# Optimization of Breast Cancer Prediction and Diagnosis using Hybrid Machine Learning Technique

Alka Chouhan
M. Tech. Scholar
Dept. of Computer Science and Engineering
Sagar Institute of Research Technology Excellence
Bhopal, India

Swati Khanve
Assistant Professor
Dept. of Computer Science and Engineering
Sagar Institute of Research Technology Excellence
Bhopal, India

Nitya Khare
Assistant Professor & HOD
Dept. of Computer Science and Engineering
Sagar Institute of Research Technology Excellence
Bhopal, India

## ABSTRACT
Breast cancer is one of the leading causes of mortality among women worldwide, and early detection plays a crucial role in improving survival rates. In recent years, machine learning techniques have shown significant potential in assisting medical diagnosis by providing accurate and efficient prediction models. This study focuses on the optimization of breast cancer prediction and diagnosis using a hybrid machine learning approach that combines K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) algorithms. The proposed methodology involves preprocessing the dataset to handle missing values, normalize features, and select the most relevant attributes for classification. Initially, individual models based on KNN and SVM are developed and evaluated in terms of accuracy, precision, recall, and F1-score. While KNN is effective in capturing local data patterns, SVM provides robust classification by maximizing the margin between different classes. However, each method has its own limitations when used independently. To overcome these limitations, a hybrid KNN+SVM model is proposed, which integrates the strengths of both algorithms. The hybrid approach enhances classification performance by improving decision boundaries and reducing misclassification rates. The optimized model is expected to achieve higher accuracy and better generalization compared to individual classifiers. The experimental results demonstrate that the hybrid model outperforms traditional methods in predicting breast cancer with improved reliability and efficiency. This approach can assist healthcare professionals in early diagnosis and decision-making, ultimately contributing to better patient outcomes and reduced mortality rates.

## Keywords
Breast Cancer Prediction, Machine Learning, Hybrid Model, Healthcare Analytics

## 1. INTRODUCTION
Breast cancer is one of the most common and life-threatening diseases affecting women across the globe. According to global health reports, early detection and timely diagnosis significantly improve the chances of successful treatment and survival. However, traditional diagnostic methods, such as manual examination and imaging techniques, often depend on the expertise of medical professionals and may be prone to human error or delayed interpretation. Therefore, there is a growing need for intelligent and automated systems that can assist in accurate and early detection of breast cancer [1].

In recent years, machine learning has emerged as a powerful tool in the field of medical diagnosis and healthcare analytics. Machine learning algorithms can analyze large volumes of medical data, identify hidden patterns, and make predictions with high accuracy. These techniques are widely used in disease classification, risk prediction, and decision support systems. In the context of breast cancer diagnosis, machine learning models can help classify tumors as benign or malignant based on features extracted from medical datasets, such as cell size, texture, shape, and other relevant parameters [2, 3].

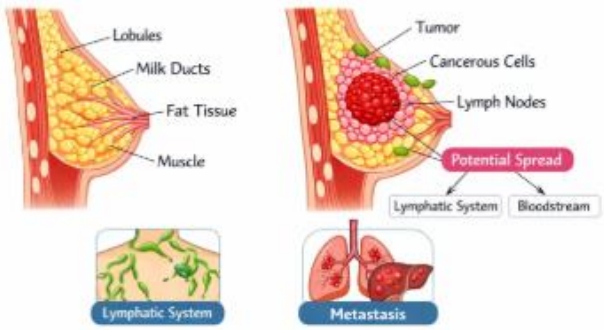

Fig. 1: Breast Cancer

Among various machine learning algorithms, KNN and SVM are widely used for classification problems. KNN is a simple and effective algorithm that classifies data points based on their similarity to neighboring data. It performs well in cases where the data distribution is locally structured. However, KNN can be sensitive to noise and may require high computational time for large datasets. On the other hand, SVM is a powerful supervised learning algorithm that constructs an optimal hyperplane to separate different classes. It is highly effective in handling high-dimensional data and provides better generalization performance, but its performance depends on proper kernel selection and parameter tuning [4].
Despite their individual strengths, both KNN and SVM have certain limitations when applied independently. To address these limitations, hybrid machine learning techniques have gained significant attention. A hybrid approach combines multiple algorithms to leverage their strengths and minimize their weaknesses. In this study, a hybrid KNN+SVM model is proposed to enhance the accuracy and reliability of breast cancer prediction and diagnosis. The hybrid model aims to improve classification performance by integrating the local

pattern recognition capability of KNN with the strong decision boundary formulation of SVM [5, 6].
The main objective of this work is to optimize the prediction and diagnosis of breast cancer by improving classification accuracy and reducing misclassification rates. The proposed hybrid approach is expected to provide better results compared to traditional single-model techniques. This research contributes to the development of intelligent healthcare systems that can assist doctors in making faster and more accurate decisions, ultimately improving patient outcomes and reducing mortality rates.

## 2. PROPOSED METHODOLOGY

The proposed methodology aims to develop an efficient breast cancer prediction system using a hybrid combination of KNN and SVM algorithms. The complete process is carried out in multiple stages as follows and flow diagram is present in fig.2:

**1. Data Collection**
The dataset (such as Wisconsin Breast Cancer Dataset) is collected, containing features like cell size, texture, radius, and diagnosis (benign/malignant).

**2. Data Preprocessing**
Raw data is processed to improve quality:

- Handling missing values
- Removing noise and outliers
- Normalizing/scaling features
- Encoding class labels

**3. Feature Selection**
Important features are selected using statistical or correlation-based methods to improve accuracy and reduce computation time.

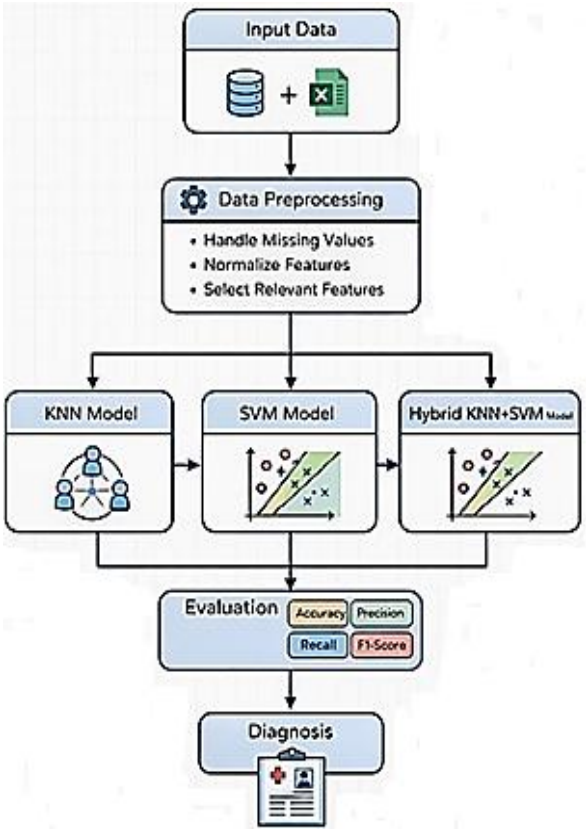


**Fig. 2: Proposed Methodology**

**4. Model Development**
Three models are developed:

- **KNN Model:**

The KNN algorithm is a simple and effective supervised machine learning technique used for classification and prediction is present in fig. 3. In this project, KNN is applied to classify breast cancer data into two categories: benign (non-cancerous) and malignant (cancerous). KNN works on the principle of similarity, where a data point is classified based on the majority class of its nearest neighbors. During the training phase, the algorithm stores all the training data points. When a new test sample is given, the algorithm calculates the distance between the test point and all training points using distance metrics such as Euclidean distance. It then selects the 'K' closest neighbors and assigns the class that is most common among them [7, 8].

In the context of breast cancer prediction, features such as cell size, texture, radius, and smoothness are used as input parameters. The value of 'K' plays a crucial role in the performance of the model. A smaller value of K may lead to overfitting, while a larger value may reduce the model's accuracy. Therefore, an optimal value of K is selected through experimentation [9].

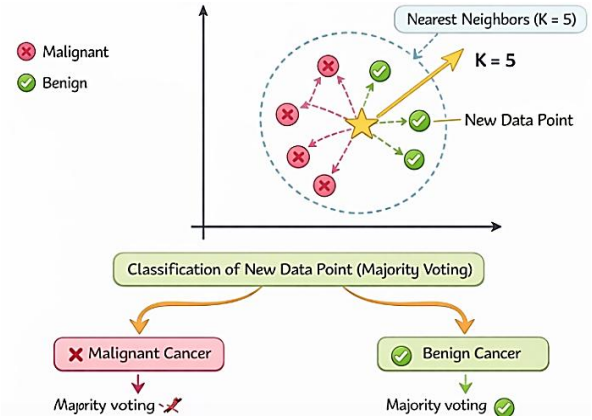


**Fig. 3: KNN**

The KNN model is easy to implement and does not require complex training, making it suitable for small to medium-sized datasets. However, it has some limitations, such as high computational cost during prediction and sensitivity to noise in the data. Despite these drawbacks, KNN provides good classification performance and is useful as a baseline model in this study.

- **SVM Model:**

The Support Vector Machine (SVM) is a powerful supervised machine learning algorithm used for classification and regression tasks. SVM is used to classify breast cancer data into two categories: benign (non-cancerous) and malignant (cancerous) is present in fig. 4.
SVM works by finding an optimal boundary, known as a hyperplane, that separates data points of different classes with maximum margin. The margin is the distance between the hyperplane and the nearest data points from each class, which are called support vectors. These support vectors play a crucial role in defining the position and orientation of the hyperplane.
In breast cancer prediction, input features such as cell size, texture, radius, and smoothness are used to plot data points in a multi-dimensional space. The SVM algorithm analyzes these features and constructs a decision boundary that best separates

benign and malignant samples. If the data is not linearly separable, SVM uses kernel functions (such as linear, polynomial, or radial basis function) to transform the data into a higher-dimensional space where separation becomes possible.

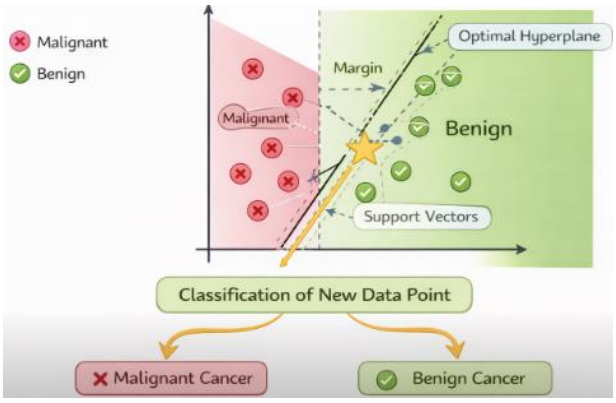


**Fig. 4: SVM**

One of the main advantages of SVM is its ability to handle high-dimensional data and provide high accuracy with good generalization. It is also effective in avoiding overfitting, especially when the margin is maximized. However, SVM requires careful selection of parameters such as kernel type and regularization factor, and it can be computationally intensive for very large datasets.

- **Hybrid KNN+SVM Model:**

The Hybrid KNN + SVM model is an advanced machine learning approach that combines the strengths of both K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) algorithms to improve the accuracy and reliability of breast cancer prediction.

In this approach, KNN is first used to analyze the local structure of the data. When a new data point (patient data) is given, the KNN algorithm identifies the 'K' nearest neighbors based on similarity (distance). This step helps in selecting a relevant subset of data points that are most similar to the input sample, reducing noise and unnecessary data.

After this, the selected neighbors are passed to the SVM model. SVM then constructs an optimal hyperplane using these refined data points to classify the input as either benign or malignant. Since SVM works best with well-structured and less noisy data, the preprocessing done by KNN improves the overall classification performance.

**5. Training and Testing**

The dataset is divided into training and testing sets. Models are trained and tested to evaluate performance.

**6. Performance Evaluation**

Models are compared using:

- Accuracy
- Precision
- Recall
- F1-Score

**7. Final Diagnosis**

The best-performing model (Hybrid KNN+SVM) is selected for predicting whether cancer is benign or malignant.

## 3. RESULTS

The proposed hybrid machine learning model combining KNN and SVM algorithms was implemented and evaluated using a standard breast cancer dataset. The dataset was preprocessed by handling missing values, normalizing features, and selecting the most relevant attributes to improve classification performance.

| | mean_radius | mean_texture | mean_perimeter | mean_area | mean_smoothness | diagnosis |
|---|---|---|---|---|---|---|
| 0 | 17.99 | 10.38 | 122.80 | 1001.0 | 0.11840 | 0 |
| 1 | 20.57 | 17.77 | 132.90 | 1326.0 | 0.08474 | 0 |
| 2 | 19.69 | 21.25 | 130.00 | 1203.0 | 0.10960 | 0 |
| 3 | 11.42 | 20.38 | 77.58 | 386.1 | 0.14250 | 0 |
| 4 | 20.29 | 14.34 | 135.10 | 1297.0 | 0.10030 | 0 |

**Fig. 5: Dataset**

The given fig. 5 represents a sample dataset used for breast cancer prediction, where each row corresponds to an individual patient or tumor sample, and each column represents a specific feature of the tumor. The features such as *mean radius*, *mean texture*, *mean perimeter*, *mean area*, and *mean smoothness* describe the physical and structural properties of the tumor cells obtained from medical imaging or biopsy analysis. For instance, mean radius indicates the average size of the tumor, while mean texture reflects the variation in pixel intensity, showing how uniform or irregular the cells appear. Similarly, mean perimeter and mean area provide information about the size and shape of the tumor, and mean smoothness indicates how regular or irregular the cell boundaries are.

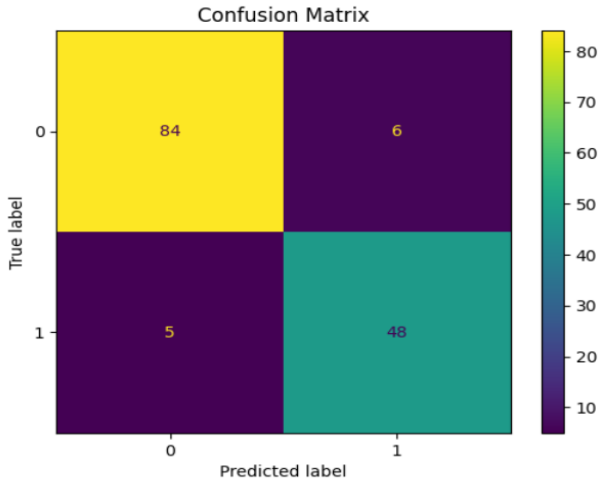


**Fig. 6: CM for KNN**

These features are important because malignant (cancerous) tumors generally exhibit larger size and more irregular structures compared to benign (non-cancerous) tumors. The last column, *diagnosis*, is the target variable, where the value '0' represents a benign tumor and '1' represents a malignant tumor.

The given confusion matrix represents fig. 6 to fig. 8 the performance of the breast cancer prediction model by comparing the actual (true) labels with the predicted labels. It is a 2×2 matrix where the rows indicate the true class (0 = benign, 1 = malignant) and the columns indicate the predicted class. From the matrix, 84 cases are correctly predicted as benign (true negatives), while 48 cases are correctly predicted as malignant (true positives). However, there are some misclassifications: 6 cases are incorrectly predicted as malignant when they are actually benign (false positives), and 5 cases are incorrectly predicted as benign when they are actually malignant (false negatives). These values indicate that the model performs well with a high number of correct predictions and relatively few errors. The low number of false

negatives is especially important in medical diagnosis, as it means fewer cancer cases are missed. Overall, the confusion matrix demonstrates that the model has strong classification ability and can effectively distinguish between benign and malignant tumors.

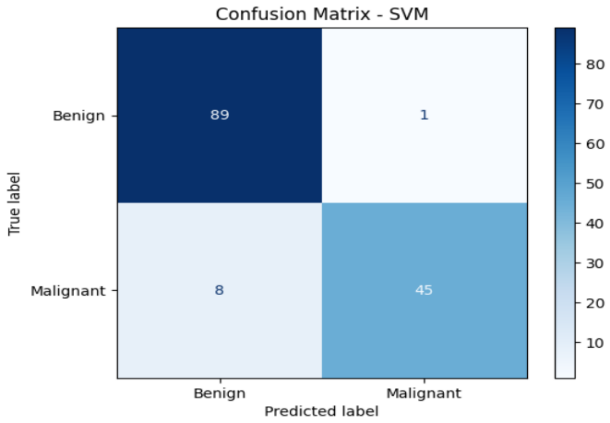


**Fig. 7: CM for SVM**

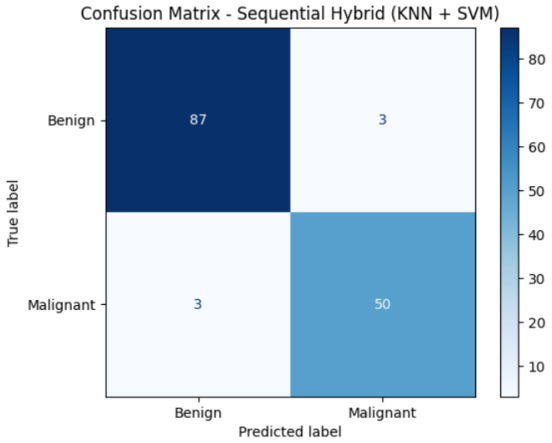


**Fig. 8: CM for SVM+KNN**

**Table 1: Results**

| ML Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| KNN | 92.31% | 88.89% | 90.57% | 89.72% |
| SVM | 93.70% | 97.82% | 84.90% | 90.90% |
| Hybrid Model | 95.80% | 97% | 97% | 97% |

The given table 1 presents the performance comparison of three machine learning models—KNN, SVM, and the Hybrid KNN+SVM model—used for breast cancer prediction based on evaluation metrics such as accuracy, precision, recall, and F1-score. The KNN model achieves an accuracy of 92.31%, with a precision of 88.89%, recall of 90.57%, and F1-score of 89.72%, indicating a balanced but slightly lower performance compared to the other models. The SVM model shows improved accuracy at 93.70% and a very high precision of 97.82%, meaning it is highly effective in correctly identifying malignant cases with fewer false positives; however, its recall is comparatively lower at 84.90%, indicating that it misses some actual cancer cases. In contrast, the Hybrid KNN+SVM model outperforms both individual models, achieving the highest accuracy of 95.80% along with balanced precision and recall values of 97%, resulting in an F1-score of 97%.

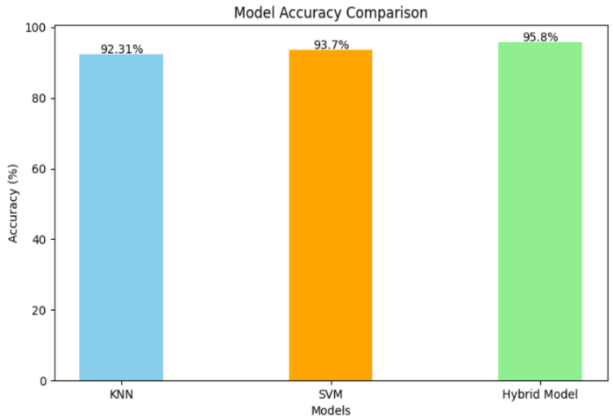


**Fig. 9: Graphically Accuracy**

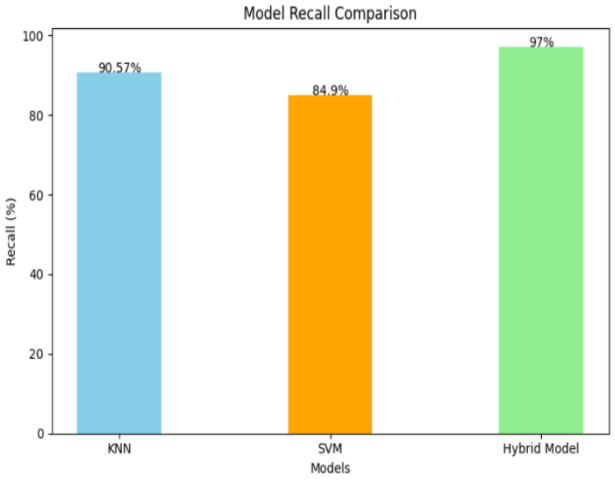


**Fig. 10: Graphically Recall**

This indicates that the hybrid model not only improves overall prediction accuracy but also maintains a strong balance between correctly identifying cancer cases and minimizing errors. Overall, the results clearly demonstrate that combining KNN and SVM enhances the classification performance and makes the hybrid model more reliable for breast cancer diagnosis.

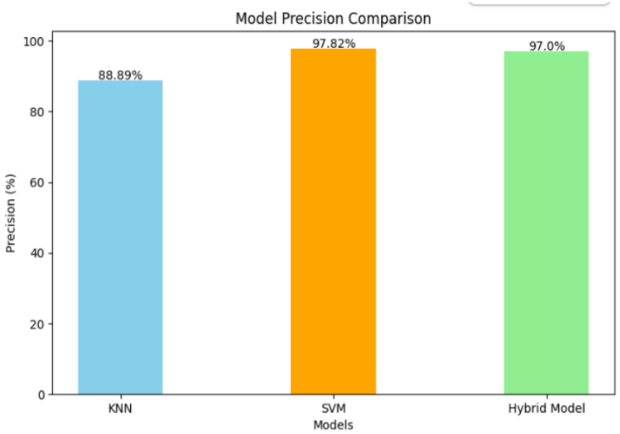


**Fig. 11: Graphically Precision**

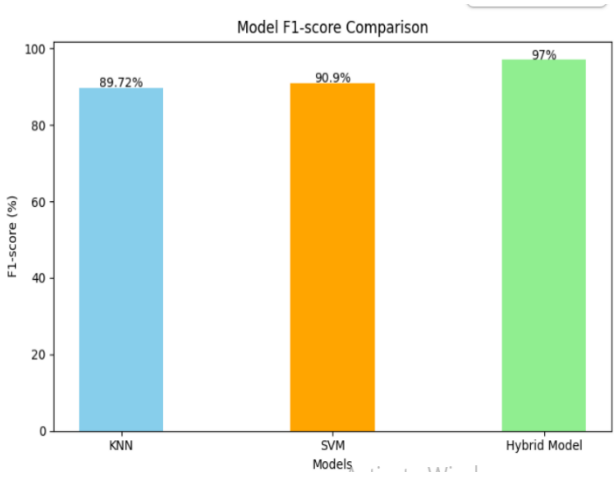


**Fig. 12: Graphically F1-score**

## 4. CONCLUSION

In this work, an optimized approach for breast cancer prediction and diagnosis will be developed using a hybrid machine learning model that combines KNN and SVM algorithms. The proposed system will focus on improving classification accuracy and reducing misclassification rates by leveraging the strengths of both techniques. KNN will be used to identify the nearest and most relevant data points, while SVM will be applied to construct an optimal decision boundary for accurate classification.

The methodology will include data preprocessing, feature selection, model training, and performance evaluation using metrics such as accuracy, precision, recall, and F1-score. It is expected that the hybrid KNN+SVM model will outperform individual models in terms of prediction accuracy and robustness.

The proposed system will assist in early detection of breast cancer by providing reliable and efficient diagnosis, which can support medical professionals in decision-making. Overall, this work will contribute to the development of intelligent healthcare systems and demonstrate the effectiveness of hybrid machine learning techniques in improving diagnostic performance.

## 5. REFERENCES

[1] C. Bista, A. M, S. Slimanzay, M. S. Sheikh and P. Srinivasa Rao, "Breast Cancer Prediction System Utilizing Machine Learning Algorithms," *2024 IEEE AITU: Digital Generation*, Astana, Kazakhstan, 2024, pp. 80-84.

[2] T. Matsuda, M. Matsuda, H. Haque *et al.*, "Diagnostic accuracy of a machine learning model using radiomics features from breast synthetic MRI," *BMC Med. Imaging*, vol. 25, art. no. 399, pp. 1–11, Sept. 2025.

[3] J. Zhang, Q. Wu, P. Lei *et al.*, "Diagnostic accuracy of machine learning-based magnetic resonance imaging models in breast cancer classification: A systematic review and meta-analysis," *World J. Surg. Onc.*, vol. 23, art. no. 231, pp. 1–13, Jun. 2025.

[4] C. F. Lee, J. Lin, Y.-L. Huang *et al.*, "Deep learning-based breast MRI for predicting axillary lymph node metastasis: A systematic review and meta-analysis," *Cancer Imaging*, vol. 25, art. no. 44, pp. **1–15**, Mar. 2025.

[5] R. Liang, F. Li, J. Yao *et al.*, "Predictive value of MRI-based deep learning model for lymphovascular invasion status in node-negative invasive breast cancer," *Sci. Rep.*, vol. 14, art. no. 16204, pp. 1–12**,** Jul. 2024.

[6] G. Houssami, M. Turner and R. Morrow, "Machine Learning for Breast MRI: Current Applications and Future Directions," *European Radiology*, vol. 31, no. 6, pp. 3752–3764, 2021.

[7] Y. Zheng, B. Liu and S. Chen, "Comparative Study of Machine Learning Techniques for Breast Cancer Classification Using MRI Images," *Biomedical Signal Processing and Control*, vol. 68, Article ID 102645, 2021.

[8] X. Zhang, Y. Chen and Z. Wang, "Breast Cancer Diagnosis Using Transfer Learning with Pretrained CNN Models on MRI," *Computers in Biology and Medicine*, vol. 115, Article ID 103498, 2019.

[9] Khurma RA, Aljarah I, Sharieh A, Elaziz MA, Damaševičius R, Krilavičius T. A review of the modifcation strategies of nature inspired algorithms for feature selection problem. Mathematics. 2022;10(3):1–45.

[10] Jain AK. Data clustering: 50 years beyond K-means. Pattern Recogn Lett. 2010;31(8):651–66.

[11] Murtagh F, Contreras P. Algorithms for hierarchical clustering: an overview. Wiley Interdiscip Rev Data Min Knowl Discov. 2012;2(1):86–97.

[12] Dabhi DP, Patel MR, Dipak MRP, Dabhi P. Extensive survey on hierarchical clustering methods in data mining. Int Res J Eng Technol, 2016; 03(11):659–665.

[13] Kriegel HP, Kröger P, Sander J, Zimek A. Density-based clustering. Wiley Interdiscip Rev Data Min Knowl Discov. 2011;1(3):231–40.

[14] Moulavi D, Jaskowiak PA, Campello RJGB, Zimek A, Sander J. Density-based clustering validation. SIAM Int Conf Data Min 2014, SDM 2014. 2014; 2(i):839–847.

[15] Aziz R, Verma CK, Srivastava N. A novel approach for dimension reduction of microarray. Comput Biol Chem. 2017; 71:161–169.