Literature Review

Breast cancer is the second most deadly cancer among women globally. According to the World Health Organization (WHO), 2.1 million new breast cancer cases were diagnosed in 2018, with an estimated 627,000 deaths. Early detection through tests such as mammograms, MRIs, and biopsies can significantly improve survival rates. The classification of tumours as benign or malignant is crucial, as benign tumours are non-cancerous, and malignant tumours are cancerous. Machine learning can aid in early diagnosis and classification, providing an automated, efficient solution for breast cancer detection. This study aims to build a prediction system for breast cancer using the Wisconsin breast cancer dataset and various supervised machine learning algorithms.

Previous research in breast cancer prediction has applied several machine learning techniques. For example, Vikas Chaurasia et al. compared Naive Bayes, RBF Network, and J48, finding Naive Bayes to be the most accurate with a classification accuracy of 97.36%. Other studies have applied algorithms like Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Decision Trees to improve prediction accuracy, with varying success rates.

he study employed five supervised machine learning algorithms for breast cancer classification:

1. **Logistic Regression**
2. **Decision Tree Classifier**
3. **Gaussian Naive Bayes**
4. **K-Nearest Neighbors**
5. **Support Vector Machine**

The dataset was preprocessed by handling missing values, encoding categorical data, normalising features, and applying Principal Component Analysis (PCA) for feature selection. The dataset was split into training (80%) and testing (20%) subsets.

The results showed that **Logistic Regression** performed the best, with the highest accuracy, precision, recall, and F-score. It was followed by

Support Vector Machine and **Decision Tree** in terms of accuracy. **K-Nearest Neighbors** and **Gaussian Naive Bayes** showed lower performance, with accuracy scores of 92.10% and 93.85%, respectively.

This study demonstrates that Logistic Regression is the most effective model for breast cancer prediction, outperforming other machine learning algorithms like Decision Trees, SVM, KNN, and Naive Bayes. The proposed methodology provides a robust approach to classifying breast cancer as benign or malignant, enabling early detection and treatment, thus improving patient outcomes.