International Journal of Engineering & Technology, 7 (4.36) (2018) 160-166

SPC

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Evaluating the Performance of Machine Learning Techniques in the Classification of Wisconsin Breast Cancer

Omar Ibrahim Obaid [1*], Mazin Abed Mohammed[2], Mohd Khanapi Abd Ghani [3], Salama A. Mostafa [4], Fahad Taha AL-Dhief [5]

[1]Department of Computer, College of Education, AL-Iraqia University, IRAQ
[2]Planning and Follow Up Department, University Headquarter, University of Anbar, Anbar, Iraq
[3] Biomedical Computing and Engineering Technologies (BIOCORE) Applied Research Group, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Malaysia
[4] Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Johor, Malaysia
[5] Faculty of Electrical Engineering, Department of Communication Engineering, Universiti Teknologi Malaysia, UTM Johor Bahru, Johor, Malaysia
**Corresponding Author Email: alhamdanyomar23@gmail.com*

Abstract

Breast cancer is a considerable problem among the women and causes death around the world. This disease can be detected by distinguishing malignant and benign tumors. Hence, doctors require trustworthy diagnosing process in order to differentiate between malignant and benign tumors. Therefore, the automation of this process is required to recognize tumors. Numerous research works have tried to apply the algorithms of machine learning for classifying breast cancer and it was proven by many researchers that machine learning algorithms act preferable in the diagnosing process. In this paper, three machine-learning algorithms (Support Vector Machine, K-nearest neighbors, and Decision tree) have been used and the performance of these classifiers has been compared in order to detect which classifier works better in the classification of breast cancer. Furthermore, the dataset of Wisconsin Breast Cancer (Diagnostic) has been used in this study. The main aim of this work is to make comparison among several classifiers and find the best classifier which gives better accuracy. The outcomes of this study have revealed that quadratic support vector machine grants the largest accuracy of (98.1%) with lowest false discovery rates. The experiments of this study have been carried out and managed in Matlab which has a special toolbox for machine learning algorithms.

Keywords: *Breast Cancer; Machine Learning; Accuracy; Classification; Support Vector Machine; Decision Tree; k-Nearest Neighbors; Wisconsin Breast Cancer (Diagnostic) Dataset*

1. Introduction

The rate of death is very high because of breast carcinoma. As per WHO (World Health Organization), the breast carcinoma affects more than 1.5 million women every year around the world [1]. Breast carcinoma was first distinguished in Egypt in nearly 1600 BC, is one of the most established known kinds of carcinoma [2]. Breast carcinoma can be diagnosed by detecting tumors. Malignant and benign are two different kinds of tumors. Doctors require an active determination technique to recognize these tumors. But mostly, it is exceptionally hard to recognize tumors even by the specialists [3].

Thus, an automatic method is required in order to detect the tumors. Numerous researchers have endeavored to apply machine learning techniques for identifying survivability of carcinoma in people and it is additionally been demonstrated by the researchers that these techniques work better in recognizing carcinoma diagnosis [3]. Normally, the detection precision of a patient relies upon a doctor's practice and his expertise [4].

Yet, this skill is developed over numerous long periods of perceptions of various patients' side effects and affirmed diagnosis. Even though, the reliability still can't be ensured. With the coming of processing advancements, it is presently moderately simple to gain and store a considerable measure of data, for instance the committed databases of electronic patient records [5]. It is unthinkable for health experts to break down these complex datasets without the guide of computer especially when undertaking complex examinations of the data. The following figures show the breast cancer tumors in two different types of images, the mammography and ultrasound images [6, 7].

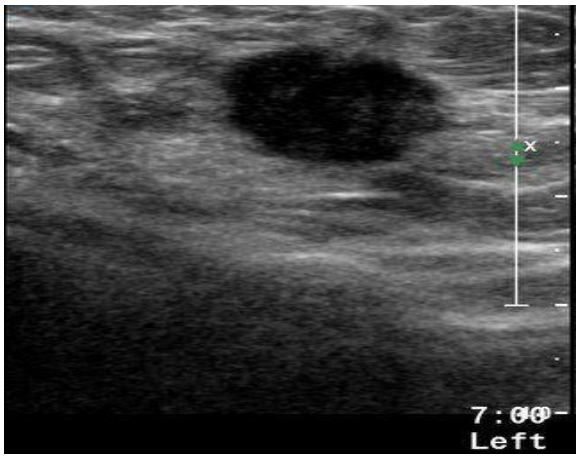

Fig1. Ultrasound Image

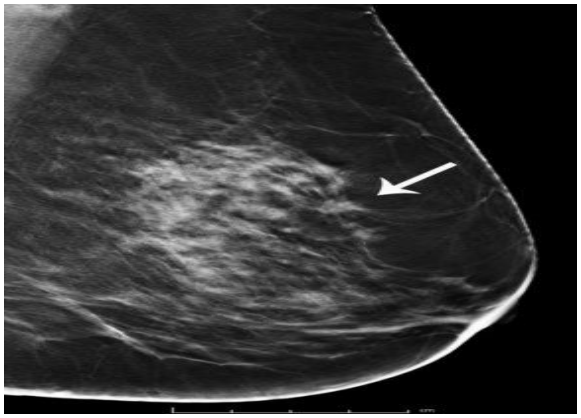

Fig 2. Mammogram Image

Furthermore, an exact classification of malignant tumor can prohibit people sustain dispensable remediation. Therefore, the right determination and classification of breast carcinoma to benign set and malignant set is the topic of frequently study. During the decenniums, ML methods were utilized vastly in order to diagnosis breast carcinoma to earn diverse notions from data patterns. Machine learning is broadly famous as a method in the classification and modeling for breast carcinoma. It is a method that can find already obscure regularities and patterns from assorted datasets. It incorporates a wide assortment of methods utilized for the disclosure of rules, paradigm and connections in groups of data and produces a speculation of these connections that can be utilized to decipher new concealed data. Figure 3 shows the main applications of machine learning in medicine.

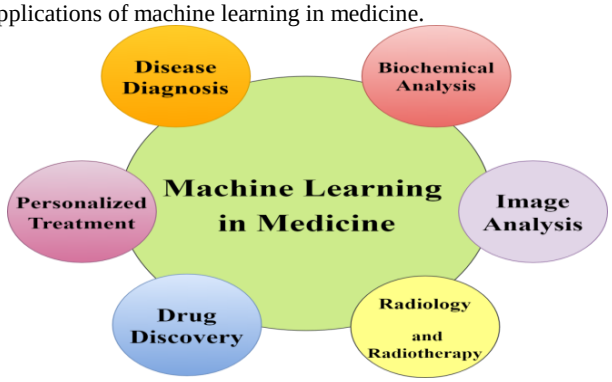

Fig 3. The main application of ML in medicine

The inspiration in the current research mentioned in this study is the outcomes acquired from expansions of a continuous research labor. The work revealed here expands on the starting work by; first, utilizing machine learning methods to think about and comprehend the precise expectation of breast carcinoma illness and it supports doctor to effortlessly distinguish suggestive cures depend on the classification plans or patterns.

Moreover, the main aim of this study is to use multi machine learning techniques for the classification of malignant and benign tumors for the Wisconsin breast cancer diagnostic. This mechanism involves the collection of the whole values of malignant and benign tumors from obtainable dataset. Another point is to develop multi-class models in order to classify the cancerous and non-cancerous tumors. The last point is to compare the performance of multi-classifiers used in this study in order to define the preferable classifier for the classification of breast cancer. The remnant of the research is systematic as follows: Section 2 presents an overview for the breast cancer disease along with anatomy of the breast. In section 3, we discuss the related previous work and show the important of ML techniques in solving the breast cancer disease. Then, section 4 describes the materials and methods used in this study and section 5 displays the experiment methodology of this work. In section 6, we present the outcome of the classifiers used in the study. Then, section 7 shows the best classifier in this study and section 8 gives an extensive discussion for the gained outcomes. Finally, section 9 summarizes the study and show the best result obtained.

2. Related work

The study presented in [8] has developed an intelligent system using support vector machine (SVM) classifier and artificial neural network (ANN) in order to automate breast carcinoma detection. The datasets of Wisconsin diagnostic were utilized in order to implement the model of support vector machine (SVM) to give detection between the benign and malignant breast clusters. The datasets used in this study include measure possessed based on Fine Needle Aspirates (FNA). Frequently work contrasting diverse conventional statistical strategies with conventional Machine Learning (ML) classification strategies were released in order to explain the merits of (ML) and its chance [9].

Recently, with the development and change of the ML methods and the expanding amount and intricacy of the data, outcomes demonstrate that (ML) methods have best classification reliability [10, 11-12]. In the study presented in [13], an ensemble method were used in order to merge various models with a method that the foretell accuracy of every classifier vary from diverse kinds of produces classes. This method merged SVM along with Naïve Bayes and J48 utilizing voting classifier strategy in order to accomplish 97.13 accuracy which is preferable than every of single classifiers.

In the study displayed in [14], a duo-phase-SVM were displayed by merging duo-phase clustering strategy with an effective probabilistic SVM in order to analyze Wisconsin Breast Cancer Diagnosis WBCD and get an accuracy of 99.10% for the classification model. This strategy unlike other methods, it can recognize the figure of the masses and give efficient analyses efficiency for huge body.

3. Materials and Methods

3.1 Materials

In this study, Wisconsin Breast Cancer (Diagnostic) dataset is used. The dataset is obtainable from UCI machine learning repository and the (CSV) file is obtainable from Ref [15]. The (CSV) file then converted into (MAT) file using Matlab. The summarized characterization for the dataset as follows. The dataset has 569

patterns (357 for benign, 212 for malignant) with three classes (ID number, benign, malignant) and 32 columns for the features. The features are registered from a digitized picture of fine needle aspirate (FNA) for the mass of the breast. The dataset does not have any missing attribute values and are coded with four considerable digits. A simulation environment (Matlab 2015a) was utilized for this study. All the experiments of the used classifiers were conducted using the machine learning toolbox (classification learner) which contains a collection of machine learning algorithms.

3.2 Methods

3.2.1 Support Vector Machine (SVM)

SVM classifier is used in this study as it is one of the most methods utilized in breast cancer diagnosis. The term of SVM was first suggested by Vapnik on the foundation of statistical learning theory [16]. It has turned into a main part of machine learning methods. It was basically created for binary sorting (classification). Yet, it can be effectively expanded for multi-class problems [17].
The main advantage of SVM classifier is to discover the improved decision border which exemplifies the greatest decisiveness (maximum margin) amidst the classes. The standard of SVM begins from resolving the problems of linear separable then expands to treat the non-linear cases. A paradigm of SVM framework [18] of breast cancer is shown in figure 4.

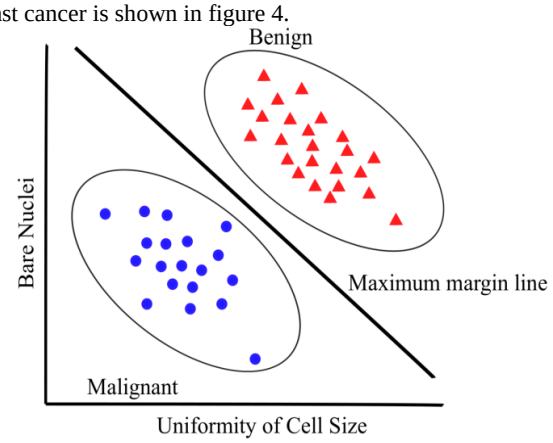

Fig 4. An example of SVM in recognized between malignant and benign cancer

SVM develops a hyperplane that isolates two classes and attempts to accomplish utmost separation between the classes. Isolating the classes with a substantial edge limits a bound on the normal speculation error.

3.2.2 K-Nearest Neighbor (k-NN)

K-Nearest neighbor classifier is used in this study and it is a standout amongst the most focal machine learning strategies in classification [19]. K-Nearest neighbor is non parametric sluggish learning method utilized for classification. This classifier assorts the things utilizing their "k" closest neighbors. It treats the neighbors round the thing, not the essential data allocation [20].

The figure 5 which exist in Ref [21] shows the k-NN for breast cancer classification. The blue color in the circle signifies the test pattern; the green color in the triangle signifies the malignant tumor and the pink color in the square signifies the benign tumor.

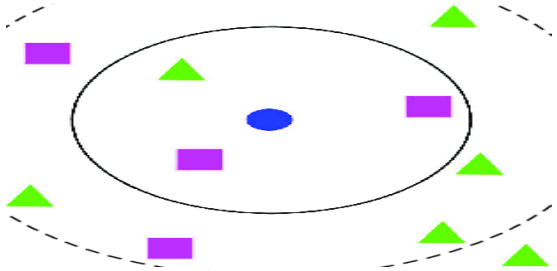

Fig5. K-NN for breast cancer classification

3.2.3 Decision Tree (DT)

The decision tree classifier is used in this study. This classifier comes across as a recursive split of the example space [22]. It is a predictive paradigm that acts as mapping amidst the features of the object and the values of the object [23]. It splits each potential result of the data repeatedly into portions. This classifier is like the flowchart, such that each node which is non-leaf denotes an experiment on special feature, each branch indicates the result of that experiment and each leaf-node comes across a decision or classification [23]. The root-node of the tree is stated at the top of the tree and it coincides to the preferable prediction model. In order to comprehend the work of this classifier in the case of breast cancer diagnosis, figure 6 is a structure of decision tree in Ref [21] and it shows an example of breast cancer diagnosis using decision tree.

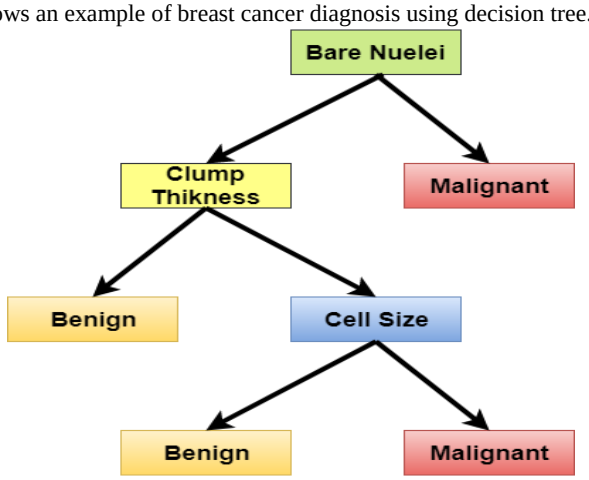

Fig 6. DT structure used in breast cancer classification

4. Experiment Methodology

As mentioned earlier, we have used Matlab R2015a in this study. Several ML classifiers have been used in this study such as SVM, K-NN, and DT. We applied these classifiers on dataset of Wisconsin Breast Cancer (Diagnostic) with 32 features and three classes (ID number, benign, malignant).
The dataset has 569 patterns (357 for benign, 212 for malignant) and it does not have any missing attribute values and are coded with four considerable digits. The figure 7 shows the flow diagram of the proposed breast cancer classification and the figure 8 shows the proposed method of the breast cancer diagnosis.

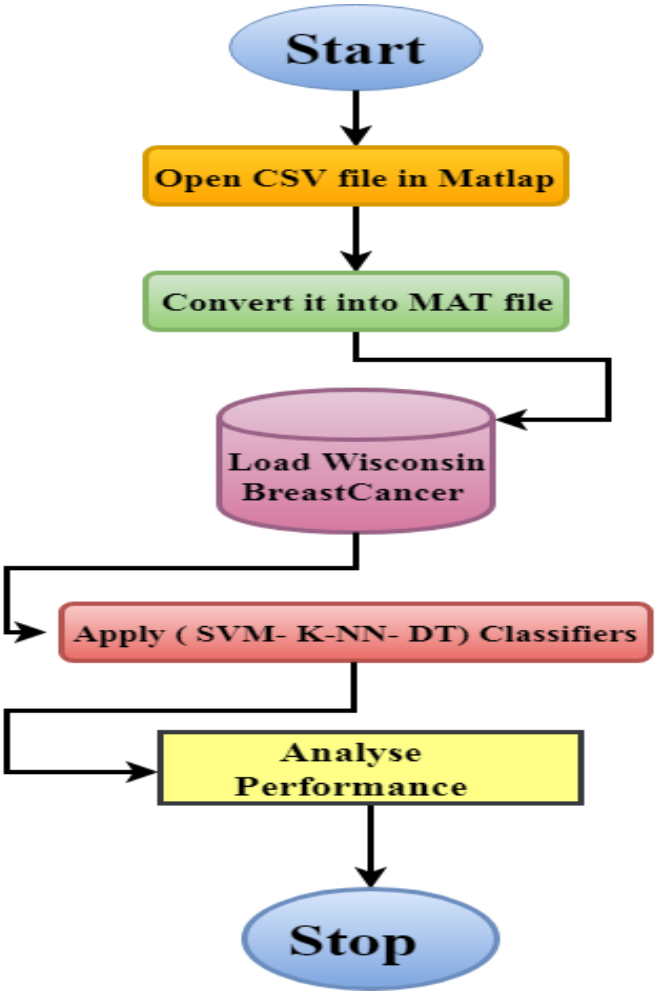

Fig 7. The flow diagram of the proposed breast cancer classification

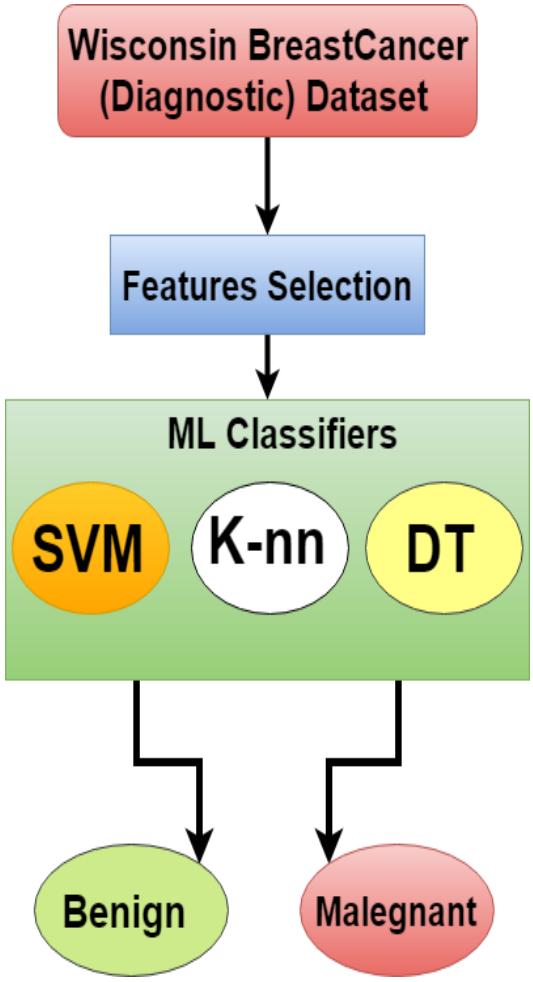

Fig 8. The proposed method of the breast cancer diagnosis

The training of dataset has been done on three classifiers. Firstly, we have trained the dataset using SVM with three types of kernel functions (linear, quadratic, and cubic). Secondly, the dataset has been trained using decision trees with three types of trees (complex, medium, and simple). Finally, the dataset is trained using nearest neighbors classifiers with three kinds (fine, medium, and coarse). On the other hand, the 15-fold cross validation has been carried out in order to test the performance of the three used models. In addition, the accuracy of the classification of the models was compared.

5. Results

The aim of this study is to get the highest accuracy for the various classifiers that we have used in this paper. Furthermore, the accuracy of the three classifiers is compared in order to recognize which classifier works better for the classification of breast cancer. All classifiers with their types are rated based on two standards, the overall accuracy and the time taken to construct the model.

5.1 SVM classifier performance

The performance of this classifier is rated with three kernel functions (linear, quadratic, and cubic). The parameters of the three kernel functions are as follows; the box constraint level is set to 1, the kernel scale mode was set to auto, and the multiclass method was set to one-vs-one. The table 3 displays the outcomes of the classification of the three kernel functions.

Table 3. The results of SVM with three kernel functions

Kernel Function	Accuracy	Time
linear	97.9%	4 sec
quadratic	98.1%	3sec
cubic	97%	4 sec

As noticed in table 3, it is obvious that quadratic kernel function performs (97.9%) accuracy which is better than other kernels. The following figure shows predicted class for quadratic SVM including positive predictive values (PPV) and false discovery rates (FDR).

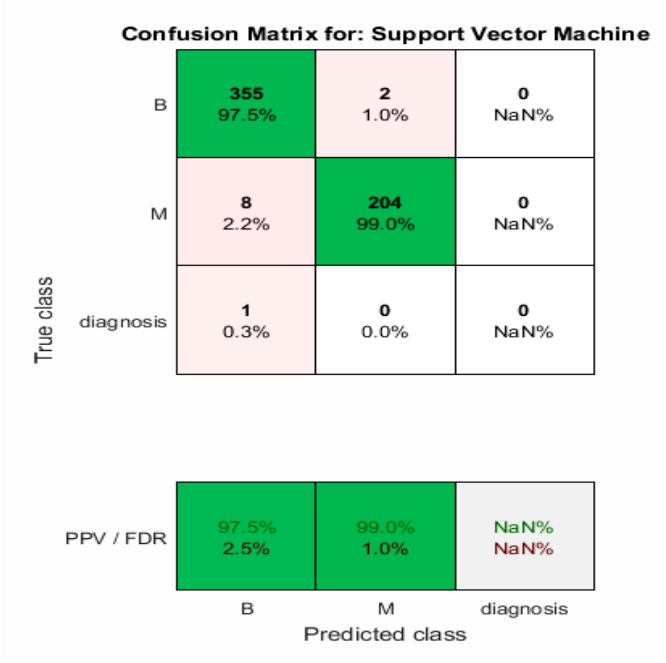

Fig 9. Predicted class for quadratic SVM with PPV and FDR

5.2 K-NN classifier performance

The performance of this classifier is rated with three k-nearest neighbors patterns (fine, medium, and coarse). The parameters of the three k-nearest neighbors are as follows; the number of neighbor was set to 1, the distance metric was set to Euclidean, and the distance weight was set to equal. The table 4 displays the outcomes of the classification of the three k-nearest neighbors.

Table 4. The results of K-NN with three k-nearest neighbors

k-NN Types	Accuracy	Time
Fine	95.4%	2 sec
Medium	96.7%	2 sec
Coarse	93.2%%	2 sec

As seen in table 4, it is clear that Medium k-nearest neighbor performs (96.7%) accuracy which is better than other kinds of k-nearest neighbors. The next figure displays predicted class for Medium k-NN including positive predictive values (PPV) and false discovery rates (FDR).

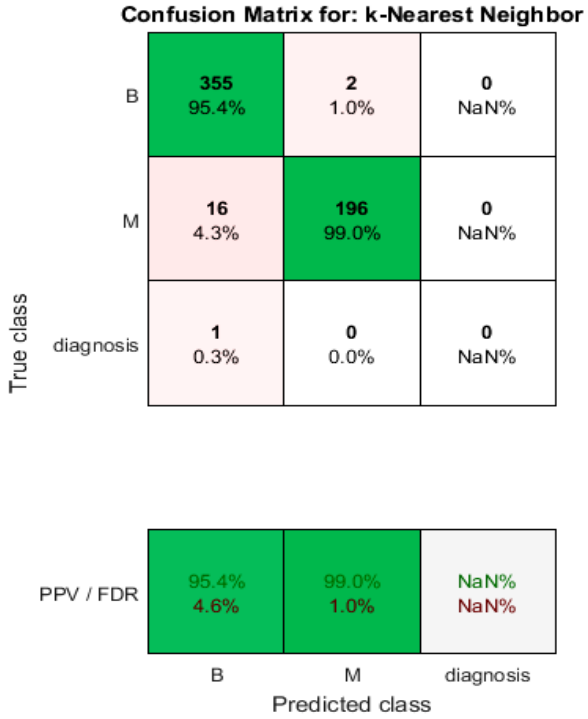

Fig 10. Predicted class for Medium K-NN with PPV and FDR

5.3 DT Classifier Performance

The performance of this classifier is rated with three kinds of trees (complex, medium, and simple). The parameters of the three decision tree are as follows; the Maximum Number of Splits was set to 100, Split Criterion was set to Ginis Diversity Index, and the Surrogate Decision Splits was set to off. Table 5 displays the outcomes of the classification of the three decision tree classifier.

Table 5. The results of DT with three decision tree classifier

DT Types	Accuracy	Time
Complex	93.7%	5 sec
Medium	93.7%	3 sec
Simple	92.3%	2 sec

As seen in table 5, it is axiomatic to notice that complex and medium decision tree performs (93.7%) accuracy but the medium DT performs this rate with less time. Hence, it is better than other DT kinds. The next figure displays predicted class for Medium DT including positive predictive values (PPV) and false discovery rates (FDR).

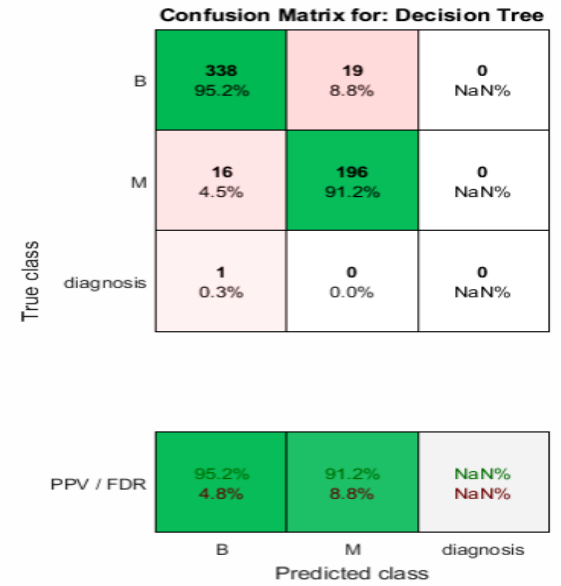

Fig 11. Predicted class for Medium DT with PPV and FDR

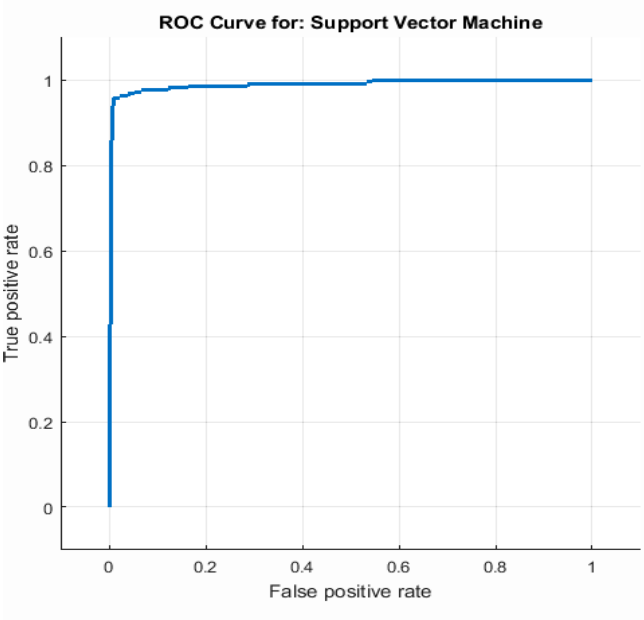

Fig 12. ROC curve with benign (B) positive class

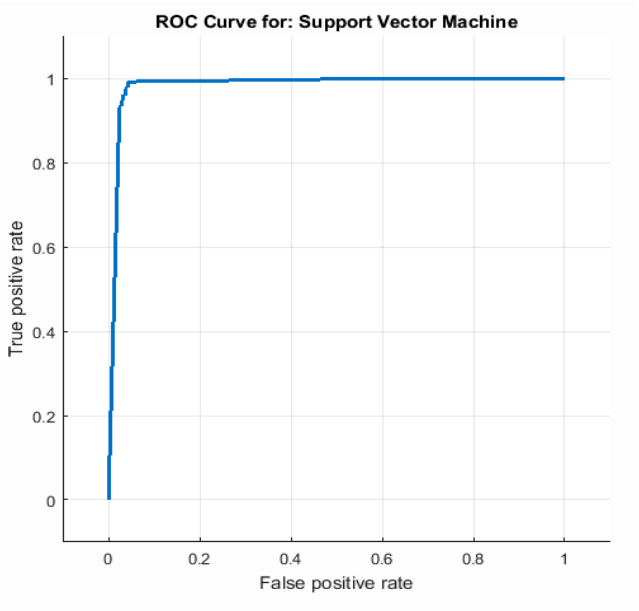

Fig 13. ROC curve with malignant (M) positive class

The following figure shows the view percentage over the entire confusion matrix for quadratic SVM.

.

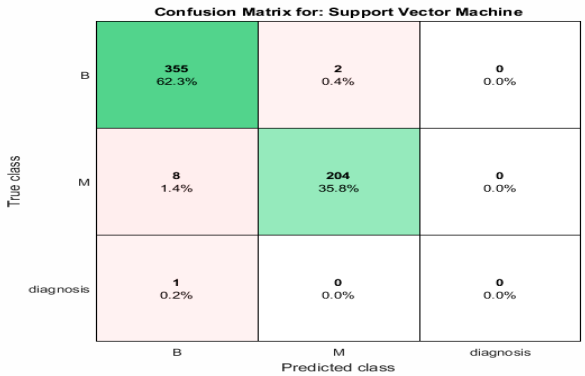

Fig 14. Overall confusion matrix for quadratic SVM

7. Disscusion The Outcomes

6. The Best Classifier

With accordance to the calculating of used classifiers accuracy in tables 3, 4, and 5, we have compared the results in order to acquire the preferable classifier for the classification of breast cancer. From the tables mentioned above, we have obtained the result as quadratic kernel based SVM grants better accuracy which is (98.1%) and the area under the ROC curve of 0.984305 for benign tumor and 0.988352 for malignant tumor which is better than other classifiers. The following figures show the ROC curve for both benign and malignant tumors for the best classifier which is quadratic SVM.

From table 3, one can easily observe from that SVM took about 4.0 seconds in order to construct its prototype which is different from k-NN which took about 2.0 seconds. The truth behind this is that k-NN is a sluggish classifier due to its little work at the training operation which is different from other classifiers which construct the models. Moreover, the overall accuracy acquired by quadratic SVM (98.1%) as seen in table 3 is greater than the overall accuracy gained by other SVM kernels and also greater than k-NN and DT classifiers. From figure 9, we can analyze acquired outcomes in order to assess the performance of ML algorithms used in this paper. The positive predictive values (PPV) and false discovery rates (FDR) of predicted class shown in figure 9 displays that the highest value of PPV which is (99.0%) goes for malignant class with FDR of (1.0%), while the lowest value of PPV which is (97.5%) belongs to benign class with FDR of (2.5%). In table 4, one can readily observe that the best k-NN is the medium k-NN classifer with (96.7%) accuracy. The positive predictive values (PPV) and false discovery rates (FDR) of predicted class shown in figure 10 displays that the

highest value of PPV which is (99.0%) goes for malignant class with FDR of (1.0%), while the lowest value of PPV which is (95.4%) belongs to benign class with FDR of (4.6%).
On the other hand, from table 5 one can easily noticed that the best DT is the medium DT classifier with accuracy of (93.7%). The positive predictive values (PPV) and false discovery rates (FDR) of predicted class shown in figure 11 displays that the highest value of PPV which is (95.2%) goes for benign class with FDR of (4.8%), while the lowest value of PPV which is (91.2%) belongs to malignant class with FDR of (8.8%). In brief, SVM in general and quadratic SVM in special have shown their ability in terms of efficiency which is depend on the accuracy and the time taken to construct the model. The outcomes have shown the highest accuracy of (98.1%) in the dataset of Wisconsin breast cancer classification.

8. Conclusion

There are several machine learning algorithms available in order to analyse diverse medical dataset. The main defiance in ML field is to construct precise classifier for medicinal usage. I this paper, three algorithms have been used; SVM, k-NN, and DT on the dataset of Wisconsin breast cancer (Diagnostic). These algorithms have been compared in order to find the best classifier in terms of the accuracy and the time taken to construct the model. Hence, quadratic SVM has reached an accuracy of (98.1%) and surpassed all other classifiers.

References

[1] Mohammed, M.A., Al-Khateeb, B., Rashid, A.N., Ibrahim, D.A., Ghani, M.K.A. and Mostafa, S.A., 2018. Neural network and multi-fractal dimension features for breast cancer classification from ultrasound images. Computers & Electrical Engineering.70,pp.871-882.

[2] Al-Hashimi MMY, Wang XJ. Breast cancer in Iraq, incidence trends from 2000-2009. Asian Pac J Cancer Prev 2014; 15(1): 281–6.

[3] B.M.Gayathri, C.P.Sumathi, and T.Santhanam. Breast Cancer Diagnosis Using Machine Learning Algorithms –A Survey, International Journal of Distributed and Parallel Systems (IJDPS) Vol.4, No.3, May 2013.

[4] Meesad, P.; Yen, G.G. Combined numerical and linguistic knowledge representation and its application to medical diagnosis. IEEE Trans. Syst. Man Cybern. 2003, 33, 206–222.

[5] Pavlopoulos, S.A.; Delopoulos, A.N. Designing and implementing the transition to a fully digital hospital. IEEE Trans. Inf. Technol. Biomed. 1999, 3, 6–19.

[6] Mohammed, M.A., Ghani, M.K.A., Arunkumar, N., Hamed, R.I., Abdullah, M.K. and Burhanuddin, M.A., 2018. A real time computer aided object detection of nasopharyngeal carcinoma using genetic algorithm and artificial neural network based on Haar feature fear. Future Generation Computer Systems, 89, pp.539-547.

[7] Radiology & Imaging. (2018). Breast Cancer Screening with 3D Mammography or Tomosynthesis - Radiology & Imaging, MA, CT. [online] Available at: https://www.rad-imaging.com/services/womens-imaging/breast-cancer-screening-3d-mammography-tomosynthesis/ [Accessed 5 Sep. 2018].

[8] Ilias Maglogiannis, E Zafiropoulos "An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers" Applied Intelligence, 2009 – Springer.

[9] Mohammed, M.A., Ghani, M.K.A., Arunkumar, N., Hamed, R.I., Mostafa, S.A., Abdullah, M.K. and Burhanuddin, M.A., 2018. Decision support system for nasopharyngeal carcinoma discrimination from endoscopic images using artificial neural network. The Journal of Supercomputing, https://doi.org/10.1007/s11227-018-2587-z.

[10] Delen, D.; Walker, G.; Kadam, A. Predicting breast cancer survivability: A comparison of three data mining methods. Artif. Intell. Med. 2005, 34, 113–127.

[11] Mostafa, S.A., Mustapha, A., Khaleefah, S.H., Ahmad, M.S. and Mohammed, M.A., 2018, February. Evaluating the Performance of Three Classification Methods in Diagnosis of Parkinson's Disease. In International Conference on Soft Computing and Data Mining (pp. 43-52). Springer, Cham.

[12] Abdulhay, E., Mohammed, M.A., Ibrahim, D.A., Arunkumar, N. and Venkatraman, V., 2018. Computer aided solution for automatic segmenting and measurements of blood leucocytes using static microscope images. Journal of medical systems, 42(4), p.58.

[13] Kumar, U.K.; Nikhil, M.B.S.; Sumangali, K. Prediction of breast cancer using voting classifier technique. In Proceedings of the IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials, Chennai, India, 2–4 August 2017.

[14] Osman, A.H. An enhanced breast cancer diagnosis scheme based on two-step-SVM technique. Int. J. Adv. Comput. Sci. Appl. 2017, 8, 158–165.

[15] "Breast Cancer Wisconsin (Diagnostic) Data Set | Kaggle", Kaggle.com, 2018. [Online]. Available: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data. [Accessed: 06- Sep- 2018].

[16] Vapnik, V.N. An overview of statistical learning theory. IEEE Trans. Neural Netw. 1999, 10, 988–999.

[17] Lee, Y.-J.; Mangasarian, O.L.; Wolberg, W.H. Breast cancer survival and chemotherapy: A support vector machine analysis. DIMACS Ser. Discret. Math. Theor. Comput. Sci. 2000, 55, 1–20.

[18] Cortes, C.; Vapnik, V. Support-vector networks. Mach. Learn. 1995, 20, 273–297.

[19] Moreno-Seco, F.; Micó, L.; Oncina, J. A modification of the LAESA algorithm for approximated k-NN classification. Pattern Recognit. Lett. 2003, 24, 47–53.

[20] Mohammed, M.A., Ghani, M.K.A., Hamed, R.I. and Ibrahim, D.A., 2017. Review on Nasopharyngeal Carcinoma: Concepts, methods of analysis, segmentation, classification, prediction and impact: A review of the research literature. Journal of Computational Science, 21, pp.283-298.

[21] Wenbin Yue, Zidong Wang, Hongwei Chen, Annette Payne, Xiaohui Liu. "Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis". 9 May 2018.

[22] Mohammed, M.A., Ghani, M.K.A., Hamed, R.I. and Ibrahim, D.A., 2017. Analysis of an electronic methods for nasopharyngeal carcinoma: Prevalence, diagnosis, challenges and technologies. Journal of Computational Science, 21, pp.241-254.

[23] De Mántaras, R.L. A distance-based attribute selection measure for decision tree induction. Mach. Learn. 1991, 6, 81–92.