2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)
Amity University, Noida, India. June 4-5, 2020

Predictive Analysis of Machine Learning Algorithms for Breast Cancer Diagnosis

Mudit Arora
Amity Institute of Information Technology
Amity University
Noida, India
muditarora23456@gmail.com

Subhranil Som
Amity Institute of Information Technology
Amity University
Noida, India
ssom@amity.edu

Ajay Rana
Amity Institute of Information Technology
Amity University
Noida, India
ajay_rana@amity.edu

***Abstract*— Cancer is one of the fastest growing disease around the world and subpart of it Breast Cancer that is growing rapidly and mostly affecting women. Early treatment of this disease is helpful and can act as an early prevention to the upcoming major cure. However this can only be possible only if women are able to know that they are suffering with such disease and this can be only diagnosed if they come up with it and openly sharing with family and Doctors about the disease. This can lead to be a bit challenging task as to detect this disease among women using mammography as patient communication can affect mammography performance. This disease has had many ideas and myths as to how we can diagnosed it but Machine Learning the subset of Artificial Intelligence that can help Doctors and Surgeons to learn from past experiments. To treat upcoming patients with similar anomalies has had the major help of saving many patients with its set of algorithms and set of applications it provides. This paper will be focusing on five of the popular supervised Machine Learning algorithms for Diagnosing Breast Cancer this will be K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes (NB) and Decision Tree (DT) and the algorithm Random Forest gave the best results and the K Nearest Neighbor was the second best performing algorithm that produce desired results and the algorithm Naïve Bayes was the least performing algorithm**

Keywords- Machine Learning; Breast Cancer; K-Nearest Neighbour; Naïve Bayes; Support Vector Machine; Random Forest; Decision Tree

I. INTRODUCTION

Machine Learning is branch of AI that has helped many sectors with its accuracy and the amount of performance it delivers be it Automobile, Healthcare and Education sector but the sector that has largely been helped is Healthcare sector as Doctors and surgeons find it easier to diagnose a patient by learning from the previous observations and they have observed from a patient or by viewing his her similar reports that matches there previous surgery. Most problems in healthcare like Diabetes, Cardiac Arrests and Cancers have seen a keen observations of the Doctors. Belgium is the country with the highest rate of breast cancer in women with an Age-standardized rate of 113.2 per 100,000[1] According to statistic 25.8 per 100,000 women were affected by Breast Cancer.

In India were the survival rate were as low as 66.6% [2] and every 4 minutes we get a new patient diagnosed with this disease. The survival rate were increasingly shocking in Australia and New Zealand were there are high chances of women getting diagnosed with such disease but there survival rate is almost 90% of the females getting diagnosed due to modern equipment's and more health awareness . Accurately predicting a cancerous tumor remains a challenging task for many physicians. The emergence of new medical technologies and the enormous amount of patient data have motivated the path for the development of new strategies in the prediction and detection of cancer. Although data assessment that is collected from the patient and a physician's intake greatly contributes to the diagnostic process, supportive tools could be added to help facilitate accurate diagnoses. These tools aim to eliminate possible diagnostic errors and provide a fast way for analyzing large chunks of data.

It is not just a disease that needs few tests and you are able to predict that you have discovered the signs of breast cancer most women can discover the signs of suffering a breast cancer my early signs of certain disease developed in them that they do not even bother to take seriously some have a hereditary problem or some may have developed in one of their breast earlier and might also detect in later stages of their life as well or more related signs of this can be when menopause stops and women having period at an early stage of less than 12 years and even getting pregnant at very later stage normally after 30 years of age and being addicted to alcohol and other nicotine substances that might affect their inner body adversely and can lead to high chances of getting affected with breast cancer. Early signs of health problems are not always the cause of it certain tissues that can be malignant that affects the nearby tissues that in turns destroys other tissues as well.

Machine Learning (ML), is a subfield of Artificial Intelligence (AI) that lets in machines to analyze without specific programming by using statistical strategies and utilizing algorithms to study from positive statistics through exposing them to sets of facts allowing them to learn a particular undertaking via experience[3].

Fig. 1. A visual of how Machine Learning works.[4]

Machine learning contains set of calculations and functions that instruct PCs to perform tasks that human want them to do. Machine learning make systems capable to learn from data, find the possible patterns and make decisions with minimum human interference. The machines are trained by learning the inputs or through experience, the more inputs we provide the more accurate the result is or we can say the more info a system get the more accurate result is. It grants computational models that are involved diverse taking care of layers to learn depictions of data with various degrees of reflection and these strategies helped a great deal to improve the image classification, visual object recognition, image retrieval, human pose estimation, speech recognition etc. The system continually modifies its model based on new information and assign a value to each model, so that in future whenever the system encounters the object it can differentiate between the objects easily and use the values that were assigned earlier.

The main problem with machine learning is that, machine learning techniques requires huge set of data to train them and more the input is fed more chances of better and accurate results are achieved but as we fed them with small and inaccurate data we may end up getting biased predictions as data that we used was irrelevant and inaccurate.

A. Machine learning tasks

Supervised learning – In this type of learning the inputs are given in the form of examples in machine so to get desired outputs. It means training the machines to achieve target by providing all the possible inputs. It is the spot you have info factors (X) and a yield variable (Y) and you use a computation to take in the mapping limit from the commitment to the yield

$$Y = f(X)$$

The idea behind using this methodology is to use the capacity so well that whenever we input new contents (x) then we can get the desired information depending upon the info we provided.

It is called inclusive learning in light of the fact that the procedure of a calculation gaining from the preparation dataset can be thought of as an educator regulating the learning procedure. When the answers matches what we exactly want then, the calculation iteratively makes prediction from the information and the results are achieved depending on inputs deployed and satisfaction achieved with the calculation. Machine Learning has always been a very popular sign of helping medical industry with the desired set of tools and techniques it has to offer be it having algorithms and helping surgeons and researchers. In Healthcare industry to know the coming trends of upcoming similar diseases that were somehow similar to the previous diseases that were diagnosed with same practice and with similar health tools.

Machine Learning algorithms provide a large set of techniques that help to find out the real problem behind that can be using supervised learning, unsupervised learning or even through reinforcement learning as well.

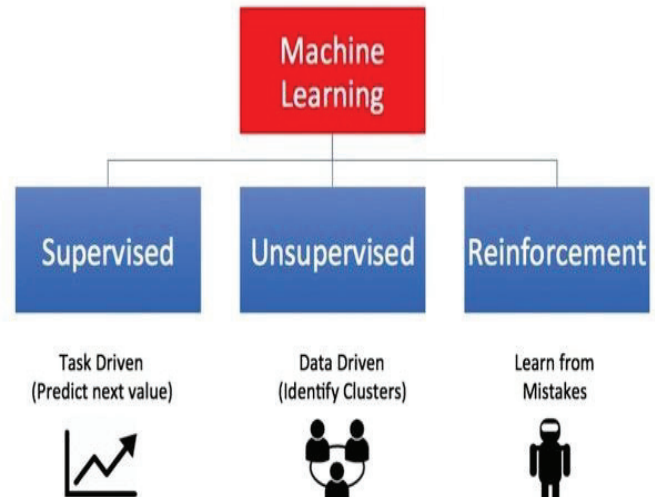


Fig. 2. Machine Learning Types[5]

II. .MATERIALS AND METHODS

In this paper, four of the popular machine learning algorithms used under supervised learning that are K-Nearest Neighbors, Random Forest, Support Vector Machine and Naïve Bayes are going to be used

A. K-Nearest Neighbors

K-Nearest Neighbors (KNN) is certainly one of the handiest algorithms used in Machine Learning for regression and class problem. KNN algorithms use facts and classify new facts new facts points primarily based on similarity measures. Classification is achieved by means of a majority vote to its neighbors [6].

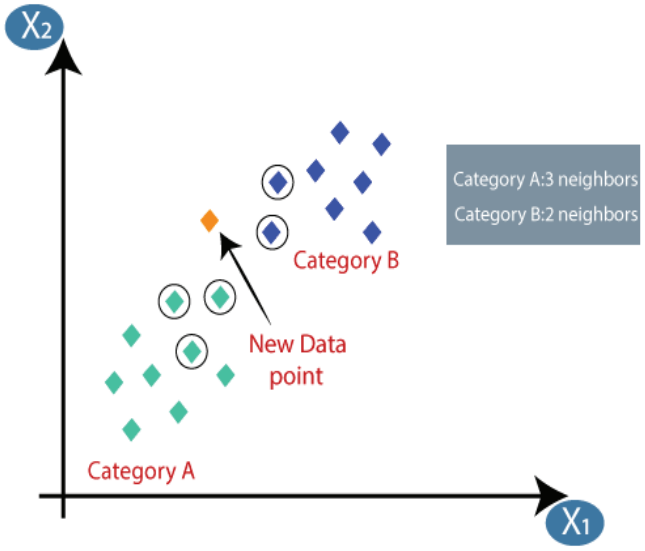


Fig. 3. KNN Algorithm[7]

B. Random Forest

It is used by constructing multiple decision trees. The end results are based on the majority of the trees and is chosen by Random Forest[8].

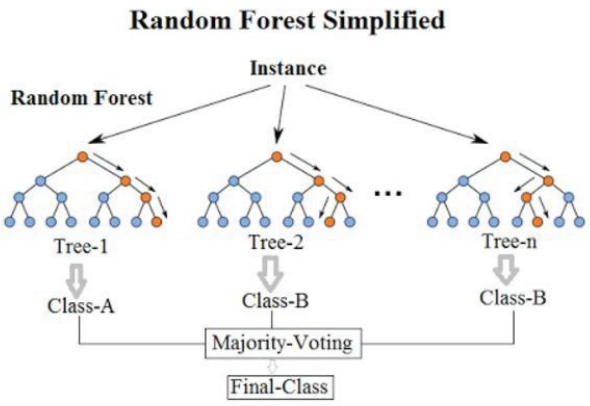
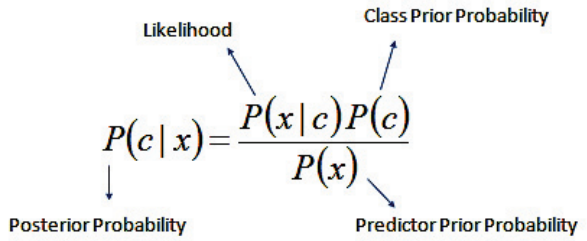


Fig. 4. Random Forest Algorithm[9]

C. Naïve Bayes

It is based on the probabilistic methods by comparing the particular feature of a class unrelated to any other feature which is having different assumptions and properties that may completely differ from each other and may. The resulting class with highest possibility is assumed to be of highest accuracy [10].



$$P(c|X) = P(x_1|c)\times P(x_2|c)\times\cdots\times P(x_n|c)\times P(c)$$

Fig. 5. Naïve Bayes Algorithm[11]

D. Support Vector Machine

To achieve classification and regression problems SVM is highly used whenever unlabeled data and we used the hyperplane which divides the two data points and help to classify the problem more related to which of the section[12].

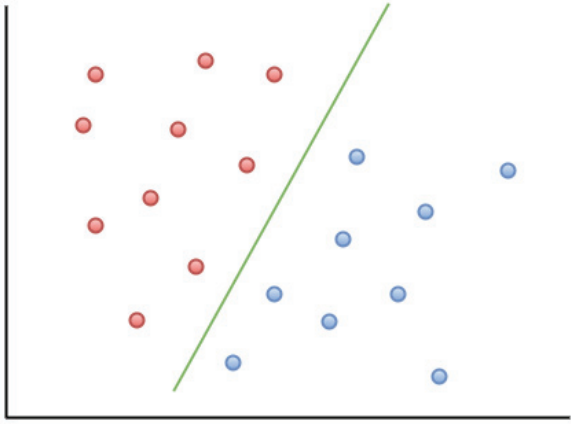


Fig. 6. Support Vector Machine Algorithm[13]

E. Decision Tree

Decision tree is a type of supervised learning algorithm (having a pre-defined target variable) that is mostly used in classification problems. It works for both categorical and continuous input and output variables. The author has tried to use the medical dataset of Brest cancer on the Decision Tree algorithm[14].

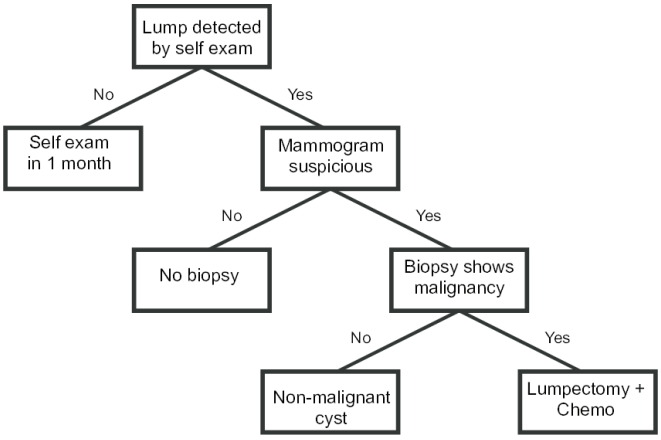


Fig. 7. Decision Tree Algorithm[15]

F. Dataset

The dataset used is collected from Wisconsin which helps to get a brief idea of the columns that are provided in the dataset so that they can act as a means of information. The Dataset is divided into 32 columns and is created by Dr. William H. Wolberg [16] and this dataset can help us to predict whether the following breast problem caused is benign or malignant and can help as an aid to the upcoming patients by discovering the desired accuracy and getting the desired results so that they do not affect the patients coming up with similar anomalies and can help to treat they by learning from the past observations and getting better results each time and make better observations using the algorithms on them can help us to find accuracy so that they might act as a sign as to which algorithm is better and by which margin and who gives better accuracy

The datasets shall not provide any value to be none that disturbs the dataset from producing the desired results and getting better outputs. We shall predict clear and accurate results so that they help in better understanding.

III. RELATED WORK

Williams et al., [17] "Breast cancer risk prediction using data mining classification techniques", in this two popular data mining techniques J48 and Naïve Bayes were used to check the accuracy and J48 perform better results accuracy error and mean value in order to check the results.

Senturk et al.,[18] "Breast cancer Diagnosis via Data Mining: Performance Analysis of Seven Different Algorithms ",in this seven different Algorithms were used and in this the best algorithms was Support vector Machine scored the best results and Decision Tree was the second best algorithm to perform with a slight decrease in the accuracy percentage.

E. Venkatesan et al.,[19] "Performance analysis of decision tree algorithms for breast cancer classification", in this four different algorithms J48, CART, AD TREE, BF TREE were used on certain parameters and J48 performed the best results with 99% accuracy preceded by BF TREE with 98% accuracy.

Animesh et al.,[20] "Study and analysis of Breast cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms", in this the algorithm Naïve Bayes performed the best results and gave 98% accuracy and even gave least time complexity compared to other algorithms which is the best among the three.

Dana Bazazeh et al.,[21] **"** Comparative Study of Machine Learning Algorithms for Breast Cancer Detection and Diagnosis" used Random Forest, Support Vector Machine and Bayesian Network and Support Vector Machine stood out with best results and Random Forest have the highest possibility of correctly satisfying the tumour.

Habib Dhahri et al.,[22] "Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms" in this machine learning algorithms perform poor than the automated breast cancer diagnosis technique that was adopted in a three stage model in this paper.

Abdelghani Bellaachia, Erhan Guven et.al,[23] **"** Predicting Breast Cancer Survivability Using Data Mining Techniques**" ,** this paper has outlined the various survival techniques in Breast cancer using seer database and how to survive with certain Cause Of Death(COD) and how it outperforms the certain Techniques used for testing the data.

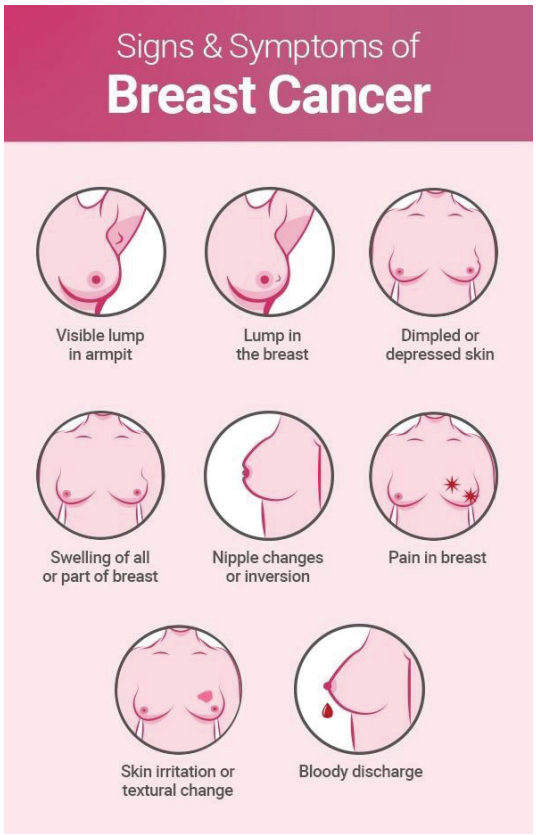


Fig. 8. Signs of Breast Cancer[24]

IV. EXPERIMENTAL WORK

In this dataset certain algorithms were used and to check whether results obtained were sufficient to produce desire conclusions and to evaluate whether these are sufficient for future researchers to evaluate their proceedings and carry on the work created and know whether these algorithms were producing results as per the desired output demanded certain python libraries like numpy, pandas , matplotlib , csv and math were used to train the dataset and split the data and then compare the results with the testing data upon knowing certain results were achieved which prove that algorithm could perform well on certain situations producing desired results and giving accuracy which sums up the desired problems that were put on the data. A dataset from Wisconsin was used and was split on the basis on 80 is to 20 percentage 80% data was used on all the five algorithms these were Random Forest, Naïve Bayes, Support Vector Machine, Decision Tree and K-Nearest Neighbor and 20% testing was used to compare the training datasets and results achieved were.

Python libraries were used in this work and also their code was created in jupyter notebook and comparison was drawn were certain algorithms performed according to the accuracy and algorithms performed according to the features deployed as parameters. Random Forest performed best among the all other four algorithms used and KNearestNeighbor was the second best performing algorithm with almost 95.61% accuracy.

The classifier accuracy is a measure of how well the classifier can correctly predict cases into their correct category. It is he number of correct predictions divided by the total number of instances in the data set and also certain graphical ideas were also used to better create a visual representation of the resultant data and also eliminating certain columns of the dataset which were not much help of producing desired results and formulating ideas It is worth noting that the accuracy is highly dependent on the threshold chosen by the classifier and can therefore change for different testing sets.

Accuracy= (No.of Correct Predictions/ Total No. of Predictions)*100

A. Accuracy Results

Algorithms	SVM	RF	NB	KNN	DT
Accuracy	94.73	96.49	92.10	95.61	93.86

Fig. 9. Representation of accuracy results[25]

In This algorithms the four used the K-Nearest Neighbor produced the best results and giving a higher accuracy rate and giving a high chances of predicting the results and deriving results from the dataset easily.

V. CONCLUSION

Different researchers have proposed their own set of theories depending on what they believe as to how healthcare as a domain can help analyze future anomalies these theories are covered using supervised learning algorithms as to get desired results some may have found their ideal output using. Some have found better results in Naïve Bayes, SVM, Random

Forest KNN and Decision Tree algorithms based on the dataset they deployed so as per results obtained we may find out at these days technology using smart prediction methods can be used in almost every area whether healthcare or any other it depends upon how they are being used to deploy.

The thing that is notable is human dependency has increased as we rely upon only smart methods we can just predict results we don't have any surety towards what is the best comparison technique for our analysis and sometimes we may also found the difference in ideologies with these due to absence of an expert in this area sometimes we may only find a doctor helpful to help diagnose disease and relying upon machines for our work because sometimes a machine fails because it .has loop holes and bugs to get recovered conclusion lies to analyzing these techniques better and using them with some expert of both the areas likewise a Doctor from healthcare industry and a trained IT Professional could better understand both of these techniques and use it better to diagnose patients because just giving a machine to a surgeon and asking him to help future coming patients is not his job and same a IT Professional can't predict an outcome until and unless he is not known to the roots of the disease.

This is all about coming to the same page and helping mankind with safer hands because it is not a competition among two different areas to emerge the winner but it is MAN WITH MACHINE to help human but not HUMAN VS MACHINE.

Out of these algorithms, the best algorithm that worked on all the conditions was Random forest and the KNN performed second best whereas other performed slightly up or somewhat less.

REFERENCES

[1] https://www.wcrf.org/dietandcancer/cancer-trends/breast-cancer-statistics

[2] https://www.livemint.com/

[3] T. J. Cleophas and A. H. Zwinderman et al., "Machine Learning in Medicine," pp. 1-271, 2013

[4] https://miro.medium.com/max/2048/0*V0GyOt3LoDVfY7y5.png

[5] https://miro.medium.com/max/1204/0*-068ud_-o3ajwq_z.jpg

[6] K. Sivakami et al., "Mining Big Data: Breast Cancer Prediction using DTSVM Hybrid Model", International Journal of Scientific Engineering and Applied Science vol.1, pp.418-429, 2015

[7] https://static.javatpoint.com/tutorial/machine-learning/images/k-nearest-neighbor-algorithm-for-machine-learning5.png

[8] Cuong Nguyen, Yong Wang, Ha Nam Nguyen et al.," Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic" , J. Biomedical Science and Engineering, 2013, 6, 551-560

[9] https://miro.medium.com/max/592/1*i0o8mjFfCn-uD79-F1Cqkw.png

[10] Megha Rathi , Arun Kumar Singh et al.," Breast Cancer Prediction using Naïve Bayes Classifier" ,international journal of information technology and systems,Vol. 1;No .2: ISSN: 2277-9825(July-Dec.2012)

[11] https://mlalgorithm.files.wordpress.com/2016/06/screenshot-19.png?w=700

[12] Y.Ireaneus Anna Rejani et al., "early detection of breast cancer using svm classifier technique" International Journal on Computer Science and Engineering Vol.1(3), 2009, 127-130

[13] https://eight2late.files.wordpress.com/2016/12/svm-fig-1.png

[14] Al-Salihy, N. K., & Ibrikci, T. (2017) et al.,"Classifying breast cancer by using decision tree algorithms. Proceedings of the 6th International Conference on Software and Computer Applications - ICSCA '17. doi:10.1145/3056662.3056716

[15] https://www.researchgate.net/publication/24442935/figure/fig2/AS:339550913220617@1457966654793/An-example-of-a-simple-decision-tree-that-might-be-used-in-breast-cancer-diagnosis-and.png

[16] Dr. William H. Wolberg, General Surgery Dept. University of Wisconsin, Clinical Sciences Center Madison, WI 53792

[17] Williams et al., "Breast cancer risk prediction using data mining classification techniques", Transactions on Networks and Communicatons,vol.2, pp.1-11,2015.

[18] Senturk et al., "Breast cancer Diagnosis via Data Mining: Performance Analysis of Seven Different Algorithms ", Computer Science & Engineering", vol.1, pp.1-10, 2014.

[19] E. Venkatesan et al., "Performance analysis of decision tree algorithms for breast cancer classification", Indian Journal of Science and Technology, vol.8, pp.1-8, 2015.

[20] Animesh et al., "Study and analysis of Breast cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms", International Journal of Computer Applications, vol.2, 2016

[21] Dana Bazazeh and Raed Shubair et.al., "Comparative Study of Machine Learning Algorithms for Breast Cancer Detection and Diagnosis" ,

[22] Habib Dhahri et al., "Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms" Volume 2019 |Article ID 4253641

[23] Abdelghani Bellaachia, Erhan Guven " Predicting Breast Cancer Survivability Using Data Mining Techniques"

[24] onco.com/about-cancer/wp-content/uploads/2020/01/breast-canc-1.jpg

Accuracy Results of the dataset with the provided Dataset.

[25] Walia, H., Rana, A., Kansal, V., "A Decision Tree Based Supervised Program Interpretation Technique for Gurmukhi Language" in Communications in Computer and Information Science, pp 356-365 (2020).

[26] Ghosh, S., Rana, A., Kansal, V., "A benchmarking framework using nonlinear manifold detection techniques for software defect prediction" in International Journal of Computational Science and Engineering, Vol. 21, No. 4, pp 593-614 (2020).

[27] Tyagi N., Rana A., Kansal V. (2020) Load Distribution Challenges with Virtual Computing. In: Solanki V., Hoang M., Lu Z., Pattnaik P. (eds) Intelligent Computing in Engineering. Advances in Intelligent Systems and Computing, vol 1125. Springer, Singapore

[28] Gupta, S., Rana, A., Kansal, V., "Optimization in wireless sensor network using soft computing" in Advances in Intelligent Systems and Computing, pp 801-810 (2020).

[29] S. Gupta, A. Rana and V. Kansal, "Comparison of Heuristic techniques:A case of TSP," 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2020, pp. 172-177, doi: 10.1109/Confluence47617.2020.9058211.

[30] S. Ghosh, A. Rana, V. Kansal, "Evaluating the Impact of Sampling-Based Nonlinear Manifold Detection Model on Software Defect Prediction Problem" in Smart Innovation, Systems and Technologies, Vol. 159, pp 141-152 (2020).

[31] N. Agarwal, A. Rana, J.P. Pandey, A. Agarwal, "Secured sharing of data in cloud via dual authentication, dynamic unidirectional PRE, and CPABE" in International Journal of Information Security and Privacy, Vol 14, Issue 1, pp 44-66 (2020)