Journal of Information Systems Engineering and Management
2025, 10(17s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

Breast Cancer Prediction Using Machine Learning on Parallel Computing

Hana Omar Mokhtar Amar
Ph.D. student, Computer Engineering Department, Istanbul Medipol University- Turkey.
Lecturer at Higher Institute of Science and Technology /souk- Al Juma/ Tripoli, Libya.
hanaomarkanan1@gmail.com

ARTICLE INFO

Received: 05 Dec 2024
Revised: 28 Jan 2025
Accepted: 09 Feb 2025

ABSTRACT

Breast cancer remains one of the most common and significant health challenges for women worldwide. Early detection is crucial for improving survival rates, yet traditional methods often fail to offer timely and accurate predictions. Recent advancements in artificial intelligence (AI) and machine learning (ML) have provided promising tools for enhancing diagnostic processes. This paper presents a practical application of machine learning techniques to predict breast cancer, leveraging parallel computing for improved processing efficiency. By utilizing a dataset of diagnostic features, the study demonstrates how ML algorithms, implemented through a parallel computing framework, can offer accurate predictions of breast cancer incidence. The proposed system utilizes Python and the Ray framework to implement a distributed approach for model training and evaluation, showing substantial potential for scalable, real-time prediction systems in healthcare.

Keywords: Breast cancer prediction, machine learning, parallel computing, artificial intelligence, healthcare diagnostics, distributed computing.

INTRODUCTION

In the world, breast cancer stands as the second most common cancer among women. It is less common among men than it is among ladies. Every year, over 400 000 women are diagnosed as having breast cancer, which is responsible for over 32% of all cancer diagnoses in women. It is the second most frequent cause of cancer–related death in women in several developed nations. Currently, Fatalities from breast cancer have recently risen by 12% as per current data, which is alarming and warrants immediate attention [1]. While mammography, used for more than 30 years, is still the most trusted test for early breast cancer, this test uses low-dose X-rays to make two-dimensional images of the breasts. Therefore, early diagnosis of breast cancer is vital in enhancing public awareness of the disease, facilitating detection at an early stage, and allowing treatment of affected patients. In those diagnosed early, the five-year survival rate is nearly 100 percent [2].

This study presents a new approach to developing a clinical data-based breast cancer detection system using enhanced parallel computing techniques [13]. It hopes to address the problems of computational sizes when it trains a complicated machine learning model and illustrates how parallel computing can improve traditional healthcare systems. This is the first time parallelization has been used with machine learning methods to build a breast cancer detection and prediction model [3]. We used a supervised logistic regression classification algorithm to predict the tumor category accurately. The results show great improvements in speed, scalability, reliability, accuracy, and timely diagnostics with the most effective fusion of these technologies.

PREVIOUS STUDIES

Recent research in breast cancer diagnosis has underscored the potential of artificial intelligence (AI) and machine learning (ML) to enhance prediction accuracy. Various studies have applied ML models to breast cancer datasets to improve early detection using structured clinical and imaging data. While traditional methods relied on rule-based systems and statistical techniques, recent studies have integrated deep learning and distributed computing to optimize prediction efficiency and scalability.

Previous studies have focused on several key objectives, including developing machine learning (ML) systems to classify breast cancer as benign or malignant, assessing the performance of various ML algorithms in breast cancer detection, investigating feature selection techniques to enhance prediction accuracy, and implementing cloud-based and distributed frameworks to manage large datasets efficiently.

These studies use various machine-learning algorithms, including logistic regression, decision trees, support vector machines (SVM), random forests, and deep-learning models. They utilize clinical datasets such as the Wisconsin Breast Cancer dataset and hospital-based diagnostic records. Computational approaches include using distributed computing frameworks like Hadoop and Spark for large-scale data processing. Performance is evaluated using metrics such as accuracy, precision, recall, F1-score, and confusion matrices.

The results from these studies indicate that machine learning models achieve high accuracy, with deep learning approaches, particularly convolutional neural networks (CNNs), demonstrating superior performance in image-based diagnostics [19]. Additionally, distributed computing enhances processing speed and efficiency, enabling real-time data processing in healthcare applications. However, challenges remain, including issues related to data quality, feature selection, and computational cost.

e.g., the most significant prior studies conducted on breast cancer prediction using machine learning were mentioned:

1. International evaluation of an AI system for breast cancer screening (McKinney, S. M. et al., 2020):

This study aimed to assess the effectiveness of an artificial intelligence (AI) system in detecting breast cancer through mammography screening on an international scale. The methodology involved training the AI system using a large dataset of mammographic images from different healthcare institutions, applying deep learning techniques for feature extraction and classification, and comparing performance with radiologists to measure accuracy and false-positive rates. The results showed that the AI system achieved comparable or better accuracy than human radiologists in cancer detection, significantly reduced false positives and false negatives, and improved efficiency in screening processes without compromising accuracy. This study focused on AI-based image analysis, whereas the current paper leverages machine learning with parallel computing to enhance data processing efficiency [1].

2. Key steps for effective breast cancer prevention (Britt, K. L. et al., 2020):

The objective of this study was to outline a strategic framework for the prevention and early detection of breast cancer through clinical interventions and lifestyle modifications. The methodology involved reviewing epidemiological data and clinical trial findings, analyzing risk factors such as genetic predisposition, lifestyle choices, and hormonal influences, and assessing the effectiveness of existing screening techniques. The results identified key preventive measures, including lifestyle changes and targeted screening programs, emphasized the importance of personalized healthcare strategies, and recommended policy-level interventions to enhance early detection rates. While this study focused on prevention strategies, the current research emphasizes predictive modeling through machine learning and distributed computing [2].

3. Global patterns of breast cancer incidence and mortality (Lei, S., et al., 2021):

This study aimed to analyze global trends in breast cancer incidence and mortality over two decades and identify factors contributing to regional disparities. The methodology involved using population-based cancer registry data from multiple countries between 2000 and 2020, applying statistical models to evaluate incidence and mortality rates, and examining healthcare infrastructure and the effectiveness of screening programs. The results indicated rising incidence rates globally, with regional variations based on healthcare accessibility, decreased mortality rates in high-income countries due to improved screening and treatment, and disparities in survival rates based on socioeconomic factors. While this study provided epidemiological insights, the current research focuses on technical advancements in prediction methods using machine learning and parallel computing [3].

4. A Combined Deep CNN: LSTM with a Random Forest Approach for Breast Cancer Diagnosis (Begum, A., et al., 2022):

This research focused on developing a hybrid machine learning model combining deep learning and traditional classifiers to improve breast cancer diagnosis accuracy. The methodology involved using a deep convolutional neural network (CNN) for feature extraction from medical images, long short-term memory (LSTM) networks for sequence analysis, and a random forest classifier for final prediction. The results showed that the hybrid approach achieved high accuracy, outperforming individual models, with improved feature extraction capabilities leading to better classification outcomes and enhanced robustness in handling noisy data. Both studies focus on ML-based cancer prediction; however, the current research distinguishes itself by integrating distributed computing to optimize performance [4].

5. Predicting breast cancer via supervised machine learning methods on class imbalanced data (Rajendran, K., et al., 2020):

This paper aimed to address the challenges posed by imbalanced breast cancer datasets using supervised machine-learning techniques. The methodology involved applying oversampling and under-sampling techniques to balance the dataset, comparing classification algorithms such as SVM, random forests, and k-nearest neighbors, and using cross-validation to validate model performance. The outcomes indicated that balanced data led to improved classification accuracy and reduced bias, with random forest outperforming other models regarding precision and recall. Identification of key features contributing to predictive accuracy. The current paper extends beyond dataset balancing by incorporating parallel computing for enhanced processing efficiency [5].

6. Comparative study of machine learning algorithms for breast cancer prediction (Sengar P. P., et al., 2020):

This paper focused on comparing the performance of various machine learning algorithms in predicting breast cancer outcomes. The methodology involved implementing algorithms such as logistic regression, decision trees, and neural networks, evaluating performance using accuracy, precision, recall, and F1-score, and analyzing feature importance with different feature selection techniques. The results showed that neural networks provided the highest accuracy, while logistic regression demonstrated better interpretability. Feature selection significantly improved model performance, and trade-offs were observed between complexity and interpretability. This study provided insights into ML algorithm performance, while the current research focuses on parallelizing model training to enhance scalability [6].

7. Breast cancer prediction and detection using data mining classification algorithms (Keleş, M. K., 2019):

This study assessed the effectiveness of data mining techniques in detecting breast cancer using clinical datasets. The methodology involved applying classification algorithms, including decision trees and support vector machines, performing feature selection through statistical correlation analysis, and evaluating models using k-fold cross-validation. The outcomes indicated that decision tree algorithms achieved high sensitivity and specificity. Also, improvement in prediction accuracy through appropriate feature selection, and highlighted the importance of data preprocessing in improving model outcomes. While this paper emphasized data mining techniques, the current research incorporates distributed computing for faster processing and scalability [7].

8. Classification of breast cancer data using machine learning algorithms (Akbugday, B., 2019):

This study used standard datasets to compare the performance of various machine learning classifiers for breast cancer diagnosis. For this study, logistic regression, k-nearest neighbors, and support vector machines were employed, their performances by various evaluation metrics were compared and some experiments with different kinds of data preprocessing were performed. Classification performance revealed support vector machines (SVM) as the most accurate classifier, and the study demonstrated the severe impact of data normalization on model performance and identified limitations such as computational efficiency and scalability for large datasets. whereas the current study leverages parallel computing to address the data computational limitations [8].

9. Breast Cancer Risk Prediction Using Machine Learning: A Comprehensive Review:

This study aimed to systematically review the use of deep learning (DL) techniques in predicting breast cancer risk by integrating various data types, including imaging, radionics, genomics, and clinical information. The methodology involved reviewing 600 articles, with 20 meeting the inclusion criteria, and analyzing imaging features (like digital mammography) and non-imaging features (such as genetic variants and clinical data). The study focused on advanced DL methods applied to these features for risk assessment. The results showed that combining imaging and non-imaging features improved the accuracy of breast cancer risk prediction models, with advanced DL models outperforming traditional statistical approaches. The study also highlighted the potential of natural language processing (NLP) and parallel benchmarking of DL models to enhance risk prediction. While this review emphasized integrating diverse data types and advanced DL techniques for risk prediction, my study uniquely focuses on improving computational efficiency through parallel computing using the Ray framework [9].

10. Deep Learning-Based Breast Cancer Detection Using Decision Fusion:

This paper proposed a new AI-based system using a three-parallel-channel method to improve early breast cancer detection accuracy. The first channel used a Support Vector Machine (SVM) with Local Binary Pattern (LBP) features to identify tumor types. The second channel used a pre-trained Convolutional Neural Network (CNN) for feature extraction and then SVM for tumor identification. The third channel developed a new CNN specifically for classifying mammogram images. The results from the three channels were combined using various fusion rules to enhance overall system performance. The decision fusion-based system achieved an overall accuracy of 99.1% using the product rule, outperforming existing methods. The multi-channel approach improved reliability across different diagnostic scenarios. While this research focused on a multi-channel AI system for breast cancer detection, my research emphasizes using parallel computing to optimize machine learning model training and evaluation processes [10].

Unique Contribution of the Current Study:

My paper introduces parallel computing via the Ray framework, which enhances the processing of machine learning models; unlike prior studies, it focuses on the following:

- Parallelized model training and evaluation.
- Scalability and real-time healthcare applications.
- Optimized resource utilization through distributed computing.

Breast Cancer System Architecture

The breast cancer prediction system is a collection of machine-learning algorithms based on actual samples. The system tries to make a model that learns a pattern from the set of data it is given. When presented with enough information, it can predict even with high-dimensional data. This paper will utilize the Ray framework tool through distributed systems for the mentioned processes [4].

Limitations in Traditional Healthcare Systems:

1. Traditional Healthcare Systems (Without Programming)

Errors and inefficiencies arise from reliance on manual data entry and record-keeping. This reliance creates limitations that hinder decision-making, which should be guided by data analysis and bolstered by clinical judgment. A lack of automation leads to delays in diagnosis and the commencement of treatment. Furthermore, sequential batch processing presents challenges related to scalability. Additionally, this scenario demands more resources and increased human effort, resulting in higher costs and a greater dependence on medical professionals for accurate diagnoses.

2. Healthcare Systems Using Traditional Programming

Rule-based approaches, by nature, rely on set logic that cannot easily adapt to new forms of data. As a result, it will be difficult to adjust to altering medical protocols or the emergence of new disease patterns. Intentional sharing of information across the country will be limited [28]. Moreover, processing complex data does not fully attempt to analyze unstructured medical data such as images and genetic data. Such manual updates depend on developers frequently intervening to change algorithms and systems based on new medical findings.

Healthcare Systems Using Machine Learning

Data-Driven Decision Making, Where ML-based systems examine big datasets to recognize patterns and help predict accurately. Adaptability: the model can learn and adapt, improving with new data and increasing diagnostic accuracy over time. Timeliness will be enhanced by minimizing the period of time that an expert must wait onsite to intervene after an anomaly is detected. The system is scalable, with vast amounts of medical data, including medical images, patient records, and genetic data, from multiple sources that could be easily managed. All have their usual benefits and pitfalls to consider. Still, machine learning, in particular, is the most capable but bears the highest cost regarding implementation and validation in use cases [29].

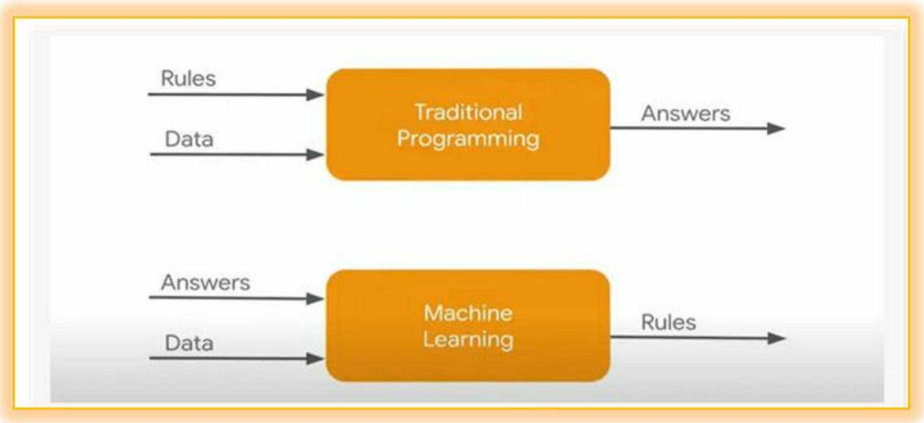


Fig. 1 Machine learning VS traditional programming

The system dataset

This dataset containing such diagnostic features is taken from the UCI Machine Learning Repository (https://data.world/health/breast-cancer-wisconsin) by Nick Street, Mangasarian, and Wolberg. It is used in this research to train and evaluate machine learning models for correctly predicting unseen data.

Features Information as described by Wisconsin Data Creators [35]*:*

1- ID number (1st column).

2- Diagnosis (M = malignant, B = benign – 2nd Column).

3- Columns from 3th to 32th:

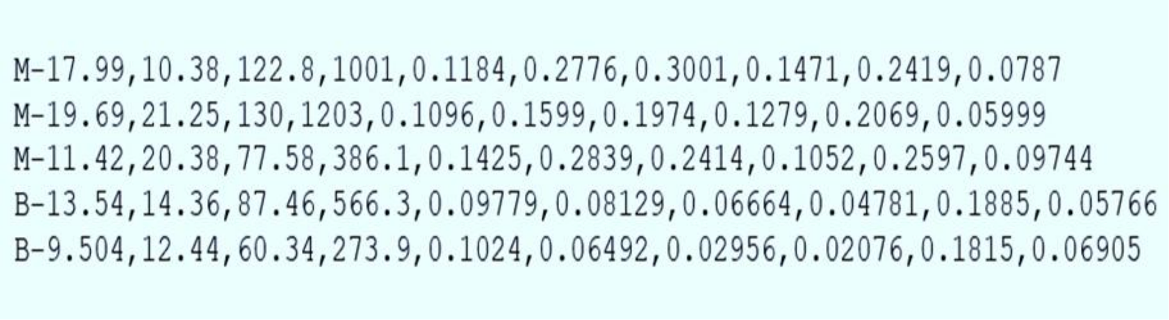
Ten real-valued features (mean M, standard error SE, and worst value) were extracted for each cell nucleus and calculated for all 10 features in each image, resulting in 30 features with 569 samples.

1- Radius: the mean distance from a center to each point on the perimeter of the tumor.
2- Texture: refers to the variation of pixel intensity patterns within the tumor.
3- Perimeter: the length of the tumor's boundary.
4- Area: refers to the total number of pixels in the tumor.
5- Smoothness (regularity): quantifies the variations in the contour or surface texture of the tumor.
6- Compactness: indicates how round the tumor shape or irregular the tumor boundary is.
7- Concavity: measures the degree or depth of inward curves along the tumor boundary.
8- Concave points: refer to the specific locations (countable points) on the tumor boundary where the concavities occur.
9- Symmetry: measures the shape’s balance and how similar the tumor appears when divided into equal parts.
10- Fractal dimension: quantifies the complexity of the tumor boundary.

For example, columns 3 and 4 are the mean radius and the mean texture. The SE radius and SE texture are found in columns 13 and 14, respectively. Column 23 has the worst radius, and Column 24 has the worst texture. The class distribution is 357 benign and 212 malignant, as shown in the data set file below.

data-New.csv

The next figure demonstrates a part of the data, which is 5 rows (cases). Each sample has 11 columns (features), where M refers to Malignant, and B indicates Benign. Other columns are part of valued features:

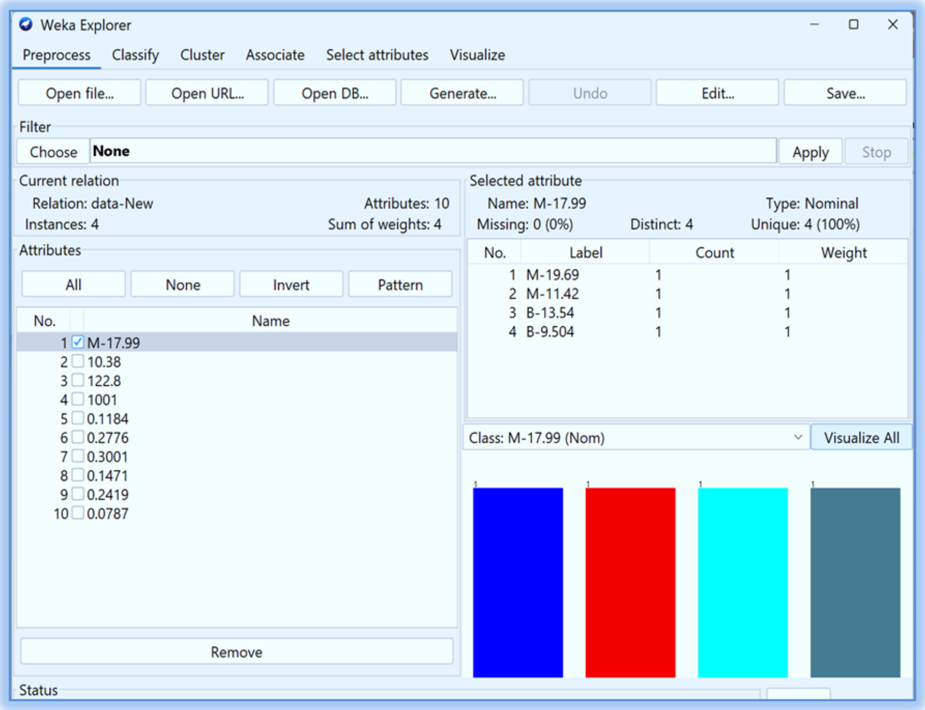


```
M-17.99,10.38,122.8,1001,0.1184,0.2776,0.3001,0.1471,0.2419,0.0787
M-19.69,21.25,130,1203,0.1096,0.1599,0.1974,0.1279,0.2069,0.05999
M-11.42,20.38,77.58,386.1,0.1425,0.2839,0.2414,0.1052,0.2597,0.09744
B-13.54,14.36,87.46,566.3,0.09779,0.08129,0.06664,0.04781,0.1885,0.05766
B-9.504,12.44,60.34,273.9,0.1024,0.06492,0.02956,0.02076,0.1815,0.06905
```

Fig. 2 An example of some patients

Weka Application to analyze and visualize data set:

K-means clustering was used on Breast Cancer data to cluster and visualize all 569 instances [32]. below is the main WEKA Explorer interface with data file classification [33] [34].



None

Current relation
Relation: data-New
Attributes: 10
Instances: 4
Sum of weights: 4

Selected attribute
Name: M-17.99
Type: Nominal
Missing: 0 (0%)
Distinct: 4
Unique: 4 (100%)

Attributes

No.	Name
1	☑ M-17.99
2	□ 10.38
3	□ 122.8
4	□ 1001
5	□ 0.1184
6	□ 0.2776
7	□ 0.3001
8	□ 0.1471
9	□ 0.2419
10	□ 0.0787

No.	Label	Count	Weight
1	M-19.69	1	1
2	M-11.42	1	1
3	B-13.54	1	1
4	B-9.504	1	1

Class: M-17.99 (Nom)

Fig. 3 Weka window to analyze data

RAY Framework

RAY is a Python framework that allows parallel and distributed computing [20]. It enables processes to run concurrently and simulates doing so with multiple cores or machines working quickly [26]. To demonstrate this, my paper used a local framework with four nodes representing four remote computers working simultaneously, as depicted in the figure below.

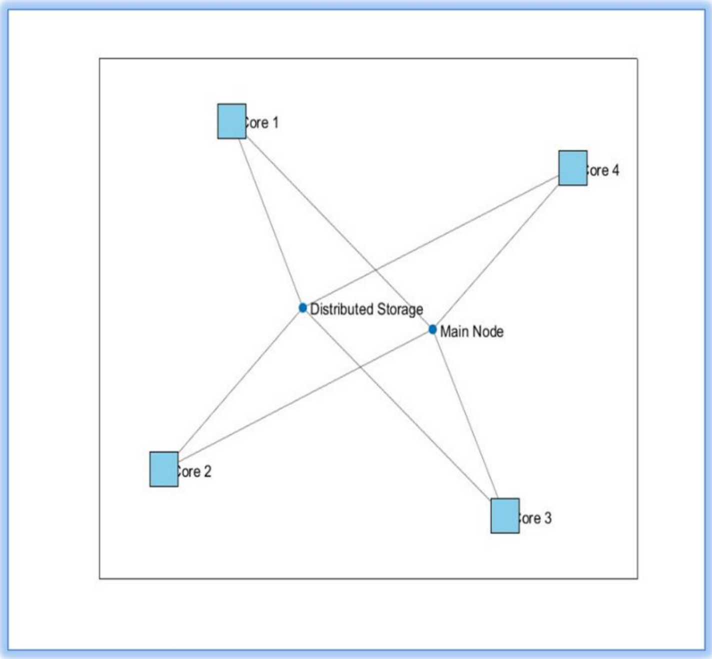


Fig. 4 RAY Framework Architecture

Scalability and Reliability

Deploying machine learning models, especially in healthcare applications where predictions must be made accurately and in real time [21], is crucial to utilizing machine learning effectively. Distributing computation significantly enhances the system's ability to process unstructured data in breast cancer prediction. By employing parallel processing, all data loading, pre-processing, training, and evaluation tasks can be executed simultaneously across multiple units, optimizing both time and computing resource usage [11].

Scalability refers to the system's ability to handle increasing data and computational demands without compromising performance. As the volume of patient data grows [12], scalable machine learning systems can distribute workloads across multiple processing cores or cloud-based resources, ensuring smooth operation even under heavy loads. This enables healthcare providers to make data-driven decisions swiftly and accurately, improving patient outcomes.

Reliability ensures that the system produces accurate and reliable results consistently over time. Reliability in distributed machine learning applications involves resource management, fault tolerance, and cross-processing unit consistency. Distributed frameworks provide an efficient environment to balance several computing flows of data in real-time, assuring the robustness and fault tolerance in applications where this requirement is needed [22].

One of the top advantages of using parallel and distributed computing frameworks such as the Ray framework is the ability to optimize the utilization of system resources, both memory and processing power. This framework allows the background synchronization of tasks, which releases system administrative resources and thus simultaneously facilitates training and/or evaluation of multiple models. Scalability is not only confined to on-premises; cloud-based solution integrations can also be an added flexible and cost-effective solution for healthcare institutions, catering the scalability to demand [24].

In conclusion, scalable and reliable machine learning systems are key to improving healthcare diagnostics. With an appropriate infrastructure, organizations can perform computations quickly, minimize downtime, ensure accurate results, and ultimately support timely and informed medical decisions. Healthcare systems with limited resources may adopt new technology in the earliest stages of implementation by complementing existing infrastructures or completely replacing them. This is the solution that the company intends to pursue.

MACHINE LEARNING

Machine learning (ML) is a powerful branch of artificial intelligence (AI) [25]. It allows computers to learn from labeled data and make informed predictions without being explicitly programmed. It imitates human thought processes by recognizing patterns and learning from data to make better predictions. For example, machine learning helps doctors diagnose diseases in healthcare with their experience and data-driven insights [23].

In contrast to traditional programming, where fixed instructions are written to do specific tasks, machine learning creates predictive models from labeled data and the patterns observed in enormous datasets. Using a database management system, we can improve those models by feeding them back and getting good predictions [6].

Hence, standalone machine learning can be applied in augmentative systems complementary to domain expertise, i.e., where the systems can improve accuracy and efficiency. By separating traditional processing and distribution processing, ML can be scaled well, improve performance, and reduce processing time. The ability of machine learning to continually improve its performance makes it a revolutionary tool in sectors such as healthcare, finance, and autonomous vehicles [7], providing solutions that can grow and adapt as more data becomes available.

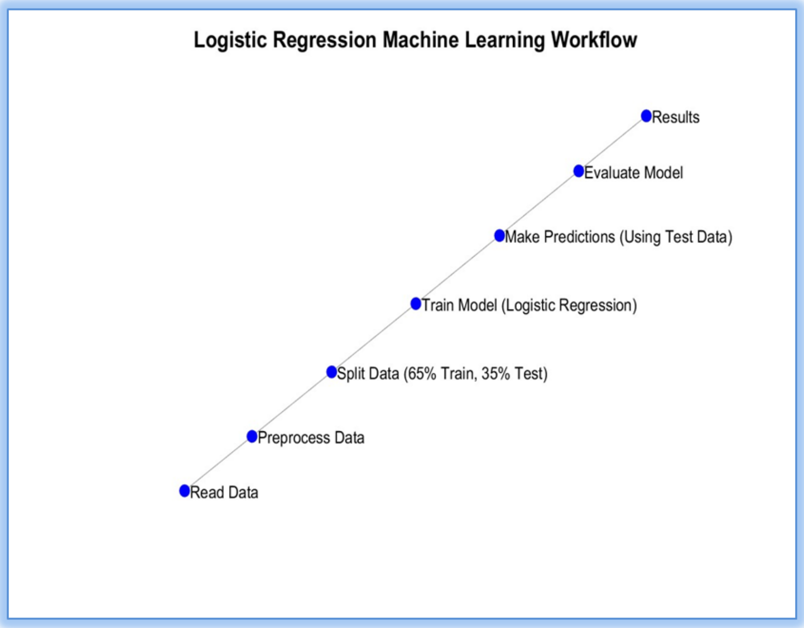

Logistic Regression Machine Learning Workflow
Results
Evaluate Model
Make Predictions (Using Test Data)
Train Model (Logistic Regression)
Split Data (65% Train, 35% Test)
Preprocess Data
Read Data

Fig. 5 The Breast Cancer Prediction System Architecture

Logistic Regression Classifier

Logistic regression (LR) is a standard statistical and machine-learning method for binary classification problems. It aims to predict one of two possible outcomes given several independent variables. These models are suitable in cases where the target variable is categorical, i.e., they help separate benign vs. malignant tumors in medical datasets [8].

Logistic regression is a statistical method for predicting binary classes. This function (Sigmoid Function) takes an arbitrary output to a number in the 0-1 range, so it is suitable for classification issues where the outcome is a particular class rather than a continuous value. Logistic regression aims to relate independent features to the dependent variable to get accurate predictions.

LR — A linear combination of input features is obtained and run through the sigmoid function to yield an output probability. If the predicted probability exceeds a certain threshold (commonly 0.5), the model assigns the observation to one category; otherwise, it is classified into another. This method makes logistic regression very interpretable and useful for medical diagnosis, fraud detection, and risk assessment applications [9].

If the data is linearly separable, logistic regression will perform well for accuracy. This simple algorithm works well in many practical use cases and is a building block to create complex machine learning models. Moreover, logistic regression can be generalized to multiple categories (multinomial logistic regression) or hierarchical structures, making it applicable to broader contexts [10].

In conclusion, logistic regression is a key element of supervised learning. It provides a trade-off between interpretability and predictive performance and is commonly used in domains ranging from healthcare to finance to social sciences.

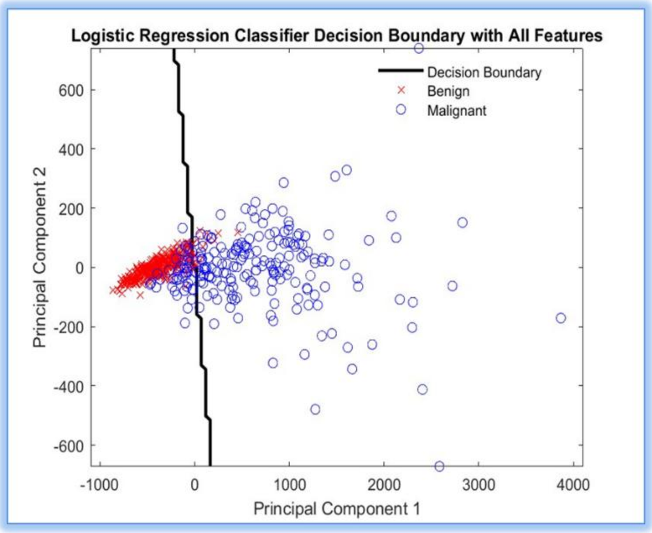
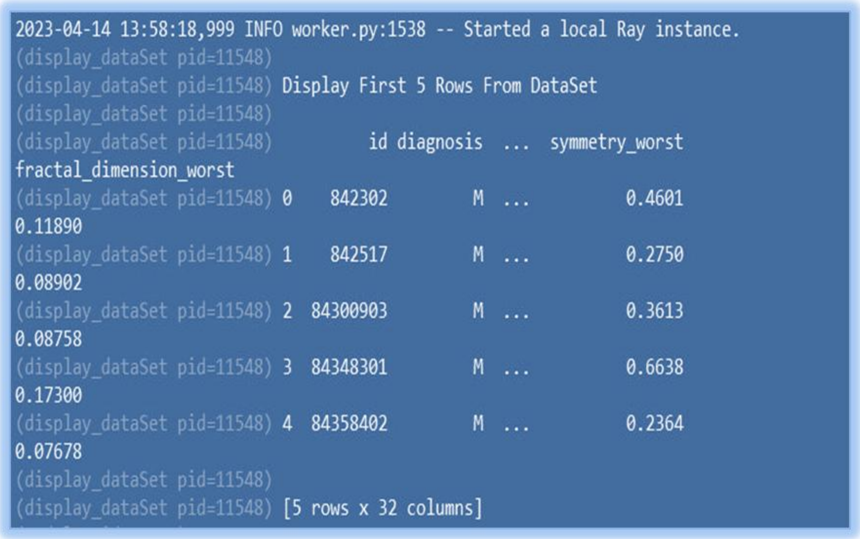


Fig. 6 Logistic Regression Classifier

This figure shows how we apply Principal Component Analysis (PCA) to reduce our high-dimensional dataset into Principal Components 1 and 2. The decision boundary is now a vertical straight line, which means logistic regression relies on the first principal component for separation. In contrast, the benign data points (red) form a dense cluster, while the malignant data points (blue) are more spread out. The average method generates more information on all features and is thus a better choice for generalizing well on new and large datasets.

The GUI implementation

When running the program, the RAY infrastructure gets initialized, so four cores are ready to act as separate computers [15]. The first core fetches the data set from the source, and the second one performs the pre-processing process and then displays the retrieved data. The third core does the training and testing step and applies logistic regression algorithms. The optimization [20], evaluation, and visualization of the model's performance are done within the fourth core [17], as they are done in the background and explained in the following figure.



```
2023-04-14 13:58:18,999 INFO worker.py:1538 -- Started a local Ray instance.
(display_dataSet pid=11548)
(display_dataSet pid=11548) Display First 5 Rows From DataSet
(display_dataSet pid=11548)
(display_dataSet pid=11548)          id diagnosis  ...  symmetry_worst
fractal_dimension_worst
(display_dataSet pid=11548) 0    842302         M  ...          0.4601
0.11890
(display_dataSet pid=11548) 1    842517         M  ...          0.2750
0.08902
(display_dataSet pid=11548) 2  84300903         M  ...          0.3613
0.08758
(display_dataSet pid=11548) 3  84348301         M  ...          0.6638
0.17300
(display_dataSet pid=11548) 4  84358402         M  ...          0.2364
0.07678
(display_dataSet pid=11548)
(display_dataSet pid=11548) [5 rows x 32 columns]
```

Fig. 7 A pre-processing step and work in the background.

After completing all the previous operations [18], the GUI of Predicting breast cancer using machine learning techniques depending on a distributed system application will create:

Fig. 8 GUI of Predicting Breast Cancer Using Machine Learning Techniques

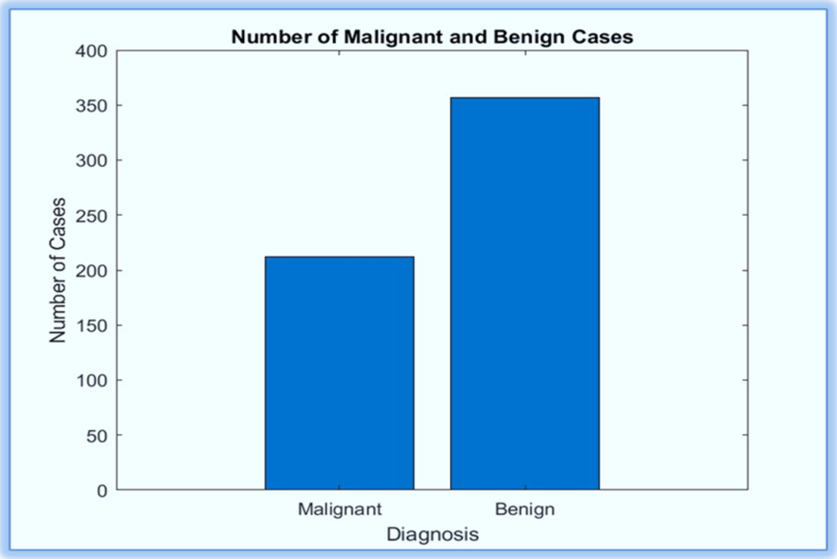


Fig. 9 Logistic Regression Model Performance

The Evaluation

After the optimization step, reliable breast cancer prediction system results will be reached [30]. The critical step comes after applying the model to unseen data to evaluate the model performance using some matrixes, such as precision-recall and confusion, to get accuracy [27], which is evident in the following plots [31].

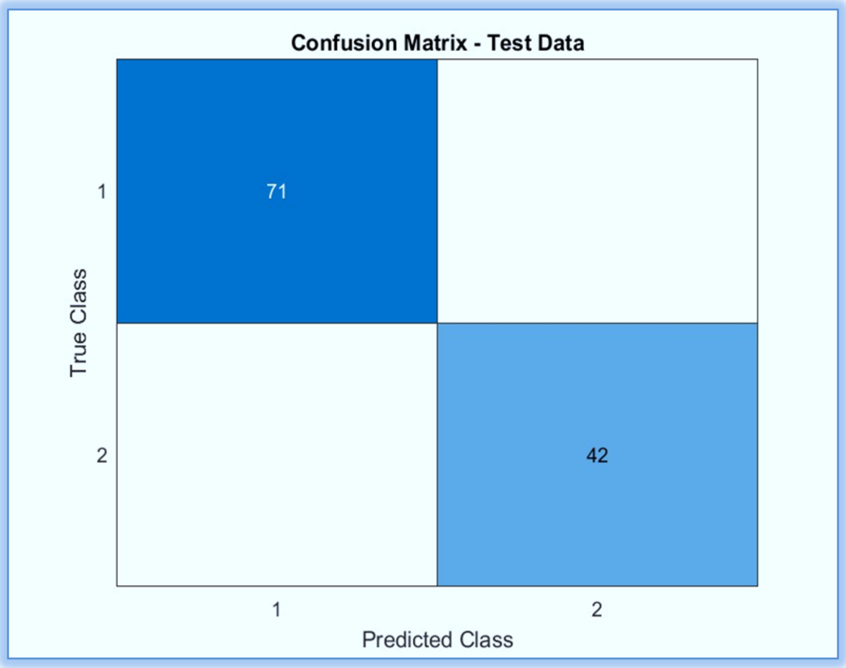


Fig. 10 Confusion Matrix

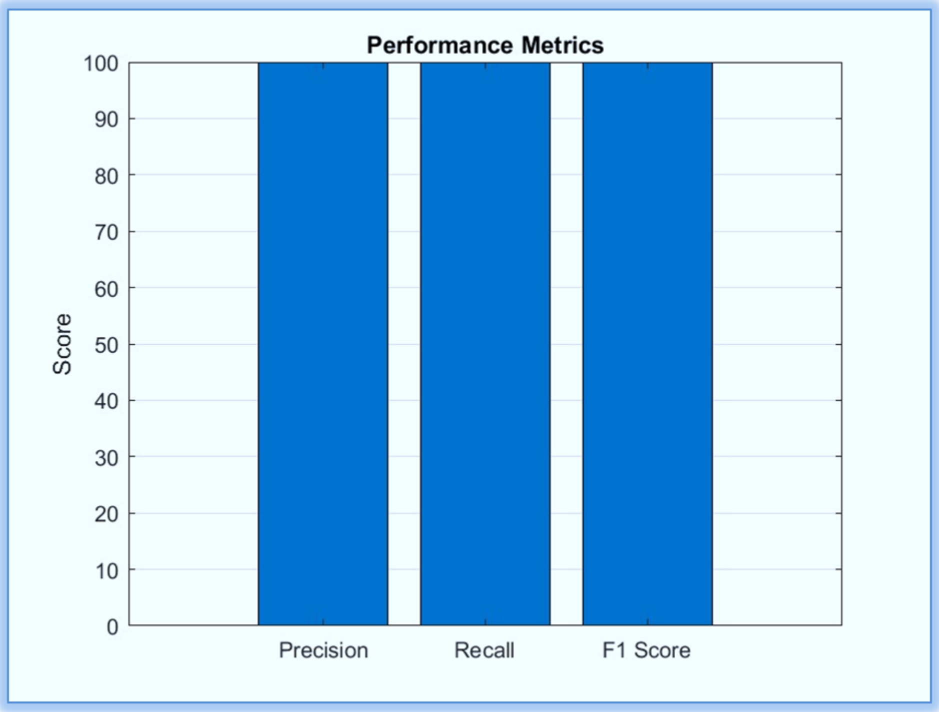


Fig.11 Precision, Recall, F1 Matrixes

Future Enhancement

The improvements using distributed computing environments could engage deep learning as part of a breast cancer prediction system [14]. Also, utilizing advanced imaging modalities (X-rays, MRI, CT, FMRI, etc.) with developed image processing techniques allows for the enhanced diagnosis of diseases [16]. Widely used frameworks such as cloud computing will assist artificial intelligent algorithms, enabling trainers to create stronger and larger models. These advances may enable novel techniques for feature extraction and pattern recognition. This will allow us to select better strategies for evolving this system, maximizing clinical diagnosis performance, reliability, and accuracy.

CONCLUSION

Implementing the breast cancer prediction system using machine learning techniques on a distributed computing framework has demonstrated significant efficiency, scalability, and accuracy advantages. By leveraging the Ray framework, the system effectively utilized four processing cores, each acting as an independent computational unit to handle various data processing, training, and evaluation stages. This parallelized approach has enabled optimal utilization of computational resources, improved memory management, and reduced processing time.

This has dramatically developed the performance of systems. Ensuring predictions are generated on time and improved system resource utilization by defined procedures such as the preprocessing step, training the model, and evaluating the model can be done in separate processing units. Since parallel computing leads to more robust and flexible systems, they work well with large data sets. Furthermore, the system merges the automation of complicated calculations and machine learning applications to create a comprehensive strategy that assists in the early diagnosis and detection of health issues, improved decisions, and better outcomes for patients in need.

The approaches presented here are not the last word on using parallel computing and machine learning algorithms; they lead to growth, acceleration, and enhanced time efficiency in medical diagnosis. The models could be expanded and improved with bigger unstructured data sets, novel models, and tools like cloud computing, artificial intelligence, and deep learning, creating more accurate, scalable, and dependable healthcare solutions.

Acknowledgment

I would like to express my sincere gratefulness to **Dr. *Mohamed Khalifa***, my supervisor, and ***Prof. Dr. Selim Akyakus***, my advisor and Manager of the Computer engineering department at Istanbul Medipol University in Turkey, for their understanding, support, and valuable advice during this study. I am very thankful for their professionalism, support, and guidance in this research's extensive development and quality. Additionally, I extend my deepest appreciation to the original creators of the dataset: *Dr. William H. Wolberg, W. Nick Street, and Olvi L. Mangasarian*. Their efforts in collecting and maintaining a quality medical database have enabled such Research to significantly contribute to improving breast cancer diagnosis and prediction. I am finally thankful to my colleagues, family, and friends for their encouragement and continued support throughout this entire process. Their motivation inspires me to reach my goal every day.

Thanks to all participants for contributing and making this paper meaningful work!

REFERENCES

[1] McKinney, S. M., et al. (2020). International evaluation of an AI system for breast cancer screening. Nature.

[2] Britt, K. L., et al. (2020). Key steps for effective breast cancer prevention. Nature Reviews Cancer.

[3] Lei, S., et al. (2021). Global patterns of breast cancer incidence and mortality. Cancer Communications.

[4] Begum, A., et al. (2022). A Combined Deep CNN: LSTM with a Random Forest Approach for Breast Cancer Diagnosis. Complexity.

[5] Rajendran, K., et al. (2020). Predicting breast cancer via supervised machine learning methods. International Journal of Advanced Computer Science and Applications.

[6] Sengar, P. P., et al. (2020). Comparative study of machine learning algorithms. IEEE Conference Proceedings.

[7] Keleş, M. K. (2019). Breast cancer prediction using data mining. Tehnički vjesnik.

[8] Akbugday, B. (2019). Classification of breast cancer data using ML. IEEE Medical Technologies Congress.

[9] Hussain S, Ali M, Naseem U, Nezhadmoghadam F, Jatoi MA, Gulliver TA and Tamez-Peña JG (2024) Breast cancer risk prediction using machine learning: a systematic review. Frontiers in Oncology. Frontiers.

[10] Doğu M, Hasan D, Alaa E. (2024). Deep Learning-Based Breast Cancer Detection Using Decision Fusion. MDPI Computers. MDPI

[11] Momenimovahed, Z., & Salehiniya, H. (2019). Epidemiological characteristics of and risk factors for breast cancer in the world. Breast Cancer: Targets and Therapy, 151-164.

[12] Yedjou, C. G., Sims, J. N., Miele, L., Noubissi, F., Lowe, L., Fonseca, D. D., ... & Tchounwou, P. B. (2019). Health and racial disparity in breast cancer. Breast cancer metastasis and drug resistance: Challenges and progress, 31-49.

[13] Hanker, A. B., Sudhan, D. R., & Arteaga, C. L. (2020). Overcoming endocrine resistance in breast cancer. Cancer Cell, 37(4), 496-513.

[14] Yin, L., Duan, J. J., Bian, X. W., & Yu, S. C. (2020). Triple-negative breast cancer molecular subtyping and treatment progress. Breast Cancer Research, 22, 1-13.

[15] Michael, E., Ma, H., Li, H., & Qi, S. (2022). An optimized framework for breast cancer classification using machine learning. BioMed Research International, 2022.

[16] Ravikumar, A., & Sriraman, H. (2023). Real-time pneumonia prediction using pipelined spark and high-performance computing. PeerJ Computer Science, 9, e1258.

[17] Datta, D., Agarwal, R., & David, P. E. (2020). Performance enhancement of customer segmentation using a distributed Python framework, Ray. International Journal of Scientific & Technology Research, 9(11), 130-139.

[18] Król, M., Mastorakis, S., Oran, D., & Kutscher, D. (2019, September). Compute first networking: Distributed computing meets in Proceedings of the 6th ACM Conference on Information-Centric Networking (pp. 67-77).

[19] Nentvich, O., Urban, M., & Hudec, R. (2023). PyXLA: Python x-ray-tracing for Lobster-Eye application. Journal of Optics, 25(5), 053501.

[20] Bauer, M., & Garland, M. (2019, November). Legate NumPy: Accelerated and distributed array computing. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (pp. 1-23).

[21] Liang, E., Wu, Z., Luo, M., Mika, S., Gonzalez, J. E., & Stoica, I. (2021). RLlib Flow: Distributed Reinforcement Learning is a Dataflow Problem. Advances in Neural Information Processing Systems, 34, 5506-5517.

[22] Li, S., & Avestimehr, S. (2020). Coded computing: Mitigating fundamental bottlenecks in large-scale distributed computing and machine learning. Foundations and Trends® in Communications and Information Theory, 17(1), 1-148.

[23] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32.

[24] Nassef, O., Sun, W., Purmehdi, H., Tatipamula, M., & Mahmoodi, T. (2022). A survey: Distributed Machine Learning for 5G and beyond. Computer Networks, 207, 108820.

[25] Anisetti, M., Ardagna, C. A., Bena, N., & Foppiani, A. (2021, September). An assurance-based risk management framework for distributed systems. In 2021 IEEE International Conference on Web Services (ICWS) (pp. 482-492). IEEE.

[26] Stoller, S. D., Carbin, M., Adve, S., Agrawal, K., Blelloch, G., Stanzione, D., ... & Zaharia, M. (2019, October). Future directions for parallel and distributed computing: Spx 2019 workshop report. In NSF Workshop Reports.

[27] Han, X., He, H., Wu, J., Peng, J., & Li, Y. (2019). Energy management based on reinforcement learning with double deep Q-learning for a hybrid electric tracked vehicle. Applied Energy, 254, 113708.

[28] O'Mahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G. V., Krpalkova, L., ... & Walsh, J. (2020). Deep learning vs. traditional computer vision. In Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1 1 (pp. 128-144). Springer International Publishing.

[29] Alam, A. (2022, April). A digital game-based learning approach for effective curriculum transaction for teaching-learning of artificial intelligence and machine learning. In 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS) (pp. 69-74). IEEE.

[30] Naji, M. A., El Filali, S., Aarika, K., Benlahmar, E. H., Abdelouhahid, R. A., & Debauche, O. (2021). Machine learning algorithms for breast cancer prediction and diagnosis. Procedia Computer Science, 191, 487-492.

[31] Tseng, Y. J., Huang, C. E., Wen, C. N., Lai, P. Y., Wu, M. H., Sun, Y. C., ... & Lu, J. J. (2019). Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies. International journal of medical informatics, 128, 79-86.

[32] Rashid, A., Farhad, S. S. B., Bhuyian, A., Yeasmin, N., Azim, M. A., & Alom, Z. (2022, December). A Comparative Analysis of Machine Learning techniques on Breast Cancer diagnosis using WEKA. In 2022 25th International Conference on Computer and Information Technology (ICCIT) (pp. 663-668). IEEE.

[33] Srikanth, K., Zahoor, S., Huq, U., & Kumar, A. P. S. (2019). Analysis, implementation, and comparison of machine learning algorithms on breast Cancer dataset using the WEKA tool. International Journal of Recent Technology and Engineering (IJRTE), 7.

[34] Kirola, M., Memoria, M., Dumka, A., & Joshi, K. (2022). A comprehensive review study on optimized data mining, machine learning, and deep learning techniques for breast cancer prediction in a significant data context. Biomedical and Pharmacology Journal, 15(1), 13-25.

[35] O.L. Mangasarian, W.N. Street and W.H. Wolberg. Breast cancer diagnosis and prognosis via linear programming.

[36] Operations Research, 43(4), pages 570-577, July-August 1995.