ANVESAK
ISSN: 0378–4568
UGC Care Group1 Journal

DETECTION AND PREVENTION OF BREAST CANCER USING ML AND SVM ALGORITHMS

[1]Dr. N. Chandra Mouli, [2]S. Sateesh Reddy, [3]Anusha Nalla
[1]Associate Professor, [2]Assistant Professor, [3]Student, [1,2,3] Dept. of Computer Science Engineering,
[1,2,3]Vaageswari College of Engineering, Karimnagar, Telangana.
E-Mail: [1]cmnarsingoju@gmail.com, [2]sateesh.singireey@gmail.com

ABSTRACT
According to the current research piece, cancer is a global concern that affects people of all ages and socioeconomic backgrounds. The most common malignancy in women is specifically breast cancer. Therefore, every advancement in cancer detection and prognosis poses a risk to leading a healthy life. The early detection and forecasting of cancer could benefit greatly from the use of machine learning techniques. The Wisconsin Cancer dataset is classified using two of the most popular machine learning algorithms in this work, and the categorization presentation of this technique was compared using exactness, precision, recall, and ROC Area scores. The Support Vector Machine method delivered the best results with the highest level of precision. SVM is the most accurate method for prophetic analysis with a 99.1% accuracy rate. We tend to conclude by this study is that SVM is the complementary algorithm for forecasting, and on the entire DT conferred well next to SVM.
Keywords: Breast cancer Detection, benign, malignant, NB, KNN, SVM, RF, DT

1. INTRODUCTION
In 2018, cancer was the second-leading cause of mortality worldwide, accounting for almost 9.6 million fatalities. One out of every six fatalities in the world are brought on by cancer. In low- and middle-income nations, cancer-related deaths account for over 70% of all fatalities [1]. The three malignancies that affect women the most frequently are colon, lung, and breast cancers, which together account for 50% of all cancer cases. Furthermore, cancer accounts for 30% of all new cancer diagnoses in women [2]. ML techniques ensure that data is evaluated and that information and important correlations are extracted from a dataset. Additionally, it creates a computational illustration for the most effective data interpretation. Particularly, according to examiners about cancer disease, it tells ML methods handles on premature exposure and forecasting of cancer [3]. Asri et al. evaluated many machine learning techniques for breast cancer risk forecasting and diagnosis. Wisconsin Breast Cancer (Original) dataset was used to apply SVM, kNN, NB, and DT (C4.5). When the experimental data were compared, the SVM classification approach had the best exactness (97.13 percent) and the lowest error rate. The dataset for this study was Breast Cancer, and the Machine Learning tool was Weka. In conditions of exactness, recall, precision, and ROC area, the key presentation parameters of machine learning categories are evaluated. They claim that BN have the greatest remind and exact values, and that the RF approach is the best ROC area [5]. Ahmad t al. has implement ML techniques for calculating the velocity of 2 years repetition of breast cancer disease. The data was taken from the ICBC programmed and spans the years 1997 to 2008. The dataset includes population characteristics and 22 input factors, as well as cases from 1189 women are diagnosed by breast cancer. ANN, SVM, and DT are utilized by SVM presenting the top consequences in conditions of exactness and error rate.

2. LITERATURE REVIEW

To determine which machine learning methods perform better, we used SVM and ANN methodologies to forecast the categorization of breast cancer. Vladimir Vapnik was the first to explain Support Vector Machines (SVMs), and their good performance has been recognized in a variety of pattern recognition challenges. While comparing a lot of other categorization algorithms, SVMs could imply greater categorization results. SVM is mainly widely utilized machine learning categorization technique for cancer diagnosis and forecasting. The module is split by means of SVM's hyper plane that is through support vectors to represent significant samples by all modules. The hyper plane is a divider among the 2 example clusters to serve as a conclusion boundary. Depending on the patient's age and tumour size, SVM can be used to categories tumors as benevolent or malevolent. The biological neuron system can be used to describe the ANN. It is particularly comparable to the human brain's processing structure. It is prepared up of a huge number of nodes that are connected to one another [12]. Modeling distinctive and influential non-linear utility is possible with ANN. It's made up of a system of many artificial neurons.

Forecasting of Breast Cancer Pragya Chauhan and Amit Swami [2] created a method for discovering the Breast cancer forecasting is an open topic of investigate using a Genetic Algorithm Based Ensemble method in this paper. This research uses a variety of machine learning methods to notice Breast Cancer Forecasting. DT, random forests, SVM, neural networks, linear models, adabost, and naive bayes are examples of forecasting algorithms. An ensemble technique is worn to improve the exactness of breast cancer forecasting. The classification dataset is used to create a GA-based average ensemble strategy that overcomes the boundaries of the traditional weighted average method. A genetic algorithm based weighted average was worn. A variety of representation the genetic algorithm exceeds slanted regular methods in a assessment of particle swarm optimization (PSO), deferential evolution (DE), & genetic algorithm (GA) and added assessment is made among the conventional ensemble technique and the GA based weighted average technique and it is accomplished that GA based weighted average technique outputs.

Priyanka Gandhi and Pro. Shalini L [4] of VIT, Vellore, and examined machine learning methods for breast cancer calculation and in this investigation; machine learning methods are investigated in arrange to improve diagnosis exactness. CART, RF, and KNN are among the methods compared. The dataset is attained by the University of California, Irvine Machine Learning Repository. The KNN algorithm is found to perform significantly better than the other strategies tested. K-Nearest Neighbor is mainly exact model. Random Forest and Boosted Trees, two categorization models, had similar exactness. As a result, the most accurate classifier can be used to diagnose the tumor and find a cure at an premature period.

Applications of Machine Learning in Breast Cancer WY and Z Wang's Diagnosis and Prognosis [9] they explicate several ML approaches and their purpose in BC analysis and forecasting in this study, and they used the benchmark to analyze the data. The unexpected ability of 49 It has been established that there are ways to increase categorization and forecasting exactness. Even though several algorithms in WBCD have achieved exceptionally high exactness, new methods are still needed. Although precision in classification is crucial, it is not the only one. Dissimilar techniques employ dissimilar techniques and take into account dissimilar factors. For decades, ANNs have conquered BC diagnosis and prognosis, but this is changing as machine learning techniques are used to intellectual healthcare systems to present a mixture of possibilities to physicians.

Usage of Machine Learning for Breast Cancer prophecy

We are by means of ML approach to identify, categorize, recognize, or differentiate tumor and dissimilar malignancies. In alternative terms ML is mainly utilized as AN aid to cancer identification and recognition. It has solely be comparatively in recent times that cancer researchers are tried to use machine learning towards cancer prophecy and forecasting. As a result the body of prose within the field of machine learning and cancer forecasting/prophecy is comparatively little. In Wisconsin Diagnostic Breast Cancer (WDBC), Machine learning technique is utilized to predict Breast cancer tumour category (benevolent or malevolent) via analyzing mixtures of various standards in key attributes. Due to the variations among entity patients, although they comprise a

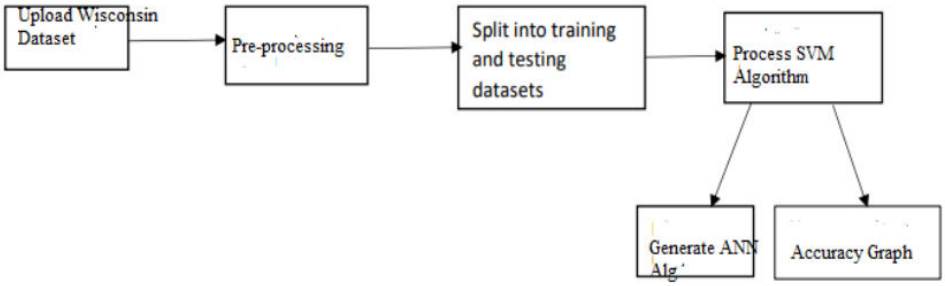
similar Breast cancer category, the values of various attributes won't be identical. Depending on these values, identical Breast cancer tumor category is separated into dissimilar serious circumstances by physicians. Therefore to help cancer forecasting, like calculate cancer inclination, survivability, development and shortly. Consecutively, efficient treatments are often given to the patients.

3. PROPOSED SYSTEM

This system compares the following ML techiques: SVM, DT, RF, NB, and KNN search. The Wisconsin datasets are utilised to create the data collection. The dataset is separated by preparing testing sets in order to implement the machine learning methods. There will be a comparison of all five algorithms. The technique that produces the finest consequences would be provided as a representation to PyCharm, a Python tool. The UCI Machine Learning Repository has this data set available. It is made up of 32 multivariate real-world properties. The total number of cases in this data collection is 569, and there are no missing values.

Design Methodology

System Architecture The system architecture for identifying the user as genuine is shown in the below Fig.1



In this the input register data is pre processed to eliminate any inconsistencies or null values. The data is then divided keen on training and testing segments in order to fit it keen on the classifier. The SVM Algorithm is trained using training data and then experienced with assessted data to determine correctness. As a result, the user is classified as authentic.

On the Wisconsin Diagnostic Breast Cancer (WDBC) dataset that is derived by a digital image of an MRI this study compares five machine learning (ML) techniques: Naive Bayes (NB), Random Forest (RF), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Decision Tree (DT). For the execution of the Machine Learning techniques, the dataset is separated into the training section and also the testing section. The techniques by the most effective consequences are accustomed to categorize the cancer as benevolent or malevolent. Concerning the results of exactness the potency of every algorithm is measured and compared. These techniques are coded in python and executed using a python experimental tool pyCharm.

Upload Dataset

Upload Dataset is a method of introducing raw data sets keen on your investigative proposal is the procedure. Databases, remote data , scripting languages, NoSQL storage and other sources can all provide it. The process of uploading a dataset entails identifying data sets, retrieving data, and querying data from the dataset. The dataset used in the project is collected from Wiconsin Data set . We used additional tools to get other information, such as, server country with Who is. The final dataset consists of around 1780 values which are capable of give out as a training set for Machine Learning models

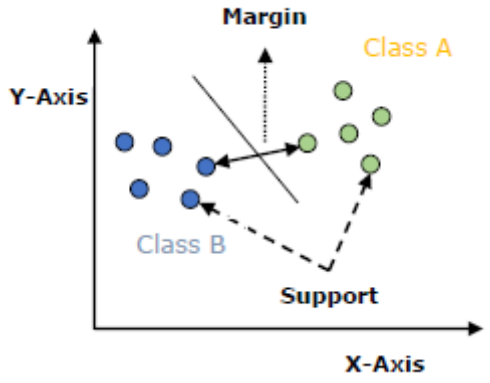
No	Attribute	Value
1	Sample code number	Sample code number
2	Clump Thickness	1 – 10
3	Uniformity of Cell Size	1 – 10
4	Uniformity of Cell Shape	1 – 10
5	Marginal Adhesion	1 – 10
6	Single Epithelial Cell Size	1 – 10
7	Bare Nuclei	1 – 10
8	Bland Chromatin	1 – 10
9	Normal Nucleoli	1 – 10
10	Mitoses	1 – 10
11	Class	2 = benign, 4 = malignant

Table.1. Wiconsin Breast cancer Data sets

Wisconsin Breast Cancer Dataset was used in this investigation (WBCD). Dr. William H. Wolberg of the University of Wisconsin compiled this dataset. It was taken from the Machine Learning Repository at UC Irvine. A digital image of a fine needle aspirate (FNA) of a breast crowd is worn to analyze the dataset's features. They describe the features of the image's cell nuclei. There are 699 cases in this dataset, two classes (benevolent and malevolent), and nine integer-valued attributes (observe Table 1). The value 10 denotes the mainly out-of-the-ordinary situation. We clean up the data by deleting 16 occurrences when a value is absent. As a result, we now have a new dataset of 683 cases. The revised dataset's class distribution is as follows: benevolent = 444 (65.01 percent), malevolent = 239 (34.99 percent). observe Figure 3 for distribution of class and attributes.

4. SUPPORT VECTOR MACHINE (SVM)

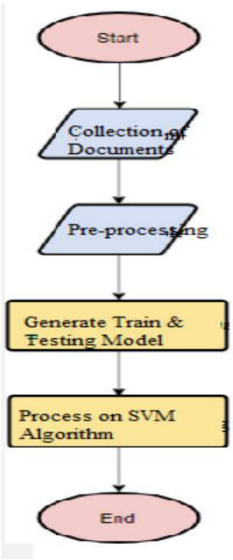
SVMs (Support Vector Machines) are a type of education representation so as to commonly utilized in conjunction with other learning systems for categorization and weakening investigation. In a multidimensional setting, SVM operates by acting as a linear divider among 2 data points to recognize 2 dissimilar modules. By determining a linear classifier, this approach creates a line to divide the 2 modules. The optimum hyper plane separator is the name given to this separation. That is chosen by the set of hyper planes in order to categorize outlines that maximize the hyper plane boundary, or the space between the hyper plane and the pattern's nearest end. On both sides of the ideal hyper plane separator, SVM creates 2 parallel hyper planes. It is assumed to facilitate the superior the margin or distance among the parallel hyper planes gives enhanced simplification error for catogarizer.



SVM representation

4.1 FLOW CHART

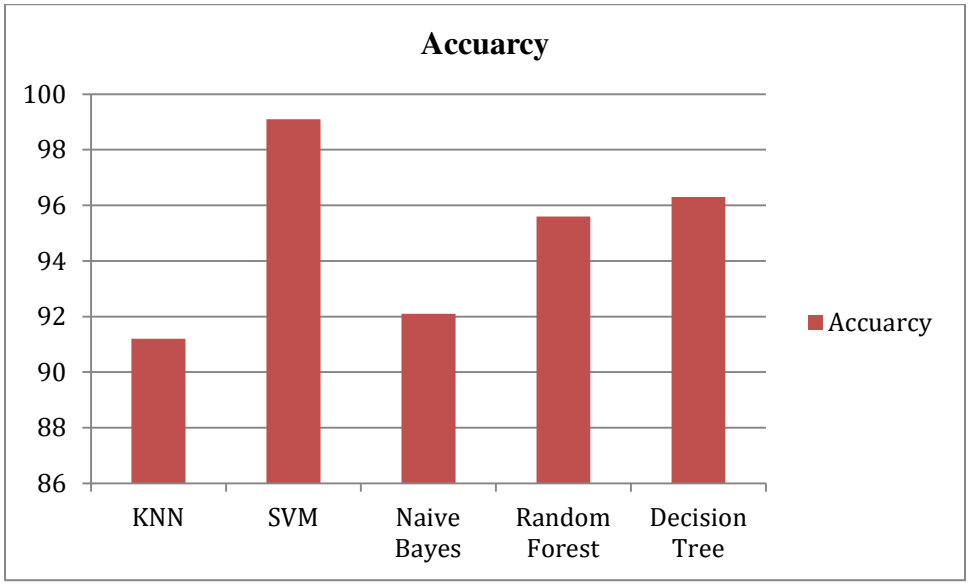
A flowchart is figures that demonstrate how a method, system, or computer method works. Here we utilize in a wide assortment of meadow to text, analyze, sketch, get better, and correspond frequent composite method in clear, straightforward figure. Flowcharts, moreover recognized as flow charts, are figures that utilize rectangles, ovals, diamonds, and perhaps new figures to represent the sort of step, plus linking arrows to show flow and progression.



4.2 EXPECTED OUTCOMES AND CONVERSATION:

Present work assessment and evaluation of every technique is executed depending on the amount of exactness, exactness and evoke.

Algorithm	Exactness
KNN	91.2
SVM	99.1
Naive Bayes	92.1
Random Forest	95.6
Decision Tree	96.3



The amount of sample correctly forecasted separated by the entire amount of samples in the data set yields exactness, which is determined as the ratio of the amount of tested models properly forecasted separated by the whole amount of tested models in the data set.

CONCLUSION

This research examines two popular machine learning techniques for categorising Wisconsin breast cancer. The use of analytical data to predict and detect breast cancer is becoming more and more appealing to develop in the current sector of health. Early identification can reduce the mortality rate from breast cancer. According to recent studies, machine learning algorithms have a critical role to play in the detection of breast cancer. This article uses three well-known machine learning techniques to detect breast cancer. Three examples are MLP, SVM, and Random Forest. The Wisconsin Breast Cancer Diagnostic dataset is used to compare the suggested tactics. The findings of this study show that when using the k-fold cross justification technique, the SVM algorithm provides the best exactness.

FUTURE ENHANCEMENT

In the WEKA tool, ML approaches like ANN and SVM are exploited to categorize the WBC (Original) dataset. In terms of significant presentation parameters counting exactness, exactness, evokes, and ROC region, the utility of appropriate ML methods is compared. SVM (Sequential Minimal Optimization method) is revealed the paramount presentation in the exactness of 96, 9957 percent for analysis and prophecy by the WBC dataset, depending on the presentation metrics of the utilized ML methods.

REFERENCES

1. Siegel, R. L., Miller, K. D., & Jemal, A. (2018). Cancer statistics, Ca-a Cancer Journal for Clinicians, 68 (1), pp. 7-30.
2. Maity, N. G., & Das, S. (2017). Machine learning for improved diagnosis and prognosis in healthcare. In 2017 IEEE Aerospace Conference, pp. 1-9.
3. Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk forecasting and diagnosis. Procedia Computer Science, 83, pp. 1064-1069.
4. Bazazeh, D., & Shubair, R. (2016). Comparative study of machine learning algorithms for breast cancer detection and diagnosis. In 2016 5th International Conference on Electronic Devices, Systems and Applications, pp. 1-4

5. Huang, M. W., Chen, C. W., Lin, W. C., Ke, S. W., & Tsai, C. F. (2017). SVM and SVM ensembles in breast cancer forecasting. PloS one, 12 (1).
6. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and forecasting, Computational and structural biotechnology Journal, 13, pp. 8-17.
7. Umadevi, S., & Marseline, K. J. (2017). A survey on data mining classification algorithms. In 2017 International Conference on Signal Processing and Communication, pp. 264-268.
8. Padmapriya S., Devika M., Meena V., Dheebikaa S. B., & Vinodhini R. (2016). Survey on Breast Cancer Detection Using Weka Tool. Imperial Journal of Interdisciplinary Research (IJIR), Vol 2, no. 4.
9. S. Turgut, M. Dagtekin, T. Ensari, Microarray Breast Cancer Data Classification Using Machine Learning Methods, Int. Conf. on Electric Electronics, Computer Science, Biomedical Engineerings' Meeting, DOI: 10.1109/EBBT.2018.8391468, April 18-19, 2018.
10. M. Amrane, S. Oukid, I. Gagaoua, T. Ensari, Breast Cancer Classification Using Machine Learning, Int. Conf. on Electric Electronics, Computer Science, Biomedical Engineerings' Meeting, DOI: 10.1109/EBBT.2018.8391453, April 18-19, 2018.
11. K. Ncibi, T. Sadraoui, M. Faycel, and A. Djenina, "Multilayer perceptron artificial neural networks based a preprocessing and hybrid optimization task for data mining and classification," Int. J. Econom. Financ. Manag., vol. 5, no. 1, pp. 12–21, 2017, doi: 10.12691/ijefm-5- 1-3.
12. S. J. Livingston, B. S. T. Selvi, M. Thabeetha, C. P. Grena, and C. S. Jenifer, "A neural network based approach for sentimental analysis on amazon product reviews," Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 6S, pp. 469–473, 2019,
13. H. Ramchoun, M. A. J. Idrissi, Y. Ghanou, and M. Ettaouil, "Multilayer perceptron: architecture optimization and training," Int. J. Interact. Multimed. Artif. Intell., vol. 4, no. 1, pp. 26–30, 2016, doi: 10.9781/ijimai.2016.415.
14. W. Castro, J. Oblitas, R. Santa-Cruz, and H. Avila-George, "Multilayer perceptron architecture optimization using parallel computing techniques," PLoS One, vol. 1
15. M. F. Ak, "A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications," Healthc., vol. 8, no. 2, p. 111, 2020, doi: 10.3390/healthcare8020111.
16. M. Akram, M. Iqbal, U. Daniyal, and A. U. Khan, "Awareness and current knowledge of breast cancer," Biol Res, vol. 50, no. 1, p. 33, 2017, doi: 10.1186/s40659-017-0140-9.
17. A. Idri, I. Chlioui, and B. El Ouassif, "A systematic map of data analytics in breast cancer," in Proceedings of the Australasian Computer Science Week Multiconference, 2018, p. 10, doi: 10.1145/3167918.3167930.
18. Y. Li and Z. Chen, "Performance evaluation of machine learning methods for breast cancer forecasting," Appl. Comput. Math., vol. 7, no. 4, pp. 212–216, 2018, doi: 10.11648/j.acm.20180704.15.
19. H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using machine learning algorithms for breast cancer risk forecasting and diagnosis," Procedia Comput. Sci., vol. 83, pp. 1064–1069, 2016, doi: 10.1016/j.procs.2016.04.224.