IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

REVIEW ON MACHINE LEARNING MODELS FOR EARLY DETECTION OF BREAST CANCER

[1]Uzma Nazir, [2]Er.Mandeep Kaur

[1]M.Tech Student, [2]Assistant Professor,
[1]Computer Science Engineering,
[1]Desh Bhagat University, Mandi Gobindgarh, Fatehgarh Sahib, Punjab- 147301, India

Abstract: Automatic learning (Machine Learning) is part of a branch of intelligence artificial that makes use of a significant number of algorithms developed to give it the possibility for the machine to learn and respond to problematic situations, of which the machine has been trained, using a data set. Machine learning applications abound in pattern identification and data classification, which have been well received in the scientific and industrial community for the implementation of solutions ranging from the detection of anomalies in industrial processes, detection and prediction from fraud to the analysis of texts and their classification. This document presents the development of the implementation of a Machine Learning model to carry out the diagnosis of breast cancer, where, in detail, the antecedents will be shown Regarding the development of existing Machine Learning models in the area of health and especially for the diagnosis of breast cancer.

***Index Terms* – Breast Cancer, Machine Learning, Random Forest, Linear Regression, Logistic Regression, Decision Tree, Support Vector Machine, KNN, K-Means Clustering.**

I. INTRODUCTION

The number of cancer cases in India is very high, according to the 2022 National Cancer Registry Programme, India In India, one in nine people are likely to develop cancer in his/her lifetime. Breast cancer is the leading sites of cancer in females resulted in death [1]. The early prognosis of a breast cancer has become a research necessity since it can facilitate preventive treatment to avoid its lethality in an advanced stage. Therefore, the idea of this investigation arose with the aim of diagnosing the possible condition of breast cancer by comparing the clinical histories of patients with mild symptoms with the of patients with an advanced stage of cancer [2].

The technological and research field, automatic learning or Machine Learning (ML) hat is within the branch of artificial intelligence provides tools and methods that allow analyzing a large amount of data and by means of a regression and classification arrive at a telling result. These techniques have been used by different investigators to model the progression and treatment of cancerous conditions due to their ability to detect significant features in complex data sets [3].

This paper seeks to carry out a review of the different Models used in Machine-Learning for the detection of cancer suggested by different researchers to carry out the implementation of a model in ML to diagnose the possible suffering of breast cancer based on the clinical histories of patients with Malignant or Benign conditions.

II. LITERATURE REVIEW

2.1 Machine Learning

The central axis of this review is based mainly on the prediction techniques used in Machine Learning (ML). ML is the design and study of computer tools that use past experience to make future decisions; It is the study of programs that can learn from data. The fundamental objective of ML is to generalize, or induce an unknown rule from examples where that rule is applied [4]. In order to understand how ML works, it is necessary to know the following variants that make it up [5]:

Supervised learning: In supervised learning problems, the algorithm is taught or trained from data that is already labeled with the correct answer. The larger the data set, the more the algorithm can learn about the subject. After the training is complete, you are provided with new data. Without the labels of the correct answers, the learning algorithm uses the past experience acquired during the training stage to diagnose a result[6].

Unsupervised learning: In unsupervised learning problems the algorithm is trained using a data set that does not have any labels; in this case, the algorithm is never told what the data represents. The idea is that the algorithm can find by itself patterns that help to understand the data set. Unsupervised learning is similar to the method we use to learn to speak when we are babies, at first we listen to our parents talk and we do not understand anything; but as we listen to thousands of conversations, our brain will begin to form a model of how language works and we will begin to recognize patterns and expect certain sounds[7].

Reinforcement learning: in reinforcement learning problems, the algorithm learns by observing the world around it. Your input information is the feedback you get from the outside world in response to your actions. Therefore, the system learns based on trial and error[8].

The algorithms that are most often used in Machine Learning problems are the following:

1. Linear Regression: It is used to estimate actual values (home costs, number of calls, total sales, etc.) based on continuous variables. The idea is to try to establish the relationship between the independent and dependent variables by means of fitting a better straight line with respect to the points[9].

2. Logistic Regression: Linear models can also be used for classifications; that is, we first fit the linear model to the probability that a certain class or category will occur, and then we use a function to create a threshold at which we specify the outcome of one of these classes or categories. ias. The function that uses this model is called logistic regression [10].

3. Decision Trees: Decision Trees are diagrams with logical constructions, very similar to rule-based prediction systems, which serve to represent and categorize a series of conditions that occur successively, for resolution. of a problem. Decision Trees are composed of interior nodes, terminal nodes, and branches emanating from interior nodes [11].

4. Random Forest: The central idea behind the Random Forest algorithm is to build a large number of very shallow trees, and then take the class that each tree chose[12].

5. Support Vector Machines(SVM): The idea behind SVM is to find a plane that separates the groups within the data in the best possible way. Here, spacing means that the choice of plane maximizes the margin between the closest points in the plane; these points are called support vectors[13].

6. K-Nearest Neighbors(KNN): This is a non-parametric classification method, which estimates the value of the posterior probability that an element xx belongs to a particular class from the information provided for the set of prototypes. The KNN regression is calculated simply by taking the average of the point k closest to the point being tested[14].

7. K-Means Clustering: This is probably one of the best known clustering algorithms and, in a broader sense, one of the best known unsupervised learning techniques. K-means is actually a very simple algorithm that works to minimize the sum of the squared distances from the mean within the clustering[15].

2.2 Related References

1. Putting breast cancer patients on the chemotherapy drug: the instance selection approach, the oversampling approach, and the hybrid approach. Investigators evaluate the performance of their approaches and compare them to a reference approach using the Area Under the ROC Curve (AUC) on data from clinical trials, as well as testing the stability of the approaches. The experimental results show the stability of the proposed approaches give the highest AUC with statistical significance[16].

2. The article Machine learning applications in cancer prognosis and prediction presents a review of recent Machine Learning (ML) approaches used in modeling cancer progression. The predictive models discussed in the research are based on various supervised ML techniques, as well as different input features and data samples. Given the increasing trend in the application of ML methods in cancer research, the article presents the most recent publications that employ these techniques as a target for modeling the risk of cancer. cancer or patient outcomes[17].

3. The article Mobile Personal Health Record (mPHR) for Breast Cancer using Prediction Modeling provides predictive decision support tools to assess the level of malignancy of a tumor, the data was obtained from Surabaya Cancer Hospital (Indonesia). . In this research, a mobile personal health record (mPHR) for breast cancer disease is also developed using predictive Machine Learning methods. To achieve this, the researchers perform logistic regression and compare it with the Naive Bayes method used to diagnose the risk of breast cancer-related tumor malignancy[18].

4. The article Breast Cancer Intelligent Diagnosis based on Subtractive Clustering Adaptive Neural Fuzzy Inference System and Information Gain uses a new intelligent method for the diagnosis of breast cancer. The method combines the information gain method and the subtractive clustering adaptive neural fuzzy inference system (IG-SCANFIS). The information gain method is applied to reduce the dimension of the attributes and then applies the selected attributes as input to SCANFIS. The SCANFIS model uses the subtractive clustering algorithm to cluster the input data to obtain the fuzzy rules and establish the neural fuzzy reasoning system. The data sets used for training and testing were obtained from the University of California Machine Learning Irvine (UCI) library. The simulation result shows that the proposed method has a precision of 99.44 %[19].

5. The article On the Scalability of Machine-Learning Algorithms for Breast Cancer Prediction in Big Data Context addresses the problem of breast cancer prediction in the context of Big Data. In it, two varieties of data are considered: Gene expression (GE) and DNA methylation (DM). The objective of the research was to extend the Machine Learning algorithms used for classification by applying each data set separately and jointly. For this purpose, they linked Apache Spark as a platform. In the article three different classification algorithms are selected: Support Vector Machine (SVM), Decision Tree and Random Forest, to create nine models that help diagnose breast cancer. The experimental results showed that the Scaling SVM classifier in the Spark environment outperforms the other classifiers, as it achieved the highest precision[20].

6. In the article Post-Surgical Survival forecasting of breast cancer patient: a novel approach, it is proposed to reduce the percentage of deaths due to breast cancer. In addition to being able to make a diagnosis, it is necessary to keep track of the surgeries that are performed on patients. Tumorectomy and mastectomy are the most commonly used surgical procedures for treatment. This article proposes the use of survival prediction using the support vector machine efficient distributed communication double coordinate climb (SVM)[21].

7. In the article Predicting cancer outcomes from histology and genomics using convolutional networks they implement convolutional neural networks (SCNN) based on histological images and genomic biomarkers for cancer diagnosis[22].

8. The Predicting Invasive Disease-Free Survival for Early Stage Breast Cancer Patients Using Follow-Up Clinical Data investigation seeks to diagnose breast cancer survival in early-stage patients with the analysis of existing clinical data, using the MP4Ei framework, which is a proposal by the authors[23].

9. In the article Prediction Models for Estimation of Survival Rate and Relapse for Breast Cancer Patients, the Naive Bayes classifier is presented as a model for the prognosis of cancer survival based on the 5-to-1 survival rate. years, while the Neural Network, according to the authors, showed better performance in the prognosis of cancer recurrence. They present one of the advantages of using classifiers to perform diagnosis: “Classification systems can help to minimize possible errors that can occur due to inexperienced experts, and also provide medical data for be examined in a shorter and more detailed time”[24].

10. In the article Probabilistic Graphical Models and Deep Belief Networks for Prognosis of Breast Cancer, they implement a (PGM), which is a Bayesian classifier and Manifold Learning for dimensional reduction for the prognosis and diagnosis of c breast cancer that can help physicians make better decisions about the best treatment for a patient[25].

11. The research Prognosis Prediction of Human Breast Cancer by integrating Deep Neural Network and Support Vector Machine integrates a deep neural network with a support vector machine for prognosis of breast cancer using information found in https://gdac.broadinstitute.org where there are genetic transcription datasets of different types of cancer, including breast cancer[26].

12. In the article Using Random Forest Algorithm for Breast Cancer Diagnosis, the Random Forest algorithm for the diagnosis of breast cancer is implemented. It shows how the Random Forest can have better performance among more than 179 tested algorithms. In the investigation, use is made of the Scikit-Learn library for the Python programming language in its version 3.6. The Dataset used is published on the UC Irvine Machine Learning website[27].

13. In the article A Multimodal Deep Neural Network for Human Breast Cancer Prognosis Prediction by Integrating Multi-Dimensional Data, use is made of a Deep Neural Network for the prediction of breast cancer implemented multidimensional data, making use of the library TensorFlow 1.0 and a dataset showing genomic information. For the comparison of Machine Learning models, four parameters are used: Acc (accuracy), Pre (Precision), Sn, and Mcc[28].

14. In the article Predicting Breast Cancer Recurrence using effective Classification and Feature Selection technique three methods were investigated: Support Vector Machine, Decision Tree and Na¨ıve Bayes to diagnose breast cancer recurrence using WEKA, a tool that contains various Machine Learning algorithms. The Dataset used was extracted from UC Irvine Machine Learning, which has 34 attributes. As a result, it can be seen that the Support Vector Machine algorithm has better results than its two competitors[29].

15. The Stacked Regression Ensemble for Cancer Class Prediction research presents a model called Stacked Regression Ensemble (SRE) for the prediction of different cancer classes, comparing its performance with other models such as: SVM (Support Machine of vectors) and GRNN (General regression neural network) focused on the classification and ordering of data referring to cancer[30].

16. For the reduction of dimensionality of the data, in the article A Hybridized Feature Selection and Extraction Approach for Enhancing Cancer Prediction Based on DNA Methylation they propose the implementation of the F-score filter and by showing the results that the methods yield: Na¨ıve Bayes, Random Forest and Support Vector Machine, to classify the different types of cancer including: breast, colon, head, kidney, lung, thyroid, and uterine with and without hybridized approach[31].

17. The article Breast Cancer Prognosis via Gaussian Mixture Regression compares the performance of classification and regression trees (CART), multivariate adaptive regression (MARS) splines, and a regression method Gaussian mixture test (GMR) to diagnose breast cancer recurrence time. It is shown that the GMR-based algorithm demonstrates a performance improvement compared to CART and MARS. In addition, the performance of GMR is comparable to that of a reference predictor with the advantage of performing automatic feature selection and model optimization. The Wisconsin Database (Prognosis Breast Cancer) was used, which is made up of 253 records of patients classified as malignant and are publicly available[32].

18. For the classification of benign and malignant tumors based on a support vector machine (SMV) in the article SVM Approach to Breast Cancer Classification they use a database offered by the University of Wisconsin with images of biopsies. The SVM suite successfully classified more than 99% of the test data and in the process yielded a prediction of benign tumor with 100% accuracy[33].

III. CONCLUSION

The relevance of this research is to analyze the different algorithmic models based on existing Machine Learning for the early diagnosis of breast cancer that make it possible to broaden this field of research and that are a complement in the various studies that are available, this with the aim of helping to reduce deaths from breast cancer. The implementation is mainly focused on the diagnosis of breast cancer based on the information collected through a data set, which after being analyzed and classified by the different Machine Learning models, will have as The result is a prototype with which analysis and subsequent decision-making on the particular case being studied can be carried out. The most important result of this research is to study several existing models that will help significantly in the diagnosis of breast cancer and thus contribute in matters technology to the area of oncology. The result of the investigation, in addition to implementing a model to diagnose the suffering from breast cancer will allow contrasting with the other models suggested by various researchers and may be used in other subsequent investigations to to obtain an increasingly close approximation for the treatment of the various types of cancer that have as reference the use of tools based on Machine learning.

REFERENCES

[1] Gaitonde, Ritika & Lankeshwar-Gajbhiye, Nilima. (2022). EPIDEMIOLOGY OF BREAST CANCER IN INDIA. Asian Journal of Microbiology, Biotechnology and Environmental Sciences. 204-207. 10.53550/AJMBES.2022.v24i01.032.

[2] Harbeck, Nadia & Penault-Llorca, Frédérique & Cortés, Javier & Gnant, Michael & Houssami, Nehmat & Poortmans, Philip & Ruddy, Kathryn & Tsang, Janice & Cardoso, Fatima. (2019). Breast cancer. Nature Reviews Disease Primers. 5. 10.1038/s41572-019-0111-2.

[3] Shaveta,. (2023). A review on machine learning. International Journal of Science and Research Archive. 9. 281-285. 10.30574/ijsra.2023.9.1.0410.

[4] Gaikwad, Anjali. (2023). The Fundamentals of Machine Learning.

[5] Chopra, Deepti & Khurana, Roopal. (2023). Introduction To Machine Learning. 10.2174/9789815124422123010004.

[6] Talukdar, Jyotismita & Singh, Thipendra & Barman, Basanta. (2023). Supervised Learning. 10.1007/978-981-99-3157-6_4.

[7] Sen, Rituparna & Das, Sourish. (2023). Unsupervised Learning. 10.1007/978-981-19-2008-0_21.

[8] Bai, Hui & Cheng, Ran & Jin, Yaochu. (2023). Evolutionary Reinforcement Learning: A Survey.

[9] Maronna, Ricardo & Martin, R. & Yohai, Victor & Salibian-Barrera, Matias. (2019). Linear Regression 1. 87-114. 10.1002/9781119214656.ch4.

[10] Backhaus, Klaus & Erichson, Bernd & Gensler, Sonja & Weiber, Rolf & Weiber, Thomas. (2023). Logistic Regression. 10.1007/978-3-658-40411-6_5.

[11] El Morr, Christo & Jammal, Manar & Ali-Hassan, Hossam & El-Hallak, Walid. (2022). Decision Trees. 10.1007/978-3-031-16990-8_8.

[12] Calhoun, Peter & Su, Xiaogang & Spoon, Kelly & Levine, Richard & Fan, Juanjuan. (2021). Random Forest. 10.1002/9781118445112.stat08287.

[13] Chopra, Deepti & Khurana, Roopal. (2023). Support Vector Machine. 10.2174/9789815124422123010006.

[14] El Morr, Christo & Jammal, Manar & Ali-Hassan, Hossam & El-Hallak, Walid. (2022). K-Nearest Neighbors. 10.1007/978-3-031-16990-8_10.

[15] Easttom, Chuck. (2023). k-Means Clustering. 10.1201/9781003230588-15.

[16] Turki, Turki & Wei, Zhi. (2016). Learning approaches to improve prediction of drug sensitivity in breast cancer patients. Conference proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference. 2016. 3314-3320. 10.1109/EMBC.2016.7591437.

[17] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, and Dimitrios I. Fotiadis. Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal, 13:8–17, 2015.

[18] Badriyah, Tessy & Fauzyah, Rimawanti & Syarif, Iwan & Kristalina, Prima. (2017). Mobile personal health record (mPHR) for Breast Cancer using prediction modeling. 1-4. 10.1109/IAC.2017.8280639.

[19] Liang, Miaomiao & Huang, Lican & Ahmad, Waheed. (2017). Breast Cancer Intelligent Diagnosis Based on Subtractive Clustering Adaptive Neural Fuzzy Inference System and Information Gain. 152-156. 10.1109/ICCSEC.2017.8446831.

[20] AlGhunaim, Sara & Al-Baity, Heyam. (2019). On the Scalability of Machine-Learning Algorithms for Breast Cancer Prediction in Big Data Context. IEEE Access. PP. 1-1. 10.1109/ACCESS.2019.2927080.

[21] Kaushik, Devender & Prasad, Bakshi & Sonbhadra, Sanjay & Agarwal, Sonali. (2018). Post-Surgical Survival Forecasting of Breast Cancer Patient: A Novel Approach. 37-41. 10.1109/ICACCI.2018.8554745.

[22] Mobadersany, Pooya & Lucas, Justin & Govind, Darshana & Aguilar-Bonavides, Clemente & McCarthy, Sharon & Brookman-May, Sabine & Yu, Margaret & Tian, Ken & Hutnick, Natalie & Zamalloa, Jose & Greshock, Joel & Khan, Najat & Yip, Stephen. (2022). Abstract 5053: Artificial intelligence (AI)-based multimodal framework predicts androgen-deprivation therapy (ADT)

outcomes in non-metastatic castration resistant prostate cancer (nmCRPC) from SPARTAN. Cancer Research. 82. 5053-5053. 10.1158/1538-7445.AM2022-5053.

[23] Fu, Bo & Liu, Pei & Lin, Jie & Deng, Ling & Hu, Kejia & Zheng, Hong. (2018). Predicting Invasive Disease-Free Survival for Early Stage Breast Cancer Patients Using Follow-Up Clinical Data. IEEE Transactions on Biomedical Engineering. PP. 1-1. 10.1109/TBME.2018.2882867.

[24] Andjelkovic Cirkovic, Bojana & Cvetkovic, Aleksandar & Ninkovic, Srdjan & Filipovic, Nenad. (2015). Prediction Models for Estimation of Survival Rate and Relapse for Breast Cancer Patients.

[25] Khademi, Mahmoud & Nedialkov, Nedialko. (2015). Probabilistic Graphical Models and Deep Belief Networks for Prognosis of Breast Cancer. 727-732. 10.1109/ICMLA.2015.196.

[26] Sun, Dongdong & Wang, Minghui & Feng, Huanqing & Li, Ao. (2017). Prognosis prediction of human breast cancer by integrating deep neural network and support vector machine: Supervised feature extraction and classification for breast cancer prognosis prediction. 1-5. 10.1109/CISP-BMEI.2017.8301908.

[27] Huang, Zexian & Chen, Daqi. (2021). A Breast Cancer Diagnosis Method Based on VIM Feature Selection and Hierarchical Clustering Random Forest Algorithm. IEEE Access. PP. 1-1. 10.1109/ACCESS.2021.3139595.

[28] Sun, Dongdong & Wang, Minghui & Li, Ao. (2018). A Multimodal Deep Neural Network for Human Breast Cancer Prognosis Prediction by Integrating Multi-Dimensional Data. IEEE/ACM Transactions on Computational Biology and Bioinformatics. PP. 1-1. 10.1109/TCBB.2018.2806438.

[29] Pritom, Ahmed & Munshi, Md. Ahadur & Sabab, Shahed & Zaman, Shihabuz. (2016). Predicting Breast Cancer Recurrence using effective Classification and Feature Selection technique. 10.1109/ICCITECHN.2016.7860215.

[30] Sehgal, M.S.B. & Gondal, Iqbal & Dooley, Laurence. (2005). Stacked regression ensemble for cancer class prediction. 2005 3rd IEEE International Conference on Industrial Informatics, INDIN. 2005. 831 - 835. 10.1109/INDIN.2005.1560481.

[31] Raweh, Abeer & Nassef, Mohammad & Badr, Amr. (2018). A Hybridized Feature Selection and Extraction Approach for Enhancing Cancer Prediction Based on DNA Methylation. IEEE Access. PP. 1-1. 10.1109/ACCESS.2018.2812734.

[32] Falk, Tiago & Shatkay, Hagit & Chan, Wai-Yip. (2006). Breast Cancer Prognosis via Gaussian Mixture Regression. Proc. Canadian Conf. Electrical and Computer Eng. (CCECE '06). 987-990. 10.1109/CCECE.2006.277570.

[33] Sewak, Mihir & Vaidya, Priyanka & Chan, Chien-Chung & Duan, Zhong-Hui. (2007). SVM Approach to Breast Cancer Classification. Proceedings - 2nd International Multi-Symposiums on Computer and Computational Sciences, IMSCCS'07. 32-37. 10.1109/IMSCCS.2007.46.