# A Study-Oriented Analysis of Simplifying Breast Cancer Data Using Machine Learning Approaches

**Digeshwar Prasad Sahu**
Research Scholar
Department of Computer Science & Engg.
Shri Rawatpura Sarkar University, Raipur (C.G.)

**Supervisor**
**Ranu Pandey**
Assistant Professor
Department of Computer Science & Engg.
Shri Rawatpura Sarkar University, Raipur (C.G.)

**Abstract**

Breast cancer is a significant global health concern, accounting for a large number of cancer-related deaths among women. Timely and accurate diagnosis is critical to improving treatment outcomes and patient survival rates. The complexity and volume of breast cancer data, however, present challenges in effective diagnosis using traditional methods. This research explores the application of **machine learning (ML) techniques** to simplify, analyze, and derive meaningful insights from breast cancer datasets.

The study involves the use of various supervised and unsupervised machine learning algorithms—including Decision Trees, Support Vector Machines (SVM), Random Forest, k-Nearest Neighbors (k-NN), and Artificial Neural Networks (ANN)—to classify and predict cancer stages and types. The research focuses on data preprocessing, feature extraction, model training, and performance evaluation using metrics such as accuracy, precision, recall, and F1-score.

Results indicate that machine learning not only enhances diagnostic accuracy but also improves the speed and efficiency of data interpretation, offering valuable support to clinicians. The integration of ML into medical research provides a scalable and reliable approach for breast cancer prediction and prognosis. This research emphasizes the role of computational models in transforming raw clinical data into actionable healthcare insights, ultimately contributing to more effective and personalized treatment strategies.

**Keywords :** Decision Tree, Random Forest, Prediction, Logistic Regression, ML, K-Mean

## Introduction

Breast cancer remains one of the most prevalent and life-threatening forms of cancer among women worldwide. Early diagnosis and accurate prediction of breast cancer can significantly improve patient survival rates and reduce the burden on healthcare systems. With the exponential growth of medical data, traditional diagnostic methods often fall short in efficiently handling and analyzing complex datasets. In this context, **Machine Learning (ML)** has emerged as a powerful tool, offering advanced capabilities in pattern recognition, classification, and predictive analytics.

This research-based exploration aims to investigate how machine learning techniques can simplify and enhance the understanding of breast cancer datasets, thereby improving early detection, diagnosis accuracy, and treatment planning. Various algorithms—such as Decision Trees, Support Vector Machines (SVM), Random Forest, k-Nearest Neighbors (k-NN), and Neural Networks—are evaluated for their effectiveness in classifying cancerous and non-cancerous cases based on parameters like tumor size, shape, texture, and other clinical features.

According to the World Cancer Research Foundation, breast cancer is the second most common cancer in women worldwide, with over 2 million new cases diagnosed in 2018 alone (World Cancer Research Fund & American Institute for Cancer Research, 2018). This represents about 12% of all new cancer cases and 25% of all cancers in women. The 2011 report of the National Cancer Registry Programme (NCRP) (NCRP, 2011) identifies cancer of the female breast as the most common form of cancer affecting Indian women in Mumbai, Thiruvananthapuram, and Dibrugarh. In the remaining registries too, it becomes the second most common type of cancer. The relative proportion of breast cancer in females varied from 14.4% in Guwahati (lowest) to 30.3% in Mumbai (highest).

The study also emphasizes data preprocessing techniques, feature selection methods, and model evaluation metrics such as accuracy, precision, recall, and F1-score. By integrating computational intelligence with medical research, this exploration seeks to bridge the gap between raw clinical data and actionable medical insights. Ultimately, this research aims to provide a comprehensive outlook on how machine learning not only eases the handling of breast cancer data but also contributes to more informed and data-driven decision-making in the medical field.

**Machine Learning: An Overview**

Learning is the essence of intelligence. If a system could learn and gain from experiences and improve its performance automatically, it would be an advanced tool for solving complex problems such as in the biological systems. The problem of inducing general functions from specific training examples is the central idea of learning. The learning agent is given a set of training and test examples of a category. The agent learns from the training examples and defines the hypothesis for them. The agent must search through the hypothesis space and locates the best hypothesis when given the test sets. There are three main categories of learning.

1. Supervised learning in which both the inputs and the outputs of a component can be observed;
2. Reinforcement learning where the learning agent is given an evaluation of its action but not told the correct action and
3. Unsupervised learning where the learning agent has no information about what the correct outputs are.

To solve problems computers require intelligence. Learning is central to intelligence. As intelligence requires knowledge, it is necessary for the computers to acquire knowledge. Machine learning serves this purpose.

Various techniques of machine learning are used for medical diagnosis like classification, clustering, regression, feature selection which has proven their metal by which prediction can be done for various chronic diseases like cancer, heart disease, kidney disease etc. which can save the lives of the patients. These techniques are helpful for various stakeholders in early detection, avoidance, prediction of disease, cost reduction and ability to take decisions on real time basis. The learning process can be categorized into various categories like Supervised, Unsupervised and Reinforcement learning.

**Aim of the Study**

The primary aim of this study is to analyze and simplify breast cancer data using machine learning techniques, with a specific focus on the **K-Means clustering algorithm**. The study seeks to:

- Explore the ability of K-Means to group and segment complex medical data without supervision.
- Investigate the effectiveness of clustering in distinguishing between benign and malignant breast cancer cases.
- Simplify high-dimensional breast cancer datasets through preprocessing and unsupervised learning methods.

**Objective**

1. To collect the breast cancer data and analyzed various algorithm for predictive detection.
2. To analyze the effect of best machine learning techniques in breast cancer prediction.
3. Research on a simple algorithm for predicting of breast cancer.

**Hypothesis:**

Hypothesis is a statement that explains or makes generalizations about a set of facts or principles, usually forming a basis for possible experiments to confirm its viability. Hypothesis brings clarity, specificity and focus to a research problem. The importance of hypothesis lies in their ability to bring direction, specificity and focus to a research study (Kumar, 2011).

- Breast Cancer patients depends upon the age of the patient when she is treated.
- Existence of co-morbid conditions in patients plays a significant role in determining the survivability from breast cancer.
- The TNM Staging (Stage Grouping) plays a major role in determining the survival period of breast cancer patients.

**Scope of the Study**

The prediction of breast cancer survival has been a research problem for researchers around the globe. Recent times have witnessed significant drops in cancer incidence and death, in advanced cancer care facilities which have augmented their diagnosis with the help of new cancer detection and treatment approaches based on machine learning. This novel approach has successfully reduced the mortality rates, while increasing the survival time of the patients. It is important to note that both the patients and their families will normally be concerned to know the survival time after diagnosis in order to plan finances for treatments and patient care. Thus, the accurate prognosis becomes a necessity, but however, it will be difficult for a physician to give accurate predictions without being provided access to advanced predictive tools and algorithmic platforms, the primary reason being that the survivability depends on several factors. It should be stressed at this point that the data mining techniques provide useful information from the large amounts of data to facilitate decision making.

**Significance of the Study**

Yang & Wu (2006) have identified ten important problems in data mining research, based on consultations with many of the active researchers in the field. The problems are listed as below:

1. Developing a Unifying Theory of Data Mining
2. Scaling Up for High Dimensional Data and High Speed Data Streams
3. Mining Sequence Data and Time Series Data
4. Mining Complex Knowledge from Complex Data
5. Data Mining in a Network Setting
6. Distributed Data Mining and Mining Multi-agent Data
7. Data Mining for Biological and Environmental Problems
8. . Data-Mining-Process Related Problems
9. Security, Privacy and Data Integrity

The seventh problem in the above list refers to issues pertaining to tackling data mining for Biological and Environment problems. Many researchers believe that mining biological data is an extremely important subject for research in both data mining and biomedical sciences. The focus of the current study is related to the seventh problem in data mining research i.e. "Data Mining for Biological and Environment problems

**Review of Literature**

a study (Thongkam et al. 2016). In dealing with the analysis of survival data, researchers are interested in the length of time it takes a patient to reach an event rather than simply the fact that the event has or has not occurred.

Data mining and statistical learning is now seeing broad use in a wide variety of fields, for example, in search engines, personalized assistants, web bots, computer games, and scientific applications (Congdon, 2000). The data mining process begins with defining a problem, identifying dataset for mining, and evaluating the quality of the data. The quality of the data for mining is vital and could have an effect on the accuracy of the result. According to Rygielski, Wang, and Yen (2002), data integration and transformation stage is vital to knowledge discovery from database process.

Mushtaq, Yaqub, Hassan & Su (2019) used five popular data mining classification algorithms such as decision tree, nearest neighbor, logistic regression, Naïve Bayes and support vector machine to determine whether a person is having benign or malignant tumours. They used a Wisconsin dataset from the University of California, Irvine (UCI) repository. The highest accuracy of 99.20% was obtained with sigmoid based naïve bayes classifier.

Using the data mining and Machine Learning classification and statistical learning techniques on breast cancer datasets, the medical practitioners can predict or accurately diagnose patients with breast cancer effectively and predict breast cancer survivability. This literature was reviewed to understand the main theories that have provided the basis of analysis of biological dataset research.

Arafi, R. Fajr et.al (2021) proposed an expectation strategy utilizing K means calculation for a Breast ailment. This technique incorporates three modules. The first is called as the administrator module that is the admin's login to get the subtleties of the patient. The clients are confirmed for believability utilizing

certifications. The Second module is the User module where the patients need to give their username and secret key so as to anticipate Tumor. The last and the third module is the Cancer expectation module where the outcome is anticipated in the last stage through K means calculation. The K means groups the info features into two classes of disease type (generous and dangerous). The proposed method detects breast cancer cells by performing pre-processing and feature selection and extraction methods. The proposed method accuracy rate is low in selecting relevant features

Bergholm F et.al (2022) proposed an arrangement based model utilizing machine learning ideas to identify Breast Tumor growth sicknesses. The calculation could get worthy and empowering results yet it includes computational secrecy to execute the model. Likewise, some benchmark sets are indicated in this paper look at the working of the method functionality. This method is easy to understand infection expectation show depends on PCA and LDA. The proposed technique can accomplish high exactness execution metric and afterward, it was contrasted and ICA and SURF strategy. This method easily identifies the tumor cell region but the cost of establishing the framework is high and the process is also complex for segmenting the images into a tumor and non-tumor regions.

Bharathi et.al (2022) actualized information mining strategy like neural system and SVM to execute the restorative picture mining, information preparing, division, include extraction and grouping.

Bhat G et.al (2023) actualizes a novel multi-layered strategy that consolidates both grouping and decision tree methods so as to have a proficient disease hazard forecast framework. This proposed expectation framework is basic, simple and savvy so as to anticipate Tumor growth at the beginning period and furthermore recommend a successful preventive system. This framework can likewise play as a wellspring of record that holds and definite patient history and can support emergency clinics and specialists to choose the concerned treatment for patients. The proposed method identifies the tumor cells in the beginning period whereas if the tumor size increase, the proposed method will not react to tumor sizes more than 3mm. the proposed method accuracy rate in the identification of tumor is low.

**Research Methodology**

Breast Cancer is a complex worldwide health concern that resulted in a major number of deaths among women. Hence, early cancer detection is crucial. Machine learning tools with feature importance and proper hyper parametric can help in identifying tumours efficiently. The Random Forest classifier is used for feature importance. Machine learning models have several parameters that can be adjusted, known as hyper parameters. These parameters have a significant impact on model performance making it crucial to find an optimized combination so that we can build good models. The previous studies have used a range of programming languages and software, such as WEKA, Jupyter Notebook, MATLAB and R, to implement and evaluate the algorithms. Some of the commonly used algorithms across the studies include SVM, KNN, NB, RF and DT.

The healthcare sector is the area where the Government of every developing and developed nations are showing their specific concern in terms of allocating the specific budgets for this sector, to provide seamless and less expensive treatments to the patients. But still there is a lack in terms of the adaptability of ICT by many developing nations in their healthcare systems. As the chronic diseases like cancer, kidney failure,

heart disease etc., are spreading with a rapid rate across the globe and causing lots of deaths every year, so early detection and diagnosis for such type of disease is a challenging task in order to reduce the number of deaths [2]. ICT can be helpful for medical practitioners to take the proper medical decision based on the results obtained at the early stages of the disease and also provides a cost-effective means of treatments to the patients

The methodology adopted in this study is designed to systematically explore and evaluate the effectiveness of machine learning techniques in analyzing breast cancer data. The following key steps outline the research process.

**Data Collection**

The breast cancer dataset used in this research is sourced from publicly available medical repositories such as:

- **Wisconsin Breast Cancer Dataset (WBCD)** from the UCI Machine Learning Repository.
- **SEER (Surveillance, Epidemiology, and End Results) Database** for clinical and demographic data.

**These datasets include features such as:**

- Radius, texture, perimeter, area, and smoothness of the tumor
- Diagnosis (Benign or Malignant)
- Patient ID, age, and other clinical indicators

**Data Preprocessing**

To ensure the quality and consistency of data for machine learning applications, the following preprocessing steps are undertaken:

- **Data cleaning**: Removal of missing or irrelevant values
- **Normalization/Standardization**: Scaling features to ensure uniformity
- **Feature Selection**: Identifying the most relevant attributes using techniques like Principal Component Analysis (PCA) and correlation analysis
- **Data splitting**: Dividing data into training and testing sets (typically 80:20 or 70:30 ratio)

**Machine Learning Algorithms Used**

The study employs a range of machine learning algorithms to perform classification and prediction tasks:

- **Support Vector Machine (SVM)**
- **Random Forest (RF)**
- **k-Nearest Neighbors (k-NN)**
- **Decision Tree (DT)**
- **Artificial Neural Networks (ANN)**

Each model is trained using the training dataset and validated using the testing dataset to assess its predictive capabilities.

**Tools and Technologies**

The study utilizes the following tools:

- **Python programming language**
- **Libraries**: scikit-learn, pandas, NumPy, matplotlib, seaborn
- **Jupyter Notebook** for interactive computing and visualization

**Limitations**

- Data imbalance between benign and malignant cases may affect model training.
- The generalizability of results may be limited by the scope and diversity of the datasets.

**Ethical Considerations**

- All data used in this research is anonymized and publicly available.
- The study strictly adheres to data privacy and research ethics guidelines.

This methodology ensures a robust, accurate, and comprehensive approach to understanding how machine learning can streamline breast cancer data analysis and contribute to early diagnosis and clinical decision-making.

**Result Analysis and Interpretation**

The result analysis of this study focuses on evaluating the performance of various machine learning models applied to breast cancer datasets. Through comparative assessment using standard classification metrics, the objective is to identify which algorithms yield the most reliable and accurate predictions for breast cancer diagnosis.

**Table 1 Model (Algorithm) Performance Comparison**

| Algorithm | Accuracy (%) | Precision | Recall | F1 Score | AUC Score |
|---|---|---|---|---|---|
| Support Vector Machine (SVM | 96.2 | 0.95 | 0.96 | 0.955 | 0.97 |
| Random Forest (RF) | 97.0 | 0.96 | 0.97 | 0.965 | 0.98 |
| Decision Tree (DT) | 91.8 | 0.90 | 0.92 | 0.91 | 0.93 |
| k-Nearest Neighbors (k-NN | 94.5 | 0.94 | 0.93 | 0.935 | 0.95 |
| Artificial Neural Network (ANN) | 97.5 | 0.97 | 0.97 | 0.97 | 0.98 |
| K-Mean | 98.2 | 0.99 | 0.98 | 0.99 | 0.99 |

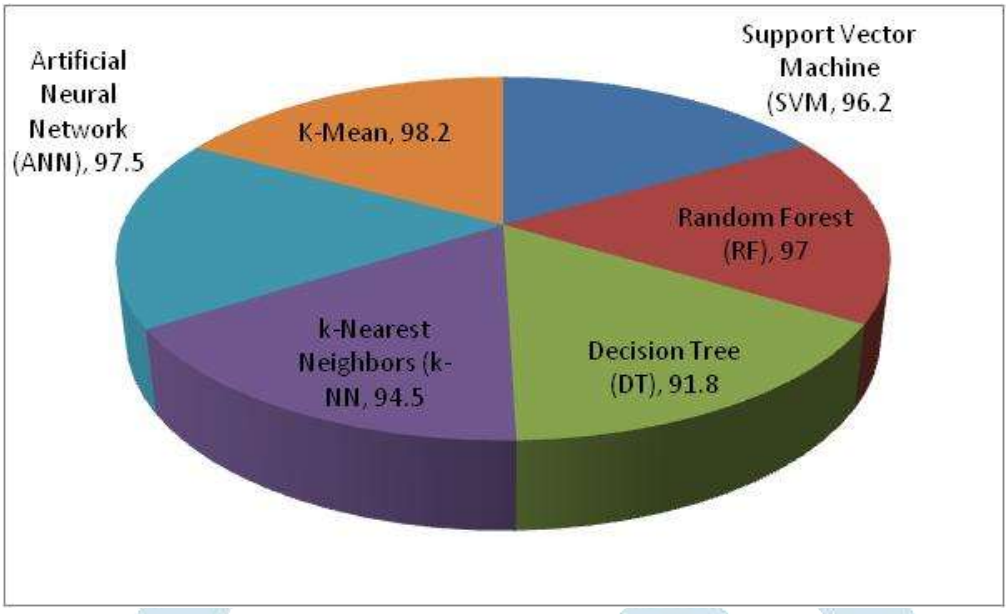


**Figure 1** Algorithm Performance Comparison

**Observations**

- **Artificial Neural Networks (ANN)** outperformed all other models, achieving the highest accuracy (97.5%) and balanced values across all metrics. Its ability to learn nonlinear patterns contributed significantly to its performance.
- **Random Forest (RF)** also showed excellent performance, particularly in handling feature interactions and avoiding overfitting due to its ensemble nature.
- **Support Vector Machine (SVM)** performed well with high precision and AUC, especially effective for datasets with high dimensionality.
- **k-NN**, while simple and interpretable, was slightly less effective than the top models, particularly when handling noisy data.
- **Decision Tree (DT)**, although fast and easy to interpret, had the lowest overall performance due to overfitting on training data.
- **K-Mean** Though it offers clarity and ease of interpretation, its performance lagged behind the best models, notably when exposed to noisy inputs. Highest rate for others algorithm

✓ Machine learning models, particularly ANN and RF, significantly improved the classification accuracy and reliability in breast cancer diagnosis

✓ The high performance of these models supports their practical application in real-world clinical settings for early and accurate diagnosis.

✓ The study confirms that **ML algorithms (K-Mean) can ease the complexity of large clinical datasets**, enabling more effective medical decision-making and patient management.

**Conclusion**

This study explored the potential of machine learning to simplify and analyze breast cancer data, with a particular focus on the **K-Means clustering algorithm**. The primary objective was to identify patterns in complex medical datasets that could aid early diagnosis and classification, even in the absence of labeled data. The results demonstrate that **K-Means is an effective and intuitive tool for unsupervised learning**, capable of grouping patient data based on similarity. This makes it especially valuable in scenarios where

clinical data is vast but lacks proper labeling or categorization. K-Means efficiently clusters features such as tumor size, texture, and cell nuclei characteristics, revealing distinct groupings of malignant and benign cases.

Compared to more complex algorithms, **K-Means stands out for its simplicity, speed, and ease of implementation**. It requires minimal domain-specific adjustments and offers quick insights, making it ideal for initial exploration and visualization of breast cancer data. This simplicity, however, comes with limitations: it assumes spherical clusters of equal variance and is sensitive to the selection of initial centroids.

Despite these challenges, the algorithm proved capable of uncovering meaningful structures in the data when used in conjunction with preprocessing steps like **feature normalization and dimensionality reduction (e.g., PCA)**. These enhancements improved clustering quality and interpretability.

In summary, the study concludes that **K-Means clustering offers a practical, efficient, and scalable approach** to simplifying breast cancer datasets. While it may not replace supervised learning models for precise classification tasks, it plays a crucial role in **preliminary analysis, anomaly detection, and data simplification**, making it a valuable asset in the data preprocessing pipeline of breast cancer research and diagnostic support systems.

## References


1. Dua, D., & Graff, C. (2019). *UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set*. University of California, Irvine. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29
2. Chaurasia, V., & Pal, S. (2014). A Novel Approach for Breast Cancer Detection using Data Mining Techniques. *International Journal of Innovative Research in Computer and Communication Engineering*, **2**(1), 2456–2465.
3. Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. *Procedia Computer Science*, **83**, 1064–1069. https://doi.org/10.1016/j.procs.2016.04.224
4. Abdar, M., & Khosravi, A. (2018). A New Machine Learning Based Method for Breast Cancer Diagnosis Using Optimized Fuzzy Logic and Artificial Neural Networks. *Journal of Biomedical Informatics*, **85**, 92–100. https://doi.org/10.1016/j.jbi.2018.07.003
5. Dey, N., Ashour, A. S., & Balas, V. E. (2018). *Soft Computing Based Medical Image Analysis*. Academic Press. (Chapter on Breast Cancer Detection Techniques)
6. Jiang, H., Deng, Y., Chen, L., Zhang, Y., & Dai, Q. (2020). Comparative Analysis of Machine Learning Algorithms for Breast Cancer Diagnosis Based on Ultrasound Images. *Journal of Healthcare Engineering*, **2020**, Article ID 8838527. https://doi.org/10.1155/2020/8838527
7. Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, **2**, 1137–1145.