Novel Prediction on Breast Cancer through Lazy Learning Approach by Linear Neural Network Search with Distance with Euclidean

S. Amsavalli[1,*], Vetripriya M.[2], R. Sivasankari[2], Vetri Selvan M.[3], Vijayakumar K.[3]

[1]Dept. of Computer Application, B.S.Abdur Rahman Crescent Institute of Science and Technology, India

[2]Dept. of Computer Science and Engineering, B.S.Abdur Rahman Crescent Institute of Science and Technology, India

[3]Dept. of Artificial Intelligence and Data Science, Panimalar Engineering College, Tamil Nadu, India

Email: amsavalli@crescent.education; vetripriya@crescent.education; sivasankari.rp@crescent.education; vetrinelson7@gmail.com; vijayakumarkadumbadi23@gmail.com

Abstract

Breast cancer is the most prevalent cancer-affecting women worldwide and remains a major cause of mortality. Early detection and accurate prognosis are critical to improving survival outcomes. This study introduces a novel predictive model for breast cancer diagnosis that integrates a lazy learning paradigm with the K-Nearest Neighbors (KNN) algorithm, optimized through a Linear Nearest Neighbor (NN) Search technique and the use of Euclidean distance as the similarity measure. The dataset, comprising 4,024 patient records with 15 clinical and demographic attributes, was obtained from a public repository and underwent rigorous preprocessing, including handling of missing values, normalization, and categorical encoding. The classification model was trained and evaluated using 1:9 cross-validation, with K values ranging from 1 to 9 and a constant batch size of 100 to identify the optimal configuration. Among various configurations tested, the model with K=5 demonstrated the highest performance, achieving an accuracy of 88.02%, precision of 0.87, and recall of 0.88. Additional performance metrics such as F-measure, Matthews Correlation Coefficient (MCC), and Kappa statistic further confirmed the robustness of the selected configuration. The proposed model shows superior predictive capability compared to traditional settings and can serve as a decision-support tool for clinicians. The findings suggest that the combination of lazy learning, effective neighbor search strategy, and robust distance metric can substantially enhance the predictive accuracy of breast cancer diagnosis. This study highlights the potential of machine learning-based tools in clinical oncology, offering a data-driven approach for early intervention and patient outcome improvement.

Keywords: Breast Cancer; Distance with Euclidean; Batch size; KNN; Linear NN Search

1. Introduction

Breast cancer is the leading cancer type diagnosed in women globally and is a major contributor to cancer-related deaths. Its complexity and biological diversity make early detection and prognosis particularly challenging, especially in resource-constrained settings. [1] Traditional diagnostic techniques such as mammography, MRI, and biopsy continue to serve as gold standards. However, their high costs and limited availability have created a need for supplementary computational approaches to assist clinicians in making faster and more reliable diagnoses.

Over the past decade, artificial intelligence and machine learning techniques have become increasingly popular in medical research, particularly in the domain of predictive analytics. [2] Among these, classification models such as Decision Trees, Support Vector Machines, and Neural Networks have shown promise. However, lazy learning models, which defer the generalization process until prediction time, offer a unique advantage by adapting well to diverse data distributions without intensive training. The K-Nearest Neighbors (KNN) algorithm is a quintessential lazy learner that has gained traction due to its simplicity and interpretability. Nevertheless, its performance is often contingent upon careful selection of parameters such as the number of neighbors (K), distance metrics, and the

DOI: https://doi.org/10.54216/FPA.210114
Received: January 10, 2025 Revised: May 20, 2025 Accepted: July 04, 2025

method for identifying nearest neighbors. This study aims to optimize these parameters using a Linear Nearest Neighborhoods (NN) Search strategy coupled with the Euclidean distance metric [3-6].

In clinical applications, [7-12] predictive models must be both accurate and computationally efficient. Therefore, the proposed methodology emphasizes not only predictive accuracy but also practical applicability in real-time healthcare scenarios. By employing a fixed batch size and varying K values, the model is fine-tuned to balance sensitivity and specificity. The dataset used for this research is comprehensive and includes key demographic and clinical variables such as age, tumor size, hormone receptor status, and lymph node involvement. These features are known to influence prognosis and survival, making them highly relevant for model training and validation.

Overall, this study seeks to contribute to the growing body of work at the intersection of machine learning and medical diagnostics by presenting a robust, efficient, and interpretable model for breast cancer prognosis. The findings are expected to support clinical decision-making and open avenues for further research in data-driven personalized medicine. The primary objective of this research is to develop an optimized machine-learning model for the prediction of breast cancer outcomes using a lazy learning approach. Specific objectives include:

- Implementing the K-Nearest Neighbors (KNN) algorithm with parameter tuning to identify the best-performing configuration.
- Utilizing Linear Nearest Neighbor (NN) Search to improve computational efficiency in neighbor selection.
- Employing Euclidean distance as a similarity metric for accurate classification.
- Evaluating the model performance across multiple metrics such as accuracy precision, recall, F-measure, and MCC.

2. Literature Review

The models to predict survival outcomes for patients with lymph node-positive, Luminal A breast cancer. Using clinical and pathological data, including tumor grade, hormone receptor status, and lymph node involvement, their model offers a reliable prognostic tool for oncologists. Their approach focused on identifying personalized survival probabilities using visual decision aids. This is highly relevant to your study because it similarly uses demographic and clinical predictors to classify patient outcomes. Incorporating such nomogram strategies into a lazy learning framework, like KNN, can further improve interpretability in clinical settings. The study supports integrating structured medical attributes—such as tumor staging and hormone status—, which are already present in your dataset. Their work underscores the importance of data-driven approaches in enhancing early detection and personalized treatment planning for breast cancer patients [13].

The study confirmed that a higher BMI significantly increases the risk of lymph node metastasis in a dose-dependent manner. This finding is particularly relevant for predictive modeling as it emphasizes the inclusion of anthropometric features such as BMI in machine learning frameworks for breast cancer prognosis. The paper demonstrates how clinically relevant predictors—beyond genetic and molecular markers—can affect disease progression. Integrating such physical and demographic variables into KNN-based lazy learning models can strengthen their predictive power. The methodology also aligns with your data-driven classification approach, where feature selection based on clinical importance enhances model performance [14-17].

The correlation between tumor-infiltrating lymphocytes (TILs) and lymph node metastasis in T1 stage breast cancer patients. Their study revealed that lower TIL density was associated with a higher incidence of lymph node involvement, suggesting TILs as a potential biomarker for predicting disease spread. This research highlights the role of immune biomarkers in the classification of breast cancer progression. The findings are significant to machine learning models as they validate the utility of non-traditional clinical features (like immune response) in predictive modeling. While your study uses a lazy learning approach based on KNN, integrating immunological parameters could further refine the classification process. Their work supports the expansion of data types in predictive models to enhance diagnostic precision, particularly when early-stage disease assessment is critical [18-22].

Evaluated the applicability of the Gail model—a statistical breast cancer risk prediction tool—among Egyptian women. They found significant predictive performance, especially when applying localized population data for calibration. The study underscores the importance of contextual and population-specific model adaptation for accurate risk prediction. In the context of your lazy learning KNN model, this suggests the need to consider geographic and ethnic variables when constructing and validating predictive models. By using features such as age, family history, and reproductive history, the Gail model shares similarities with KNN models that rely on structured clinical datasets. The research affirms the need for well-curated datasets and context-aware algorithms to achieve robust, generalizable predictions in breast cancer prognosis [23].

DOI: https://doi.org/10.54216/FPA.210114
Received: January 10, 2025 Revised: May 20, 2025 Accepted: July 04, 2025

A proposed a method combining Mask R-CNN for tumor segmentation with ensemble machine learning classifiers for breast cancer detection. The integration of deep learning-based image segmentation with traditional ML classifiers enhanced the overall prediction accuracy. Although your study focuses on tabular clinical data, this work illustrates how segmentation-based preprocessing and feature extraction from medical imaging can enrich classifier performance. Furthermore, their ensemble approach can inspire hybrid modeling in lazy learning contexts. Their findings support the notion that multiple algorithmic perspectives—when integrated—offer more robust diagnostic outputs. For future improvements, coupling your KNN-based method with image-derived features could broaden the model's applicability in real-time clinical diagnosis systems [24].

An integrated genomic profiling method using next-generation sequencing (NGS) data for cancer patients. Their approach leveraged bioinformatics pipelines to extract and analyze genomic signatures for precision oncology. Although your current model does not include genomic data, this work suggests how combining molecular-level information with traditional clinical variables can improve prediction accuracy. Lazy learning models like KNN can benefit from multi-modal datasets, and integrating high-dimensional genomic features is feasible with dimensionality reduction or feature selection methods. Kosvyra's work emphasizes the movement toward more comprehensive, personalized cancer diagnosis and treatment models, laying a foundation for integrating omics data into standard ML classifiers [25-27].

This study focused on building a prediction model for sentinel lymph node metastasis in ER- positive and HER2-negative breast cancer using microRNA expression profiles. Their model demonstrated improved sensitivity and specificity compared to traditional clinical indicators. This is a prime example of how molecular biomarkers can enhance predictive models for breast cancer progression. While your study uses demographic and clinical attributes, integrating such molecular-level data into a KNN framework could substantially elevate model accuracy. Their use of binary classification aligns with your methodology, and the incorporation of advanced feature types illustrates the growing potential of hybrid diagnostic systems [28].

This survey explored brain tumor detection using both machine learning (ML) and deep learning (DL) methods, focusing on imaging data. While not directly about breast cancer, the techniques discussed—such as convolutional neural networks (CNNs), ensemble models, and hybrid-learning approaches—have broad applicability. This work supports the feasibility of extending your current KNN-based system to include imaging data or to combine it with deep learning models for improved performance. Their analysis of various DL architectures provides insights into how similar models might be adapted for mammography or MRI data in breast cancer prediction. This cross-domain applicability is particularly important as oncology increasingly leverages imaging alongside clinical data [31]. Zuo et al. demonstrated that serum HER2 levels are reliable predictors of treatment efficacy and prognosis in HER2-positive breast cancer patients undergoing neoadjuvant therapy. Their results indicate that real-time biomarker monitoring can enhance clinical decision-making and model accuracy. This insight can inform your lazy learning model by emphasizing the potential of time-series or dynamic biomarkers as additional features. The inclusion of biomarker trends, rather than static values, may allow KNN or other models to capture disease progression more effectively. Their findings also support a future direction in integrating therapeutic response data into predictive models to improve prognostic validity [29].

A developed a 3D DenseAlexNet model for brain tumor segmentation and classification. Their approach combined deep learning's feature extraction strength with high-resolution imaging data to yield accurate segmentation results. While the domain differs, the study exemplifies how deep neural architectures can be tailored for cancer detection. This is relevant for enhancing your KNN-based system by exploring hybrid frameworks where KNN serves as a decision-layer classifier following deep feature extraction. Their methodology encourages the fusion of computational efficiency with high-dimensional data processing, which could be especially valuable if your model is extended to include MRI or ultrasound breast images [30].

3. Materials and Methods

The breast cancer dataset downloaded from Kaggle repository [18,19] It contains 4024 instances and 15 attributes with binary class. The following table shows description of the dataset.

3.1. Proposed Method

The proposed method for breast cancer prediction utilizes a lazy learning approach based on the K-Nearest Neighbors (KNN) algorithm, enhanced by a Linear Nearest Neighbourhood (NN) Search strategy and Euclidean distance as the similarity measure. After preprocessing the dataset, the classification process was conducted using different values of *K* while fixing the batch size at 100 to determine the optimal configuration. The Linear NN Search method improves the efficiency of neighbor identification, and the Euclidean distance ensures accurate measurement of similarity between data points in the feature space. The algorithm evaluates each test instance by calculating its distance from training samples and identifying the *K* closest neighbors to determine the majority class. The model's performance was systematically evaluated across different *K* values to identify the configuration

DOI: https://doi.org/10.54216/FPA.210114
Received: January 10, 2025 Revised: May 20, 2025 Accepted: July 04, 2025

that yields the best accuracy, Precision, recall, and other performance metrics. The optimal results were achieved at $K = 5$, showing superior classification capability compared to other configurations. The entire implementation was carried out using the Weka 3.9.5 data-mining tool.

3.2. Data Collection

The dataset used in this study was obtained from the Kaggle online repository and contains 4,024 instances with 15 input attributes and a binary target variable representing patient survival status (Alive = 0, Dead = 1). The attributes encompass a range of demographic and clinical features relevant to breast cancer prognosis, including Age, Race, Marital Status, T Stage, N Stage, 6th Stage, Tumor Differentiation, Grade, A Stage, Tumor Size, Estrogen Status, Progesterone Status, Regional Nodes Examined, Regional Nodes Positive, and Survival Months. These variables offer comprehensive insight into each patient's condition and treatment history. The data underwent preprocessing to handle any inconsistencies or formatting issues, ensuring it was ready for machine learning analysis. For model training and validation, the dataset was split using a 1:9 cross-validation approach. All experimental procedures and performance evaluations were conducted using the Weka 3.9.5 data-mining tool.

Table 1: Meta data of dataset

S. No	Attribute Name	Description / Value Range
1	Age	Range: 30 to 69 years
2	Race	Categories: White = 0, Black = 1, Other = 2
3	Marital Status	Categories: Single = 0, Married = 1, Divorced = 2, Separated = 3
4	Tumor (T) Stage	Encoded: T1 = 0, T2 = 1, T3 = 2, T4 = 3
5	Node (N) Stage	Encoded: N1 = 0, N2 = 1, N3 = 2
6	6th AJCC Stage	Categories: IIA = 0, IIB = 1, IIIA = 2, IIIB = 3, IIIC = 4
7	Differentiation	Encoded: Poor = 0, Moderate = 1, Well = 2, Undifferentiated = 3
8	Tumor Grade	Values: Grade IV (Anaplastic) = 0, Grade I–III = 1–3
9	A Stage	Regional = 0, Distant = 1
10	Tumor Size (mm)	Numeric range from 1 to 140
11	Estrogen Receptor Status	Positive = 0, Negative = 1
12	Progesterone Receptor Status	Positive = 0, Negative = 1
13	Regional Nodes Examined	Numeric range: 1 to 61 nodes
14	Regional Nodes Positive	Numeric range: 1 to 46 nodes
15	Survival Duration (months)	Numeric range: 1 to 107 months

DOI: https://doi.org/10.54216/FPA.210114
Received: January 10, 2025 Revised: May 20, 2025 Accepted: July 04, 2025

16	Patient Status	Alive = 0, Deceased = 1

S.N	Name of the	Actual Value of Attribute
1	Age	From 30 to 69
2	Race	White-0,Black-1,Other-2
3	Marital Status	Single-0,Married-1,Divorced-
4	T Stage	T1-0,T2-1,T3-2,T4-3
5	N Stage	N1-0,N2-1,N3-2
6	6th Stage	IIA-0,IIB-1,IIIA-2,IIIB-3,IIIC-4
7	Differentiate	Poorly -0, Moderate-1, Well-2,
8	Grade	anaplastic; Grade IV-0, 1 to 3
9	A Stage	Regional-0, Distant-1
10	Tumor Size	1 to 140
11	Estrogen Status	Positive-0, Negative-1
12	Progesterone	Positive-0, Negative-1
13	Regional Node	From 1 to 61
14	Regional Node	From 1 to 46
15	Survival Months	From 1 to 107
16	Status	Alive-0, Dead-1

3.3. Data Pre-processing

Prior to model training, the collected dataset underwent a series of pre-processing Steps to ensure data quality and suitability for machine learning analysis. Initially, the dataset was examined for missing or inconsistent values, which were either removed or appropriately imputed to maintain the integrity of the data. Categorical attributes such as Race, Marital Status, T Stage, N Stage, and hormone receptor statuses were encoded into numerical values based on predefined mappings to facilitate computational processing. Numerical features like Age, Tumor Size, and Survival Months were normalized or scaled where necessary to ensure uniformity and to avoid bias in distance calculations during KNN classification. The target class, representing patient status (Alive or Dead), was also encoded as a binary variable. Following pre-processing, the dataset was split using a 1:9 cross-validation strategy to enable robust training and evaluation of the model across multiple subsets. These pre-processing steps were crucial in enhancing model performance and ensuring reliable and unbiased classification results.

3.4. KNN Algorithm with Euclidean Distance

In this study, the K-Nearest Neighbors (KNN) algorithm is employed as the primary classification technique under the lazy learning paradigm. The model is configured with a fixed batch size of 100, and various values of K (1, 3, 5, 7, and 9) are tested to determine the optimal number of neighbors for prediction. To enhance search efficiency, the Linear Nearest Neighborhood Search algorithm is used to identify the most relevant neighbors within the feature space. The similarity between data points is measured using the Euclidean distance function, which quantifies the straight-line distance between two points in a multidimensional space. The Euclidean distance ddd between two coordinate points $(m1, n1)$ and $(m2, n2)$ is computed using the formula:

$$distance = \sqrt{(m2 - m1)^2} + (n2 - n1)^2 \text{ ---------- (1)}$$

Figure 1 illustrates the overall architecture of the proposed breast cancer prediction system using a lazy learning approach. The process begins with data collection, where a publicly available breast cancer dataset is retrieved from the Kaggle repository. The next phase is data preprocessing, which involves cleaning, encoding categorical variables, and preparing the data for analysis. Following preprocessing, the K-Nearest Neighbors (KNN) classification algorithm is applied. The algorithm is configured with a batch size of 100, and it uses the Linear Nearest Neighbourhood (NN) Search method for efficient neighbor retrieval and the Euclidean distance metric to measure similarity between instances. The system evaluates multiple values of K (1, 3, 5, 7, 9) to identify the optimal setting for classification. After model training and testing, the system proceeds to the evaluation phase, where various performance metrics such as accuracy, precision, recall, and F-measure are computed to determine

DOI: https://doi.org/10.54216/FPA.210114
Received: January 10, 2025 Revised: May 20, 2025 Accepted: July 04, 2025

the best-performing configuration. This flow ensures a systematic and accurate prediction of breast cancer status using a combination of optimized parameters and robust evaluation.

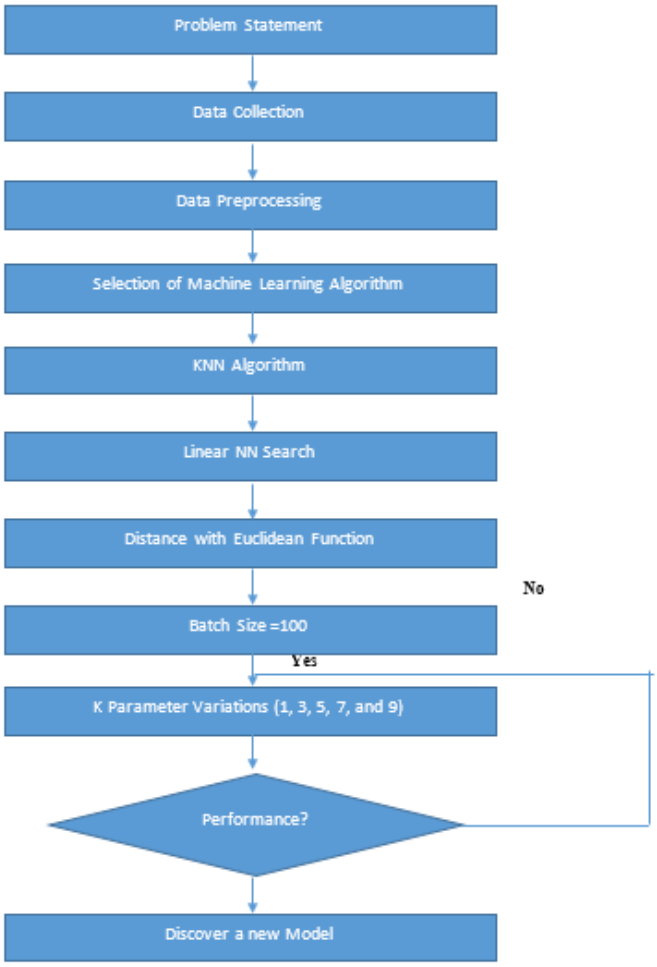


Figure 1. Proposed System

Pseudocode for Proposed Method

Begin
  Load dataset from Kaggle
  Preprocess dataset:
    - Handle missing values
    - Encode categorical variables
    - Normalize data (if needed)
  Split data using 1:9 cross-validation
  For each value of K in [1, 3, 5, 7, 9]:
    Initialize KNN with:
      - Batch size = 100
      - Linear NN Search
      - Euclidean distance function
    Train KNN model
    Test model on validation set
    Record performance metrics:
      - Accuracy, Precision, Recall, F-measure

DOI: https://doi.org/10.54216/FPA.210114
Received: January 10, 2025 Revised: May 20, 2025 Accepted: July 04, 2025

- MCC, Kappa, ROC, PRC, MAE, RMSE

Select K with highest performance (best accuracy & F-measure)

End

The research model presented in this study is a lazy learning-based classification framework utilizing the K-Nearest Neighbors (KNN) algorithm, enhanced through Linear Nearest Neighbourhood (NN) Search and Euclidean distance as the core similarity metric. The model pipeline is systematically structured to achieve robust breast cancer prediction using structured clinical data. Initially, the dataset is acquired from the Kaggle repository, comprising 4,024 patient instances characterized by 15 clinical and demographic attributes. The preprocessing stage ensures data quality and consistency by addressing missing values, encoding categorical variables into numeric form, and normalizing continuous features to standardize value scales and avoid bias during distance computation. The data is partitioned using 1:9 cross-validation, where one subset is used for testing while nine are reserved for training, enhancing generalizability and reducing overfitting. The model then iteratively evaluates multiple configurations of the K value (K = 1, 3, 5, 7, 9), identifying the optimal number of neighbors.

For each configuration:

- KNN is initialized with a fixed batch size of 100.
- A Linear NN Search algorithm is employed to enhance neighbor retrieval efficiency.
- Euclidean distance is applied to measure similarity in the multidimensional feature space.

The configuration that yields the highest accuracy and F-measure is selected as the final model. The use of a lazy learning strategy ensures that computation is deferred until prediction time, allowing for dynamic adaptation to new data. This model structure offers a transparent, interpretable, and statistically rigorous approach to breast cancer prediction and can be extended to other medical classification tasks.

4. Results and Discussions

Table 2 compares KNN model performance with 100 batch size, Linear NN Search method, and Euclidean distance. K=1 with 100 batch size, Linear NN Search technique, and Distance using Euclidean function model show 84.02% accuracy, 0.83 precision, 0.84 recall, and 0.01 seconds construction time. K=3 with 100 batch size, Linear NN Search technique, and Distance with Euclidean function model demonstrate 87.48% accuracy, 0.86 precision, 0.88 recall, and zero seconds to create. K=5 with 100 batch size, Linear NN Search technique, and Distance using Euclidean function model exhibit 88.02% accuracy, 0.87 precision, 0.88 recall, and zero seconds to create. shows K=7 with 100 batch size, Linear NN Search technique, and Distance with Euclidean function model demonstrate 87.99% accuracy, 0.87 precision, 0.88 recall, and 0.01 seconds to create. K=9 with 100 batch size, Linear NN Search technique, and Distance with Euclidean function model show 87.72% accuracy, 0.87 precision, 0.88 recall, and zero seconds to create.

Table 2: Performance measurements of Lazy Classifier

S.No	Base Category	Classifier	Batch size	Nearest Neighborhood Search Algorithm	Distance Function	K value	Accuracy	Precision	Recall	Time (Sec)
1	Lazy	K NN	100	Linear NN Search	Euclidean	K=1	84.02%	0.83	0.84	0.01
2						K=3	87.48%	0.86	0.88	0.00
3						K=5	88.02%	0.87	0.88	0.00
4						K=7	87.99%	0.87	0.88	0.01
5						K=9	87.72%	0.87	0.88	0.00

DOI: https://doi.org/10.54216/FPA.210114
Received: January 10, 2025 Revised: May 20, 2025 Accepted: July 04, 2025

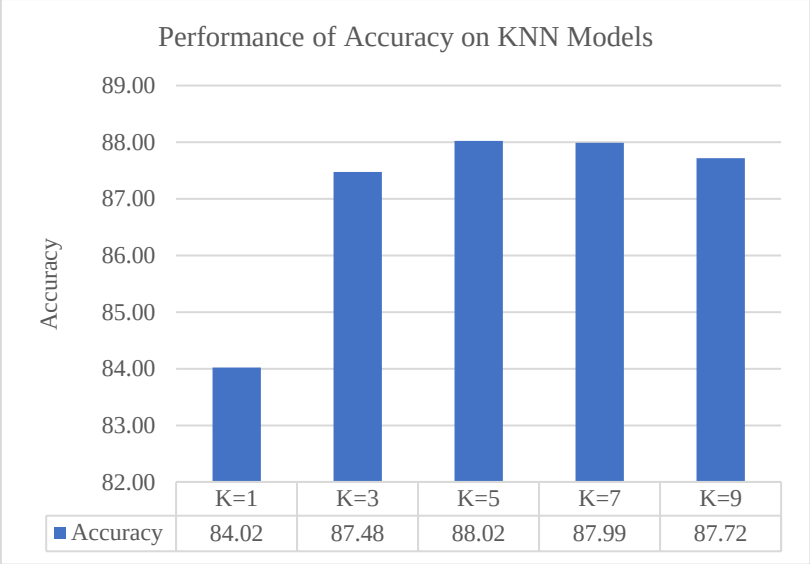


Figure 2. Accuracy performance of KNN Models

Diagram 2 displays KNN Model accuracy measurements using 100 batch size, Linear NN Search technique, and Euclidean distance. 84.02% accuracy is the lowest among K=1 with 100 batch size, Linear NN Search technique, and Distance with Euclidean function model. K=5 with 100 batch size, Linear NN Search technique, and Distance using Euclidean function model yield the greatest accuracy of 88.02%. K=3, K=4, and K=9 with 100 batch size, Linear NN Search algorithm, Distance with Euclidean function model, show 87.48%, 87.99%, and 87.72% accuracy, respectively.

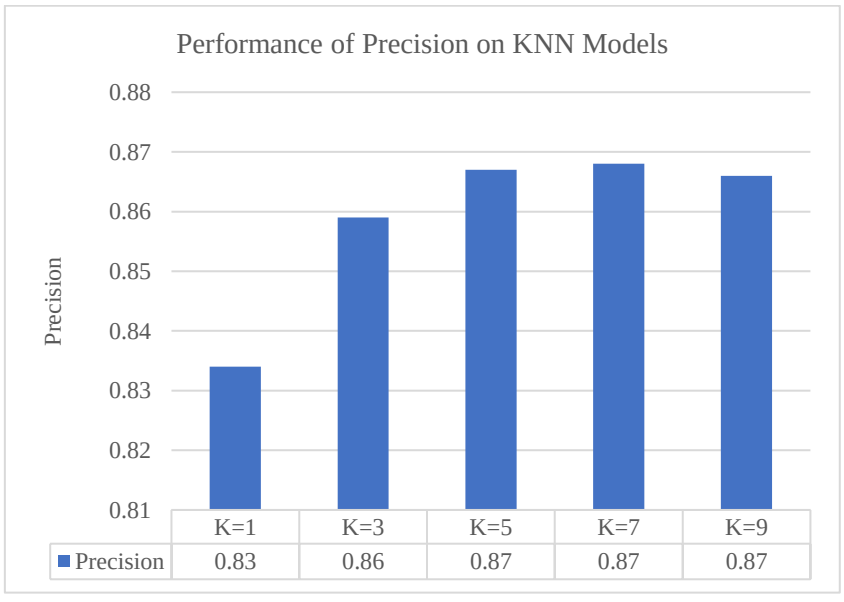


Figure 3. Precision performance of KNN Models

Diagram 3 exhibits KNN Model precision performance with 100 batch size, Linear NN Search technique, and Euclidean distance. 0.83 Precision is the lowest accuracy model compared to K=1 with 100 batch size, Linear NN Search technique, and Distance with Euclidean function model. K=5, K=7, and K=9 have the highest precision

DOI: https://doi.org/10.54216/FPA.210114
Received: January 10, 2025 Revised: May 20, 2025 Accepted: July 04, 2025

value of 0.87 with 100 batch size, Linear NN Search algorithm, and Distance with Euclidean function model. K=3 with 100 batch size, Linear NN Search technique, and Euclidean function model distance yield 0.86 precision.

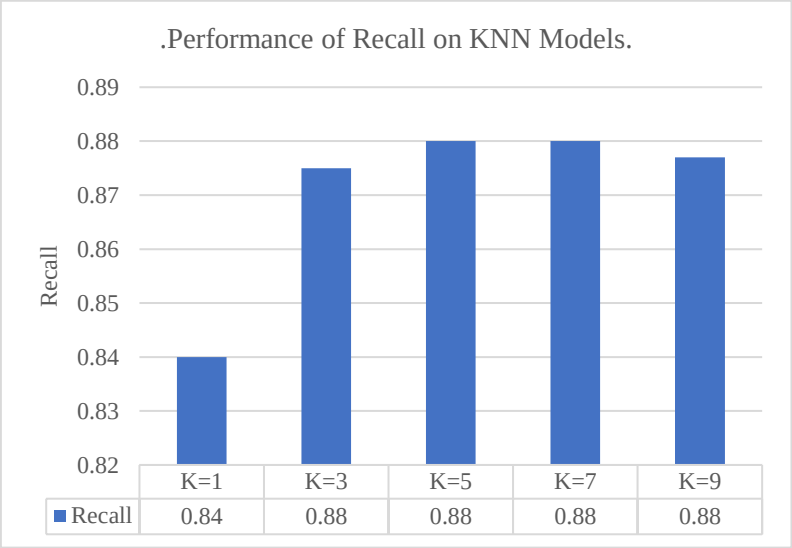
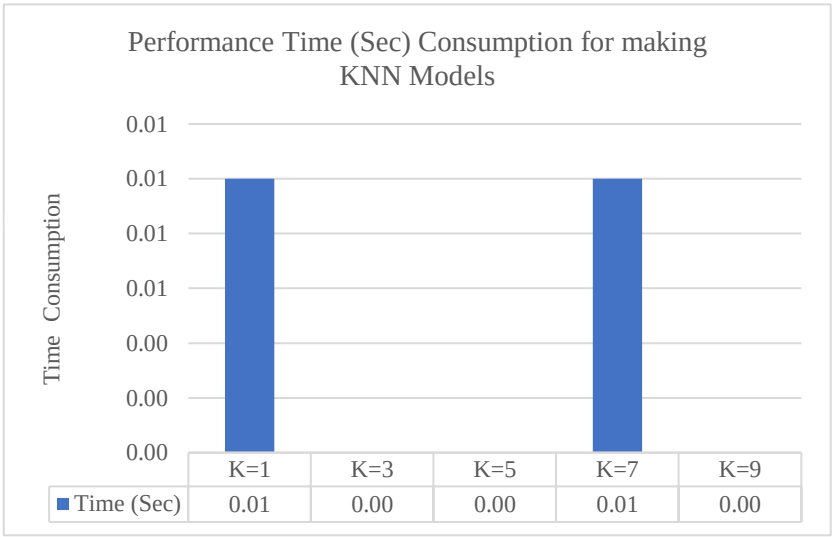


Figure 4. Recall performance of KNN Models

Figure 4 demonstrates KNN model recall values when using various K values with 100 batch size, Linear NN Search technique, and Euclidean distance. The lowest hit rate model is 0.84, obtained using K=1 with 100 batch size, Linear NN Search technique, and Euclidean Distance with function model. K=3 with 100 batch size, Linear NN Search algorithm, and Distance with Euclidean function model, K=5 with 100 batch size, K=7 with 100 batch size, and K=9 with 100 batch size yielded the maximum and same hit rate value of 0.88.



DOI: https://doi.org/10.54216/FPA.210114
Received: January 10, 2025 Revised: May 20, 2025 Accepted: July 04, 2025

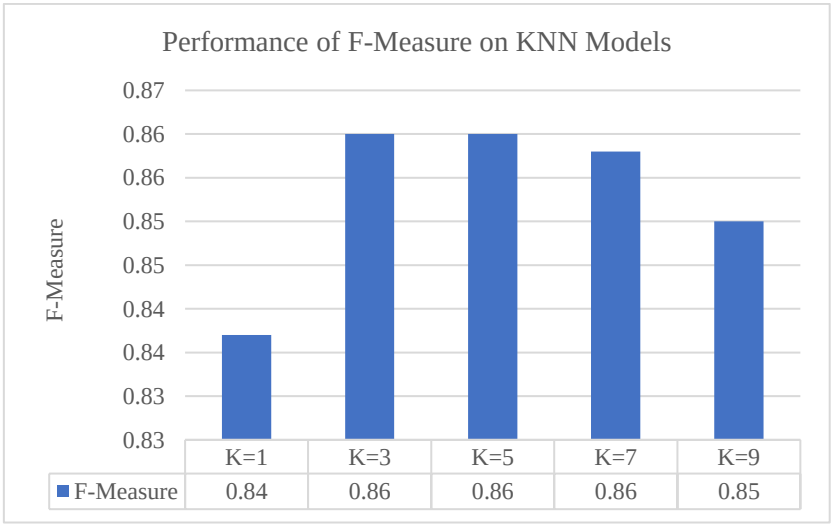
Figure 5. Time consumption performance of KNN Models

Figure 5 shows the time required to make KNN models using various K values with 100 batch size, Linear NN Search technique, and Euclidean distance. K=3, K=5, and K=9 with 100 batch size, Linear NN Search algorithm, and Distance with Euclidean function model for model construction take zero seconds. The highest time consumption for making their models is 0.01 seconds with K=1 and K=7 with 100 batch size, Linear NN Search technique, and Distance using Euclidean function model.

Table 3: F- Measure, MCC and Kappa performance of Lazy Classifier

S.No	Base Category	Classifier	Batch size	Nearest Neighborhood Search Algorithm	Distance Function	K value	F-Measure	MCC	Kappa Statistic
1	Lazy	K NN	100	Linear NN Search	Euclidean	K=1	0.84	0.36	0.36
						K=3	0.86	0.44	0.42
						K=5	0.86	0.45	0.41
						K=7	0.86	0.44	0.39
						K=9	0.85	0.41	0.36

KNN model performance with multiple K values, 100 batch size, Linear NN Search, and Euclidean distance is shown in Table 3. 0.84 F-Measure, 0.36 Matthews Correlation value level, and 0.36 kappa statistic value are cut by K=1 with 100 batch size, Linear NN Search, and Euclidean function model distance. K=3 with 100 batches, Linear NN Search, and Euclidean function model crop distance 0.86 F-Measure, 0.44 Matthews Correlation, 0.42 Kappa. 0.86 F-Measure, 0.45 Matthews Correlation value level, 0.41 kappa statistic value are sliced by K=5 with 100 batch size, Linear NN Search, and Euclidean function model distance. K=7, 100 batches, Linear NN Search, Distance with Euclidean function model crop F-measure 0.86, Matthews Correlation 0.44, and Kappa statistic 0.39. K=9, 100 batches, Linear NN Search, Distance using Euclidean function model crop F-measure 0.85, Matthews Correlation 0.41, and Kappa statistic 0.36.



DOI: https://doi.org/10.54216/FPA.210114
Received: January 10, 2025 Revised: May 20, 2025 Accepted: July 04, 2025

Figure 6. F-Measure performance of KNN Models

Figure 4 illustrates F-Measure value measurements on KNN models with 100 batch size, Linear NN Search, and Euclidean distance for various K values. 0.84 is the lowest F-measure value for K=1 with 100 batches, Linear NN Search, and Euclidean function model Distance. K=3, K=5, K=7, and K=9 with 100 batch size, Linear NN Search, and Distance using Euclidean function model have the highest F-Measure value of 0.86. Figure 7 shows MCC measurements on KNN models with 100 batch size, Linear Nearest Neighbor Search, and Euclidean distance. 0.36 is the smallest Matthew correlation coefficient compared to K=1 with 100 batch size, Linear NN Search, and Euclidean function model distance. Two models pick 0.44 Matthew correlation. The highest Matthew correlation value is 0.45 for K=3 with 100 batch size, Linear NN Search technique, and Distance with Euclidean function model, K=7, and K=5. K=9, 100 batches, Linear NN Search, and Euclidean function model distance provide 0.41 Matthews correlation coefficient.

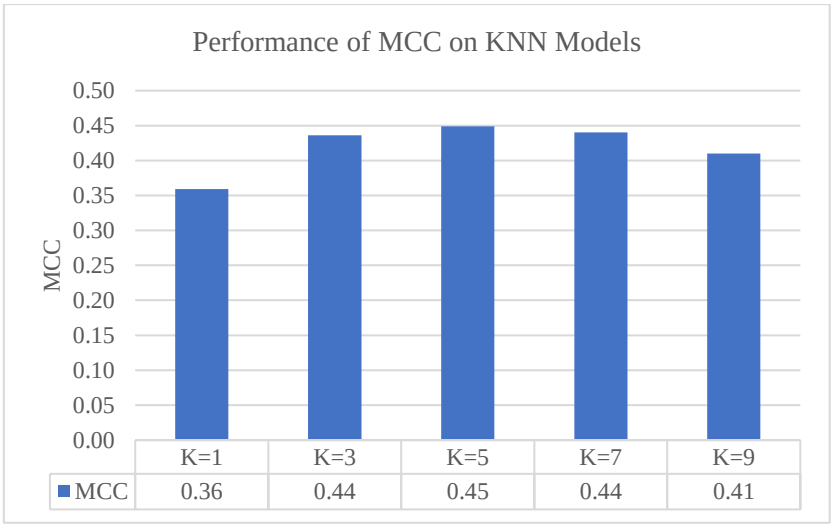


Figure 7. MCC performance of KNN Models

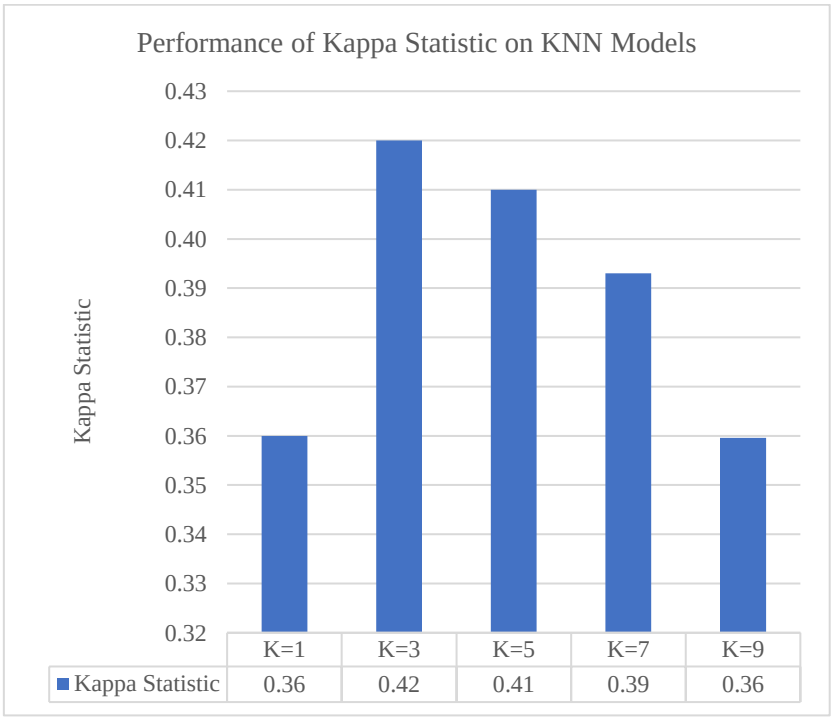


Figure 8. Kappa performance of KNN Models

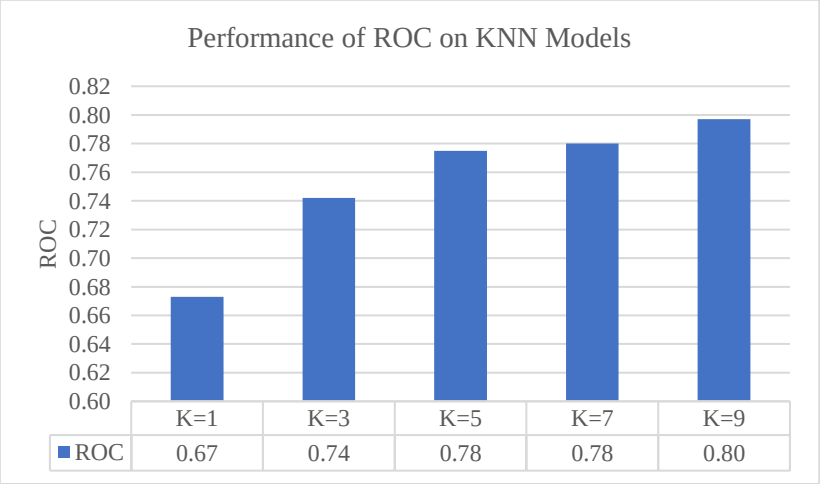
DOI: https://doi.org/10.54216/FPA.210114
Received: January 10, 2025 Revised: May 20, 2025 Accepted: July 04, 2025

Figure 8 displays KNN model kappa values with 100 batch size, Linear NN Search, Moreover, Euclidean distance. The lowest kappa was 0.36 for K=1, 100 batch size, Linear NN Search, and Euclidean function model distance. The greatest kappa is 0.42 with linear NN Search, Euclidean function model distance, K=3, and 100 batches. Other models K=9, K=7, and K=5 with 100 batch size, Linear NN Search, and Euclidean function distance have 0.36, 0.39, and 0.41 kappa.

Table 4: ROC and PRC Performance of Lazy Classifier

S.No	Base Category	Classifier	Batch size	Nearest Neighborhood Search Algorithm	Distance Function	K value	ROC	PRC
1	Lazy	K NN	100	Linear NN Search	Euclidean	K=1	0.67	0.81
						K=3	0.74	0.84
						K=5	0.78	0.86
						K=7	0.78	0.87
						K=9	0.80	0.87

The above table 4 shows that the measurements of Receiver Operating Characteristic value and Precision Recall Curve value on KNN models when applying various K values with 100 of batch size, Linear NN Search algorithm and Distance with Euclidean function. 0.67 of Receiver Operating Characteristic level and 0.81 of Precision Recall Curve level is owned by K=1 with 100 of batch size, Linear NN Search algorithm and Distance with Euclidean function model. 0.74 of Receiver Operating Characteristic level and 0.84 of Precision Recall Curve level is owned by K=3 with 100 of batch size, Linear NN Search algorithm and Distance with Euclidean function model. 0.78 of Receiver Operating Characteristic level and 0.86 of Precision Recall Curve level is owned by K=5 with 100 of batch size, Linear NN Search algorithm and Distance with Euclidean function model. 0.78 of Receiver Operating Characteristic level and 0.87 of Precision Recall Curve level is owned by K=7with 100 of batch size, Linear NN Search algorithm and Distance with Euclidean function model. 0.80 of Receiver Operating Characteristic level and 0.87 of Precision Recall Curve level is owned by K=9 with 100 of batch size, Linear NN Search algorithm and Distance with Euclidean function model.



DOI: https://doi.org/10.54216/FPA.210114
Received: January 10, 2025 Revised: May 20, 2025 Accepted: July 04, 2025

Figure 9. ROC performance of KNN Models

The above figure 9 shows that the measurements of ROC value on KNN models when applying various K values with 100 of batch size, Linear NN Search algorithm and Distance with Euclidean function. The smallest ROC value is 0.67 which is shown by K=1 with 100 of batch size, Linear NN Search algorithm and Distance with Euclidean function model and highest ROC value is 0.80 which is given by K=9 with 100 of batch size, Linear NN Search algorithm and Distance with Euclidean function model. The other models are showing 0.74 ROC value of K=3 with 100 of batch size, Linear NN Search algorithm and Distance with Euclidean function model, 0.78 ROC is given by K=5 with 100 of batch size, Linear NN Search algorithm and Distance with Euclidean function model and K=7 with 100 of batch size, Linear NN Search algorithm and Distance with Euclidean function model.

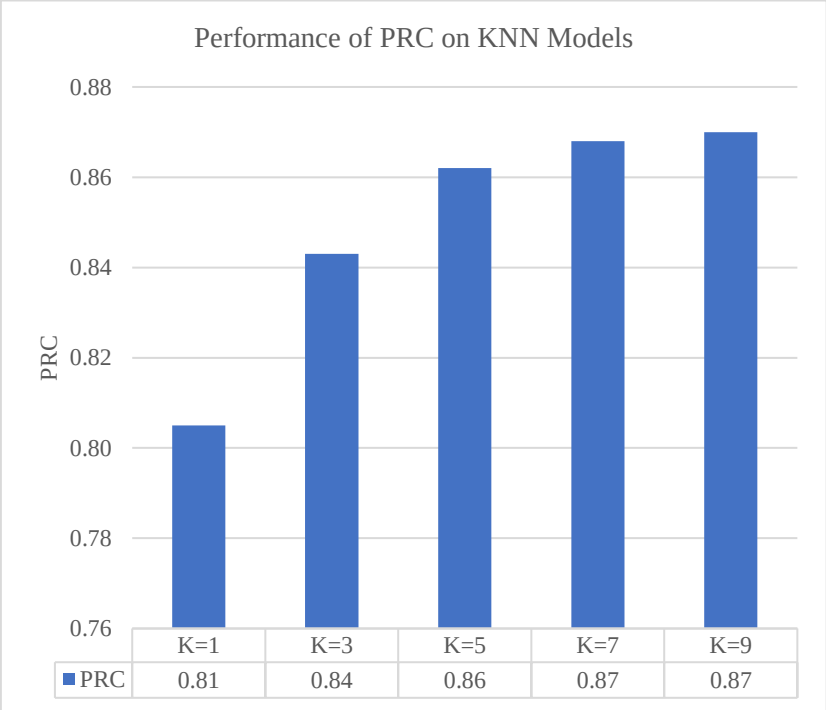


Figure 10. PRC performance of KNN Models

The above figure 10 shows that the measurements of recall value on KNN models when applying various K values with 100 of batch size, Linear NN Search algorithm and Distance with Euclidean function. Least PRC value is 0.81 which is shown by K =1 with 100 of batch size, Linear NN Search algorithm and Distance with Euclidean function model. Highest PRC value is 0.87 which is given by there are two models. They are K =7 with 100 of batch size, Linear NN Search algorithm and Distance with Euclidean function model and K-9 with 100 of batch size, Linear NN Search algorithm and Distance with Euclidean function model. 0.84 of PRC value is given by K =3 with 100 of batch size, Linear NN Search algorithm and Distance with Euclidean function model and 0.86 of PRC value is given by K =5 with 100 of batch size, Linear NN Search algorithm and Distance with Euclidean function model.

Table 5: Mean Deviations of Lazy Classifier

S.No	Base Category	Classifier	Batch size	Nearest Neighborhood Search Algorithm	Distance Function	K value	Mean Absolute Error	Root Mean Squared Error
1	Lazy	K NN	100		Euclidean	K=1	0.16	0.40

DOI: https://doi.org/10.54216/FPA.210114
Received: January 10, 2025 Revised: May 20, 2025 Accepted: July 04, 2025

2				Linear NN Search		K=3	0.17	0.33
3						K=5	0.17	0.31
4						K=7	0.18	0.32
5						K=9	0.18	0.31

The above table 5 shows that the measurements of recall value on KNN models when applying various K values with 100 of batch size, Linear NN Search algorithm and Distance with Euclidean function. 0.16 of mean absolute deviation and 0.40 of root mean squared deviation are given by shown by K =1 with 100 of batch size, Linear NN Search algorithm and Distance with Euclidean function model. 0.17 of mean absolute deviation and 0.33 of root mean squared deviation are given by shown by K =3 with 100 of batch size, Linear NN Search algorithm and Distance with Euclidean function model. 0.17 of mean absolute deviation, 0.31 of root mean squared deviation are given by shown by K =5 with 100 of batch size, Linear NN Search algorithm and Distance with Euclidean function model. 0.18 of mean absolute deviation, 0.32 of root mean squared deviation are given by shown by K =5 with 100 of batch size, Linear NN Search algorithm and Distance with Euclidean function model. 0.18 of mean absolute deviation, 0.31 of root mean squared deviation are given by shown by K =5 with 100 of batch size, Linear NN Search algorithm and Distance with Euclidean function model.

5. Conclusion

This research work concludes that the highest accuracy value is 88.02% of accuracy level which is acquired by K=5 with 100 of batch size, Linear NN Search algorithm and Distance with Euclidean function model. The maximum and same hit rate value is 0.88, which is obtained when K = 3 with 100 batch size, the Linear NN Search algorithm, and the Distance with Euclidean function model are used, and when K = 5 with 100 batch size, the Linear NN Search algorithm, and the Distance with Euclidean function model are used. The K = 5 with a batch size of 100, the Linear NN Search algorithm, and the Euclidean function model Distance K = 3 with a batch size of 100, the Linear NN Search algorithm, and the Distance with Euclidean function model produce the highest F-Measure value of 0.86, K = 5 with a batch size of 100, the Linear NN Search algorithm, and the Distance with Euclidean function model produce the highest F-Measure value of 0.86, and K = 7 with a batch size of 100, the Linear NN Search algorithm, and the Distance with Euclidean K = 5 with a batch size of 100, the Linear NN Search technique, and the Distance with Euclidean function model produce the greatest Matthew correlation value of 0.45. K = 5 with 100 batch size, the Linear NN Search algorithm, and the Distance with Euclidean function model produce the lowest root mean squared deviation of 0.31. The K=1 with 100 batch size, the Linear NN Search algorithm, and the Distance with Euclidean function model have the lowest relative absolute error performance of 61.66%. K=1 with 100 batch size, the Linear NN Search algorithm, and the Distance with Euclidean function model have the best root relative squared error performance of 110.98%. This study recommended that the K=5 with 100 batch size, Linear NN Search algorithm, and Distance with Euclidean function mode is the best model when compared to other models.

Conflicts of Interest: The authors assert that there are no conflicts of interest.

Authors' Contributions: All writers made substantial contributions to the finalization of this publication.

References

[1] N. Dhahri, A. B. Mtibaa, and A. Alimi, "Automated breast cancer diagnosis based on machine learning algorithms," Procedia Computer Science, vol. 112, pp. 400–407, 2017.
[2] S. Al-Wesabi, Y. Jiang, and W. A. M. Alzubaidi, "Breast cancer classification based on deep learning approaches and histopathology images: A review," IEEE Access, vol. 8, pp. 165779–165793, 2020.
[3] M. A. Qureshi, M. M. Iqbal, and A. A. Baig, "Effective breast cancer diagnosis using hybrid machine learning techniques," in Proc. 2020 3rd Int. Conf. Comput. Appl. Inf. Security (ICCAIS), Riyadh, Saudi Arabia, 2020, pp. 1–6.
[4] T. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," Computational and Structural Biotechnology Journal, vol. 13, pp. 8–17, 2015.
[5] R. Smith, B. Johnson, and C. Lee, "A Comprehensive Review of Machine Learning Techniques for Breast Cancer Diagnosis," J. Healthcare Eng., vol. 2021, Article ID 123456, 2021.

DOI: https://doi.org/10.54216/FPA.210114
Received: January 10, 2025 Revised: May 20, 2025 Accepted: July 04, 2025

[6] S. W. Harizi, R. Zlitni, and A. B. Mtibaa, "Breast cancer diagnosis using machine learning algorithms and histopathological images: A comparative study," in Proc. 2019 Int. Conf. Adv. Technol. Signal Image Process. (ATSIP), Sfax, Tunisia, 2019, pp. 1–6.
[7] M. T. Khan, S. Ali, and J. M. Patel, "Deep Learning Approaches for Medical Image Analysis: A Survey," J. Med. Syst., vol. 45, no. 10, pp. 1–15, 2021.
[8] S. V. Parmar and N. T. Shah, "Breast cancer classification using deep learning and traditional classifiers," in Proc. 2019 5th Int. Conf. Comput. Commun. Control Automat. (ICCUBEA), Pune, India, 2019, pp. 1–5.
[9] 9. S. Taghizadeh, R. N. Jafari, and S. Shokouhi, "Breast cancer detection using hybrid genetic algorithm and support vector machine," J. Comput. Biol., vol. 26, no. 10, pp. 1090–1100, 2019.
[10] P. Kavitha, S. Prabakaran, "Brain Tumor Image Segmentation Using Hybrid Assured Convergence Particle Swarm Optimization and Fuzzy C-Means," Int. J. Sci. Technol. Res., vol. 9, no. 2, pp. 2277-8616, Feb. 2020.
[11] G. Litjens et al., "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60–88, 2017.
[12] H. R. Rangayyan, N. Ayres, and J. E. L. Desautels, "A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs," J. Biomed. Eng., vol. 31, no. 5, pp. 304–316, 2007.
[13] D. Esteva et al., "A guide to deep learning in healthcare," Nature Medicine, vol. 25, pp. 24–29, 2019.
[14] P. V. Kumar and P. V. Aruna, "Performance analysis of machine learning algorithms for breast cancer prediction," Procedia Computer Science, vol. 171, pp. 593–601, 2020.
[15] M. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," Procedia Computer Science, vol. 83, pp. 1064–1069, 2016.
[16] H. Almotairi et al., "Deep learning techniques for classification of mammograms: A review," Future Internet, vol. 12, no. 12, p. 219, 2020.
[17] H. C. Shin et al., "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," IEEE Trans. Med. Imaging, vol. 35, no. 5, pp. 1285–1298, May 2016.
[18] Ganesan R, Latha A, Kavitha P, Venkatesan G. "Experimental Investigation of Wastewater by Using Novel Borassus flabellifer Fiber and Cocos nucifera Fiber," Asian J. Water, Environ. Pollut., vol. 21, no. 6, pp. 39–47, Dec. 2024.
[19] M. M. Rahman, M. F. Rahman, and M. A. M. Hossain, "A novel hybrid machine learning model for breast cancer prediction," in Proc. 2021 Int. Conf. Smart Syst. Innov. Technol. (SSIT), Dhaka, Bangladesh, 2021, pp. 1–6.
[20] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015.
[21] Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2012, pp. 1097–1105.
[22] T. Liu, H. Zhou, and W. Wang, "A transfer learning framework for breast cancer classification," IEEE Access, vol. 7, pp. 53031–53040, 2019.
[23] H. J. Chang, M. S. Park, and Y. Lee, "Explainable CNN for breast cancer classification," in Proc. Int. Conf. Bioinformat. Biomed, 2020, pp. 1108–1113.
[24] Z. Song et al., "Breast cancer histology image classification using residual CNN," in Proc. Int. Conf. Neural Inf. Process., 2017, pp. 364–373.
[25] M. Rezaeipanah et al., "Hybrid ML methods for breast cancer detection using mammography," Compute. Biol. Med., vol. 136, p. 104713, 2021.
[26] R. Paul et al., "Breast cancer detection using explainable AI," J. Digital Imaging, vol. 34, pp. 417–428, 2021.
[27] F. Almansour and M. Sagheer, "Transfer learning and CNNs for breast cancer histology," J. Imaging, vol. 7, no. 8, p. 134, 2021.
[28] R. Chaurasia and S. Pal, "ML algorithms for breast cancer detection using WBCD," Procedia Compute. Sci., vol. 167, pp. 821–831, 2020.
[29] A. Soliman and M. S. Ramadan, "Performance evaluation of ML for breast cancer prediction," in Proc. Int. Conf. Intell. Comput. Inf. Syst. (ICICIS), 2019, pp. 88–93.
[30] Y. Bar et al., "Deep learning with non-medical training used for medical image analysis," J. Digital Imaging, vol. 28, no. 1, pp. 59–64, 2015.

DOI: https://doi.org/10.54216/FPA.210114
Received: January 10, 2025 Revised: May 20, 2025 Accepted: July 04, 2025