Diagnosis of Cancer Using Machine Learning

[1] Nitin Mishra,[2] Saumya Chaturvedi,[3] Garima Pandey

[1][2][3] School of computing science and Engineering, Galgotias University Greater Noida, India.

Abstract

Cancer is causing the deaths of millions of people worldwide. Artificial intelligence (AI) aims to mimic human cognitive functions. It is bringing a paradigm shift to healthcare, powered by the increasing availability of healthcare data and rapid progress of analytic techniques. AI can be applied to various types of healthcare data (structured and unstructured). Popular AI techniques include machine learning methods for structured data, such as the classical support vector machine and neural network, and the modern deep learning, as well as natural language processing for unstructured data. Major disease areas that use AI tools include cancer, neurology and cardiology. In this paper, we are proposing a method of detection of cancer using machine learning. In this paper, we have used 4 different machine learning algorithms to predict the cancer type. The Accuracy has also been analysed.

Keywords: *Machine learning,Breast cancer,Logistic regres- sion,Naive bayes,Classification trees*

1. Introduction

Globally around 70 percent death are caused due to cancer with 9.6 million death. Major causes of cancer are : High BMI index, poor fruit and vegetable intake, less physical work, tobacco alcohol usage. [1] Women which contribute to 50% of our humanity are also much affected by this disease. Figure1 shows 611,625 women have died due to breast cancer, which5th largest count of death caused by cancer. In this paper we have taken a case on breast cancer which is very prevalent in women.This has attracted lot of medical and applied science re-searchers in this domain. Due to this suddenly lot of funding has also been infused in this area.The goal of studying cancer is to develop safe and effectivemethods to prevent, detect, diagnose, treat, and, ultimately, cure the many diseases we call cancer. The better we under- stand these diseases, the more progress we will make toward diminishing the tremendous human and economic tolls of cancer. [2]With computation cost reducing and cloud computing beingeasily accessible by various researchers across the world, Machine learning methods also have been tried to fight with this monster disease.With easily available data sets and sharing by variousresearchers across the world, are able to do research in this field with much ease ever before.

Cancer prediction/ prognosis consist of 3 predictive analysis

1) The prediction of cancer vulnerability (i.e risk assess- ment)

2) Prediction of Cancer intermittence

3) The prediction of cancer survival The diagnosis for cancer is important but the most important point is at what time it is detected. Our main concern is to avoid the possibility of occurrence of cancer by detecting it at pre-early stage. If malignant cancerous cells are found in a patient then whether it is recurrent or not it should be detected as early as possible.

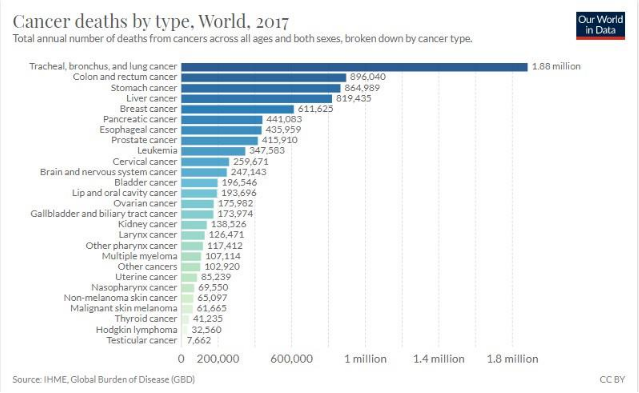

Fig. 1. Graph of deaths by cancer

A good prediction analysis can be done if all possible attributes and decision making algorithms are tested rigor- ously. Our research paper has concluded/compared the best prediction model after continues and parallel assessment of different Machine Learning algorithms.

2. Related Work

As cancer has been one of the most challenging diseases on earth, it prompts research to be done for detection of disease so that it can be treated with minimum resources in terms of time and money.

Artificial neural networks and decision trees have been used in cancer detection and diagnosis for nearly 20 years. [3] [4] Many papers have been published in recent times in using machine learning in Cancer. As computation became cheaper and the use of machine learning algorithms became prevalent machine learning has been used in health care for a wide variety of things. With advantage of big data, it has become even more convenient to apply and collect data in Healthcare. [5] Now, this data can be utilised for learning about diseases more than ever before. One of the important uses of Machines is to predict the disease. Many machine learning algorithms are used for the same purpose. [6]

Structural health monitoring patient can also be done using machine learning. Many researchers have done a lot of work for structural health monitoring of patient using machine learning. [7] [8]

Rapid progress in machine learning is enabling oppor- tunities for improved clinical decision support. Importantly, however, developing, validating and implementing machine learning models for healthcare entail some particular consider- ations to increase the chances of eventually improving patient care. [9]

Almost all methods of machine learning have been tried on the prediction of different kinds of diseases. Researchers have used classification, Logistic regression, neural networks, deep learning, clustering and other methods. Some people have used Association rule mining. Research is being done on the diagnosis and prognosis of the diseases. [10]

A trend which has been seen in Health care research in past few years includes

- Integration of mixed data such as clinical and genomic.
- ML methods for prediction of diseases
- ML methods for identification of diseases
- Identification from images

Several studies have been reported in the literature that can enable early cancer diagnosis. [11] [12]

As a result of efforts by various researchers The prediction accuracy of algorithms have improved by 15%-20%. [13]

3. Our Methodology

3.1 Breast Cancer Wisconsin (Diagnostic) Data Set

1) Attribute Information: The Data set contains 32 columns.It contains sixty three percent benign and thirty seven percent malignant records of diagnosis of breast tissues.Breast Cancer Wisconsin (Diagnostic) Data Set is used in our Experiment..

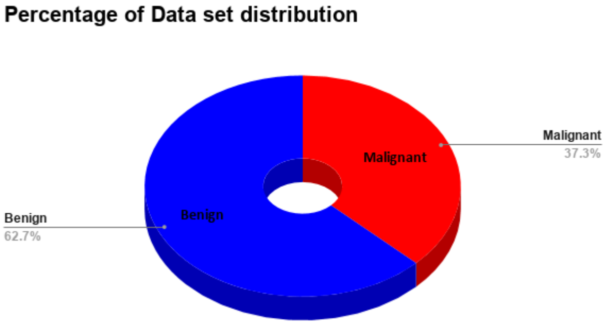

Fig. 2. Breast Cancer Wisconsin (Diagnostic) Data Set

1) ID number
2) Diagnosis –The diagnosis of breast tissues (M = malig- nant, B = benign)
3) Radius mean –Mean of distances from center to points on the perimeter
4) Texture mean –Standard deviation of gray-scale values
5) Perimeter mean–Mean size of the core tumor
6) area mean
7) smoothness mean –Mean of local variation in radius lengths
8) Compactness mean –Mean of perimeter2ˆ / area - 1.0
9) concavity mean –mean of severity of concave portions of the contour
10) concave points mean –Mean for number of concave portions of the contour
11) symmetry mean
12) fractal dimension mean –Mean for ”coastline approxi- mation” - 1
13) radius se –Standard error for the mean of distances from center to points on the perimeter
14) texture se –Standard error for standard deviation of gray-scale values
15) perimeter se
16) area se
17) smoothness se –Standard error for local variation in radius lengths
18) compactness se –standard error for perimeter2ˆ / area - 1.0
19) concavity se–standard error for severity of concave por- tions of the contour
20) concave points se –standard error for number of concave portions of the contour
21) symmetry se
22) fractal dimension se –standard error for ”coastline ap- proximation” - 1
23) radius worst –”worst” or largest mean value for mean of distances from center to points on the perimeter
24) texture worst –”worst” or largest mean value for stan- dard deviation of gray-scale values
25) perimeter worst
26) area worst
27) smoothness worst –”worst” or largest mean value for local variation in radius lengths
28) compactness worst –”worst” or largest mean value for perimeter2ˆ / area - 1.0

29) concavity worst –"worst" or largest mean value for severity of concave portions of the contour
30) concave points worst –"worst" or largest mean value for number of concave portions of the contour
31) symmetry worst
32) fractal dimension worst –"worst" or largest mean value for "coastline approximation" – 1

3.2 Architecture of Methodology

In the paper, we have followed an intervention based methodology through experimental analysis. The first step towards this is loading the raw data from the data source. The second step contains the pre-processing of data. So we have performed data cleaning in this step and the preprocessed data than split into training data and testing data. The splitting of data is very significant for cross-validation of machine learning classifier. The architecture has been shown in the figure. 3

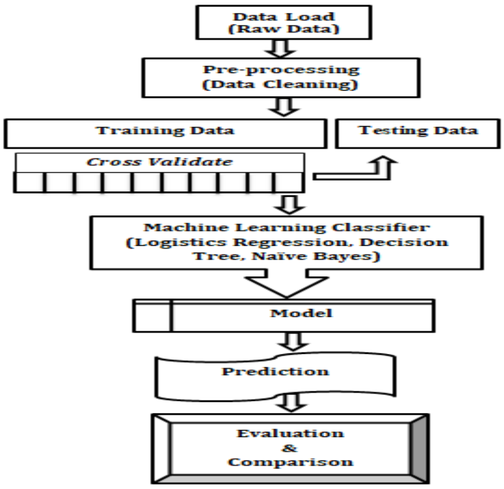

Figure. 3

In our methodology, we made sure that for every run of the experiment the data splitting is performed randomly for a specified train test ratio. In our approach, we have fed the split data to four classifiers viz Logistic Regression, Decision Tree (Gini-Index), Decision Tree(Entropy) and Naive Bayes. The end result of the aforementioned steps is four machine learning models, which is now can be used for prediction. Once the prediction is done the evaluation of models and comparison of models is carried out based on certain metrics viz prediction accuracy, Precision, Recall and F-score id done.

4. Experimental Setup

Breast Cancer Wisconsin (Diagnostic) Data Set has been used. [14] The data used has 569 instances using 32 features. After removing missing values operations, no missing values were found. The data contains 212 cases of Malignant cancer whereas 357 cases of Benign cancer cases.
The data is not proportionate hence partition to test and train data has to be done in such a way so that percentage of Malignant and benign cases remains the same in every test/train split.

5. Results

We have executed four Machine Learning Algorithms to create a learning model. These Algorithms are

1) Logistic Regression
2) Naive Bayes
3) Classification using Gini index
4) Classification using Entropy
The experiment was run 72 times per model. For each partition ratio experiment was run 18 times. The results are given in the different upcoming sections.

5.1 Logistic Regression

In the dataset all the features were numerical so we decided to train the learner using Logistic regression.The results of experiment has been shown in Figure 4.
We have chosen this algorithm for our dataset because it has some very interesting properties.
Logistic Regression performs well when the dataset is lin- early separable.Logistic Regression not only gives a measure of how relevant a predictor (coefficient size) is, but also its direction of association (positive or negative). Logistic regression is less prone to over-fitting but it can overfit in high dimensional datasets. You should consider Regularization (L1 and L2) techniques to avoid over-fitting in these scenarios. Logistic regression is easier to implement, interpret and very efficient to train. [15]
Test result were carried out with the following partition- ing/split point 0.1,0.2,0.3,0.4(i.e training data set was 90% and testing data-set was 10% for 0.1) . Figure 4 shows minimum accuracy was 28.07 and maximum accuracy value was 71.93. The average accuracy is 60.14. Since the variation of this algorithm is high therefore other methods need to be explored.

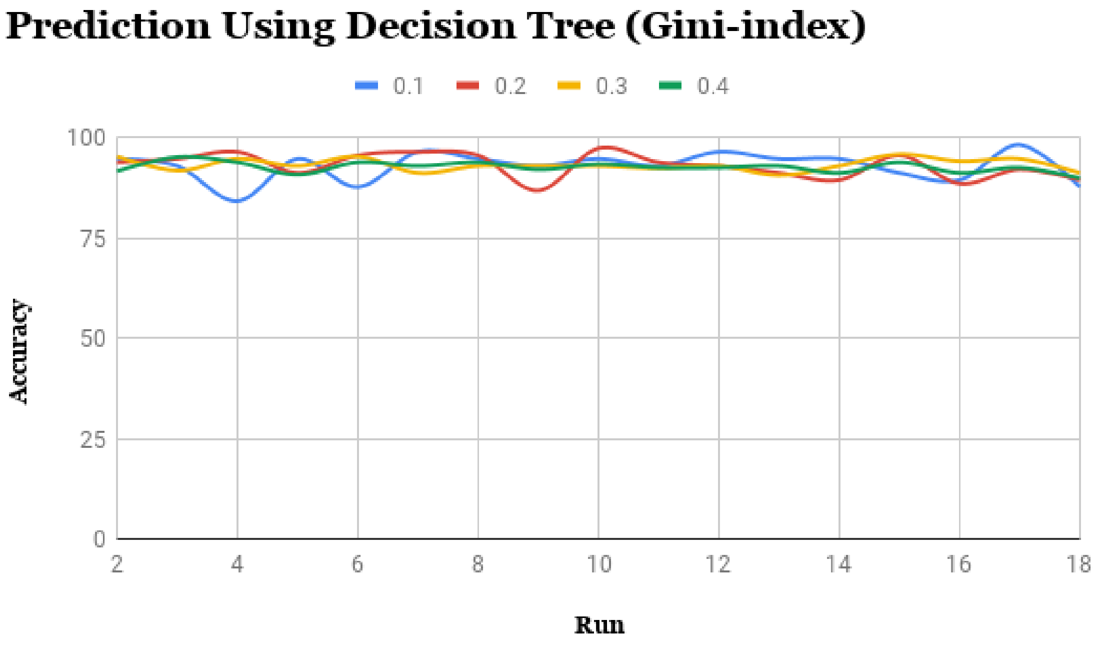

Fig. 4. Prediction Accuracy using Logistic Regression using different ratios of Testing on Dataset

5.2 Naive Bayes

The results of experiment has been shown in Figure 5. Naive bayes has been chosen because of following properties that match with the naive bayes learner.
If training set is small, high bias/low variance classifiers (e.g., Naive Bayes) have an advantage over low bias/high variance classifiers (e.g., kNN), since the latter will overfit. But low bias/high variance classifiers start to win out as your training set grows

(they have lower asymptotic error), since high bias classifiers aren’t powerful enough to provide accurate models.

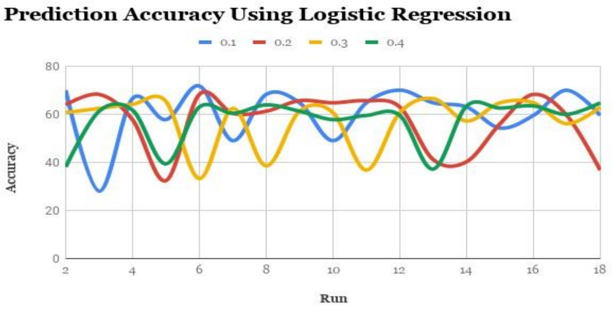

Fig. 5. Prediction Accuracy using Naive Bayes using different ratios of Testing on Dataset

Figure 5 shows test results Using Naive Bayes machine learner. The partitioning/split point are 0.1,0.2,0.3 and 0.4. Minimum prediction accuracy value is 50.00 and maximum accuracy value is 70.18. In this graph we can see that maxi- mum accuracy is 70.18 and average accuracy is 62.17. A new machine learner can be explored which gives better accuracy than this method.Figure 5 shows test results Using Naive Bayes machine learner. The partitioning/split point are 0.1,0.2,0.3 and 0.4. Minimum prediction accuracy value is 50.00 and maximum accuracy value is 70.18. In this graph we can see that maxi- mum accuracy is 70.18 and average accuracy is 62.17. A new machine learner can be explored which gives better accuracy than this method.

5.3 Classification using Decision tree

The results of experiment has been shown in Figure 6 and
Figure 7.

For Gini Index we have obtained these values shown in Figure 6. The partitioning/split point are 0.1,0.2,0.3 and 0.4. Minimum prediction accuracy value is 84.21 and maximum accuracy value is 98.25. The average accuracy is 92.99

For Entropy Index we have obtained these values shown in
Figure 7. The partitioning/split point are 0.1,0.2,0.3 and 0.4. Minimum prediction accuracy value is 87.72 and maximum accuracy value is 98.25 The average accuracy is 93.69.

Our research has tested different prediction and decision
tree based analysis. Further we conclude that Prediction Using Decision Tree has better test result therefore either Gini Index or Entropy can be used.

Fig. 6. Prediction Accuracy using Classification tree (Gini index) using different ratios of Testing on Dataset

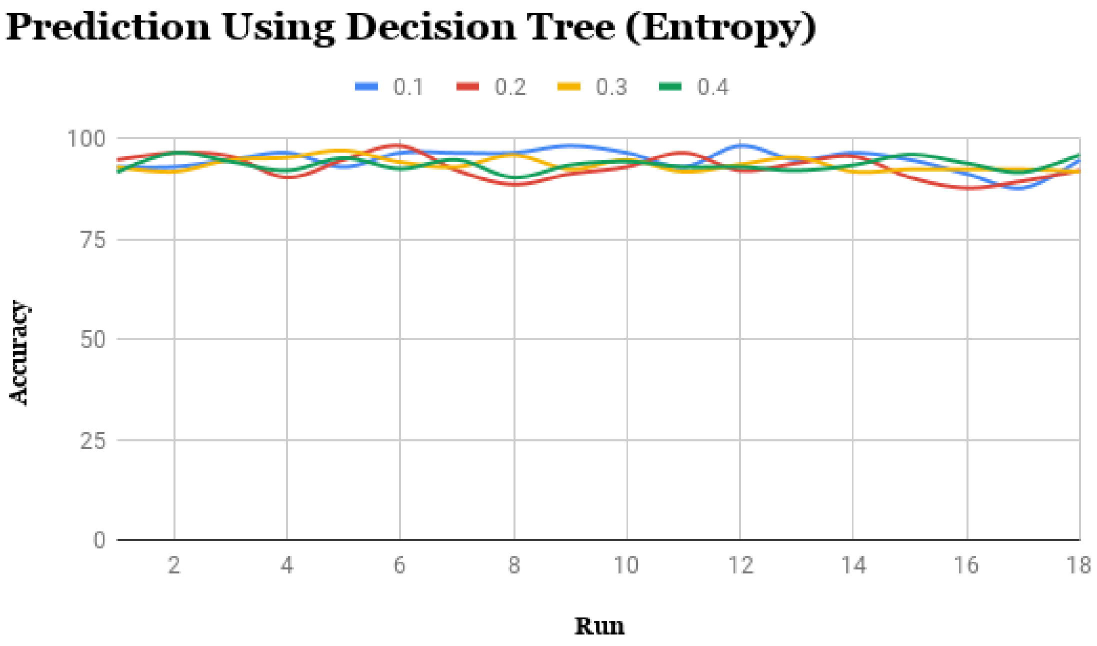

Fig. 7. Prediction Accuracy using Classification tree (Entropy) using different ratios of Testing on Dataset

ML	**AVG Accuracy**	**Min**	**Max**
LR	58.24805068	28.07071754	71.92982456
DT(Gini-index)	92.9905783	84.21052632	98.24561404
DT(Entropy)	93.6931449	87.71929825	98.24561404
NB	62.16699155	50	70.1754386

Comparison Of Results By Different Methods

5.4 Comparison of different Machine Learners

The result of comparison has been shown in Figure 8. The comparison of minimum,maximum and average accuracy has been given in the Figure 9.

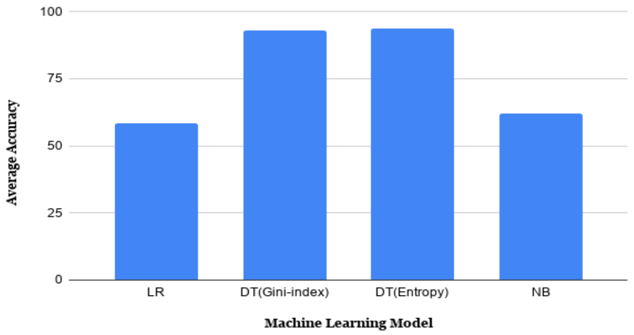

Fig. 8. Comparison of prediction accuracy of different Machine Learners

Figure 8 shows the average accuracy of different machine learners on the basis of different runs.After cross validation the values are given in Table I. We can observe in Table I that best performance among all machine learners is Decision tree (Entropy) method which gives average accuracy of 93.69% and lies in the range of 87% to 98%.

Figure 6 and Figure 7 shows test results Using Decision Tree Classification Algorithm. We have used two variations of this algorithm. Gini Index and Entropy has been used. The partitioning/split point are 0.1,0.2,0.3 and 0.4.
Decision tree learners are best suited for breast cancer datasets. [14]

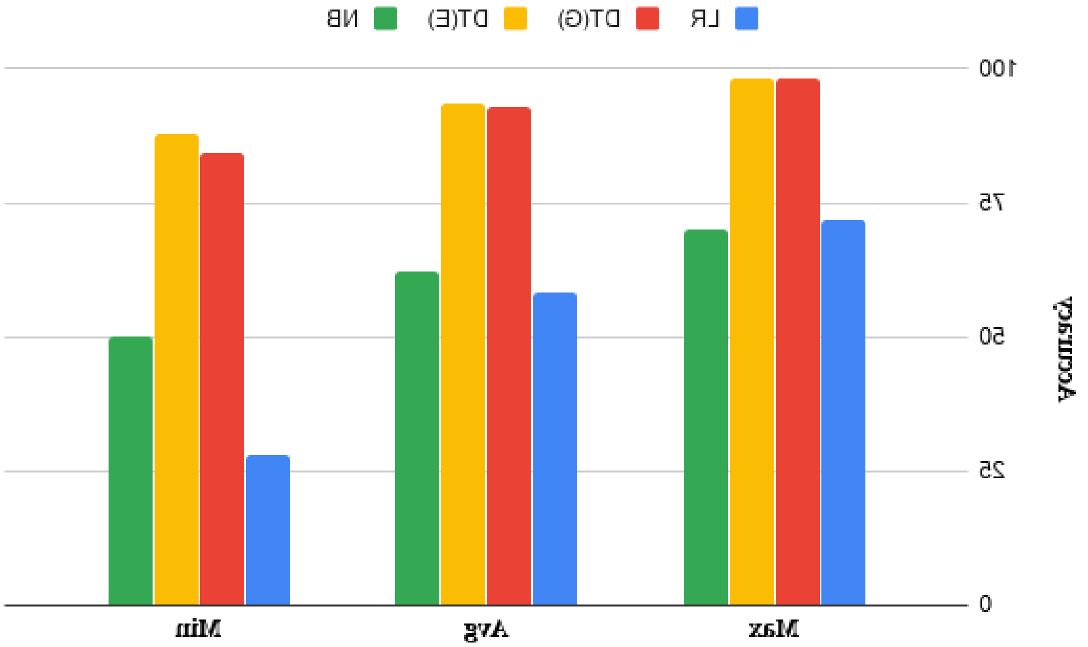

Fig. 9. Comparison of Results by different methods

7. Conclusion

In this paper we have used different machine learning algorithms to generate the learner for prediction of Malignant and Benign cases. After comparison of four Machine learners we have came to conclusion that Classification trees are more suitable and robust in terms of prediction accuracy and variance. This work can be extended to other machine learning methods. The data set can be extended to detect other cancer types.

References

1. WHO, “Cancer Death Records.” https://www.who.int/news-room/fact-sheets/detail/cancer, 2019. [Online; accessed 8-Jan-2019].
2. A. Foundation, “Cancer Death Records.” https://www.aacrfoundation.org/Pages/what-is-cancer-research.aspx, 2019. [Online; accessed 8-Jan-2019].
3. L. D. Grouse, “Has the machine become the physician?,” JAMA, vol. 250, no. 14, pp. 1891–1891, 1983.
4. K. Frolov, Methods of Machine Improvement & Contemporary Problems of Machine Science. Allerton Press, 1988.
5. Z. Obermeyer and E. J. Emanuel, “Predicting the future—big data, machine learning, and clinical medicine,” The New England journal of medicine, vol. 375, no. 13, p. 1216, 2016.
6. M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, “Disease prediction by machine learning over big data from healthcare communities,” Ieee Access, vol. 5, pp. 8869–8879, 2017.
7. C. R. Farrar and K. Worden, Structural Health Monitoring.: A Machine

8. Learning Perspective. John Wiley & Sons, 2012.
K. Worden and G. Manson, "The application of machine learning to structural health monitoring," Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 365, no. 1851, pp. 515–537, 2006.
9. P.-H. C. Chen, Y. Liu, and L. Peng, "How to develop machine learning models for healthcare," Nature materials, vol. 18, no. 5, p. 410, 2019.
10. D. C. Mohr, M. Zhang, and S. M. Schueller, "Personal sensing: under- standing mental health using ubiquitous sensors and machine learning," Annual review of clinical psychology, vol. 13, pp. 23–47, 2017.
11. O. Fortunato, M. Boeri, C. Verri, D. Conte, M. Mensah, P. Suatoni, U. Pastorino, and G. Sozzi, "Assessment of circulating micrornas in plasma of lung cancer patients," Molecules, vol. 19, no. 3, pp. 3038–
3054, 2014.
12. H. M. Heneghan, N. Miller, and M. J. Kerin, "Mirnas as biomarkers and therapeutic targets in cancer," Current opinion in pharmacology, vol. 10, no. 5, pp. 543–550, 2010.
13. J. A. Cruz and D. S. Wishart, "Applications of machine learn- ing in cancer prediction and prognosis," Cancer informatics, vol. 2, p. 117693510600200030, 2006.
14. WHO, "Breast Cancer Wisconsin (Diagnostic) Data Set." https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic), 2019. [Online; accessed 10-Jan-2019].
15. T. of P. Logistic Point, "Advantages and Disadvantages Regression in Machine Learning."http://theprofessionalspoint.blogspot.com/2019/03/advantages-anddisadvantages-of.html, 2019. [Online; accessed 8-Jan-2019].
16. Chaturvedi, V. Mishra, and N. Mishra, "Sentiment analysis using machine learning for business intelligence," in 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), pp. 2162–2166, IEEE, 2017
17. N. Mishra, S. Chaturvedi, V. Mishra, R. Srivastava, and P. Bargah, "Solving sparsity problem in rating-based movie recommendation system,"
in Computational Intelligence in Data Mining, pp. 111–117, Springer,
2017.
18. N. Mishra, V. Mishra, and S. Chaturvedi, "Tools and techniques for solving cold start recommendation," in Proceedings of the 1st International Conference on Internet of Things and Machine Learning, p. 11, ACM, 2017.
19. W. H. Wolberg, w. N Street, D. M. Heisey, and O. L. Mangasarian, "Computer-derived nuclear features distinguish malignant from benign breast cytology," Human Pathology, vol. 26:, pp. 792–796, 1995.