# A DATA-DRIVEN APPROACH TO BREAST CANCER DETECTION USING NON-NEGATIVE MATRIX FACTORIZATION

Zulfikar A. Ansari*[1], Manish M. Tripathi[2], Rafeeq Ahmed[3]

[1]Integral University, CSE, Lucknow, 226022, India
Koneru Lakshmaiah Education Foundation, CSE, Vaddeswaram, 522302, India
[2]Integral University, CSE, Lucknow, 226022, India
[3]Government Engineering College, CSE West Champaran, 845438, India

JOURNAL OF DYNAMICS AND CONTROL
VOLUME 8 ISSUE 12

 ISSN: 1672-6553

# A DATA-DRIVEN APPROACH TO BREAST CANCER DETECTION USING NON-NEGATIVE MATRIX FACTORIZATION

**Zulfikar A. Ansari*[1], Manish M. Tripathi[2], Rafeeq Ahmed[3]**

[1]Integral University, CSE, Lucknow, 226022, India
Koneru Lakshmaiah Education Foundation, CSE, Vaddeswaram, 522302, India
E-mail: zulfi78692@gmail.com
ORCID id: https://orcid.org/0000-0003-0095-8713
*Corresponding Author

[2]Integral University, CSE, Lucknow, 226022, India
E-mail: mmt@iul.ac.in
ORCID id: https://orcid.org/0000-0003-3441-5733

[3]Government Engineering College, CSE West Champaran, 845438, India
E-mail: rafeeq.amu@gmail.com
ORCID id: https://orcid.org/0000-0001-5952-611X

**Abstract:** Breast cancer (BC) remains a prevalent form of cancer among women globally, with the potential to be fatal. Timely detection is crucial for effective therapy and improved survival rates. Machine learning methodologies have demonstrated their potential to assist physicians in the early detection of breast cancer. Non-negative Matrix Factorization (NMF) is an effective method for dimensionality reduction and feature extraction. It is utilized in various domains, including medical image analysis. We examined the efficacy of Support Vector Machines (SVM), Logistic Regression (LR), Decision Trees (DT), Random Forests (RF), Multi-Layer Perceptron's (MLP), and Non-negative Matrix Factorization (NMF) in diagnosing various types of issues. The combination of MLP and NMF achieved the highest accuracy of 97.90%. The study's findings indicate that machine learning techniques, particularly MLP in conjunction with NMF, may enhance the accuracy of breast cancer diagnosis. This is crucial for the early detection of breast cancer and obtaining effective treatment, ultimately resulting in improved patient outcomes. Employing Non-negative Matrix Factorization (NMF) to diminish the dimensionality of gene expression data, in conjunction with machine learning methodologies, may enhance the precision of breast cancer detection and treatment.

**Index Terms**: Breast Cancer, Machine Learning, Non-Negative Matrix Factorization, Support Vector Machine, Multilayer Perceptron.

## 1. Introduction

Worldwide, breast cancer is one of the most pressing health concerns, with increasing rates and considerable impacts on healthcare systems globally [1]. Timely identification is crucial for successful therapy and patient survival rates. Conventional breast cancer diagnosis approaches mainly depend on histological inspection and molecular profiling; however, they may not fully capture the intricate molecular characteristics of the illness [2]. High-throughput genomic technologies in recent years have offered exceptional potential for studying cancer biology and creating more accurate diagnostic tools. NMF is a potent approach for reducing dimensions and revealing concealed patterns in extensive genomic datasets [3]. This work aims to utilize NMF to analyse breast cancer detection by examining the complex biological signatures linked to various cancer subtypes and stages of progression. By breaking down gene expression patterns into meaningful components, we aim to deepen our comprehension of breast cancer biology and boost the precision of diagnostic and prognostic models, ultimately leading to more tailored and efficient therapeutic treatments.
This study embarks on a comparative journey to assess the utility of ML Techniques with NMF in the context of BC prediction. Our primary aim is to identify the ideal number of components that yield the best accuracies for BC detection using these methodologies, while also tackling the computational difficulties linked to high-dimensional data. We investigate various NMF techniques to accomplish this objective. The results indicate that NMF attained the maximum accuracy along with MLP.

BC is a common and serious disease that impacts women worldwide [4]. The objective of this article is to investigate the potential of Non-Negative Matrix Factorization to improve breast cancer diagnosis and characterisation. This study intends to create new computational frameworks to improve the accuracy, effectiveness, and clarity of BC diagnosis by using the unique characteristics of NMF and sophisticated imaging techniques, with the objective of promoting personalized medicine and improving patient outcomes [5].
We have made important advancements in breast cancer diagnosis by utilising NMF in our research. Our research aims to focus on the following essential contributions.

- Performed a thorough comparison examination of different Machine Learning Algorithms utilizing NMF.
- We have investigated the components with which the NMF provides better results using various ML classifiers, such as LR, SVM, RF, DT & MLP.
- Our research has enabled the identification of key discriminative biomarkers that play a crucial role in distinguishing between malignant and benign breast tumors.

## 2. Related Work

Various research has investigated the effectiveness of dimensionality reduction techniques such as NMF in enhancing breast cancer diagnosis with ML algorithms. A detailed review of the ML has been done by the author [6] to forecast cancer and in- depth learning techniques are described and compared. The study in [6] organized to an examination of scholarly works that focus on the categorization of histological images related to BC through the utilization of deep learning approaches.
The authors [7] has utilized the transfer learning technique with pre-trained models such as GoogleNet, VGGNet, and ResNet, yielding a mean accuracy outcome of 97.52%. In subsequent studies, the suggestion to employ hand-crafted characteristics to potentially achieve improved outcomes was put out. The features and classifier of the Xception architecture were assessed using several classifiers, including SVMs, AdaBoost, RF, bagging, and logistic regression. Among the examined methodologies, the Xception classifier exhibited the most favorable outcomes, with accuracy, sensitivity, and F1-score all attaining a value of 0.9 by authors [8]. This author [9] key objective of this investigation is to concentrate on prior studies on ML methods for BC forecasting. Freshers interested in evaluating ML methods can get all the necessary material to fully understand deep learning. This paper contains all of the required information for novices who wish to study ML algorithms in order to obtain a solid understanding of deep learning. In this study [10] NB, RF, and KNN are three popular ML methods used to diagnose breast cancer. The WBCD data set was utilized as a training dataset to test how well different machine learning methods worked in terms of important factors like precision and accuracy. Authors [11] talked about how advances in ML have improved understanding of sizable and complicated visual data sets. With simply an image of a tumor sample as input, the main objective is to study patient outcome prediction using ML. They asserted that tumor histology can be used to derive prognostically relevant information that can be added to the commonly used BC prediction variables. The major goal of this [12] research was to offer an effective approach for identifying cancers utilizing mammography pictures of breasts and a ML algorithm. Second, based on the proposed strategy in the first phase, The aim of this investigation intended to develop a CAD program for the detection of BC. Authors [13] examined a variety of ML algorithms for detecting BC. They looked to study how CNN, SVM, and RF are similar and different. CNN outperforms the other systems now in use in terms of accuracies, precisions, and the amount of data required. In this study [14] applied K-NN, SVM, RF, LR, and DT to the BC WBCD dataset for this study. Once the data was collected, they evaluated and contrasted the performance of these different approaches. The main aim of this research is to use ML algorithms to detect and diagnose BC as well as to compare the accuracy, precision, and CM of various methods. Understanding and classifying BC, authors [15] developed an automated learning and image processing-based evolutionary technique. To aid in identifying and classifying of problems with the skin, this model includes image initial processing, extraction of features, choice of features, and ML approaches. Investigation, various classification models that are used to detect BC have been chosen. Both single and ensemble classifiers are part of this set. To choose solely weighted features and exclude other attributes from a trustworthy dataset, five distinct feature selection techniques have been applied by authors [16] Xiao conducted [17] a comparative analysis of multiple-ensemble approaches and arrived at the conclusion that the stacking of diverse classifiers exhibits greater performance. In a peer-reviewed study released in 2019 [18], various machine-learning methodologies were utilized to facilitate the identification and diagnosis of BC. RF achieved the highest classification accuracy, with the neural approach ranking second in terms of accuracy. Numerous recent studies have utilized deep learning methodologies on large-scale datasets pertaining to BC by author [19]. They have implemented an ensemble voting technique for analyzing the Wisconsin Breast Cancer (WBC) dataset to detect breast cancer. Their intention

is to explain how CNN and logistic algorithms can be utilized to detect breast cancer when the variables are reduced [20]

An analysis was performed on several methodologies used for diagnosing BC, as shown in table 1. The examination involved analyzing the methodology used, the dataset employed, and the accuracy information provided.

## 3. Material and Methods

This part will cover information regarding dataset details, preprocessing, feature extraction, and the methods to be utilized. This part displays the flow diagram illustrating the method for forecasting BC. The proposed classification approaches incorporate many techniques such as SVM, DT, RF, MLP and LR. We used NMF to assess the characteristics, as depicted in Figure-2.

### *3.1 Dataset Description:*

Irvine ML Repository, which is affiliated with UCI, provided the dataset utilized in this research. One public resource for conducting empirical research on ML methods is the UCI ML repository, which houses several datasets well-suited for such studies. The work involved analyzing publicly accessible WBCD datasets employed from the UCI ML repository [21]. The dataset, originally from the University of Wisconsin, has 569 samples, with 212 being malignant and 357 normal, each having 32 features as depicted in Figure-1. The image depicts the fundamental characteristics of breast mass cell nuclei. There are 30 real-valued input attributes, an id, and a diagnostic field included. Field 2 is labeled "diagnostic" and consists of two categories: malignant and benign. Each cell nucleus is characterized by ten real-valued characteristics. The parameters include perimeter, radius, area, texture, concavity, compactness, smoothness, concave points, symmetry, and fractal dimension [22].

### *3.2 Preprocessing*

An essential step in improving data quality and performance, data preparation seeks to reduce noise, inconsistencies, and redundancy [23]. Missing values are filled in during data pre-processing by using the meaning of the corresponding characteristic. The dataset format has been confirmed to be consistent [24]. All the features with incorrect data types were corrected to their required data type as depicted in Figure-2. We identify and address missing values in the dataset by removing duplicates and filling missing values with the median. Columns with constant values across all features have been removed. Specifically, for NMF, the preprocessing step may entail additional procedures like as selecting a suitable number of components or factors, setting initialization parameters, or applying regularization techniques to prevent overfitting. These are all examples of additional steps that may be included.

Graphical representation of a few attributes of our dataset has been shown as the heatmap [22]. The utilization of this technology proves to be advantageous in the visualization of intricate data, specifically within the realms of statistical analysis and predictive modelling. The heatmap consists of a grid of cells, where each cell is assigned, a color based on its value [25]. The values in the matrix can represent anything from the frequency of a particular event to the intensity of a particular attribute. By color-coding the cells based on their values, it can be quickly figured out which attributes have the highest or lowest values, and therefore which areas of the dataset are most important or interesting according to that we constructed a heatmap of our dataset in Figure-5 it shown that some components are highest corelated and some components are lowest corelated.  We can use a heatmap to visualize either the basis matrix (W) or the coefficient matrix (H), or we can use it to visualize both. The heatmap is organized so that each row represents a feature (for example, a gene) in the dataset, and each column represents a component or sample contained inside the dataset [26].

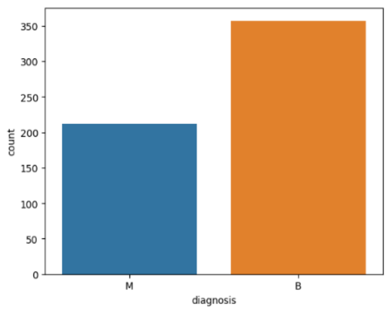


Figure 1: (Malignant & Normal data)

 ISSN: 1672-6553

Table 1: Existing Methods with Accuracy

| Author / Year | Methodology | Dataset | | Accuracy | Year |
|---|---|---|---|---|---|
| | | WBCD | OTHERS | | |
| Zexian Zeng et al.[27] | SVM with NMF | | BRCA | 80% | 2019 |
| Poonam Kathale et al. [28] | RF | | Mammography screening images | 95% | 2020 |
| Sami Ekici et al. [29] | CNN | | Thermal images (140 individuals) | 98.95% | 2020 |
| Riku Turkki et al. [30] | Utilizing tissue microarray (TMA) | | sample images of 868 patients | 95% | 2019 |
| Ramik Rawal[31] | LR, SVM & RF | ✓ | | 96.24 | 2020 |
| Jiande Wu et al. [32] | SVM, KNN, NB and DT | | 992 non 110 triple and Breast Cancer tumor samples | 90%, 87%, 85%, 87% | 2021 |
| Saeid Jafarzadeh Ghoushchi et al.[33] | CNN | | BCDRD01 | NA | 2021 |
| Sharmin Ara et al. [34] | LR, SVM, KNN, DT, NB and RF | ✓ | | 96.5% | 2021 |
| Ryuji Hamamoto et al. [35] | NA | NA | NA | NA | 2022 |
| Farouk A. K. Al-Fahaidy et al. [36] | SVM | | MIAS dataset of mammogram images | 87.10% | 2022 |
| Nasser Binsaif [37] | DT, RF, K-NN, ANN, SVM & LR | | BC Database of Coimbra (UCI) | 64% | 2022 |
| R. Sathesh Raaj [38] | hybrid CNN architecture | | (MIAS & DDSM | 97.83% | 2023 |
| Kristina lang et al.[39] | mammography screening | | 46 345 screen readings | 95% | 2023 |
| Akhil Kumar Das et al. [40] | MLISBCP | ✓ | | 97.53 % | 2024 |

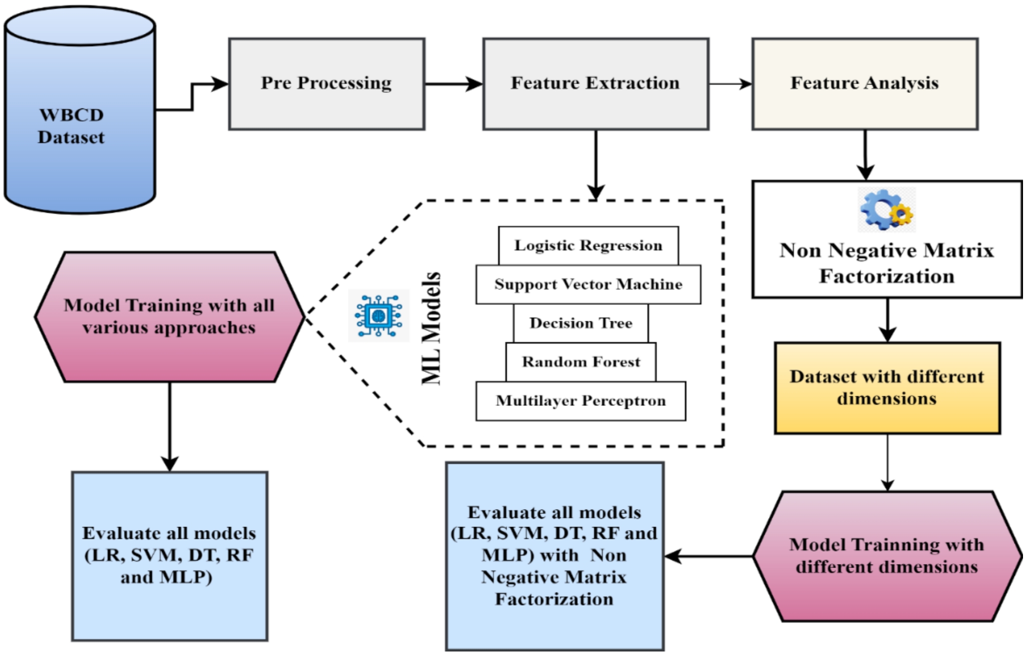

Figure 2. Flowchart of proposed model using NMF.

**DR:** The ability to reduce high-dimensional data to a lower-dimensional space while maintaining its fundamental traits makes dimensionality reduction (DR) an essential tool in machine learning [42]. It improves data utilization efficiency while reducing the computational burden on automated processes. The use of this technique not only enhances the computational efficiency of data processing, but also reduces the intricacy of the NN architecture and facilitates the effective construction of fuzzy inference rules.

**NMF:** NMF is an effective approach for decreasing the dimensionality of data and identifying important characteristics. Its adaptability is useful in a variety of fields, including bioinformatics, image processing, and text mining. The capacity to build interpretable and robust data representations highlights its importance to data scientists and researchers. Nonetheless, paying close attention to characteristics like rank selection, initialization tactics, and optimization approaches is critical for assuring meaningful and efficient data decomposition. Non-negative matrix factorization (NMF) is a method used in linear algebra as well as data analysis. NMF decomposes a matrix into two non-negative matrices, which is advantageous for certain data analysis applications such as feature extraction, image

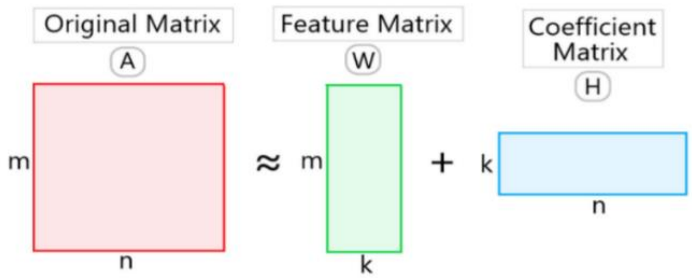


Figure 4: Non-Negative Matrix Factorization

processing, and text mining, in contrast to SVD that involves three matrices [42]. Mathematically, NMF) aims to find two non -ve matrix, W (s x t) and H (t x u), from a given non--ve matrix X of size s x u, where t is often smaller than s and u. Regarding factorization.

$$X \approx WH \quad .... (1)$$

X represents the initial non-negative matrix to be factorized. Matrix W is of dimensions s x t and contains only non-negative values. Each column in matrix W represents a fundamental vector, which is used collectively to estimate the data in matrix X. Matrix H is of size t x u and contains only non-negative values. H's columns indicate the coefficients of the basis vectors in W that are utilized to rebuild the columns of X. NMF aims to determine optimal choices for matrices W and H so that their multiplication roughly approximates the original matrix X, with the additional constraint that all members in W and H must be non-negative as shown in Figure-4.

### *3.3 Methods Incorporated.*

Over the course of this part, we proposed classification models that primarily made use of LR, SVM, DT, RF, and MLP.

**LR:** Logistic Regression is a statistics and Machine Learning model frequently utilized for binary classification applications [43]. This model may predict one of two possible outcomes by using a set of input data. Although sometimes referred to as logistic regression, it is noted that this approach is mostly utilized for classification purposes rather than regression analysis, which is typically employed for predicting continuous outcomes. Binary categorization has two potential outcomes, typically denoted as 0 and 1, or "B" and "M". Linear regression utilizes a group of input properties, often known as independent variables or predictors, to forecast outcomes. These attributes might be either numerical or category. The chance of a positive output and its association with the input attributes can be predicted using logistic regression. Making a weighted linear combination of input information using coefficients is what the procedure is all about.

$$Log\left(odds\right) = Log\left(\frac{P\left(Y=M/A\right)}{P\left(Y=B/A\right)}\right) = Logit\left(P\left(Y=M/A\right)\right) = \alpha_0 + \alpha_1 A_1 + \alpha_2 A_2 + .. + \alpha_n A_n \quad \ldots (2)$$

$$P\left(Y=M\right) = \frac{1}{1+e^{-z}} \quad \ldots (3)$$

Z represents linear programing. The sigmoid function converts the linear programing into a number ranging from 0 to 1, representing the likelihood of a positive event. When making a binary classification, must select a decision threshold, often set at 0.5. The model detects the +ve phase (1) if the anticipated probability is equal to or greater than the threshold; otherwise, it predicts the -ve phase (0). The logistic regression model is trained using a dataset with labelled examples. The goal is to identify the values of the coefficients ($\alpha_0$, $\alpha_1$, $\alpha_2$, ..., $\alpha_n$) that best fit the data and minimize the prediction error. Following the training process, the model's performance is assessed on a distinct dataset utilizing metrics such as accuracy, precision, recall, F1-score, and the ROC curve.

**SVM:** The Support Vector Machine is mainly recognized as a seminal method within the domain of Supervised Learning. It is commonly employed in assignments involving both classification and regression [44]. Its primary application is in the branch of categorization issues in the domain of machine learning. The main objective of this method is to establish an appropriate decision boundary, which can be linear or non-linear, to effectively segregate classes within an n-dimensional space. This divide allows for accurate categorization of upcoming data pieces. A hyperplane is a border that effectively separates data points optimally. This method detects the crucial points or vectors that have a substantial impact on creating the hyperplane [45]. The instances that demonstrate exceptional characteristics are frequently known as support vectors, thus leading to the acronym SVM for the technique.

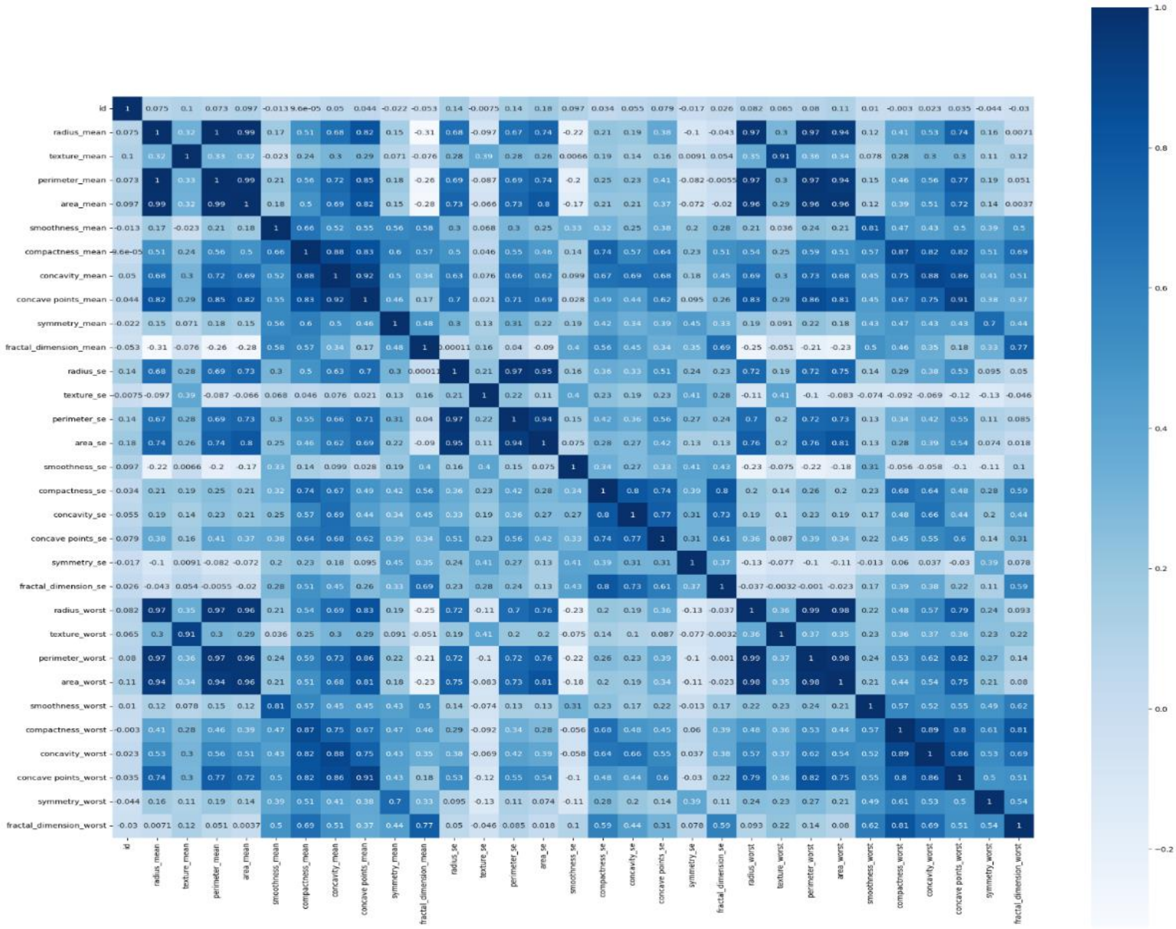

Figure 5. Heatmap (Correlation between the different features)

**DT:** Decision tree technique is mainly used in supervised machine learning to classification & regression methods. The structure is defined via a tree-structure arrangement, where the data is divided in subsets through a recursive method [46]. This segmentation is established by identifying the most influential attributes, often referred to as features, to aid in decision-making tasks.

**RF:** The Random Forest algorithm is a famous ensemble learning strategy in Machine Learning, commonly applied to classification and regression both [47]. This is a strong and flexible method that creates a set of DTs while training, and then merges their predictions to decrease overfitting and increase accuracy. Random Forest creates a "forest" out of an ensemble of decision trees. The "bagging" approach may train these. The bagging approach works by merging many learning models [48][49]. In the case of n decision trees, the final prediction is generated by collecting a vote from all n trees [50]. In machine learning, RFs are often referred to as an ensemble or bagging strategy. A decision tree is a set of rules learned from a dataset that may be used to predict future data. RFs are sometimes termed an ensemble or bagging approach in machine learning. A decision tree is a collection of rules learned from a dataset that can predict future data.

Entropy is used to measure the randomness of a system.

$$\text{Entropy } \Phi(s) = -\sum p(s)*\log(p(s)) \qquad \ldots(4)$$

$\Phi(s)$ is representation where $p(s)$ is the probability/percentage of class s in a node.

Gini index / Gini impurity measures sampling inequality. It ranges from 0 to 1. A Gini score of 0 implies the sample is entirely homogenous, while a Gini index of 1 suggests the components are maximally unequal.

$$\text{Gini Index} = 1 - \sum_{i=0}^{n} p_i^2 \qquad \ldots (5)$$

$P_i$ is the proportion of the samples that belong to class n for a particular node.

The information gain is the amount by which the system's Entropy decreases owing to the split. The tree is based on observations.

$$Gain(S, A) \equiv \text{Entropy}(S) - \sum_{v \in D_A} \frac{|S_v|}{|S|} \text{Entropy}(S_v) \qquad (6)$$

S refers to prior Entropy, $S_v$ refers to post entropy.

Each tree focuses on a specific aspect of the data by considering a significantly smaller number of parameters at each split compared to the total number of features available.

**MLP:** A Multi-Layer Perceptron, often known as an MLP, is a sort of Artificial Neural Network (ANN) that is utilized in the field of machine learning to address supervised learning objectives, such as classification and regression tasks [51].The model is adaptable and powerful, able to approximate intricate, non-linear combination inside data. The term multi-layer in its name denotes its structure, comprising numerous layers of all connected nodes, also known as neurons.

**Performance measures:** Assessment measures for categorization include Accuracy, Precision, Recall, Specificity, and F-measure [52]. CM components, which offer data on expected and real results, are utilized to create these metrics. The following equations represent performance measures in real-world values. true positive = TP, true negative = TN, false positive = FP and false negative = FN.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad \ldots \qquad (7)$$

$$Recall = \frac{TP}{TP+FN} \qquad \ldots \qquad (8)$$

$$Precision = \frac{TP+TN}{TP+TN+FP+FN} \qquad \ldots \qquad (9)$$

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \qquad \ldots \qquad (10)$$

$$Specificity = \frac{TN}{FP+TN} \qquad \ldots \qquad (11)$$

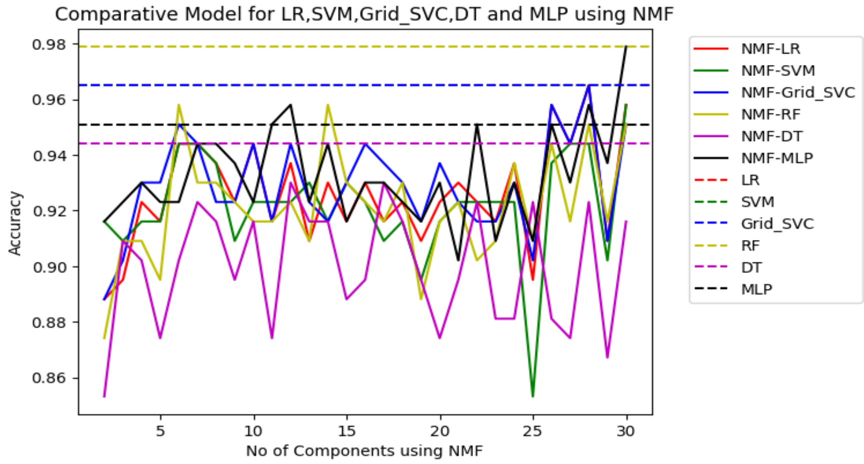


Figure 6: Result with NMF

## 4. Result Analysis

This section provides information on the experimental protocols, setup, and instruments used to compare DR technique mainly NMF for BC diagnosis. The study uses the publicly accessible WBCD dataset from the UCI ML Repository for research purposes. The dataset consists of characteristics that have been collected from a digitized picture of a fine needle aspirate obtained from a breast mass. The dataset has a total of 569 occurrences, each characterized by 30 distinct feature characteristics. These examples are classified into two categories: malignant (M) and benign (B). We employed the performance of every method using metrics such as, precisions, accuracies, recalls, F1-scores, as shown in Figure-6.
The higher performance of NMF with MLP is due to the synergy between NMF and MLP. NMF efficiently extracts relevant features from data while maintaining non-negativity, possibly improving interpretability and resilience. MLP uses these characteristics to learn complicated non-linear correlations between features and class labels, which may lead to enhanced classification accuracy.

## 5. Discussion

The study shows that dimensionality reduction approaches can greatly improve the effectiveness of ML classifiers in BC detection. Reducing the number of features simplifies the data and enhances the classifiers' capacity to detect crucial patterns. We found NMF to be highly effective in our comparison analysis, resulting in a maximum accuracy of 97.90%. This discovery highlights the significance of feature selection and NMF in enhancing the Breast Cancer (BC) detection model. Figure-8 demonstrates that our model obtained an accuracy of 97.90%, outperforming the old model. We conducted a comparison analysis using various BC datasets that represented diverse patient groups and data sources. The continual enhancement in classifier performance on various datasets underscores the resilience and adaptability of DR approaches.

### *5.1 Comparison*

Within the framework of an ensemble approach, every breast cancer detection method that makes use of ML Classifier with NMF presents both potential and challenges. The selection of the most appropriate method is influenced by several factors, including the characteristics of the data and resources available for processing that is necessary. Regarding the classifiers that are now in use, we compared DR methods such as PCA, NMF, Deep Semi NMF, and Autoencoder with Classifiers in Table-3. In comparison to the findings of past study, the work that we offered produced the best outcomes.
The proposed method outperforms all other individual methods, including:

- NMF & PCA ([51], 2018) by 5.3%
- MF & Deep semi NMF ([52], 2021) by 26.54%
- NMF ([53], 2022) by 21%
- Deep Semi NMF ([54], 2023) by 16.3%

### *5.2 Limitations and Future Scope*

While our work effectively illustrates the usefulness of dimensionality reduction (DR) methodologies for breast cancer detection utilising Non-Negative Matrix Factorization (NMF), we admit several limitations that will allow for future research developments.

**a) Interpretability of DR Models:** Our present study concentrated on the performance enhancement produced using NMF, however more research is needed to determine the interpretability of these models.

In the therapeutic arena, it is critical to explain and justify model predictions to healthcare providers and patients. This transparency builds confidence and informs therapeutic decision-making.

Future research should investigate methods for evaluating NMF models and turning their findings into clinically relevant insights. This may involve:

- Using methods such as feature significance analysis to determine which original characteristics contribute the most to the NMF components.
- Creating visualization tools to display the underlying patterns and links discovered by NMF in an understandable format for medical experts.

**b) Exploration of other Approaches:** While our study focused on NMF, a well-established DR technique, it is critical to evaluate the possibility of other approaches for breast cancer detection. This might include:

- Deep learning-based disaster risk reduction techniques: These approaches have demonstrated promising results in a variety of domains and may open up new possibilities for feature extraction and representation learning in breast cancer imaging.
- Hybrid models: Combining NMF with other DR techniques or machine learning models may allow you to take advantage of the characteristics of each approach and get even better results.

**c) Continuous Improvement for Clinical Implementation:** Our discovery lays the groundwork for future research targeted at turning NMF and other DR techniques into clinically relevant tools for breast cancer detection. Future study should concentrate on:

- Validating DR models' performance and interpretability in large-scale clinical trials.
- Creating user-friendly software solutions that include DR models into existing clinical procedures, making them available to healthcare practitioners for practical usage.

By addressing these constraints and investigating future prospects, we may improve and advance the use of DR methods for breast cancer detection, resulting in better patient care and outcomes.

Table 3. Comparation of results

| **Reference, Year** | **DR Methods** | **Result** |
|---|---|---|
| [53] , 2018 | NMF & PCA | 92.6 |
| [54], 2021 | MF & Deep semi NMF | 71.36 |
| [55], 2022 | NMF | 76.9 |
| [56], 2023 | Deep Semi NMF | 81.6 |
| [57], 2023 | PCA & Auto encoder | 95.50 |
| **Proposed, 2024** | **NMF with MLP** | **97.90** |

Table-2. Compared Result with NMF with Components

| Components | LR | SVM | Grid_SVC | RF | DT | MLP |
|---|---|---|---|---|---|---|
| 1 | 0.8881 | 0.9160 | 0.8881 | 0.8741 | 0.8532 | 0.9161 |
| 2 | 0.8951 | 0.9090 | 0.9021 | 0.9091 | 0.9091 | 0.9231 |
| 3 | 0.9440 | 0.9440 | 0.9441 | 0.9301 | 0.9231 | 0.9441 |
| 4 | 0.9370 | 0.9370 | 0.9231 | 0.9301 | 0.9161 | 0.9441 |
| 5 | 0.9230 | 0.9090 | 0.9231 | 0.9231 | 0.8951 | 0.9371 |
| 6 | 0.9440 | 0.9230 | 0.9441 | 0.9161 | 0.9161 | 0.9231 |
| 7 | 0.9160 | 0.9300 | 0.9301 | 0.9301 | 0.8881 | 0.9161 |
| 8 | 0.9300 | 0.9230 | 0.9441 | 0.9231 | 0.8951 | 0.9301 |
| 9 | 0.9160 | 0.9090 | 0.9371 | 0.9161 | 0.9301 | 0.9301 |
| 10 | 0.9230 | 0.9160 | 0.9301 | 0.9301 | 0.9161 | 0.9231 |
| 11 | 0.9230 | 0.9231 | 0.9161 | 0.9021 | 0.9231 | 0.9511 |
| 12 | 0.9370 | 0.9231 | 0.9301 | 0.9371 | 0.8811 | 0.9301 |
| 13 | **0.9650** | 0.9441 | 0.9650 | 0.9511 | 0.9231 | 0.9580 |
| 14 | 0.9090 | 0.9021 | 0.9091 | 0.9161 | 0.8671 | 0.9371 |
| 15 | 0.9580 | 0.9580 | 0.9511 | 0.9511 | 0.9161 | **0.9790** |

## 6. Conclusion

This study validates the efficacy of Non-Negative Matrix Factorization (NMF) as a dimensionality reduction technique in breast cancer screening, addressing challenges posed by high-dimensional feature spaces. A comparative analysis evaluated the performance of machine learning models, including Support Vector Machines (SVM), Decision Trees (DT), Logistic Regression (LR), Random Forests (RF), and Multi-Layer Perceptron's (MLP), when combined with NMF. The findings emphasize that the choice of NMF parameters, machine learning algorithms, and hyperparameter optimization significantly influences detection accuracy and interpretability. This research provides valuable insights into the role of NMF in enhancing diagnostic precision and supports its integration into clinical workflows for early breast cancer detection and improved patient outcomes. By offering a comprehensive analysis, this work contributes to the development of more effective diagnostic tools in the field.

*Acknowledgment*

Thank you from the bottom of our hearts to everyone who had a hand in making this research project a reality. The invaluable guidance, instruction, and supervision provided by Manish Madhava Tripathi and Rafeeq Ahmed is deeply appreciated. Both the preparation and carrying out of this task benefited greatly from their knowledge and advice. This study recognized with Integral University MCN: IU/R&D/2024-MCN0002746.

**Disclosure statement**: No authors have reported any financial or other conflicts of interest that could affect the results of this study.

**Data Availability Statement:** Available on UCI Repository

## 7. References


[1] S. S. Coughlin and D. U. Ekwueme, “Breast cancer as a global health concern,” *Cancer Epidemiol*, vol. 33, no. 5, pp. 315–318, Nov. 2009, doi: 10.1016/J.CANEP.2009.10.003.

[2] F. Pareja, C. Marchiò, and J. S. Reis-Filho, “Molecular diagnosis in breast cancer,” *Diagn Histopathol*, vol. 24, no. 2, pp. 71–82, Feb. 2018, doi: 10.1016/J.MPDHP.2018.01.001.

[3] X. Lin and P. C. Boutros, “Optimization and expansion of non-negative matrix factorization,” *BMC Bioinformatics*, vol. 21, no. 1, pp. 1–10, Jan. 2020, doi: 10.1186/S12859-019-3312-5/FIGURES/4.

[4] Y. Feng *et al.*, “Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis,” *Genes Dis*, vol. 5, no. 2, pp. 77–106, Jun. 2018, doi: 10.1016/J.GENDIS.2018.05.001.

[5] M. Bernardini, L. Romeo, P. Misericordia, and E. Frontoni, “Discovering the Type 2 Diabetes in Electronic Health Records Using the Sparse Balanced Support Vector Machine,” *IEEE J Biomed Health Inform*, vol. 24, no. 1, pp. 235–246, Jan. 2020, doi: 10.1109/JBHI.2019.2899218.

[6] X. Zhou *et al.*, “A Comprehensive Review for Breast Histopathology Image Analysis Using Classical and Deep Neural Networks,” *IEEE Access*, vol. 8, pp. 90931–90956, 2020, doi: 10.1109/ACCESS.2020.2993788.

[7] S. U. Khan, N. Islam, Z. Jan, I. Ud Din, and J. J. P. C. Rodrigues, “A novel deep learning based framework for the detection and classification of breast cancer using transfer learning,” *Pattern Recognit Lett*, vol. 125, pp. 1–6, Jul. 2019, doi: 10.1016/J.PATREC.2019.03.022.

[8] M. Parida and S. Chand, “Hermite Hadamard and Fejer type integral inequalities for harmonic convex (concave) fuzzy mappings,” *International Journal of Mathematics in Operational Research*, vol. 18, no. 1, pp. 1–20, 2021, doi: 10.1504/IJMOR.2021.112270.

[9] N. Fatima, L. Liu, S. Hong, and H. Ahmed, “Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis,” 2020, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2020.3016715.

[10] M. S. Harinishree, C. R. Aditya, and D. N. Sachin, “Detection of Breast Cancer using Machine Learning Algorithms - A Survey,” in *Proceedings - 5th International Conference on Computing Methodologies and Communication, ICCMC 2021*, Institute of Electrical and Electronics Engineers Inc., Apr. 2021, pp. 1598–1601. doi: 10.1109/ICCMC51019.2021.9418488.

[11] R. Turkki *et al.*, “Breast cancer outcome prediction with tumour tissue images and machine learning,” *Breast Cancer Res Treat*, vol. 177, no. 1, pp. 41–52, Aug. 2019, doi: 10.1007/s10549-019-05281-1.

[12] F. A. K. Al-Fahaidy, B. Al-Fuhaidi, I. Al-Darouby, F. Al-Abady, M. Al-Qadry, and A. Al-Gamal, “A Diagnostic Model of Breast Cancer Based on Digital Mammogram Images Using Machine Learning Techniques,” *Applied Computational Intelligence and Soft Computing*, vol. 2022, 2022, doi: 10.1155/2022/3895976.

[13] V. R. Allugunti, “Breast cancer detection based on thermographic images using machine learning and deep learning algorithms,” *International Journal of Engineering in Computer Science*, vol. 4, no. 1, pp. 49–56, Jan. 2022, doi: 10.33545/26633582.2022.V4.I1A.68.

[14] M. A. Naji, S. El Filali, K. Aarika, E. H. Benlahmar, R. A. Abdelouhahid, and O. Debauche, “Machine Learning Algorithms for Breast Cancer Prediction and Diagnosis,” in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 487–492. doi: 10.1016/j.procs.2021.07.062.

[15] V. D. P. Jasti *et al.*, “Computational Technique Based on Machine Learning and Image Processing for Medical Image Analysis of Breast Cancer Diagnosis,” *Security and Communication Networks*, vol. 2022, 2022, doi: 10.1155/2022/1918379.

[16] M. A. Elsadig, A. Altigani, and H. T. Elshoush, “Breast cancer detection using machine learning approaches: a comparative study,” *International Journal of Electrical and Computer Engineering*, vol. 13, no. 1, pp. 736–745, Feb. 2023, doi: 10.11591/ijece.v13i1.pp736-745.

[17] D. Zahras and Z. Rustam, “Cervical Cancer Risk Classification Based on Deep Convolutional Neural Network,” *Proceedings of ICAITI 2018 - 1st International Conference on Applied Information Technology and Innovation: Toward A New Paradigm for the Design of Assistive Technology in Smart Home Care*, pp. 149–153, Jul. 2018, doi: 10.1109/ICAITI.2018.8686767.

[18] A. SAYGILI, “Classification and Diagnostic Prediction of Breast Cancers via Different Classifiers,” *International Scientific and Vocational Studies Journal*, vol. 2, no. 2, pp. 48–56, Dec. 2018, Accessed: Sep. 22, 2023. [Online]. Available: https://dergipark.org.tr/en/pub/bilmes/issue/42033/474827

[19] S. Gupta and M. Kumar, “Prostate cancer prognosis using multi-layer perceptron and class balancing techniques,” *ACM International Conference Proceeding Series*, pp. 1–6, Aug. 2021, doi: 10.1145/3474124.3474125.

[20] “Breast Cancer Detection using Machine Learning Way,” *International Journal of Recent Technology and Engineering*, vol. 8, no. 2S3, pp. 1402–1405, Aug. 2019, doi: 10.35940/ijrte.B1260.0782S319.

[21] K. Bouchard, J. Maitre, C. Bertuglia, and S. Gaboury, “Activity Recognition in Smart Homes using UWB Radars,” *Procedia Comput Sci*, vol. 170, pp. 10–17, 2020, doi: 10.1016/j.procs.2020.03.004.

[22] M. S. Yarabarla, L. K. Ravi, and A. Sivasangari, “Breast Cancer Prediction via Machine Learning,” in *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, IEEE, Apr. 2019, pp. 121–124. doi: 10.1109/ICOEI.2019.8862533.

[23] K. Maharana, S. Mondal, and B. Nemade, “A review: Data pre-processing and data augmentation techniques,” *Global Transitions Proceedings*, vol. 3, no. 1, pp. 91–99, Jun. 2022, doi: 10.1016/J.GLTP.2022.04.020.

[24] N. A. Farooqui, A. K. Mishra, and R. Mehra, “Concatenated deep features with modified LSTM for enhanced crop disease classification,” *Int J Intell Robot Appl*, vol. 7, no. 3, pp. 510–534, Sep. 2023, doi: 10.1007/S41315-022-00258-8/METRICS.

[25] Z. Gu, "Complex heatmap visualization," *iMeta*, vol. 1, no. 3, p. e43, Sep. 2022, doi: 10.1002/IMT2.43.

[26] S. K. Verma, D. Arora, and R. Bhardwaj, "Breast Cancer Survival Rate Prediction in Mammograms Using Machine Learning," *Proceedings - IEEE 2020 2nd International Conference on Advances in Computing, Communication Control and Networking, ICACCCN 2020*, pp. 169–171, Dec. 2020, doi: 10.1109/ICACCCN51052.2020.9362741.

[27] Z. Zeng, A. H. Vo, C. Mao, S. E. Clare, S. A. Khan, and Y. Luo, "Cancer classification and pathway discovery using non-negative matrix factorization," *J Biomed Inform*, vol. 96, p. 103247, Aug. 2019, doi: 10.1016/J.JBI.2019.103247.

[28] P. Kathale and S. Thorat, "Breast Cancer Detection and Classification," *International Conference on Emerging Trends in Information Technology and Engineering, ic-ETITE 2020*, Feb. 2020, doi: 10.1109/IC-ETITE47903.2020.367.

[29] S. Ekici and H. Jawzal, "Breast cancer diagnosis using thermography and convolutional neural networks," *Med Hypotheses*, vol. 137, p. 109542, Apr. 2020, doi: 10.1016/J.MEHY.2019.109542.

[30] R. Turkki *et al.*, "Breast cancer outcome prediction with tumour tissue images and machine learning," *Breast Cancer Res Treat*, vol. 177, no. 1, pp. 41–52, Aug. 2019, doi: 10.1007/S10549-019-05281-1/TABLES/2.

[31] R. Rawal, "BREAST CANCER PREDICTION USING MACHINE LEARNING," 2020, Accessed: Mar. 04, 2024. [Online]. Available: www.jetir.org

[32] J. Wu and C. Hicks, "Breast Cancer Type Classification Using Machine Learning," *J Pers Med*, vol. 11, no. 2, p. 61, Jan. 2021, doi: 10.3390/jpm11020061.

[33] S. Jafarzadeh Ghoushchi, R. Ranjbarzadeh, S. A. Najafabadi, E. Osgooei, and E. B. Tirkolaee, "An extended approach to the diagnosis of tumour location in breast cancer using deep learning," *J Ambient Intell Humaniz Comput*, vol. 14, no. 7, pp. 8487–8497, Jul. 2023, doi: 10.1007/S12652-021-03613-Y/TABLES/6.

[34] S. Ara, A. Das, and A. Dey, "Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms," in *2021 International Conference on Artificial Intelligence (ICAI)*, IEEE, Apr. 2021, pp. 97–101. doi: 10.1109/ICAI52203.2021.9445249.

[35] R. Hamamoto *et al.*, "Application of non-negative matrix factorization in oncology: one approach for establishing precision medicine," *Brief Bioinform*, vol. 23, no. 4, pp. 1–17, Jul. 2022, doi: 10.1093/BIB/BBAC246.

[36] F. A. K. Al-Fahaidy, B. Al-Fuhaidi, I. AL-Darouby, F. AL-Abady, M. AL-Qadry, and A. AL-Gamal, "A Diagnostic Model of Breast Cancer Based on Digital Mammogram Images Using Machine Learning Techniques," *Applied Computational Intelligence and Soft Computing*, vol. 2022, pp. 1–17, Sep. 2022, doi: 10.1155/2022/3895976.

[37] N. Binsaif, "Application of Machine Learning Models to the Detection of Breast Cancer," *Mobile Information Systems*, vol. 2022, 2022, doi: 10.1155/2022/7340689.

[38] R. Sathesh Raaj, "Breast cancer detection and diagnosis using hybrid deep learning architecture," *Biomed Signal Process Control*, vol. 82, p. 104558, Apr. 2023, doi: 10.1016/J.BSPC.2022.104558.

[39] K. Lång *et al.*, "Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study," *Lancet Oncol*, vol. 24, no. 8, pp. 936–944, Aug. 2023, doi: 10.1016/S1470-2045(23)00298-X.

[40] A. K. Das, S. K. Biswas, A. Mandal, A. Bhattacharya, and S. Sanyal, "Machine Learning based Intelligent System for Breast Cancer Prediction (MLISBCP)," *Expert Syst Appl*, vol. 242, p. 122673, May 2024, doi: 10.1016/J.ESWA.2023.122673.

[41] S. Ayesha, M. K. Hanif, and R. Talib, "Overview and comparative study of dimensionality reduction techniques for high dimensional data," *Information Fusion*, vol. 59, pp. 44–58, Jul. 2020, doi: 10.1016/J.INFFUS.2020.01.005.

[42] A. Hassani, A. Iranmanesh, and N. Mansouri, "Text mining using nonnegative matrix factorization and latent semantic analysis," *Neural Comput Appl*, vol. 33, no. 20, pp. 13745–13766, Oct. 2021, doi: 10.1007/S00521-021-06014-6/FIGURES/14.

[43] A. Zaidi and A. S. M. Al Luhayb, "Two Statistical Approaches to Justify the Use of the Logistic Function in Binary Logistic Regression," *Math Probl Eng*, vol. 2023, pp. 1–11, Apr. 2023, doi: 10.1155/2023/5525675.

[44] Y. Xu, B. Klein, G. Li, and B. Gopaluni, "Evaluation of logistic regression and support vector machine approaches for XRF based particle sorting for a copper ore," *Miner Eng*, vol. 192, p. 108003, Feb. 2023, doi: 10.1016/J.MINENG.2023.108003.

[45] R. M. Devi *et al.*, "Plant type classification based on leaves using Fusion based Support Vector Machine," *2023 International Conference on Computer Communication and Informatics, ICCCI 2023*, 2023, doi: 10.1109/ICCCI56745.2023.10128587.

[46] M. Awad and S. Fraihat, "Recursive Feature Elimination with Cross-Validation with Decision Tree: Feature Selection Method for Machine Learning-Based Intrusion Detection Systems," *Journal of Sensor and Actuator Networks 2023, Vol. 12, Page 67*, vol. 12, no. 5, p. 67, Sep. 2023, doi: 10.3390/JSAN12050067.

[47] D. Datta, M. Bhattacharya, S. Suman Rajest, T. Shynu, R. Regin, and S. Silvia Priscila, "Development of predictive model of diabetic using supervised machine learning classification algorithm of ensemble voting," *Int J Bioinform Res Appl*, vol. 19, no. 3, pp. 151–169, 2023, doi: 10.1504/IJBRA.2023.133695.

[48] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *The Stata Journal: Promoting communications on statistics and Stata*, vol. 20, no. 1, pp. 3–29, Mar. 2020, doi: 10.1177/1536867X20909688.

[49] "Introduction to Random Forest in Machine Learning."

[50] "A Step by Step ID3 Decision Tree Example."

[51] A. Kurani, P. Doshi, A. Vakharia, and M. Shah, "A Comprehensive Comparative Study of Artificial Neural Network (ANN) and Support Vector Machines (SVM) on Stock Forecasting," *Annals of Data Science*, vol. 10, no. 1, pp. 183–208, Feb. 2023, doi: 10.1007/S40745-021-00344-X/METRICS.

[52] M. Botlagunta *et al.*, "Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms," *Sci Rep*, vol. 13, no. 1, p. 485, Jan. 2023, doi: 10.1038/s41598-023-27548-w.

[53] B. Wu, J. T. Smith, L. Zhang, X. Gao, and R. R. Alfano, "Characterization and discrimination of human breast cancer and normal breast tissues using resonance Raman spectroscopy," *https://doi.org/10.1117/12.2288094*, vol. 10489, pp. 109–115, Feb. 2018, doi: 10.1117/12.2288094.

[54] B. Yousefi, H. M. Sharifipour, and X. P. V. Maldague, "A Diagnostic Biomarker for Breast Cancer Screening via Hilbert Embedded Deep Low-Rank Matrix Approximation," *IEEE Trans Instrum Meas*, vol. 70, 2021, doi: 10.1109/TIM.2021.3085956.

[55] N. Vigil, B. M. Nouri, H. C. Fernandes, C. Ibarra-Castanedo, X. P. V. Maldague, and B. Yousefi, "Convex Factorization Embedding Thermography for Breast Cancer Diagnostic," *IEEE Open Journal of Instrumentation and Measurement*, vol. 1, pp. 1–8, Sep. 2022, doi: 10.1109/OJIM.2022.3203452.

[56] L. D. Buitrago, J. J. Azarnoosh, X. P. V. Maldague, and B. Yousefi, "Optimal thermomic biomarkers for early diagnosis of breast cancer," *https://doi.org/10.1117/12.2663835*, vol. 12536, pp. 245–254, Jun. 2023, doi: 10.1117/12.2663835.

[57] M. F. Kabir, T. Chen, and S. A. Ludwig, "A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction," *Healthcare Analytics*, vol. 3, p. 100125, Nov. 2023, doi: 10.1016/J.HEALTH.2022.100125.