

1

METHODS & TECHNIQUES

2

Comparative Analysis of Machine Learning Algorithms for Breast Cancer Classification: SVM Outperforms XGBoost, CNN, RNN, and Others

3

Prithwish Ghosh¹ and Debasish Chatterjee²

4

5

ABSTRACT

6

This study evaluates ten machine learning algorithms for classifying breast cancer cases as malignant or benign based on physical attributes. Algorithms tested include XGBoost, CNN, RNN, AdaBoost, Adaptive Decision Learner, fLSTM, GRU, Random Forest, SVM, and Logistic Regression. Using a robust dataset from UCI machine learning Breast Cancer, SVM emerged as the most accurate, achieving 98.2456% accuracy. While AdaBoost, Logistic Regression, Neural Networks, and Random Forest showed promise, none matched SVM's accuracy. These findings underscore the potential of machine learning, particularly SVMs, in cancer diagnosis and treatment by analyzing physical attributes for improved diagnostics and targeted therapies.

7

KEYWORDS: XGBoost, CNN, RNN, AdaBoost, Adaptive Decision Learner, LSTM Networks, GRU, Random Forest Classifier, SVM, and Logistic Regression, Breast Cancer

8

9

INTRODUCTION

10

Breast cancer is a type of cancer that develops in the cells of the breast. It is one of the most common cancers among women worldwide, but it can also affect men, though it's rare. Early detection through screening, such as mammograms, and advances in treatment have significantly improved the prognosis for many people diagnosed with breast cancer [9]. Treatment options typically include surgery, chemotherapy, radiation therapy, hormone therapy, targeted therapy, or a combination of these approaches, depending on the type and stage of the cancer [30].

11

Machine learning (ML) algorithms have revolutionized various scientific fields in recent years. [36] developed a Computer-Aided Diagnosis (CAD) system using Machine Learning (ML) and region-growing segmentation to analyze breast ultrasound images. [2] proposed using B-mode and elastography images for breast cancer detection. Their system utilized 82 ultrasound images, employing geometrical and texture features. [8] developed a CAD system based on morphological features from B-mode ultrasound images. [11] proposed a CAD system employing various classifiers to classify breast ultrasound images based on textures and

12

13

¹Department of Statistics, North Carolina State University 5109, SAS Hall, 2311 Stinson Dr, Raleigh, NC 27607, United States

14

²Department of Statistics, Visva Bharati University Siksha Bhavana (Institute of Science), Santiniketan Bolpur, WB, India, 731235

15

Authors for correspondence: Prithwish Ghosh (pghosh4@ncsu.edu)

16

Received 1 May 2018; revised 5 December 2018

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

morphological features, with Linear Discriminant Analysis (LDA) performing best [20].

Other researchers like [33], [17], [25], and [19] introduced different approaches using SVM, LDA, and Modified Neural Network (MNN) achieving notable accuracies ranging from 75.94% to 97.80%. Additionally, methods by [14] using logistic regression, [4] utilizing XGBoost, and [10] employing morphological features showed promising results with accuracies around 89.40% to 94.0%. Lastly, [15] proposed a deep learning approach combining semantic segmentation and DenseNet201 with SVM.

Objective of the Paper

This study takes aim at a critical question: which machine learning approach reigns supreme in classifying breast cancer based on physical attributes? We intend to assess the effectiveness of machine learning, Neural Networking, and Deep Learning techniques in predicting Breast Cancer Classification based on class. We unleashed a diverse arsenal of ten classification methods, including neural networks and deep learning algorithms, on a robust dataset. This finding underscores the crucial role of choosing the right tool for the job in breast cancer diagnosis. By strategically implementing these techniques, we achieved near-perfect accuracy, a significant leap forward compared to traditional machine learning methods.

THE DATASET

The dataset of breast cancer [34]. Characteristics are derived from a digitized image of a breast mass's fine needle aspirate (FNA), depicting attributes of the cell nuclei within the image. Some sample images can be accessed at <http://www.cs.wisc.edu/~street/images/>.

Attribute Information: ID number , Diagnosis (M = malignant, B = benign) , Ten real-valued features are computed for each cell nucleus: , radius (mean of distances from the center to points on the perimeter) , texture (standard deviation of gray-scale values) , perimeter , area , smoothness (local variation in radius lengths) , compactness ($\text{perimeter}^2 / \text{area} - 1.0$) , concavity (severity of concave portions of the contour) , concave points (number of concave portions of the contour) , symmetry , fractal dimension

METHODOLOGIES

This study embarked on a mission to identify the most effective machine learning warrior in the fight against breast cancer. We assembled an arsenal of ten classification algorithms, each a powerful tool for analyzing physical attributes and distinguishing between malignant and benign tumors.

57

58

59

60

61

62

63

64

65

66

67

68

69

70

71

72

73

74

75

76

77

78

79

80

81

82

83

84

85

86

87

88

89

90

91

92

93

94

95

96

97

98

99

100

101

102

103

104

105

106

107

108

109

110

111

112

For all the algorithms we choose $X = [x_1, x_2, \dots, x_n]$ which represent the input features, where each x_i represents a vector of features including the parameters: Radius (mean of distances from the center to points on the perimeter): $x_t^{(1)}$, Texture (standard deviation of gray-scale values): $x_t^{(2)}$, Perimeter: $x_t^{(3)}$, Area: $x_t^{(4)}$, Smoothness (local variation in radius lengths): $x_t^{(5)}$, Compactness (perimeter² / area - 1.0): $x_t^{(6)}$, Concavity (severity of concave portions of the contour): $x_t^{(7)}$, Concave points (number of concave portions of the contour): $x_t^{(8)}$, Symmetry: $x_t^{(9)}$, Fractal dimension ("coastline approximation" - 1): $x_t^{(10)}$

Let y denote the target variable, which is the Diagnosis (Malignant or Benign)

The Algorithm Legion:

Our ten valiant contenders included:

- (i) Support Vector Machine (SVM) [31]: A veteran classifier known for its ability to find clear boundaries between data points [13].
- (ii) Random Forest: A committee-based approach that leverages the wisdom of multiple decision trees for robust predictions [3] [23].
- (iii) Logistic Regression: A workhorse algorithm that calculates the probability of an outcome based on its features [35] [29].
- (iv) XGBoost is a popular gradient boosting algorithm known for its efficiency and performance in solving regression, classification, and ranking problems by sequentially building a series of decision trees. [22, 9] [5, 21]
- (v) AdaBoost: A champion for boosting the performance of weaker learners by strategically focusing on challenging data points [27] [12, 7].
- (vi) Adaptive Decision Learner: A dynamic approach that tailors decision trees to the specific characteristics of the data.
- (vii) Neural Network Variants: We utilized the power of three neural network architectures [1] [32]:
 - (a) Convolutional Neural Network (CNN): An expert at identifying patterns in image data, even if they're subtly hidden [28].
 - (b) Long Short-Term Memory (LSTM): A master at handling sequential data, potentially useful for capturing the progression of the disease [18].
 - (c) Gated Recurrent Unit (GRU): Another sequential data specialist, offering an alternative approach to LSTMs [6] [24].
- (d) Recurrent Neural Networks (RNNs) are a class of artificial neural networks designed to efficiently process sequential data by maintaining an internal state, allowing them to capture temporal dependencies within the input sequences. [26] [18][16]

RESULTS

Our research utilized ten distinct machine learning algorithms to discern Breast Cancer cases (Malignant or Benign), relying on their physical attributes. The algorithms employed in our investigation mentioned in the section

We utilized real-valued features for each cell nucleus: radius (mean of distances from the center to points on the perimeter), texture (standard deviation of gray-scale values), perimeter, area, smoothness (local variation in radius lengths), compactness (perimeter² / area - 1.0), concavity (severity of concave portions

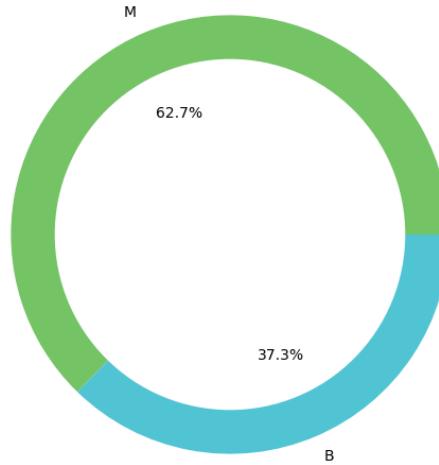


Fig. 1. A cylindrical plot concerning Breast Cancer Classification where we can say that we have 62.7% Malignant and 37.3% Benign from our data.

of the contour), concave points (number of concave portions of the contour), and symmetry. The target variable was Diagnosis (Malignant, Benign).

After rigorous training and testing, the Support Vector Machine Classifier exhibited the highest accuracy among all the algorithms, with the Random Forest Classifier closely following with an accuracy of 98.24%, surpassing the other classifiers in predictive performance.

ML Algorithm	Accuracy of the algorithm
XGboost	95.6140
Neural Networks	97.368
CNN	94.736
RNN	94.7
AdaBoost	97.368
Adaptive Decision Learner	93.859
LSTM networks	94.736
Gated Recurrent Units	93.859
Random Forest Classifier	96.4912
Support Vector Machine	98.2456
Logistic Regression	97.368

Table 1. By using the ten different Deep Learning and machine learning algorithms on the data after missing value imputations, we get that the Random Forest Classifier Algorithm gives us the best result among them per the accuracy score. All of the accuracy scores are mentioned in the table.

This outcome indicates that the Random Forest Classifier is notably effective for Breast Cancer classification based on physical characteristics. The ensemble nature of the SVC algorithm likely contributed to its superior performance in capturing intricate relationships within the data.

This has significant implications for various medical or biological studies and can aid in understanding the properties and characteristics of cancer cells in bodies.

The accuracy scores for each algorithm are presented in Table 1, where it is evident that the SVC Classifier achieved the highest accuracy at 98.2456%. Other algorithms, such as AdaBoost, Logistic Regression, Neural Networks, and Random Forest, also

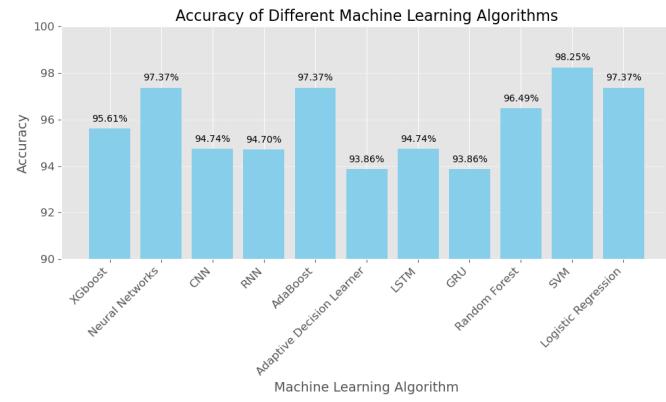


Fig. 2. Using the 11 different machine learning algorithms, we get the best result from the SVM Classifier Algorithm. A pictorial bar diagram concerning their accuracy score for 11 different Deep and Machine learning methods is given in this plot.

demonstrated commendable accuracy rates but were outperformed by the SVC Classifier. The remaining algorithms, including CNN, GRU, RNN, XGBoost, and LSTM, exhibited reasonable performance, albeit with slightly lower accuracy scores than the top-performing algorithms, as depicted in Figure 2.

CONCLUSION

The results section concludes that the Support Vector Machine (SVM) Classifier achieved the highest accuracy at 98.2456%, making it the most effective algorithm for categorizing Breast Cancer cases based on physical attributes. While other algorithms like AdaBoost, Logistic Regression, Neural Networks, and Random Forest also showed reasonable accuracy rates, they were surpassed by the SVM Classifier. The remaining algorithms, including CNN, GRU, RNN, XGBoost, and LSTM, demonstrated reasonable performance but slightly lower accuracy than the top-performing algorithms. Overall, these findings underscore the efficacy of machine learning algorithms, particularly the SVM Classifier, in accurately categorizing Breast Cancer data, which has significant implications for medical and biological studies, aiding in understanding the properties and characteristics of cancer cells within the body.

Author contributions statement

D.C. designed, conceptualized, and developed the research and synthesized interdisciplinary statistical methodologies and models. P.G. conceptualized the model, collected and prepared the datasets, wrote codes for various modified datasets, and performed code-based analysis (mainly using Python). P.G., D.C., wrote, modified, and reviewed the manuscript.

Declaration of competing interest

Conflicts of interest: none.

Data Availability

The Breast Cancer Classification data of [34] that we have used in this study is publicly available from the GitHub Machine Learning Repository [<http://www.cs.wisc.edu/~street/images/>]

Code Availability Statement

The code generated in this paper's results is publicly available on GitHub at [<https://github.com/Prithwish-ghosh/Breast-Cancer>].

We have provided the code under the public license, allowing researchers to reproduce our results and facilitate further development in this area.

References

- [1] Oludare Isaac Abiodun, Aman Jantan, Abiodun Esther Omoilara, Kemi Victoria Dada, Nachaat AbdElatif Mohamed, and Humaira Arshad. State-of-the-art in artificial neural network applications: A survey. *Helijon*, 4(11), 2018.
- [2] Mohamed Adel, Ahmed Kotb, Omar Farag, M Saeed Darweesh, and Hassan Mostafa. Breast cancer diagnosis using image processing and machine learning for elastography images. In *2019 8th International Conference on Modern Circuits and Systems Technologies (MOCAST)*, pages 1–4. IEEE, 2019.
- [3] Mariana Belgiu and Lucian Drăguț. Random forest in remote sensing: A review of applications and future directions. *ISPRS journal of photogrammetry and remote sensing*, 114:24–31, 2016.
- [4] Chi-Chang Chang and Ssu-Han Chen. Developing a novel machine learning-based classification scheme for predicting spcs in breast cancer survivors. *Frontiers in Genetics*, 10:483100, 2019.
- [5] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 785–794, 2016.
- [6] Rahul Dey and Fathi M Salem. Gate-variants of gated recurrent unit (gru) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, pages 1597–1600. IEEE, 2017.
- [7] Carlos Domingo, Osamu Watanabe, et al. Madaboost: A modification of adaboost. In *COLT*, pages 180–189, 2000.
- [8] Ahmed RM El-Azizy, Mohamed Salaheldien, Muhammad A Rushdi, Hanan Gewefel, and Ahmed M Mahmoud. Morphological characterization of breast tumors using conventional b-mode ultrasound images. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6620–6623. IEEE, 2019.
- [9] Prithwish Ghosh. Breast cancer wisconsin (diagnostic) prediction.
- [10] Wilfrido Gómez-Flores and Juanita Hernández-López. Assessment of the invariance and discriminant power of morphological features under geometric transformations for breast tumor classification. *Computer methods and programs in biomedicine*, 185:105173, 2020.
- [11] Francisco A González-Luna, Juanita Hernández-López, and Wilfrido Gómez-Flores. A performance evaluation of machine learning techniques for breast ultrasound classification. In *2019 16th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, pages 1–5. IEEE, 2019.
- [12] Trevor Hastie, Saharon Rosset, Ji Zhu, and Hui Zou. Multi-class adaboost. *Statistics and its Interface*, 2(3):349–360, 2009.
- [13] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.
- [14] Soa-Min Hsu, Wen-Hung Kuo, Fang-Chuan Kuo, and Yin-Yin Liao. Breast tumor classification using different features of

- quantitative ultrasound parametric images. *International journal of computer assisted radiology and surgery*, 14:623–633, 2019.
- [15]Rizwana Irfan, Abdulwahab Ali Almazroi, Hafiz Tayyab Rauf, Robertas Damaševičius, Emad Abouel Nasr, and Abdelatty E Abdalgawad. Dilated semantic segmentation for breast ultrasonic lesion detection using parallel feature fusion. *Diagnostics*, 11(7):1212, 2021.
- [16]Shrutika S Jadhav and Sudeep D Thepade. Fake news identification and classification using dssm and improved recurrent neural network classifier. *Applied Artificial Intelligence*, 33(12):1058–1068, 2019.
- [17]Piotr Karwat, Ziemowit Klimonda, Hanna Piotrkowska-Wróblewska, Katarzyna Dobruch-Sobczak, and Jerzy Litniewski. Quantitative ultrasound examination of peritumoral tissue improves classification of breast lesions. In *2019 IEEE International Ultrasonics Symposium (IUS)*, pages 1–3. IEEE, 2019.
- [18]Jihyun Kim, Jaehyun Kim, Huong Le Thi Thu, and Howon Kim. Long short term memory recurrent neural network classifier for intrusion detection. In *2016 international conference on platform technology and service (PlatCon)*, pages 1–5. IEEE, 2016.
- [19]Yongshuai Li, Yuan Liu, Mengke Zhang, Guanglei Zhang, Zhili Wang, and Jianwen Luo. Radiomics with attribute bagging for breast tumor classification using multimodal ultrasound images. *Journal of Ultrasound in Medicine*, 39(2):361–371, 2020.
- [20]Epimack Michael, He Ma, Hong Li, Shouliang Qi, et al. An optimized framework for breast cancer classification using machine learning. *BioMed Research International*, 2022, 2022.
- [21]Rory Mitchell and Eibe Frank. Accelerating the xgboost algorithm using gpu computing. *PeerJ Computer Science*, 3:e127, 2017.
- [22]Adeola Ogunleye and Qing-Guo Wang. Xgboost model for chronic kidney disease diagnosis. *IEEE/ACM transactions on computational biology and bioinformatics*, 17(6):2131–2140, 2019.
- [23]Mahesh Pal. Random forest classifier for remote sensing classification. *International journal of remote sensing*, 26(1):217–222, 2005.
- [24]Rajib Rana. Gated recurrent unit (gru) for emotion classification from noisy speech. *arXiv preprint arXiv:1612.07778*, 2016.
- [25]V Mary Kiruba Rani and SS Dhenakaran. Retracted article: Classification of ultrasound breast cancer tumor images using neural learning and predicting the tumor growth rate. *Multimedia Tools and Applications*, 79(23):16967–16985, 2020.
- [26]Melissa Roemmele, Sosuke Kobayashi, Naoya Inoue, and Andrew Gordon. An rnn-based binary classifier for the story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 74–80, 2017.
- [27]Robert E Schapire. Explaining adaboost. In *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, pages 37–52. Springer, 2013.
- [28]Alok Sharma and Kuldip K Paliwal. Linear discriminant analysis for the small sample size problem: an overview. *International Journal of Machine Learning and Cybernetics*, 6:443–454, 2015.
- [29]Sandro Sperandei. Understanding logistic regression analysis. *Biochemia medica*, 24(1):12–18, 2014.
- [30]Yi-Sheng Sun, Zhao Zhao, Zhang-Nv Yang, Fang Xu, Hang-Jing Lu, Zhi-Yong Zhu, Wen Shi, Jianmin Jiang, Ping-Ping Yao, and Han-Ping Zhu. Risk factors and preventions of breast cancer. *International journal of biological sciences*, 13(11):1387, 2017.
- [31]Shan Suthaharan and Shan Suthaharan. Support vector machine. *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, pages 207–235, 2016.
- [32]Volkmar Uebel, Shigeo Abe, and Ming-Shong Lan. A neural-network-based fuzzy classifier. *IEEE Transactions on Systems, Man, and Cybernetics*, 25(2):353–361, 1995.
- [33]Mengwan Wei, Yongzhao Du, Xiuming Wu, and Jianqing Zhu. Automatic classification of benign and malignant breast tumors in ultrasound image with texture and morphological features. In *2019 IEEE 13th International Conference on Anti-counterfeiting, Security, and Identification (ASID)*, pages 126–130. IEEE, 2019.
- [34]Street Nick Wolberg William, Mangasarian Olvi and Street W. Breast cancer wisconsin (diagnostic). UCI Machine Learning Repository, 1995. <https://doi.org/10.24432/C5DW2B>.
- [35]Raymond E Wright. Logistic regression., 1995.
- [36]Diyar Qader Zeebaree, Habibollah Haron, Adnan Mohsin Abdulazeez, and Dilovan Asaad Zebari. Machine learning and region growing for breast cancer segmentation. In *2019 International Conference on Advanced Science and Engineering (ICOASE)*, pages 88–93. IEEE, 2019.

Accuracy of Different Machine Learning Algorithms

