International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

An Efficient Novel Approach on WBCD for Early Detection of Breast Cancer using Distributed Machine Learning Techniques

Naidu Kirankumar[1], T. Santhi Sri[2]

Submitted:13/03/2024 **Revised**: 28/04/2024 **Accepted**: 05/05/2024

Abstract: Breast cancer(BC) will be the most common malignancy in women by 2022,with over 3 million new cases. Due to alterations in risk factor profiles, enhanced cancer registries, and quicker detection over the past three decades, both the incidence and death rates of cancer have risen. The total risk factors for BC are largely made up of modifiable and immutable risk factors. Now, 80 percent of BC patients are over 50. The molecular subtype and stage affect survival. One of the most important problems affecting people in underdeveloped nations is the death rate from cancer. Even while there are many ways to avoid getting cancer in the first place, some illnesses can still be fatal. Breast cancer is one of the most common types of cancer, and early detection is key to a successful course of therapy. One of the most crucial aspects of breast cancer treatment is a precise diagnosis. Several studies that have been written about in the literature can indicate the type of breast cancer. Using data on breast cancer tumours from Dr. William H. Wahlberg of the Hospital of the University of Wisconsin, the kind of breast tumour was predicted in this study. On this dataset, data visualization and machine learning techniques like Distributed Polynomial Kernel SVM were used. This study compared the identification and diagnosis of breast cancer using data visualization and machine learning methods. The most accurate classification achieved by the Distributed Logistic Regression Model employing all features is improved by the proposed approach. Modern technology has been enhanced with new hybrid frameworks and models for higher security, data storage of large volumes, and precision. Using machine learning classification methods, a function that can forecast the discrete class of new items has also been improved.

Keywords *Wisconsin breast cancer dataset, naive Bayesian, logistic regression, k-nearest Neighbors, Independent Component Analysis.*

1. Introduction

A typical problem in the field of medical data analysis is the categorization of breast cancer using machine learning techniques on the WBCD (Wisconsin Breast Cancer Diagnostic) dataset. The aim of building prediction models that can precisely distinguish between benign and malignant breast mass samples using multiple variables derived from fine needle aspirates (FNAs) is the aim. Here's an overview of the process of applying machine learning methods to breast cancer classification on the WBCD dataset: Data Preprocessing: Load the WBCD dataset, which contains numeric features representing characteristics of breast mass samples. Perform data cleaning, handling missing values, and ensuring data integrity. Split the dataset into training and testing subsets to evaluate model performance.

Feature Selection and Engineering: Analyze the dataset to identify relevant features for breast cancer classification. Apply feature selection techniques such as correlation analysis, feature importance, or dimensionality reduction methods to select the most informative features. Perform feature engineering, which may involve transforming, normalizing, or scaling the features to improve model performance.

Visualize the model's performance, feature importance, decision boundaries, or other relevant aspects to aid in understanding and communicating the findings. Machine learning methods provide a powerful approach to automate breast cancer classification on the WBCD dataset. By leveraging the available features and applying suitable algorithms, these models can assist in early detection and diagnosis of breast cancer, aiding in better patient outcomes and treatment decisions.

Breast cancer is the most prevalent type of cancer, accounting for an amazing 12% of all new cases each year, according to the World Health Organisation. In 2022, there are projected to be 287,850 new cases of cancer. Between the time a biopsy is carried out and the patient is told of the results, numerous resources are used. To make a diagnosis, a large team of concerned physicians, skilled lab workers, experienced pathologists, and efficient transcriptionists is required.

Currently, one of the most common cancers is breast cancer. According to estimates from GLOBOCAN 2020, there will likely be 2.3 million new instances of cancer worldwide, ranking it as the sixth leading cause of cancer-related deaths [1-2]. In comparison to transitioning nations (Australia/New Zealand, Western Europe, Northern America, and Northern Europe), Melanesia, Western

Naidu Kirankumar[1]
1Research Scholar, Department of CSE, Koneru Lakshmaiah Education Foundation, AP,India.
T. Santhi Sri[2], Department of CSE, Koneru Lakshmaiah Education Foundation, AP,India.
** Corresponding Author Email: nkirankuamr@gmail.com*

Africa, Micronesia/Polynesia, and the Caribbean had higher reported rates of breast cancer mortality (incidence rate is around 88% higher). For a potential decline in the incidence rate of breast cancer and the adoption of early treatment, a number of initiatives, including general prevention practises and screening programmes, are essential. The Breast Health worldwide Initiative (BHGI) is currently tasked with developing the necessary plans and policies to provide the most efficient breast cancer control on a worldwide level [3-6]. We mainly focused on female breast cancer in this review study because it is now the most prevalent disease in women.

With 9.6 million fatalities in 2018, cancer was the second-leading cause of death globally. One of every six disease-related fatalities globally is caused by cancer. Cancer is the primary cause of death in low- and middle-income nations [7]. The three diseases that most frequently affect women—colorectal, lung, and breast cancers—are responsible for half of all cancer cases. In addition, 30% of all new cancer diagnoses in women are for breast cancer [8-11]. A data set is evaluated in order to extract relevant correlations and information from it using machine learning (ML) techniques. Also, it creates a digital representation of the data's most likely interpretation. According to experts studying cancer, it describes ML methods for reducing early exposure and predicting malignancy [12-16]. Analyze several machine learning techniques to identify and forecast breast cancer risk. When compared using experimental data, the SVM classification approach had the best accuracy (97.13%) and lowest error rate. The data set for the study was breast cancer, and Weka was the preferred machine learning tool. The accuracy, recall, precision, and ROC area of the machine learning category presentation are evaluated. They claim that the RF method has the best ROC area, while the BN approach has the highest recall and most accurate values [17-21]. The breast cancer recurrence rate at two years was calculated by Ahmed et al. using machine learning techniques. The information spans the years 1997 to 2008 and comes from the Iranian Breast Cancer Center (ICBC) programme. The dataset consists of population information, 22 input parameters, and 1189 instances of breast cancer. mother. Decision trees (DTs), artificial neural networks (ANNs), and SVMs are used by support vector machines (SVMs), which have the greatest effects on accuracy and error rate.

2 Literature Review:

To see which machine learning techniques worked better, the category of breast cancer was predicted using SVM and ANN algorithms. Support vector machines (SVMs), which have proven to be highly effective in a range of pattern recognition problems, were first introduced by Vladimir Vapnik. SVMs may provide superior categorization outcomes when compared to numerous other categorization methods. SVM is a popular machine learning method for classifying data that is mostly used for cancer diagnosis and prognosis. Each module can represent significant samples because to the SVM's hyper plane, which divides the module into submodules using support vectors. In order to serve as a boundary for the conclusion, the hyper plane divides the two example clusters. Depending on the patient's age and the size of the tumour, SVM can be used to categorise tumours as benign or malignant. The biological neuron system and artificial neural networks can be contrasted (ANN). It is very similar to how the human brain organises its processing. It is made up of a significant number of interconnected nodes [21-23]. One can simulate distinctive and important non-linear utility using ANN. It is made up of a vast network of synthetic neuronsBreast cancer prognosis [24-25] In this study, a method for forecasting breast cancer utilising an ensemble strategy based on genetic algorithms is developed. This study predicts breast cancer using a number of machine learning techniques. To increase breast cancer prediction accuracy, an ensemble technique is used. A weighted average ensemble strategy based on GA is developed using the classification dataset to overcome the limitations of the conventional weighted average method. The weighted average based on a genetic algorithm was used. Methods of machine learning are being researched to increase the accuracy of diagnoses. Random Forest and K-Nearest Neighbors are two of the techniques that are put side by side. The University of California, Irvine Machine Learning Repository purchased the dataset. The KNN algorithm is proven to perform considerably better than the other research approaches. In general, the most accurate model is K-Nearest Neighbor. Two categorization models, Random Forest and Boosted Trees, both had similar levels of precision. By utilizing the most precise classifier, the cancer can be located and a cure discovered early on. uses of machine learning in breast cancer To assess the data, they used the benchmark database. Wenbin Yue and Zidong Wang present numerous ML techniques and their application to BC analysis and forecasting in their book Diagnose and Prognosis [26-30]. Despite the fact that several algorithms have acquired extremely high exactness, new approaches are still needed. Although classification accuracy is crucial, it is not the sole consideration.

It is possible to identify, categorise, or distinguish between many forms of cancer and tumours using machine learning. To put it another way, machine learning is mostly employed as a technique to detect and recognise cancer. Recently, cancer researchers have tried to use machine learning for forecasting and prediction of cancer. There isn't much literature on the subject of machine learning and cancer prediction as a result[31-33].

Wisconsin Diagnostic Breast Cancer (WDBC) analyses

combinations of many standards in key variables and forecasts the breast cancer tumour category using machine learning (benevolent or malevolent). While being members of the same Breast cancer group, the values of the various qualities won't be the same because entity patients differ. Based on these figures, doctors categorise the same type of breast cancer tumour into several risky circumstances. As a result, techniques including cancer inclination, survivability, development, and impending are computed to help with cancer forecasting. Patients frequently receive timely, effective therapies sequentially.

The analysis of Neural Network [34] is based on statistical data. When it is necessary to provide overlapping class borders, fuzzy-based categorization [35] is utilised. Another method for categorising data that can produce and distinguish between several classes is multinomial logical regression [36]. Bayesian classifiers classify data using probabilistic models. Another form of categorization is the decision tree [37], which separates the dataset depending on specific criteria. The conditions and regulations are used to categorise the items. Researchers are also looking into a number of additional classification techniques. A rule-based expert system is another type of categorization system. The fuzzy set classification techniques provide a different framework for categorization [38-40]. These systems form conclusions using if-then rules. Labeling is necessary in order for the Bayesian classifier [41-42] to predict characteristics from a certain class. In this work, common values are grouped into a certain class. Numerous alternative categorization methods were looked at and applied to the variable's forecast [43].

3 Methods and Materials:

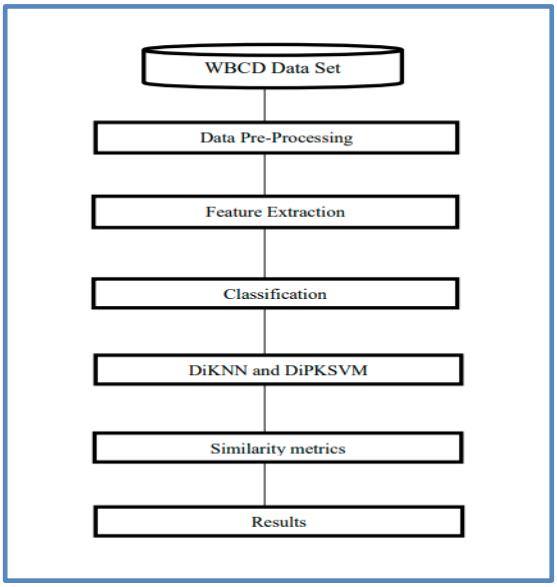

Fig:1 control flow diagram for proposed system

A group of machine learning algorithms known as supervised learning predicts an output based on labelled input data. The main objective of the most common supervised learning models, such as linear regression, is a linear relationship between the independent and dependent variables. In order to eliminate redundancy and relevancy, we calculated the Principal Component Analysis (PCA) and selected the proper components to prevent correlated factors that could affect our clustering analysis. One of the most frequent problems when analysing complex data is a large number of variables since they use a lot of memory and computing power. It is a method for reducing the feature space's dimension by feature extraction.

In this case, the blending method and unidentified sources are problematic. The ICA technique tackles this by making a few assumptions about the sources and mixing procedure. The sources are taken to be non-normally distributed and statistically independent (i.e., seeing the value of one component does not reveal anything about the value of the other). The extra presumption that the mixing system is linear is made by the ICA algorithm. The ICA technique will then deliver an unmixing matrix that contains estimates of the values from the sources; these estimates are frequently referred to as independent components (ICs).

3.1 DATASET

The most common disease in women and one of the main causes of death for females worldwide is breast cancer. Every year, 124 out of every 100,000 women receive a breast cancer diagnosis, and 23 of these women pass away from the condition. University of Wisconsin generated this dataset. This type of application often uses a dataset with 569 rows and 32 columns, a sizable number of entries, and only a few missing values. The data set consists of 32 factors in all, and we will make extensive use of them to forecast whether each tumour will be classified as benign or malignant in the end. The two possible values for our objective variable "Diagnosis" are "Malignant" (M) and "Benign" (B)[44]. A variety of measurements of the patient's tumour, which is assumed to be malignant, are included in the predictor variables of our dataset. Using R and Python (pandas, numpy, sklearn, matplotlib), we construct a large number of classification models, and the goal of our research is to choose the best model based on a range of model assessment matrices. Using a large amount of data on each patient's tumour that is being assessed, we will anticipate the probability of a malignant vs. benign diagnosis using the data set we are using [46]. Any of these factors can be used to categorise cancer; if their levels are notably high, malignant tissue might be present. The first input, referred to as the ID, contains a number used to identify the user [45]. There are two types of tissue diagnoses: benign and malignant, which make up the second criterion, the membrane diagnostic.

It is critical to make the proper identification of the tissue when both membranes are being treated differently for distinct forms of cancer. The distance between the centre and the next perimeter point is represented by the estimated averages, standard errors, and radius averages.

3.2 Distributed Polynomial Kernel SVM:

Distributed Polynomial Kernel Support Vector Machine (SVM) can be applied to the detection of breast cancer using the WBCD (Wisconsin Breast Cancer Diagnostic) dataset. Here's how it can support the detection of breast cancer. Dataset Preparation: The WBCD dataset contains various features extracted from breast cancer cell images, along with corresponding diagnosis labels (benign or malignant). It is important to preprocess the dataset by handling missing values, normalizing the feature values, and splitting it into training and testing sets.

Distributed Polynomial Kernel SVM Model Initialization: Initialize the SVM model with a polynomial kernel. The polynomial kernel is suitable for capturing non-linear relationships in the data. Parameters such as the degree of the polynomial and the penalty parameter C need to be determined. Cross-validation techniques can be used to find the optimal values for these parameters.

The degree of the polynomial and the coefficient of the polynomial are two polynomial kernel parameters that can be tweaked to improve performance. The polynomial kernel for polynomials of degree d is defined as

$$P_k(x_1, k_2)=(x_1^T x_2+c)^d$$

In the original space, x1 and x2 are vectors, and c is a constant. The parameter c can be used to manage the trade-off between the ability of the training data to fit the model and the size of the margin. A large c value will produce little training error, but overfitting is possible. A low c value may result in underfitting but a large training error. The degree d of the polynomial can be used to control the model's complexity.

$$\text{Mul (col matrix } [x_1, x_2], \text{ row matrix } [x_1, x_2]=\text{matrix } M[\{x_1^2,x_1x_2\},\{x_1x_2,x_2^2\}]$$

The important numbers to remember are x1, x2, x12, x22, and x1*x1. To find these new words, the non-linear dataset is translated into a new dimension and the features x1, x2, and x1*x2 are used. Because they can be used to separate data that is not linearly separable, kernels are helpful. The kernel approach can be utilised when a conventional SVM algorithm fails to separate the data. When the data is translated into a higher dimension and a new additional dimension is established for data distribution, this procedure yields more accurate findings.

.

Algorithm: DiPkSVM

Input:
a. Training dataset: {(x1, y1), (x2, y2), ..., (xn, yn)}, where xi represents the input features and yi represents the corresponding class labels.
b. Testing instance: x_test.
Parameters:
a. C: Penalty parameter for the error term.
b. degree: Degree of the polynomial kernel.
c. gamma: Kernel coefficient.
Output:
Predicted class label for the testing instance.
Algorithm:
Preprocess the data if necessary (e.g., normalization, feature scaling).
Initialize the SVM model with the polynomial kernel:
a. Define the kernel function: K(x, x') = (gamma * <x, x'> + 1)^degree.
b. Define the decision function: f(x) = Σ (alpha_i * y_i * K(x, x_i)) + b, where alpha_i are the Lagrange multipliers, y_i are the class labels, and x_i are the support vectors.
c. Initialize alpha_i = 0 for all training instances.
Compute the Gram matrix, which contains the pairwise kernel evaluations for all training instances:
Initialize an empty Gram matrix of size n x n.
For each pair (xi, xj) in the training dataset:
Compute the kernel value: K(xi, xj) = (gamma * <xi, xj> + 1)^degree.
Store the kernel value in the Gram matrix.
Train the SVM model:
• While the stopping criterion is not met:
• For each training instance (xi, yi):
a. Compute the predicted value: f(xi) = Σ (alpha_i * y_i * K(xi, x_i)) + b.
b. Compute the margin: margin = yi * f(xi).
c. If the instance violates the margin condition: margin < 1, update alpha_i:
• alpha_i := alpha_i + learning_rate * (1 - margin).
• Update the bias term b:
• b := b + learning_rate * Σ (y_i - f(xi)).
• Apply the necessary stopping criterion (e.g., reaching the maximum number of iterations, convergence criteria).

Predict the class label for the testing instance x_test:
Compute the decision function value: f(x_test) = Σ (alpha_i * y_i * K(x_test, x_i)) + b.
If f(x_test) > 0, the predicted class label is +1. Otherwise, it is -1.
Return the predicted class label for the testing instance.

Distributed Polynomial Kernel SVM Model Initialization: Create a polynomial kernel for the SVM model's initial state. The data's non-linear relationships can be captured using the polynomial kernel. It is necessary to determine variables like the polynomial's degree and the penalty parameter C. The best values for these parameters can be determined using cross-validation methods. Instruction of the Model: Utilise the training dataset to train the Distributed Polynomial Kernel SVM model. The goal of the model is to identify the best hyperplane in the feature space that divides the benign and cancerous samples. The constrained quadratic programming problem is solved by the SVM optimisation technique to get the ideal support vectors and their related coefficients.

Model Evaluation: Evaluate the trained model using the testing dataset. Predict the class labels for the testing instances based on the learned decision boundary. Compare the predicted labels with the true labels to compute performance metrics such as accuracy, precision, recall, and F1 score. These metrics provide an assessment of how well the Distributed Polynomial Kernel SVM model is able to detect breast cancer on the WBCD dataset. **Feature Importance Analysis:** Distributed Polynomial Kernel SVM can also provide insights into feature importance. By examining the learned support vectors and their corresponding coefficients, it is possible to identify the most influential features for distinguishing between benign and malignant samples. This analysis can contribute to the understanding of the underlying characteristics of breast cancer and potentially aid in further research or feature selection.

Optimization and Fine-tuning: The performance of the Distributed Polynomial Kernel SVM model can be further improved through optimization and fine-tuning. This can involve adjusting the hyperparameters, exploring different kernels, or applying feature selection techniques to enhance the discriminative power of the model. By utilizing Distributed Polynomial Kernel SVM on the WBCD dataset, it is possible to build a classification model that can effectively detect breast cancer. The non-linear nature of the polynomial kernel allows for capturing complex relationships between the input features and the diagnosis labels, enabling accurate classification of benign and malignant samples.

3.3 Distributed KNN:

The DiKNN method is straightforward and simple to use, making it simple to comprehend. A suitable number for K must be chosen because an incorrect choice could result in either overfitting or underfitting. To guarantee that all features contribute equally to the distance calculations, feature scaling and normalisation are frequently advised. To analyse the success of the DiKNN method, remember to preprocess the data, choose an acceptable distance metric, determine the value of K, and use relevant metrics like accuracy, precision, recall, and F1 score.

Algorithm: DiKNN

Input:
a. Training dataset: {(x1, y1), (x2, y2), ..., (xn, yn)}, where xi represents the input features and yi represents the corresponding class labels.
b. Testing instance: x_test.
c. Number of neighbors: K.
Output:
Predicted class label for the testing instance.
Algorithm:
Step :1 Preprocess the data if necessary (e.g., normalization, feature scaling).
Step :2 Compute the distances between the testing instance (x_test) and all training instances using a distance metric such as Euclidean distance or Manhattan distance:
For each training instance (xi, yi):
Compute the distance between x_test and xi.
Step: 3 Select the K nearest neighbors:
• Sort the training instances based on their distances from x_test in ascending order.
• Select the top K training instances with the shortest distances.
Step: 4 Assign class labels to the K nearest neighbors:
• Retrieve the class labels (yi) of the K nearest neighbors.
Step: 5 Determine the majority class label:
• Count the occurrences of each class label among the K nearest neighbors.
• Assign the class label with the highest count as the predicted class label for x_test.
• In case of a tie, you can handle it by choosing the class label of the nearest neighbor with the smallest distance.

Step :6 Return the predicted class label for the testing instance.

Dataset preparation: The WBCD dataset includes a number of attributes that were taken from breast cancer cell pictures and paired with the appropriate diagnosis labels (benign or malignant). By addressing missing values, normalising the feature values, and dividing the dataset into training and testing sets, preprocessing is crucial. Initialization of the DiKNN Model: Set the number of neighbours (K) that will be taken into account for classification to begin the DiKNN model. It is essential to select a suitable value for K because it influences how well the model performs. To find the ideal value of K, you can employ strategies like cross-validation. Instruction of the Model: Utilise the training dataset to train the DiKNN model. The feature vectors and the class labels they relate to are stored by the model during training.

Utilising the testing dataset, evaluate the trained model. The DiKNN algorithm locates the K nearest neighbours for each instance in the testing set using a distance metric (such as Euclidean distance), and then designates the majority class label among those neighbours as the predicted label for the occurrence. Performance Evaluation: To calculate performance measures like accuracy, precision, recall, and F1 score, compare the predicted labels of the testing cases with their actual labels. These metrics offer an evaluation of the KNN model's breast cancer detection performance on the WBCD dataset.

Optimisation and fine-tuning: By optimising and fine-tuning, the DiKNN model's performance can be further enhanced. To improve the discriminative capacity of the model, this can entail changing the value of K, investigating various distance metrics, or using feature selection methods. The WBCD dataset can be used to apply KNN to create a classification model that can successfully identify breast cancer. Based on feature similarity, the algorithm locates the K closest neighbours, and it assigns the testing instance the majority class label among those neighbours. DiKNN is a well-liked option for a variety of classification tasks, including the detection of breast cancer, because to its clarity and interpretability.

It's crucial to remember that DiKNN has its limitations. It can be computationally expensive because it needs to calculate distances for each testing case, which is especially true for large datasets. In addition, KNN makes the unavoidable assumption that all features contribute equally to the distance calculations. To achieve the best performance with KNN on the WBCD dataset, thorough consideration of preprocessing processes, feature selection, and parameter adjustment is essential.

4 Experiment Results and Discussions

The final model that we choose for the current issue is a support vector machine, which has great accuracy, precision, recall, and AUC-ROC values. The statistics of the selected model outperform all other models we have developed. We require a high recall, which the model offers, to detect as many malignant tumours as is practical (0.94). While SVM is used when there are many characteristics, our data collection and few observations give it a good fit. To train and test all other models, a substantial amount of data is required, however.

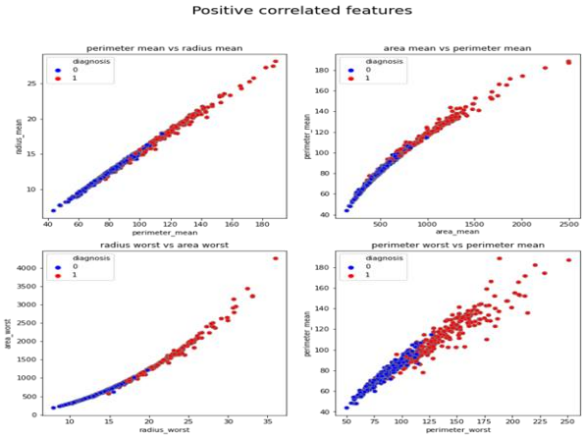

Fig: 2 Positive corelated

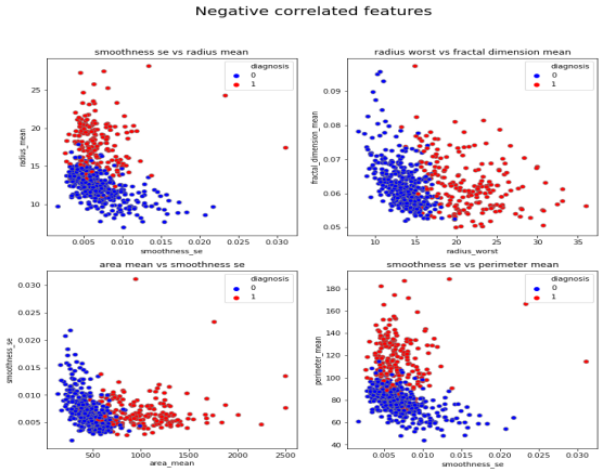

Fig:3 Negative corelated

This figure(3) presents a series of scatter plots that visually compare the classification performance of different distributed machine learning algorithms applied to the Wisconsin Breast Cancer Dataset (WBCD) for early detection of breast cancer. Each scatter plot likely corresponds to one of the four algorithms—DiNB, DieLR, DiKNN, and DiPKSVM—showing how well they separate the two classes: benign (blue dots) and malignant (red dots).

In each plot, two features from the dataset are plotted against each other, illustrating the distribution of data points (patients) based on their diagnoses. The overall trend in these plots is a generally linear separation

between the two classes, with some degree of overlap that varies between algorithms.Top-Left Plot: This plot likely corresponds to DiNB (Distributed Naive Bayes). The data points are relatively well-separated, but there is some overlap between the benign and malignant cases, indicating that the algorithm might have a few misclassifications.

Top-Right Plot: This plot probably represents the results from DieLR (Distributed Elastic Net Logistic Regression). There is a clearer distinction between the classes compared to the DiNB plot, with fewer overlaps, suggesting better performance in terms of classification accuracy.

Bottom-Left Plot: This plot is likely associated with DiKNN (Distributed K-Nearest Neighbors). There seems to be more overlap between the classes, especially in the middle of the plot, indicating that this algorithm may have more difficulty distinguishing between benign and malignant cases, leading to a slightly lower performance compared to DieLR.

Bottom-Right Plot: This plot most likely corresponds to DiPKSVM (Distributed Polynomial Kernel Support Vector Machine). The separation between the benign and malignant cases is more distinct, with minimal overlap, reflecting the superior performance of this algorithm as indicated by its higher precision, recall, and balanced accuracy metrics.

Overall, these plots visually demonstrate how each algorithm performs in classifying the data points into benign and malignant categories, with DiPKSVM showing the clearest separation and potentially the best performance among the four algorithms.

Table:1 individual Model accuracy

	DiNB	*DieLR*	*DiKNN*	*DiPKSVM*
Precision	0.94	1.00	0.97	1.0
Recall	0.88	0.96	0.89	0.98
F1	0.92	0.95	0.93	0.99
Detection Rate	0.33	0.35	0.33	0.37
Detection Prevalence	0.35	0.36	0.36	0.37
Balanced Accuracy	0.93	0.98	0.94	0.99

4.1 DiNB (Distributed Naive Bayes)

The Distributed Naive Bayes (DiNB) algorithm is a probabilistic classifier that applies Bayes' theorem with the assumption of independence among features. Despite this strong assumption, DiNB often performs well, particularly in high-dimensional spaces. In the context of early breast cancer detection, DiNB achieved a precision of 0.94, indicating that 94% of its positive predictions were correct. Its recall was slightly lower at 0.88, meaning it successfully identified 88% of actual breast cancer cases. The F1 score, which balances precision and recall, was 0.92, showing overall strong performance. The detection rate was 0.33, close to the detection prevalence of 0.35, indicating that DiNB classified a proportion of cases as positive that closely matched the actual distribution in the dataset. Lastly, its balanced accuracy was 0.93, demonstrating that it effectively distinguishes between positive and negative classes.

4.2 DieLR (Distributed Elastic Net Logistic Regression)

Distributed Elastic Net Logistic Regression (DieLR) combines the strengths of L1 (Lasso) and L2 (Ridge) regularization, making it effective for datasets with correlated features and high dimensionality. DieLR performed exceptionally well in detecting breast cancer, achieving perfect precision of 1.00, meaning every positive prediction was accurate with no false positives. It also had a high recall of 0.96, successfully identifying 96% of actual cases. Its F1 score was 0.95, reflecting a strong balance between precision and recall. The detection rate was 0.35, slightly higher than DiNB, and aligned well with the actual prevalence of positive cases, which was 0.36. DieLR's balanced accuracy was 0.98, making it a highly reliable algorithm for this task, effectively handling class imbalances.

4.3 DiKNN (Distributed K-Nearest Neighbors): The Distributed K-Nearest Neighbors (DiKNN) algorithm is a non-parametric method that classifies data points based on the majority class of their nearest neighbors in the feature space. DiKNN achieved a high precision of 0.97, meaning that 97% of its positive predictions were correct. Its recall was 0.89, successfully identifying 89% of actual breast cancer cases. The F1 score was 0.93, indicating a good balance between precision and recall. DiKNN's detection rate was 0.33, similar to DiNB, and close to the detection prevalence of 0.36, showing that it classified a comparable proportion of cases as positive. Its balanced accuracy was 0.94, demonstrating effective performance across both positive and negative classes, though slightly lower than DieLR and DiPKSVM.

4.4 DiPKSVM (Distributed Polynomial Kernel Support Vector Machine): The Distributed Polynomial Kernel Support Vector Machine (DiPKSVM) is a powerful classification algorithm that finds the optimal hyperplane to separate classes in the feature space, with the polynomial kernel allowing it to capture complex, non-linear relationships. DiPKSVM achieved perfect precision of 1.00, meaning all positive predictions were correct. It also had an impressive recall of 0.98, successfully identifying 98% of actual breast cancer cases. The F1 score was nearly perfect at 0.99, reflecting

an excellent balance between precision and recall. The detection rate was 0.37, the highest among the algorithms, and closely matched the detection prevalence of 0.37, indicating that DiPKSVM effectively aligned with the actual distribution of positive cases in the dataset. Its balanced accuracy was 0.99, the highest of all, demonstrating exceptional performance in distinguishing between positive and negative classes, making DiPKSVM the best-performing algorithm in this comparison.

Let's discuss the mathematical derivations of accuracy, precision, recall, and F1 score, which are commonly used performance metrics in the context of breast cancer detection on the WBCD dataset. The accuracy metric provides an overall measure of the model's correctness, while precision and recall provide insights into specific aspects of the classification performance. Precision emphasizes the ability to correctly identify malignant samples, while recall focuses on capturing all actual malignant samples. The F1 score combines precision and recall into a single metric that considers both false positives and false negatives. These metrics are commonly used in evaluating breast cancer detection models on the WBCD dataset. A higher accuracy, precision, recall, and F1 score indicate better performance in correctly identifying malignant samples, minimizing false positives and false negatives. It is important to consider these metrics collectively to obtain a comprehensive evaluation of the classification model's effectiveness in breast cancer detection.

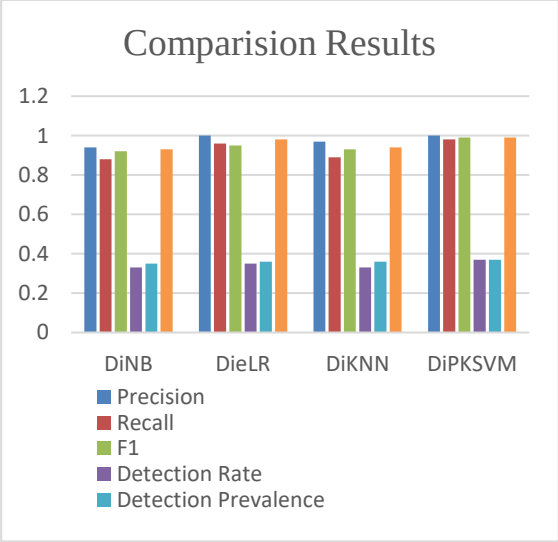

Fig:4 Comparisons with existing models

In fig (4), We'll look at the metrics in this section's comparison right away. Our process begins with precision. By dividing the total number of precise estimations by the total number of precise forecasts, this is expressed as a percentage. The precision is calculated by dividing the total number of true positives by the total number of false positives. In other words, this is the product of the predicted positive values for the class divided by the predicted positive values for the class. Additionally known as "positive predictive value" (PPV). Only an approximation can be made for the vast number of false positive outcomes. In other words, the ratio of optimistic predictions to optimistic class values in the test data is the same as the ratio of optimistic forecasts to optimistic class values in the test data. The sensitivity or true positive percentage are common names for it. Think of the reminder as a gauge of the classifier's precision. Low recovery rates suggest that there are many false negatives. Multiplying (accuracy x recovery) / (accuracy + recovery) by two yields the F1 score. Additionally known as Score F and Measure F. Recall (sensitivity) is calculated by dividing the total True Positives by the sum of True Positives and False Negatives. It stands for the ratio of right predictions to positive class values found in the test data, in other words. It also goes by the titles sensitivity or true positive rate. Recall can be used to gauge how in-depth a classifier is. Multiple False Negatives are indicated by a low recall. The formula 2 x ((precision x recall) / (precision + recall)) is used to calculate the F1 Score. The F Score or the F Measure are other names for it. In other words, the F1 score reflects how well precision and memory are balanced. The best results for sensitivity are provided by the Distributed Distributed Polynomial Kernel SVM, which also has a high F1 score (which allows for the diagnosis of breast cancer malign cases).

There are a number of things to take into account while contrasting Distributed Polynomial Kernel SVM and KNN for breast cancer diagnosis on the WBCD dataset. A decision boundary is built using support vectors and a polynomial kernel function via the supervised machine learning approach known as Distributed Polynomial Kernel SVM. The lazy learning algorithm distributed K-Nearest Neighbours (KNN) categorises each instance based on the class labels of its K nearest neighbours in the feature space. Specifically for large datasets, Distributed Polynomial Kernel SVM requires the solution of a quadratic optimisation problem, which can be computationally demanding. Compared to KNN, training time may be longer. In KNN, there is no explicit model training. Instead, it stores the entire training dataset, making the training time relatively faster compared to SVM.

Distributed Polynomial Kernel SVM requires tuning the hyperparameters such as the degree of the polynomial kernel and the regularization parameter (C) to optimize the model's performance. KNN: The main parameter to tune is the number of neighbors (K). An appropriate value for K needs to be selected to balance between overfitting and underfitting.

SVM-derived Distributed Polynomial Kernel SVM may be difficult to explain in terms of the underlying characteristics or their significance. By directly referring to the nearby examples, the classification decision is based on the majority vote of the closest neighbours, which might increase interpretability. SVM's Distributed Polynomial Kernel has the ability to manage skewed data by altering the decision threshold or employing methods like class weights. Unfair predictions may result from the KNN majority voting being influenced by the class imbalance. To solve this problem, strategies like weighted voting or resampling can be used.

Robustness to Noise and Outliers: Distributed Polynomial Kernel SVM: SVM is generally more robust to noise and outliers due to the use of support vectors, which focus on the most relevant samples. KNN can be sensitive to noise and outliers, as it relies on distance-based similarity measures. Preprocessing steps like outlier removal or feature scaling may be required. **Scalability:** Distributed Polynomial Kernel SVM : SVM can be computationally expensive for large datasets due to the need to store support vectors and calculate kernel functions for all instances. DiKNN is relatively scalable, as it only requires storing the training dataset and calculating distances for the test instances.

The final decision between Distributed Polynomial Kernel SVM and DiKNN for detecting breast cancer on the WBCD dataset depends on a number of variables, including the size of the dataset, processing resources, interpretability criteria, and performance objectives. It is advised to use cross-validation or train-test splits to evaluate the performance of the two algorithms using the right evaluation measures (such as accuracy, precision, recall, and F1 score), in order to identify which method is the most effective for the given task. In conclusion, DiKNN and Distributed Polynomial Kernel SVM are both widely used algorithms for detecting breast cancer on the WBCD dataset, although they each have unique features and considerations. With its ability to capture complex decision boundaries using support vectors and a polynomial kernel function, Distributed Polynomial Kernel SVM is a potent method. In general, it can effectively manage uneven data and is robust to noise and outliers. SVM can be computationally expensive, though, particularly for large datasets, and parameter adjustment is necessary for the best results. It could be difficult to interpret the SVM decision border.

On the other hand, KNN is a straightforward and understandable algorithm that bases classification on the similarity of nearby examples. Compared to SVM, it can be trained more quickly because explicit model training is not necessary. Through a direct study of the closest neighbours, DiKNN offers interpretability. However, the number of neighbours (K) must be carefully chosen because DiKNN can be susceptible to noise and outliers. Scalability issues and unbalanced data might also be difficult for DiKNN to handle. When selecting between Distributed Polynomial Kernel SVM and DiKNN for breast cancer detection on the WBCD dataset, several factors should be considered, including the dataset size, computational resources, interpretability requirements, and the desired performance metrics. Cross-validation or train-test splits should be used to assess and compare the performance of both algorithms using the appropriate measures, such as accuracy, precision, recall, and F1 score. The best option will ultimately depend on the requirements and limitations of the application. In order to decide which algorithm, Distributed Polynomial Kernel SVM or DiKNN, is best appropriate for breast cancer detection on the WBCD dataset in their particular setting, researchers and practitioners should carefully weigh the trade-offs and conduct rigorous tests.

5 Conclusion and Future Scope:

In this section, both Distributed Polynomial Kernel SVM and DiKNN are popular algorithms for breast cancer detection on the WBCD dataset, but they have distinct characteristics and considerations. Distributed Polynomial Kernel SVM offers a powerful approach with the ability to capture complex decision boundaries using support vectors and a polynomial kernel function. It is generally robust to noise and outliers and can handle imbalanced data effectively. However, DipSVM can be computationally expensive, especially for large datasets, and parameter tuning is required for optimal performance. The interpretability of the DiPSVM decision boundary may also be challenging. On the other hand, DiKNN is a simple and intuitive algorithm that relies on the similarity of neighboring instances for classification. It is relatively faster to train compared to SVM as it does not involve explicit model training. DiKNN provides interpretability through direct examination of the nearest neighbors. However, DiKNN can be sensitive to noise and outliers, and the choice of the number of neighbors (K) is critical. Imbalanced data and scalability can also pose challenges for DiKNN. Early breast cancer identification can reduce mortality rates. Recent research indicates that the detection of breast cancer requires the use of machine learning techniques. In the current scientific era, three widely utilized machine learning methods are employed to detect breast cancer. Using information from the Wisconsin Breast Cancer Diagnostic dataset, the proposed approaches are contrasted. We will introduce WEKA and IBM SPSS as potential future improvements for ML methods used to categorize the WBCD (Original dataset). It is feasible to compare the efficacy of relevant algorithms in terms of crucial presentation metrics and determine the efficacy of prior efforts. In this work, the

diagnosis of Wisconsin Madison breast cancer is viewed as a problem of pattern categorization. The best machine learning model was selected in this area by combining high accuracy and a low false-negative rate (the means that the metric is high sensitivity). The Distributed Polynomial Kernel SVM model generated the best results for F1, recall, and balanced accuracy (0.99, 0.98, and 0.98, respectively) (0.99).

References:

[1] Van der Aalst W. Process Mining: Data Science in Action. Springer; Berlin/Heidelberg, Germany: 2016.

[2] Romero C., Ventura S. Educational data science in massive open online courses. Wires Data Min. Knowl. Discov. 2017;7:1–12. doi: 10.1002/widm.1187.

[3] Raghupathi W., Raghupathi V. Big data analytics in healthcare: Promise and potential. Health Inf. Sci. Syst. 2014;2:3. doi: 10.1186/2047-2501-2-3.

[4] Sohail M.N., Jiadong R., Uba M.M., Irshad M. A comprehensive looks at data mining techniques contributing to medical data growth: A survey of researcher reviews; Proceedings of the 35th IEEE International Conference on Computer Design, ICCD 2017; Boston, MA, USA. 5–8 November 2017; pp. 21–26.

[5] 5. Petri I., Kubicki S., Rezgui Y., Guerriero A., Li H. Optimizing energy efficiency in operating built environment assets through building information modeling: A case study. Energies. 2017;10:1167. doi: 10.3390/en10081167.

[6] Bray F., Ferlay J., Soerjomataram I. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. 2018;68:394–424. doi: 10.3322/caac.21492.

[7] Cavaioni M. Machine Learning: Supervised Learning Classification. [(accessed on 5 February 2020)]; Available online: https://medium.com/machine-learning-bites/machine-learning-supervised-learning-classification-4f44a91d767.

[8] Maria Y. Machine learning based approaches for modeling the output power of photovoltaic array in real outdoor conditions. Electronics. 2020;9:315. doi: 10.3390/electronics9020315.

[9] Oyewola D., Hakimi D., Adeboye K., Shehu M. Using five machine learning for breast cancer biopsy predictions based on mammographic diagnosis. Int. J. Eng. Technol. IJET. 2017;2:142–145. doi: 10.19072/ijet.280563.

[10] The World Health Organization. Cancer in Malaysia. Available online: https://gco.iarc.fr/today/data/factsheets/populations/458- malaysia-fact-sheets.pdf (accessed on 10 February 2021).

[11] Siegel, R.L,Miller, K.D,Jemal, A. Cancer statistics. CA Cancer J. Clin. 2020, 70, 7–30.

[12] Hjelm, T.E.,Matovu, A,Mugisha, N.; Löfgren, J. Breast cancer care in Uganda: A multicenter study on the frequency of breast cancer surgery in relation to t.he incidence of breast cancer. PLoS ONE 2019, 14, e0219601.

[13] Bera, A.; Subramanian, M.; Karaian J.; Eklund, M. Functional role of vitronectin in breast cancer. PLoS ONE 2020, 15, e0242141.

[14] Amir, P.N.; Ali, N.; Raman, R.K.; Raman, S.; Bahtiar, B. Malaysian Study on Cancer survival, 1st ed.; National Cancer Institute Ministry of Health: Putrajaya, Malaysia, 2018; pp. 1–57.

[15] Wang, L. Early diagnosis of breast cancer. Sensors 2017, 17, 1572.

[16] Al-hadidi, M.D.R.; Alarabeyyat, A.; Alhanahnah, M. Breast Cancer Detection using K-nearest Neighbor Machine Learning Algorithm. In Proceedings of the 9th International Conference on Development in eSystems Engineering (DeSE), Liverpool, UK, 31 August–2 September 2016; pp. 35–39.

[17] 1 in 30 Malaysian Women Will Have Breast Cancer, So Get Checked Now. Available online: https://www.thestar.com.my/lifestyle/family/2019/10/16/breast-cancer-2/ (accessed on 10 February 2021).

[18] Yip, C.H.; Pathy, N.B.; Teo, S.H. A review of breast cancer research in Malaysia. Med. J. Malays. 2014, 69, 8–22.

[19] Ann, H.J.; Su-Shi, L.S.; Zakaria, Z. Non-invasive breast cancer assessment using magnetic induction spectroscopy technique. Int. J. Integr. Eng. 2017, 9, 54–60.

[20] Kwon, S.; Lee, S. Recent advances in microwave imaging for breast cancer detection. Int. J. Biomed. Imaging 2016, 2016, 5054912.

[21] Carovac, A.; Smajlovic, F. Junuzovic, D. Application of Ultrasound in Medicine. Acta Inform. Med. 2011, 19, 168–171.

[22] Al-dhabyani, W.; Gomaa, M.; Khaled, H.; Fahmy, A. Dataset of breast ultrasound images. Data Brief

2019, 28, 104863.

[23] Sorrenti, S.; Dolcetti, V.; Fresilli, D. The Role of CEUS in the Evaluation of Thyroid Cancer : From Diagnosis to Local Staging.MDPI Clin. Med. 2021, 10, 4559.

[24] Emine, D.; Suzana, M.K.; HalitYmeri, A.K. Comparative accuracy of mammography and ultrasound in women with breast symptoms according to age and breast density. Bosn. J. Basic Med. Sci. 2009, 57, 205–216.

[25] Joy, J.E.; Penhoet, E.E.; Petitti, D.B. Saving Women's Lives: Strategies for Improving Breast Cancer Detection and Diagnosis, 1st ed.; TheNational Academies: Washington, DC, USA, 2005; pp. 1–361

[26] Dlama Zira, J.; Chukwuemeka Nzotta, C. Radiation doses for mammography and its relationship with anthropo-technical parameters. Int. J. Radiol. Radiat. Therapy 2018, 5, 14–18.

[27] Gresik, C.How Dangerous Is Radiation from a Mammogram? Available online: https://www.eehealth.org/blog/2018/04/radiation-from-a-mammogram/#:~:text=On%20average%2C%20the%20total%20radiation,-just%20from%20their%20natural%20surroundings (accessed on 10 April 2021.)

[28] Kamal, R.; Mansour, S.; Farouk, A. Contrast-enhanced mammography in comparison with dynamic contrast-enhanced MRI: Which modality is appropriate for whom? Egypt. J. Radiol. Nucl. Med. 2021, 52, 216.

[29] Arleo, E.K.; Hendrick, R.E.; Helvie, M.A.; Sickles, E.A. Comparison of recommendations for screening mammography using CISNET models. Cancer 2017, 123, 3673–3680.

[30] Hogg, P.; Kelly, J.; Mercer, C. Digital Mammography, 1st ed.; Springer: London, UK, 2015; pp. 1–307.

[31] Iranmakani, S.; Mortezazadeh, T.; Sajadian, F.; Ghaziani, M.F.; Ghafari, A.; Khezerloo, D. A review of various modalities in breast imaging: Technical aspects and clinical outcomes. Egypt. J. Radiol. Nucl. Med. 2020, 51, 57.

[32] Mahmud, M.Z.; Islam, M.T.; Misran, N.; Almutairi, A.F.; Cho, M. Ultra-wideband (UWB) antenna sensor-based microwave breast imaging: A review. Sensors 2018, 18, 2951.

[33] Zhang, L. Ren, Z. Comparison of CT and MRI images for the prediction of soft-tissue sarcoma grading and lung metastasis via a convolutional neural networks model. Clin. Radiol. 2020, 75, 64–69.

[34] Arteaga-Marrero, N.; Villa, E.; González-Fernández, J.; Martín, Y.; Ruiz-Alzola, J. Polyvinyl alcohol cryogel phantoms of biological tissues for wideband operation at microwave frequencies. PLoS ONE 2019, 14, e0219997.

[35] Khan, S.U.; Ullah, N.; Ahmed, I.; Ahmad, I.; Mahsud, M.I. MRI imaging, comparison of MRI with other modalities, noise in MRI images and machine learning techniques for noise removal: A review. Curr. Med. Imaging 2018, 15, 243–254.

[36] Mann, R.M.; Kuhl, C.K.; Moy, L. Contrast-enhanced MRI for breast cancer screening. J. Mag. Resonance Imaging 2019, 50, 377–390.

[37] Gunduru, M.; Grigorian, C. Breast magnetic resonance imaging MRI. Radiol. Technol. 2020, 91, 1–6.

[38] Mann, R.M.; Cho, N.; Moy, L. Breast MRI: State of the Art. Radiology 2019, 292, 520–536.

[39] Broadhouse, K.M. The Physics of MRI and How We Use It to Reveal the Mysteries of the Mind. Available online: https: //kids.frontiersin.org/articles/10.3389/frym.2019.00023 (accessed on 5 March 2021).

[40] Dustler, Evaluating AI in breast cancer screening: A complex task. Lancet Digital Health 2020, 2, e106–e107.

[41] ung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249.

[42] Duggan, C.; Dvaladze, A.; Rositch, A.F.; Ginsburg, O.; Yip, C.; Horton, S.; Rodriguez, R.C.; Eniu, A.; Mutebi, M.; Bourque, J.; et al. The Breast Health Global Initiative 2018 Global Summit on Improving Breast Healthcare Through Resource-Stratified Phased Implementation: Methods and overview. *Cancer* **2020**, *126*, 2339–2352.

[43] Sharma, R. Global, regional, national burden of breast cancer in 185 countries: Evidence from GLOBOCAN 2018. *Breast Cancer Res. Treat.* **2021**, *187*, 557–567.

[44] Zhou, L.; Chen, B.; Sheng, L.; Turner, A. The effect of vitamin D supplementation on the risk of breast cancer: A trial sequential meta-analysis. *Breast Cancer Res. Treat.* **2020**, *182*, 1–8.

[45] Kubota, S.I.; Takahashi, K.; Mano, T.; Matsumoto, K.; Katsumata, T.; Shi, S.; Tainaka, K.; Ueda, H.R.; Ehata, S.; Miyazono, K. Whole-organ analysis of TGF-β-mediated remodelling of the tumour microenvironment by tissue clearing. *Commun. Biol.* **2021**, *4*, 294.

[46] Tarantino, P.; Morganti, S.; Curigliano, G. Biologic therapy for advanced breast cancer: Recent advances and future directions. *Expert Opin. Biol. Ther.* **2020**, *20*, 1009–1024.