegetables or milk. Children identified with cancer may have extra copy of chromosome. Data set of children suffering from cancer are retrieved. The data set may have missing values. The missing values are identified through co-clustering Bayesian principal component analysis method. Then the occurrence percentage of cancer is predicted through Naive Bayesian method.

Key words: Bayesian Principal Component Analysis (BPCA), co-clustering, computation of DNA missing values, retrieved, occurrence

INTRODUCTION

Cancer is the result of growth of abnormal cells. Cancer is also referred to as Carcinoma or Malignant tumour. Most of the cancer are caused by the viruses. The viruses may be DNA virus or RNA virus. Those who are suffering from cancer will undergo the usual treatment of chemotherapy and radio therapy that may gradually decrease the immune system of human beings day by day. In fact childhood cancer can be curable if it is diagnosed correctly in the earlier stage. In hospitals a biopsy test will be carried out in which some portion of tissues are removed from the tumour and carried over to lab test. The types of cancer that occur frequently in children are Leukemia, Brain and Spinal Cord tumor, Brain and Central Nervous System Tumors, Neuroblastoma, Wilms, Tumors, Encometrial, Lymphoma, Rhabdomyosarcoma, Retinoblastoma, Bone cancer, Cervial cancer, Esophagus cancer, Gall bladder cancer, etc. Generally each and every hospital will maintain a database of patients records. In this work we have received those data set of childhood cancer from the hospitals. Usually database may have some missing values. Those missing values are computed by co-clustering Bayesian principal component analysis method. After finding the missing values the same dataset is used for predicting the existence percentage of cancer through Naive Bayesian Method.

Literature review: Data mining has the great ability to deal with medical care systems or health protection systems. This plays an important role in the cost reduction. This study specifies about various classficiation and clustering techniques that are used in the medical field. It suggests that KNN is the simple classifier, decision tree is the popular approach that helps to choose the best alternative method, SVM gives the exact result, neural network was considered as the best classification algorithm, it has the primary elements as neurons but this approach is expensive. For huge data sets Bayesian methods works fastly and gives the exact result (Ahmad *et al*., 2015). Medical and health care fields are having enormous amount of information but they are not properly used for decision making. This paper focuses on the data that are not mined and they uses K-means and

Corresponding Author: M. Sangeetha, Department of IT, Sri Krishna College of Technology, Kovaipudur, 641 0042 Coimbatore, Tamil Nadu, India

Naïve Bayesian approach for predicting the heart attack (Akash Jarad, 2015). In today's environment commonly spreading disease is swine flu. (Borkar and Deshmukh, 2015) This study predicts the swine flu through Naive Bayesian approach through specified parameters. As the approach is easier to interpret Naïve Bayesian is selected as the choice. In present scenario, lot of informations are hidden in the database (Tewary, 2015). The data can be extracted from the database through various data mining techniques. If the extraction process is accurate then the hospitals will get benefit from the proper decision making. Neural network is best for solving the problems of data mining because of its properties of good robustness, high degree of fault tolerance, distributed storage, parallel processing. This study provides a literature review of various classification techniques. Naive Bayesian is the simple classifier technique that uses Bayesian theorem. SVM is popularly known for fast and accurate results. K-nearest neighbour is mainly used in pattern recognition and text categorization. Neural network produces a relationship between input and output (Gomathi and Narayani, 2014). Lupus is a disease that can affect any part of the body. It is a chronic disease which can be due to environment or genetic factors. In this study, ID3 algorithm is used to predict the disease in the early stage. Khaleel *et al*. (2013) data Mining techniques are used to enhance the quality of prognosis of disease. Various approaches of data extraction are used to find the frequent patterns. Kumar and Padmapriya (2012) data Extraction is the process of finding hidden information from the large amount of data set. Neural network and ID3 algorithm are utilized to predict the occurrence of any common disease. Data Mining Techniques are utilized to analyze the large amount of information and present to the user in a useful manner. In this paper, as CART is the best classifier, it is used to detect the breast cancer.

Types of childhood cancer: Following are the types of cancer that occur most frequently in children:

- Leukemia: cancers of the bone marrow and blood are referred to as leukemia
- Brain and central nervous system tumors: these are the second most probable cancer in children. These begins in the lesser portion of brain
- Neuroblastoma: cancer that are found in the fetus are referred to as neuroblastoma
- Wilms tumors: These are also referred to as nephro blastoma. If the cancer is found in any one of the kidney or both of the kidneys then they are referred to as nephroblastoma
- Lymphoma: cancer that occur in particular cells of immune system are referred to as lymphoma
- Rhabdomyosarcoma: Cancer that starts in head, neck, belly, arm or leg are referred to as Rhabdomyosarcoma
- Retinoblastoma: cancer that occur in eye are referred to as Rectinoblastoma
- Bone cancers: cancer that occur in bones are referred to as bone cancer. They start in the hip bones or in the ribs or shoulders
- Cervical cancer: cancer that is found in the lower part of womb are referred to as cervical cancer. Womb are also known as uterus
- Colon/rectum cancer: cancer that begins in the colon or rectum are referred to as colorectal cancer
- Encometrial cancer: cancer that occur in the inner side of uterus is referred to as Endometrial cance
- Esophagus cancer: cancer that occur in the esophagus are referred to as Esophagus cancer
- Gall bladder cancer: cancer that starts in Gall Bladder are referred to as Gall Bladder cancer
- Gastric cancer: gastric cancer is also known an stomach cancer. Cancer that occurs in stomach are referred to as gastric cancer

Motivation: Hospitals are maintaining patients details in a databank. When we shift the data from one place to another place, information may be lost. These missed data are computed through Co-Clustering Bayesian Principal Component Analysis. Then these data set are used to forecast the existence percentage of cancer through Naïve Bayesian Method.

Proposed method: The proposed method uses Bi-BPCA strategy for computing the missing information and Naïve Bayesian Approach for estimating the occurrence percentage of cancer. Cancer is one of the fatal diseases throughout all the nations. This scheme focus on guiding the medical people to identify the patients at the preliminary stage so that the death rate will be reduced. BPCA stands for Bayesian Principal Component analysis. Bayesian approach is an important method in statistics. Statistics is nothing but planning, analysing and presentation of data.

In statistics, missing data means a variable is not possessing any value. Data created from any experiments or data collected from hospitals may have missing values. These missing values have to be evaluated. BPCA is the popular method used for finding the missing values but its performance is not satisfied on similar type of data set. Missing values can be estimated by four methods:

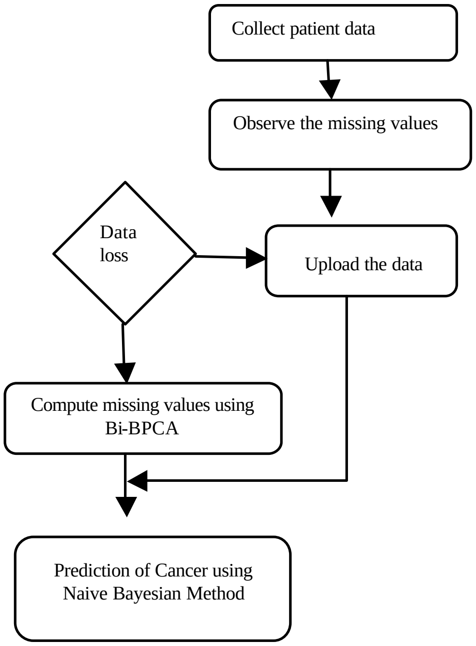

Collect patient data
Observe the missing values
Data loss
Upload the data
Compute missing values using Bi-BPCA
Prediction of Cancer using Naive Bayesian Method

Fig. 1: Overall working of system

- Local approach
- Global approach
- Hybrid approach
- Knowledge approach

Two widely used Global approaches are Singular Value Decomposition and Bayesian Principal Component Analysis. Bayesian approach is commonly used in various fields such as face identification and decision making. Bayesian principal component analysis considers the T-dimensional micro array expression vector z can be represented as a linear combination of L which is given as:

$$Z = \sum_{i=1}^{L} Y_i w_i + \varepsilon$$

Generally, data are in the form of matrices. A row and a column in a matrix represents a gene and an experimental condition. As a first step, the incomplete matrix is converted into a complete matrix by BPCA. Missed values in every row are filled with the average values of every row. Then, Bi-clustering is applied. Bi-clustering is also known as Co-clustering or Two Mode Clustering or Block Clustering. Bi-clustering is a data mining approach that supports parallel grouping of rows and columns of a matrix. Then the same data set is used to predict the percentage of cancer through Naive Bayesian Method. Figure 1 shows the overall working of the system.

Advantages of naïve bayesian method: Among various data extracting methods, naïve bayesian have various benefits:

- Performance is goodp
- Works well for huge amount of data
- Gives fast result
- Helps to get exact result
- Various features does not depend each other
- Computation process is easy

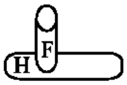
MATERIALS AND METHODS

Bi-BPCA algorithm: It uses ILLS technique from Local method to evaluate the missing information and finds the local figure and also evaluate the missing information by aligning it in the bi-cluster BPCA.The bi-cluster based BPCA scheme is used to deal with the neighbour similarity figures of matrix where the most mutually rows and columns with the missing information are chosen to evaluate the missing value. This method eliminates the disadvantages of BPCA and decreases the estimation error. Decision tree learning is a strategy popularly used in data mining. Bi-BPCA is one of the schemes of decision tree algorithm and the objective is to create a model that forecasts the value of a target variable depending on various input variables (Fig. 2).

Naive Bayesian algorithm: Naive Bayesian is considered as the most efficient data mining classification algorithm on behalf of the following reasons: Calculation process is easier, performance is so good, works well for huge amount of data, features are independent with each other, works fastly and gives the accurate result. For events A and B, provided that P(B)≠0,

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

P(A|B) = the probability of A being true, given that we know that B is true:



Where:
H = Excess number of WBC
F = Occurrence of cancer

If children have abnormal count of WBC then children will have cancer.

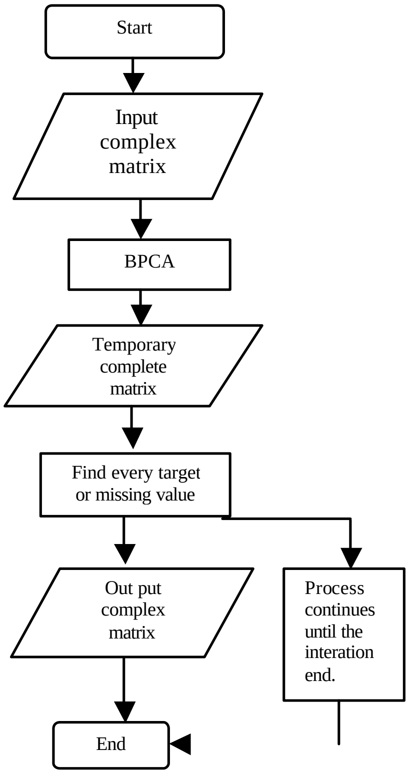

Start
Input complex matrix
BPCA
Temporary complete matrix
Find every target or missing value
Out put complex matrix
Process continues until the interation end.
End

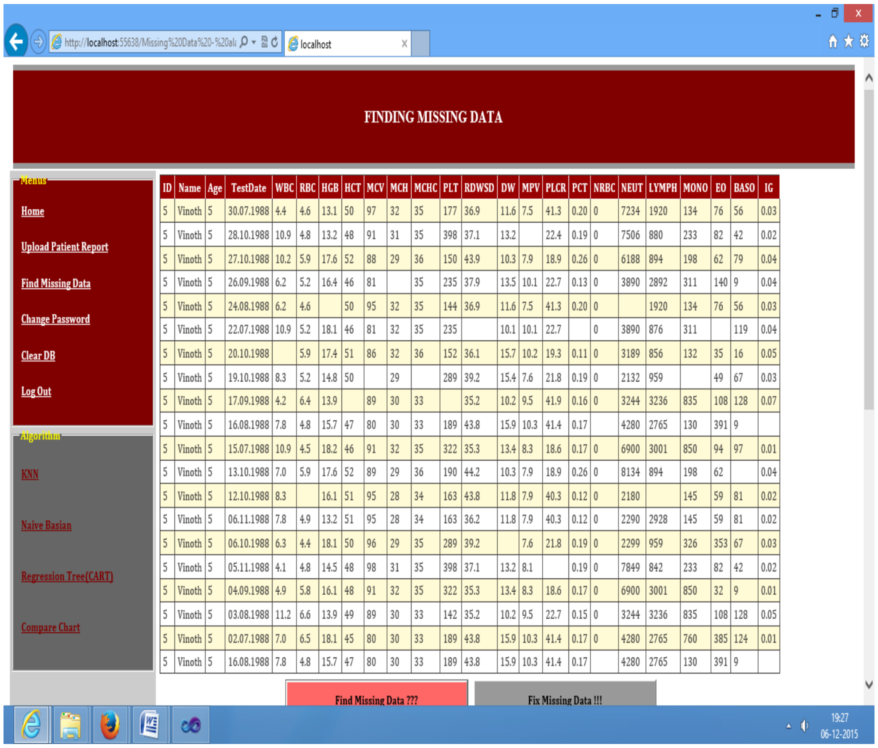
Fig. 2: Finding missing values

RESULTS AND DISSCUSSION

Finding the missing information: Generally hospitals maintain patients details in a data bank. Those details has to be properly mined for proper decision making. If there is good decision making, then proper treatment can be made. Here we have collected the details of children those who are suffering from blood cancer. Several parameters are tested to check the existence of blood cancer. Figure 3 illustrates how the patient details are having missing values. It also indicates the twenty parameters that a patient has to undergo during the examination process.

Fixing missed value using Bi-BPCA: As a first step the incomplete matrix is converted into a complete matrix by BPCA. Missed values in every row are filled with the average values of every row. Then Bi-clustering is applied. Bi-Clustering is also known as co-clustering or Two Mode Clustering or Block Clustering. Bi-clustering is a data mining approach that supports parallel grouping of rows and columns of a matrix (Fig. 4).

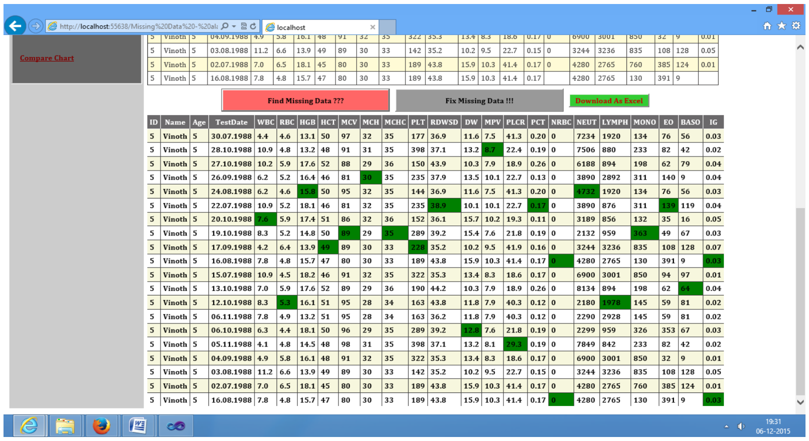
Predicting occurrence percentage of cancer using naïve bayesian approach: Even though, there are lot of classification techniques are available for diagonising diseases, here we have chosen Naive Bayesian approach



FINDING MISSING DATA

ID	Name	Age	TestDate	WBC	RBC	HGB	HCT	MCV	MCH	MCHC	PLT	RDWSD	DW	MPV	PLCR	PCT	NRBC	NEUT	LYMPH	MONO	EO	BASO	IG
5	Vinoth	5	30.07.1988	4.4	4.6	13.1	50	97	32	35	177	36.9	11.6	7.5	41.3	0.20	0	7234	1920	134	76	56	0.03
5	Vinoth	5	28.10.1988	10.9	4.8	13.2	48	91	31	35	398	37.1	13.2		22.4	0.19	0	7506	880	233	82	42	0.02
5	Vinoth	5	27.10.1988	10.2	5.9	17.6	52	88	29	36	150	43.9	10.3	7.9	18.9	0.26	0	6188	894	198	62	79	0.04
5	Vinoth	5	26.09.1988	6.2	5.2	16.4	46	81		35	235	37.9	13.5	10.1	22.7	0.13	0	3890	2892	311	140	9	0.04
5	Vinoth	5	24.08.1988	6.2	4.6		50	95	32	35	144	36.9	11.6	7.5	41.3	0.20	0		1920	134	76	56	0.03
5	Vinoth	5	22.07.1988	10.9	5.2	18.1	46	81	32	35	235		10.1	10.1	22.7		0	3890	876	311		119	0.04
5	Vinoth	5	20.10.1988		5.9	17.4	51	86	32	36	152	36.1	15.7	10.2	19.3	0.11	0	3189	856	132	35	16	0.05
5	Vinoth	5	19.10.1988	8.3	5.2	14.8	50		29		289	39.2	15.4	7.6	21.8	0.19	0	2132	959		49	67	0.03
5	Vinoth	5	17.09.1988	4.2	6.4	13.9		89	30	33		35.2	10.2	9.5	41.9	0.16	0	3244	3236	835	108	128	0.07
5	Vinoth	5	16.08.1988	7.8	4.8	15.7	47	80	30	33	189	43.8	15.9	10.3	41.4	0.17		4280	2765	130	391	9	
5	Vinoth	5	15.07.1988	10.9	4.5	18.2	46	91	32	35	322	35.3	13.4	8.3	18.6	0.17	0	6900	3001	850	94	97	0.01
5	Vinoth	5	13.10.1988	7.0	5.9	17.6	52	89	29	36	190	44.2	10.3	7.9	18.9	0.26	0	8134	894	198	62		0.04
5	Vinoth	5	12.10.1988	8.3		16.1	51	95	28	34	163	43.8	11.8	7.9	40.3	0.12	0	2180		145	59	81	0.02
5	Vinoth	5	06.11.1988	7.8	4.9	13.2	51	95	28	34	163	36.2	11.8	7.9	40.3	0.12	0	2290	2928	145	59	81	0.02
5	Vinoth	5	06.10.1988	6.3	4.4	18.1	50	96	29	35	289	39.2		7.6	21.8	0.19	0	2299	959	326	353	67	0.03
5	Vinoth	5	05.11.1988	4.1	4.8	14.5	48	98	31	35	398	37.1	13.2	8.1		0.19	0	7849	842	233	82	42	0.02
5	Vinoth	5	04.09.1988	4.9	5.8	16.1	48	91	32	35	322	35.3	13.4	8.3	18.6	0.17	0	6900	3001	850	32	9	0.01
5	Vinoth	5	03.08.1988	11.2	6.6	13.9	49	89	30	33	142	35.2	10.2	9.5	22.7	0.15	0	3244	3236	835	108	128	0.05
5	Vinoth	5	02.07.1988	7.0	6.5	18.1	45	80	30	33	189	43.8	15.9	10.3	41.4	0.17	0	4280	2765	760	385	124	0.01
5	Vinoth	5	16.08.1988	7.8	4.8	15.7	47	80	30	33	189	43.8	15.9	10.3	41.4	0.17		4280	2765	130	391	9	

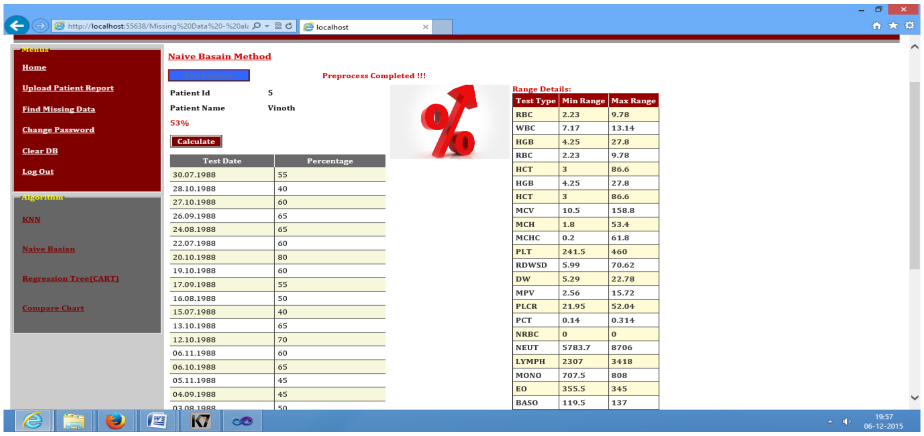
Fig.3. Twenty attribute test for a patient



5	Vinoth	5	04.09.1988	4.9	5.8	16.1	48	91	32	35	322	35.3	13.4	8.3	18.6	0.17	0	6900	3001	850	32	9	0.01
5	Vinoth	5	03.08.1988	11.2	6.6	13.9	49	89	30	33	142	35.2	10.2	9.5	22.7	0.15	0	3244	3236	835	108	128	0.05
5	Vinoth	5	02.07.1988	7.0	6.5	18.1	45	80	30	33	189	43.8	15.9	10.3	41.4	0.17	0	4280	2765	760	385	124	0.01
5	Vinoth	5	16.08.1988	7.8	4.8	15.7	47	80	30	33	189	43.8	15.9	10.3	41.4	0.17		4280	2765	130	391	9	

ID	Name	Age	TestDate	WBC	RBC	HGB	HCT	MCV	MCH	MCHC	PLT	RDWSD	DW	MPV	PLCR	PCT	NRBC	NEUT	LYMPH	MONO	EO	BASO	IG
5	Vinoth	5	30.07.1988	4.4	4.6	13.1	50	97	32	35	177	36.9	11.6	7.5	41.3	0.20	0	7234	1920	134	76	56	0.03
5	Vinoth	5	28.10.1988	10.9	4.8	13.2	48	91	31	35	398	37.1	13.2	8.7	22.4	0.19	0	7506	880	233	82	42	0.02
5	Vinoth	5	27.10.1988	10.2	5.9	17.6	52	88	29	36	150	43.9	10.3	7.9	18.9	0.26	0	6188	894	198	62	79	0.04
5	Vinoth	5	26.09.1988	6.2	5.2	16.4	46	81	30	35	235	37.9	13.5	10.1	22.7	0.13	0	3890	2892	311	140	9	0.04
5	Vinoth	5	24.08.1988	6.2	4.6	15.8	50	95	32	35	144	36.9	11.6	7.5	41.3	0.20	0	4732	1920	134	76	56	0.03
5	Vinoth	5	22.07.1988	10.9	5.2	18.1	46	81	32	35	235	38.9	10.1	10.1	22.7	0.17	0	3890	876	311	139	119	0.04
5	Vinoth	5	20.10.1988	7.6	5.9	17.4	51	86	32	36	152	36.1	15.7	10.2	19.3	0.11	0	3189	856	132	35	16	0.05
5	Vinoth	5	19.10.1988	8.3	5.2	14.8	50	89	29	35	289	39.2	15.4	7.6	21.8	0.19	0	2132	959	363	49	67	0.03
5	Vinoth	5	17.09.1988	4.2	6.4	13.9	49	89	30	33	228	35.2	10.2	9.5	41.9	0.16	0	3244	3236	835	108	128	0.07
5	Vinoth	5	16.08.1988	7.8	4.8	15.7	47	80	30	33	189	43.8	15.9	10.3	41.4	0.17	0	4280	2765	130	391	9	0.03
5	Vinoth	5	15.07.1988	10.9	4.5	18.2	46	91	32	35	322	35.3	13.4	8.3	18.6	0.17	0	6900	3001	850	94	97	0.01
5	Vinoth	5	13.10.1988	7.0	5.9	17.6	52	89	29	36	190	44.2	10.3	7.9	18.9	0.26	0	8134	894	198	62	64	0.04
5	Vinoth	5	12.10.1988	8.3	5.3	16.1	51	95	28	34	163	43.8	11.8	7.9	40.3	0.12	0	2180	1978	145	59	81	0.02
5	Vinoth	5	06.11.1988	7.8	4.9	13.2	51	95	28	34	163	36.2	11.8	7.9	40.3	0.12	0	2290	2928	145	59	81	0.02
5	Vinoth	5	06.10.1988	6.3	4.4	18.1	50	96	29	35	289	39.2	12.8	7.6	21.8	0.19	0	2299	959	326	353	67	0.03
5	Vinoth	5	05.11.1988	4.1	4.8	14.5	48	98	31	35	398	37.1	13.2	8.1	29.3	0.19	0	7849	842	233	82	42	0.02
5	Vinoth	5	04.09.1988	4.9	5.8	16.1	48	91	32	35	322	35.3	13.4	8.3	18.6	0.17	0	6900	3001	850	32	9	0.01
5	Vinoth	5	03.08.1988	11.2	6.6	13.9	49	89	30	33	142	35.2	10.2	9.5	22.7	0.15	0	3244	3236	835	108	128	0.05
5	Vinoth	5	02.07.1988	7.0	6.5	18.1	45	80	30	33	189	43.8	15.9	10.3	41.4	0.17	0	4280	2765	760	385	124	0.01
5	Vinoth	5	16.08.1988	7.8	4.8	15.7	47	80	30	33	189	43.8	15.9	10.3	41.4	0.17	0	4280	2765	130	391	9	0.03

Fig. 4: Finding missing values



Naive Basain Method

Preprocess Completed !!!

Patient Id 5

Patient Name Vinoth

53%

Test Date	Percentage
30.07.1988	55
28.10.1988	40
27.10.1988	60
26.09.1988	65
24.08.1988	65
22.07.1988	60
20.10.1988	80
19.10.1988	60
17.09.1988	55
16.08.1988	50
15.07.1988	40
13.10.1988	65
12.10.1988	70
06.11.1988	60
06.10.1988	65
05.11.1988	45
04.09.1988	45
03.08.1988	50

Range Details:

Test Type	Min Range	Max Range
RBC	2.23	9.78
WBC	7.17	13.14
HGB	4.25	27.8
RBC	2.23	9.78
HCT	3	86.6
HGB	4.25	27.8
HCT	3	86.6
MCV	10.5	158.8
MCH	1.8	53.4
MCHC	0.2	61.8
PLT	241.5	460
RDWSD	5.99	70.62
DW	5.29	22.78
MPV	2.56	15.72
PLCR	21.95	52.04
PCT	0.14	0.314
NRBC	0	0
NEUT	5783.7	8706
LYMPH	2307	3418
MONO	707.5	808
EO	355.5	345
BASO	119.5	137

Fig. 5: Occurrence of cancer

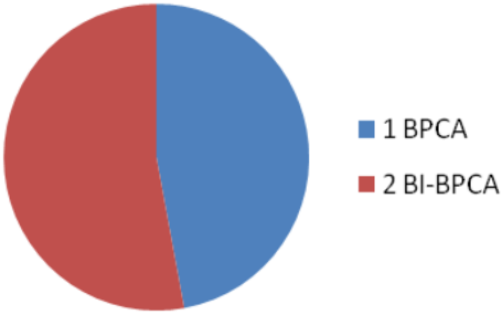


Fig. 6: Comparison of accuracy

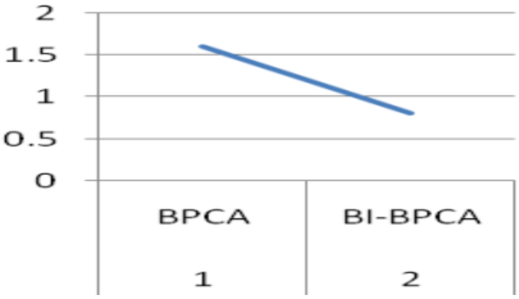


Fig. 7: Comparison of execution time

due to its efficiency. For huge data sets Naive Bayesian method work fastly and accurately. Figure 5 indicates the existence percentage of cancer through Naive Bayesian Method.

Comparison chart: A comparative study of the above two algorithms has revealed that Bi-BPCA is the most effective method for finding the missing values. The comparison chart of BPCA and Bi-BPCA is followed below. Bi-BPCA gives better performance than BPCA by producing accuracy of 92.14% (Table 1). The following chart illustrates that Bi-BPCA provides more accuracy than BPCA shown in Fig. 6.

Bi-BPCA gives better performance than BPCA by producing an accuracy of 92.14%. Table 2 shows the execution time of BPCA and Bi-BPCA. Figure 7 that Bi-BPCA provides less execution time than BPCA shown in Fig. 7.

Table 1: Comparison of algorithms

Algorithms	Accuracy (%)
Based on accuracy	
BPCA	82.00
BI-BPCA	92.14

Table 2: Comparison of algorithms

Algorithms	Execution time (sec)
Based on execution time	
BPCA	1.6
BI-BPCA	0.8

CONCLUSION

Results of BPCA and Bi-BPCA are compared that illustrates Bi-BPCA as an efficient method. Generally Bayesian Principal Component Analysis is used to compute the missing values. The same data set is used to predict the existence percentage of cancer through Naive Bayesain Approach. In future work various classification techniques can be utilized to predict the occurrence percentage of cancer and their performance can be compared.

REFERENCES

Ahmad, P., S. Qamar and S.Q.A. Rizvi, 2015. Techniques of data mining in healthcare: A review. Intl. J. Comput. Appl., Vol. 120.

Borkar, A.R., P.R. Deshmukh, 2015. Naive bayes classifier for prediction of swine flu. Int. J. Adv. Res. Comput. Sci. Software Eng., 5: 120-123.

Gomathi, S . and V. Narayani, 2014. A data mining classification approach to predict systemic lupus erythematosus using I d3 algorithm. Intl. J. Adv. Res. Comput. Sci. Software Eng., 4: 449-453.

Jarad, A., R. Katkar, A.R. Shaikh and A. Salve, 2015. Intelligent heart disease prediction system with mongodb. Intl. J. Emerg. Trends Technol. Comput. Sci., 4: 236-239.

Khaleel, M.A., S.K. Pradham and G.N. Dash, 2013. A survey of data mining techniques on medical data for finding locally frequent diseases. Intl. J. Adv. Res. Comput. Sci. Software Eng., 3: 149-153.

Kumar, L.S. and A. Padmapriya, 2012. ID3 algorithm performance of diagnosis for common disease. Intl. J. Adv. Res. Comput. Sci. Software Eng., 2: 57-62.

Tewary, G., 2015. Effective data mining for proper mining classification using neural networks. Int. J. Data Mining Knowl. Manage. Process, 5: 65-82.