Analysis of Breast Cancer Classification with Machine Learning based Algorithms

Abdoulaye Bah
Department of Computer Science and Engineering
Istanbul Sabahattin Zaim University
Istanbul, Turkey
0000-0002-8546-225X

Muhammed Davud
Department of Computer Science and Engineering
Istanbul Sabahattin Zaim University
Istanbul, Turkey
0000-0002-6864-2339

2022 2nd International Conference on Computing and Machine Intelligence (ICMI) | 978-1-6654-7483-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICMI55296.2022.9873696

***Abstract*—Nowadays experienced radiologists can perform successful detection of malignant tumors by examining the histological images or patients' data. However, experts may have different diagnosis or decisions about the type of cancer. Recently, breast cancer has become a trend topic because of its mortality rate affected by this disease. With the improvement of computer aided systems, specialist can benefit from more accurate results and therefore detect the cancer and apply the required treatment in early stages. Knowing the success achieved in Artificial Intelligence field, the biomedical sector has been attracted to this technology as well as its new techniques. Recent studies have proven the ability of artificial intelligence to give accurate results that help specialists make better decisions due to its ability to capture details better than Humans. In this paper, four of Machine Learning algorithms, which are Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Convolutional Neural Networks (CNN), with five different breast cancer datasets are tested and analyzed to verify their performance in a binary classification of breast cancer. The results show that CNN obtained higher accuracy than the other tested algorithms in this type of data. This study will help future researchers in breast cancer field to continue their research and focus on improving the performance of specific algorithms.**

***Index Terms*—Breast Cancer, Machine Learning, KNN, SVM, Random Forest, CNN, Computer Aided Diagnosis, Artificial Intelligence**

I. Introduction

Cancer is currently one of the major causes of mortality globally. For women, breast cancer-related deaths are more common than deaths from other types of cancer due to the disease's annual death toll of thousands of people [1], [2]. The rate of incidence for breast cancer differs by region, from 19.3 per 100,000 women in East Africa to 89.7 per 100,000 women in Western Europe, according to some statistics [3]. It is well known that the amount of new cases has been rising these years and will likely exceed 27 million in 2030 [4]. On the other hand, breast lumps enable us to recognize breast tissue that differs from that found under normal circumstances [5]. Breast ultrasonography, mammography, and biopsy—one of the best diagnostic techniques for determining whether a surface or area is cancerous—are all used in clinical screening. The traditional manual diagnosis done under the microscope needs a lot of expertise, and specialist might have different various, which may lead to diagnostic errors. To solve this, we can use the computer-aided diagnosis (CAD), which automatically classifies data by giving us more accurate results than experts do. Therefore, using Machine Learning will result to a better diagnosis hence a better treatment or early detection of the cancer. In Machine Learning we have many algorithms, but we decided to continue our work with the 4 best performing algorithms which are Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Convolutional Neural Networks (CNN).

The scope of our work is a binary classification which means that as output the algorithm only checks whether the patient has breast cancer or not. Since the invention of pattern recognition and machine learning, numerous handcrafted features-based studies for categorizing breast cancer histology images have been proposed. Nuclei segmentation has been the subject of other investigations, such as [6]. The applicability of decision trees to predict breast cancer was investigated by Shajahaan et al [7]. In [8] authors did deep learning based research on liver disease. In contrast, a recent study [9] with differing findings compared and assessed nine machine learning methods. These algorithms, like have been employed in numerous studies to classify breast cancer tumors with encouraging outcomes. Asri et al. [10], proved that Support vector Machine (SVM) is efficient regarding Breast Cancer predictions by obtaining a good accuracy of 97.13 percent.

The structure of this paper is as follows: In Section I, there is an introduction about breast cancer and the concern of this work. In Section II, we discuss about our 5 datasets obtained from Kaggle website and the algorithms used in this work. Section III explains our experiment and the obtained results. In Section IV we have conclusion and suggestions for future works.

II. Machine Learning Algorithms

Machine Learning is a subpart of artificial intelligence. With machine learning we can build and deploy models that can understand and learn from the given data as input and give output of unseen data, this includes statistics and probability, by recognizing some key patterns and relations in the dataset the machine can learn to predict or classify the data. Usually, we talk about supervised learning when the machine is given a label and trains from that label to give a output. In unsupervised learning we do not have a label, the

machine has to find relations in the data, like clustering for example. Another learning model is Reinforcement Learning where the model is rewarded thus it learns or not rewarded, then the machine can learn anytime it is rewarded, this is used in many applications such as the self-driving cars.

Although there are many Machine Learning algorithms with different properties and capabilities, the focus in this paper will be on four algorithms which are Random Forest, Support Vector Machine, K-Nearest Neighbors and Convolutional Neural Network (CNN). The reason why these algorithms are chosen is that there are the best performing for classification tasks.

A. Random Forest

Random forests is an algorithm mainly used for classification, regression and other tasks that works by constructing plenty of decision trees at training time. Random Forest approach trains a huge amount of decision trees before producing the class that represents the mean of the classes (for classification) or the mean/average prediction (for regression) of the individual trees [11], [12].

B. Support Vector Machine (SVM)

Support-vector machines is another supervised learning model that analyze data for regression and classification in machine learning [13]. Given a set of training examples, each labelled as belonging to one of two categories, an SVM training algorithm develops a model that categorizes new examples into one of two groups, resulting in a non-probabilistic binary linear classifier. SVM is preferred for its computational power and its ability to detect outliers. Apart from that the prediction's precision and its effectiveness in small datasets are among SVM advantages.

C. K-Nearest Neighbors (KNN)

The KNN approach is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. In KNN approach, a new data point is classified depending on how similar it is to previously classified data. It places the new case in the category that is most similar to the existing categories on the premise that the new case/data and previous cases are similar [14].

D. Convolutional Neural Networks (CNN)

CNNs are a subclass of artificial neural networks that use a variety of building blocks, including convolution layers, pooling layers, and fully connected layers, to learn spatial hierarchies of information using backpropagation automatically and adaptively. CNN has been utilized in numerous applications, including natural language processing, medical image analysis, and image segmentation, as a result of its promising findings [15].

III. Experimental Results

To examine the four algorithms, 5 datasets have been used in this paper. These datasets are: Wisconsin diagnostic [16], Breast cancer prediction [17], Breast cancer data [18], Breast cancer [19], and mammography [20] datasets. Data cleaning such as one-hot encoding and removing null rows are applied. The models are run on a jupyter notebook from anaconda. The scikit-learn and keras python libraries are used to implement traditional machine learning models. Tensorflow accuracy is used to evaluate the models' results. Table I shows the details of the used datasets. The numbers of attributes in the table include the class (or label) attribute.

TABLE I
DETAILS OF THE USED DATASETS

Dataset	No. of Attributes	No. of Instances
Wisconsin diagnostic	31*	569
Breast cancer prediction	11	683
Breast cancer data	6	569
Breast cancer	10	116
Mammography	7	11183

*In total, there are 32 attributes, but ID attribute is ignored.

In all experiments, datasets were splitted with 80% of training and 20% of testing. Sklearn library is used to import SVM and KNN Classifiers, while Keras from Tensorflow library is used to import CNN Classifier.

Table II shows the accuracy results of each algorithm using different datasets. It can be noticed that CNN performed better than other approaches for the 5 datasets, which implies that it may be preferred over other algorithms that have been studied in this paper for this type of applications. In line with that, but with slightly lower accuracy values in some cases, we also note that the Random Forest algoritham gives acceptable results in all datasets. To a lesser degree also comes the SVM algorithm, and finally KNN.

TABLE II
DETAILS OF THE USED DATASETS

Dataset	Algorithm			
	Random Forest	***SVM***	***KNN***	***CNN***
Wisconsin diagnostic	91%	90%	84%	95%
Breast cancer prediction	93%	89%	87%	96%
Breast cancer data	89%	87%	85%	89%
Breast cancer	92%	65%	69%	95%
Mammography	96%	96%	55%	98%

Fig.1 shows the relationship between the classification accuracy of the algorithms and the number of instances in each data set. It can be seen that both CNN and Random Forest were not affected as much by the number of instances as KNN and SVM. However, the accuracy of KNN decreases when the number of instances increases or decreases, which means the happining of overfitting or underfitting, respectively. SVM also suffers in the case of a low number of instances, which means that it may suffer from underfitting but not overfitting.

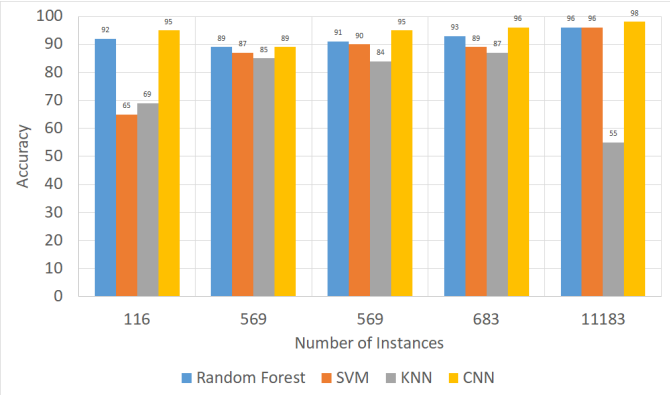

Fig. 1. Accuracy against number of instances.

Tables III-VII shows the confusion matrices for Wisconsin diagnostic, Breast cancer prediction, Breast cancer data, Breast cancer, and Mammography datasets, respectively. Confusion matrices are performed on test data. An outcome where the model accurately predicted the positive class is known as a true positive. Similar to a True Positive, a True Negative is a result when the model accurately foresees the negative class. A False Positive (also known as error type 1) is a result where the model forecasts the positive class wrongly. A False Negative (also known as error type 2) is a result where the model forecasts the negative class wrongly. As an example of the False Negative is when the model predicts that there is no cancer, but there is in fact. Hence, the specialist may not try to cure the cancer based on the model incorrect prediction, which causes a threat to the patient's life. For this reason, the focus here is on False Negative errors.

It can be concluded from Tables III-VII that with the exception of breast cancer data dataset, CNN outperforms the others in that the errors of type 2 (False Negative) are lower than in other algorithms. This supports the preference of CNN over other algorithms under study in this type of applications.

TABLE III
CONFUSION MATRICES FOR WISCONSIN DIAGNOSTIC DATASET

	Random Forest	**SVM**	**KNN**	**CNN**
True Positive	55	47	54	54
True Negative	49	56	42	55
False Positive	7	9	11	4
False Negative	3	2	7	1

TABLE IV
CONFUSION MATRICES FOR BREAST CANCER PREDICTION DATASET

	Random Forest	**SVM**	**KNN**	**CNN**
True Positive	59	48	65	57
True Negative	69	75	55	75
False Positive	6	9	10	3
False Negative	3	5	7	2

TABLE V
CONFUSION MATRICES FOR BREAST CANCER DATA DATASET

	Random Forest	**SVM**	**KNN**	**CNN**
True Positive	55	42	54	57
True Negative	47	60	44	45
False Positive	8	9	10	6
False Negative	4	3	6	6

TABLE VI
CONFUSION MATRICES FOR BREAST CANCER DATASET

	Random Forest	**SVM**	**KNN**	**CNN**
True Positive	11	8	7	12
True Negative	11	8	10	11
False Positive	2	4	4	1
False Negative	0	4	3	0

TABLE VII
CONFUSION MATRICES FOR MAMMOGRAPHY DATASET

	Random Forest	**SVM**	**KNN**	**CNN**
True Positive	980	730	750	1350
True Negative	1180	1420	485	850
False Positive	42	47	692	23
False Negative	35	40	310	14

IV. CONCLUSION

In this paper, a comparative analysis has been conducted on the performance of 4 Machine Learning algorithms (Random Forest, SVM, KNN, and CNN) for breast classification tasks. Results shows that CNN has the best performance over the other experimented algorithms for all the five used datasets. This implies that in terms of classifying breast cancer, it is obvious that using CNN would be a good choice. It should be mentioned that one of the issues in this work is that there is a lack of data for training a solid model. There are probabilities that models that perform well on the available datasets might not be that accurate because of overfitting. In Future works, the models should be validated using clinical data to see how good they are performing. Moreover, doing the same comparative analysis using subclass classification of breast cancer, which will help in knowing the level of the cancer.

REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2017," *CA Cancer J Clin*, vol. 67, no. 1, pp. 7–30, Jan. 2017.

[2] Cancer Genome Atlas Network, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, pp. 61–70, Sep. 2012.

[3] M. M. A. Rahhal, "Breast cancer classification in histopathological images using convolutional neural network," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 3, 2018. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2018.090310

[4] T. Araújo, G. Aresta, E. Castro, J. Rouco, P. Aguiar, C. Eloy, A. Polónia, and A. Campilho, "Classification of breast cancer histology images using convolutional neural networks," *PLoS One*, vol. 12, no. 6, p. e0177544, Jun. 2017.

[5] K. R and N. K, "Automated diagnosis of breast cancer using wavelet based entropy features," *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 274–279, 2018.
[6] M. Kowal, P. Filipczuk, A. Obuchowicz, J. Korbicz, and R. Monczak, "Computer-aided diagnosis of breast cancer based on fine needle biopsy microscopic images," *Comput Biol Med*, vol. 43, no. 10, pp. 1563–1572, Aug. 2013.
[7] S. S. Shajahaan, S. Shanthi, and V. Manochitra, "Application of data mining techniques to model breast cancer data," 2013.
[8] E. Mutlu, A. Devim, A. Hameed, and A. Jamil, "Deep learning for liver disease prediction," pp. 95–107, 01 2022.
[9] N. Al-Azzam and I. Shatnawi, "Comparing supervised and semi-supervised machine learning models on diagnosing breast cancer," *Annals of Medicine and Surgery*, vol. 62, pp. 53–64, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2049080120305604
[10] H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," *Procedia Computer Science*, vol. 83, pp. 1064–1069, 2016, the 7th International Conference on Ambient Systems, Networks and Technologies (ANT 2016) / The 6th International Conference on Sustainable Energy Information Technology (SEIT-2016) / Affiliated Workshops. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050916302575
[11] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324
[12] Y. Ono and Y. Mitani, "Effect of the random forests with recursive feature elimination for breast cancer classification," in *2021 6th International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, vol. 6, 2021, pp. 95–96.
[13] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995.
[14] E. Fix and J. L. Hodges, "Discriminatory analysis. nonparametric discrimination: Consistency properties," *International Statistical Review / Revue Internationale de Statistique*, vol. 57, no. 3, pp. 238–247, 1989.
[15] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into Imaging*, vol. 9, no. 4, pp. 611–629, Aug. 2018.
[16] B. C. W. D. D. Set, "Wisconsin diagnostic," 2016. [Online]. Available: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data
[17] A. Maji, "Breast cancer prediction," 2020. [Online]. Available: https://www.kaggle.com/adhyanmaji31/breast-cancer-prediction
[18] M. S. Suwal, "Breast cancer prediction dataset," 2018. [Online]. Available: https://www.kaggle.com/merishnasuwal/breast-cancer-prediction-dataset
[19] Y. H. Shakir, "Breast cancer coimbra data set," 2020. [Online]. Available: https://www.kaggle.com/yasserhessein/breast-cancer-coimbra-data-set
[20] A. Khan, "Mammography - breast cancer," 2021. [Online]. Available: https://www.kaggle.com/ashrafkhan94/mammography-breast-cancer