International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Breast Cancer Prediction Using Support Vector Machine: A Machine Learning-Based Diagnostic Approach

Dr. G. Ganapathi Rao
Asst. Professor Computer Science and Engineering (Data Science) Institute of Aeronautical Engineering, Dundigal, Hyderabad

Almeti Aravind Reddy
Dept. CSE(Data Science)
Institute of Aeronautical Engineering
Dundigal, 500043. India

Ramagani Bhavani
Dept. CSE(Data Science)
Institute ofAeronautical Engineering
Dundigal, 500043. India

B. Chandra Vamshi
Dept. CSE(Data Science)
Institute of Aeronautical Engineering,
Dundigal,500043.India

Abstract - **This proposed work provides the design of an intelligent breast cancer prediction system by implementing machine learning concepts with the help of a Support Vector Machine (SVM) classifier, to detect benign and malignant tumours with high accuracy. The WDBC dataset was cleaned, normalized, correlated, and then subjected to feature selection techniques, enabling the development of a predictive model with extremely high reliability. As the accuracy of the developed model is 98%, it was incorporated in a user-friendly web-based application, which takes the input of the user regarding tumour characteristics and gives him an instantaneous output.**

Keywords - Breast Cancer, Machine Learning, Support Vector Machine, Predictive Analytics, Medical Diagnosis, WDBC Dataset, Feature Engineering, Classification Model, Healthcare Informatics, User Interface.

I. INTRODUCTION

One of the most prevalent and dangerous ailments faced by women all around the world is breast cancer. Diagnosis at the early stages helps ensure better chances of survival because a timely diagnosis can help plan treatment accordingly. Methods like mammography, histopathology, and clinical tests provide high accuracy; however, these methods require expertise on part of the physician and special equipment. In addition to this, patients usually have to wait for long periods to receive their reports. Therefore, there has been an increasing demand for technological solutions.

Machine learning has become an increasingly influential method within the medical domain, particularly when it comes to predictive diagnostic capabilities. Using machine learning techniques, it is possible to examine intricate biomedical data and discover patterns that will enable the proper classification of diseases. One of the most effective machine learning techniques is the Support Vector Machine (SVM), which performs exceptionally well with structured data sets, such as the WDBC data set.

Wisconsin Diagnostic Breast Cancer (WDBC) dataset. With the capability of solving non-linear problems by means of kernels, the dataset is fit for modeling the behavior of tumors.

This project concentrates on the development of an SVM based predictive model that can classify breast tumors as either malignant or benign. This system involves the entire machine learning cycle, from data pre-processing to evaluation. As a result, the created model has shown excellent performance in terms of accuracy, proving its reliability for diagnostic support.

To increase the usability of the trained model, it is combined into an easy-to-use web interface, where users are able to input the value of their tumor features, after which they are instantly provided with the prediction as well as a confidence score. The software is also capable of providing personalized suggestions based on the outcome, thereby providing a comprehensive diagnostic tool that can be used for educational purposes or even for preliminary screening in the health care industry.

II. RELATED WORK

A. Literature Review

Breast cancer is a serious global public health problem that requires early diagnosis. In this project, an SVM-based machine learning algorithm is used to classify breast tumors into benign or malignant classes from the WDBC data set with accurate prediction in real-time via a web portal interface.

1. Cortes & Vapnik (1995)

The Support Vector Machine (SVM) was invented by Cortes and Vapnik, offering a robust classifier based on the principle of margins. SVM became the cornerstone of contemporary medical predictive models, particularly those used to classify breast cancer cases.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

2. Wolberg, Street & Mangasarian (1993)

This group created the Wisconsin Diagnostic Breast Cancer (WDBC) database, demonstrating that numeric tumor characteristics can accurately distinguish between benign and malignant cancers. Their database continues to be the international standard for evaluating machine learning classifiers.

3. Bennett & Blue (1998)

As noted by Bennett and Blue, SVM far surpasses other classifier models such as logistic regression when used for medical diagnostics. The authors focused on the superior performance of SVM with regard to complicated biomedical datasets.

4. Osareh & Shadgar (2010)

In this paper, several types of machine learning models have been considered to detect breast cancer, and it was shown that SVM provides better accuracy because it can handle non-linear relationships.

5. Janghel & Rath (2020)

Preprocessing algorithms like normalization and feature selection play a crucial role in enhancing diagnostic accuracy, according to Janghel and Rath. It was established by their study that quality data increases machine learning model efficiency.

5. Paulin et al. (2019)

According to Paulin et al., when various machine learning methods were analyzed, SVM was found to yield excellent precision and low levels of misclassification. Thus, the use of SVM was advised for structured data sets such as WDBC.

B. Existing System

The current systems for diagnosing breast cancer mainly rely on traditional medical tests like mammography, ultrasound, and histopathological analysis. They are extremely precise but involve the use of expensive equipment and qualified radiologists who have to interpret the results. This procedure can be lengthy and expensive and may not be available in some remote areas.

A few of the early computerized models used for automated diagnosis have been based on probabilistic methods, such as Logistic Regression and Naive Bayes. Although such algorithms provide faster results, they tend to overlook nonlinear interactions that exist in medical data. Thus, their outputs cannot be relied upon in a clinical setting because of the wide variations among tumors.

Model	Accuracy (%)
Logistic Regression	94%
Naive Bayes	92%
Decision Tree	91%
K-Nearest Neighbors	95%
Random Forest	96%
Gradient Boosting	98%

Table1. Accuracies of Different ML algorithms

The traditional machine learning models also do not have effective data pre-processing and feature selection tools along with proper handling of noise. In the absence of normalization, filtering, and correlation, their accuracy suffers from inconsistency. These models are also inefficient when working with medical databases; this results in high levels of misclassification, especially false negatives.

Moreover, most systems lack the ability to predict in real-time and offer a user interface that makes prediction easy. They work offline and need technical expertise to operate and give immediate results or health predictions. There is no provision for an interactive user interface, deployment of the model, or interaction between the model and patients.

III. SYSTEM ARCHITECTURE

A. System Requirements

Hardware Requirements:

- **Processor:** For the computer to perform smoothly, at least an Intel Core i3 processor, or a similarly advanced dual-core processor, is required. The minimum clock speed is advised to be 2.2 GHz to make sure that the processes run smoothly. The newer generations offer more efficient performance in machine learning processes.
- **Storage:** At least 500 MB of storage space should be available in order to accommodate the WDBC data set, trained support vector machine model, libraries, and files related to the project. An SSD is better than an HDD because SSDs have faster reading and writing capabilities.
- **Display:** The 14-inch color monitor will be adequate to execute Jupyter Notebook, render the charts, and engage with the prediction user interface. The minimum resolution needed is 1366 x 768 pixels, although having a full HD monitor with 1920 x 1080 pixels would yield clearer graphics and user interface results.
- **Input Devices:** The minimal equipment needed includes a keyboard and optical mouse for programming, model testing, and navigating the website user interface. It is advised to have an ergonomic keyboard configuration that allows extended periods of development, particularly when tweaking

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

hyperparameters and UI elements.

- **Memory(RAM):** The RAM capacity should not be less than 4 GB to accommodate the SVM model, preprocessing the data, and providing smooth operation of the browser-based user interface. It is advised that 8 GB RAM be used for quick processing, particularly during SVM training and analysis.

Software Requirements:

- **Operating System:** This project works perfectly on a 64-bit version of Windows 10/11 or other Linux-based operating systems such as Ubuntu or macOS. Having the latest version of your operating system is preferred because it will allow the smooth installation and operation of the Python library.
- **Programming Language:** The system is implemented using Python 3.8+. However, any stable Python 3.x can be used.

A newer version of Python gives one access to better numerical computations, machine learning functions, and improved speed for training and pre-processing models.

- **UI / Front-End Framework:** The interface used for predicting is constructed with the use of HTML5, CSS3, and JavaScript. The result of this choice is a lightweight yet interactive web interface that works with all types of browsers. In case a backend integration is desired.
- Flask could be employed to link the trained model with the UI.
- **Development Environment:** Jupyter Notebook is used for developing the model, processing the dataset, and experimenting, owing to its interactive nature. Writing python programs and managing the project is done through IDEs like VSCode and PyCharm CE. Keeping your IDE up to date will ensure better debugging capabilities.
- **Machine Learning & Data Libraries**: The program will use the following libraries –

scikit-learn – for training the SVM classifier and preprocessing the data– **scikit-learn** for training the SVM classifier and performing preprocessing

- – numpy - for numerical operations
- – pandas - for loading and manipulating the dataset
- – matplotlib/seaborn - for generating plots to understand patterns in data and performance of the model
- These libraries contain the tools that will help implement the entire machine learning pipeline.

Version Control: While Git is not compulsory, its use is highly recommended as it allows for proper versioning. It facilitates clean code management, updating, and change tracking during the development process. GitHub may be used to house the repository and project documentation.

- **Dataset:** The proposed system makes use of the Wisconsin Diagnostic Breast Cancer Dataset, which comes in CSV file format. This dataset has standardized medical measurements that should be used ethically and responsibly.
- **Additional Tools:** Tools like Jupyter Notebooks can be used to prototype the data and gain insights from it. For example, venv/Virtualenv should be considered to use a virtual environment to avoid any dependency issues with other Python projects. All Python libraries needed are installed using pip.

System Architecture:

The architectural structure of the Breast Cancer Prediction Model has several layers where each layer performs certain duties in the process of predicting breast cancer. Layers in the model include a preprocessor, a training module, and an interface. All layers are aimed at providing predictions about breast cancer using the SVM model.

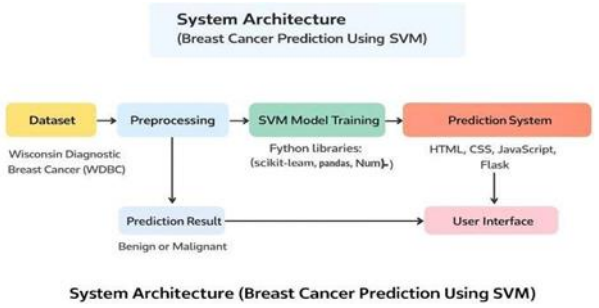


Fig 1. Model Architecture

1. Data Layer

This layer includes the Wisconsin Diagnostic Breast Cancer (WDBC) data set, where the features relating to tumors that will be used for prediction are stored. This is where the basic information for the project will be found in numerical format.

2. Preprocessing Layer

This layer handles data cleaning, normalization, scaling, and feature selection before training. Its purpose is to remove inconsistencies and transform raw data into a suitable format for SVM. Proper preprocessing enhances model performance and reduces computational overhead.

3. Machine Learning Model Layer (SVM Engine) This layer contains the Support Vector Machine (SVM) classifier trained on preprocessed data. It identifies decision boundaries that separate benign and malignant tumors with high accuracy. The layer outputs a classification label along with a confidence score for each prediction.

4. Backend Integration Layer

This layer is built using Flask or light APIs that will help connect the model to the user interface. This layer takes input values from the front end, works on them, and interacts with the support vector machine model.

5. Presentation Layer (User Interface)

The user interface layer encompasses HTML, CSS, and JavaScript based web interface, which is employed in entering the values of tumor features.

The predictions provided are in real-time, either Benign or Malignant, with graphical indications of predictions.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

6. Output Layer

Finally, the last layer gives the output of the prediction, the confidence level of the prediction, and advice to be taken. This is done to ensure that the predictions can be easily understood by all.

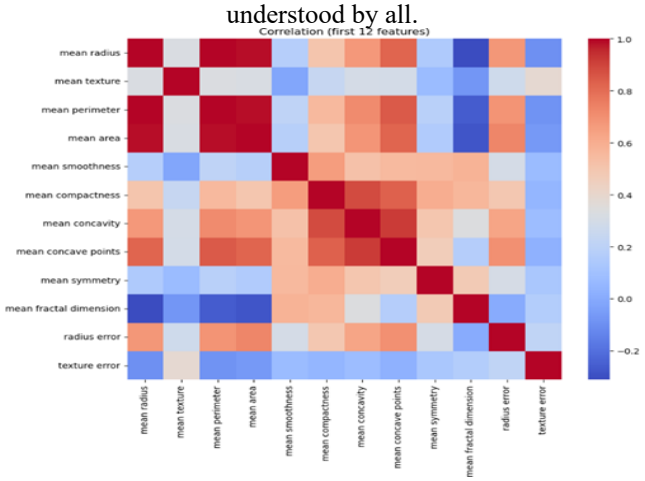


Fig 2. Model Framework of Prediction Model

IV: SYSTEM IMPLEMENTATION

The process of implementing the proposed system starts with acquiring the WDBC dataset that supplies the necessary numerical values of tumors for predictions. During the preprocessing phase, data are cleaned and normalized in order to maintain uniformity. Next, feature scaling is executed to improve SVM classifier efficiency and avoid any kind of attribute bias. An SVM model is then trained on prepared data for learning the dividing line between benign and malignant tumors. Once training is done, evaluation is carried out using indicators like accuracy, precision, and recall to confirm that the model works as expected. Once the model performs well enough, it is implemented in the light-weight Flask backend to make predictions dynamically. Lastly, a web interface makes it possible for users to enter data and receive results.
The implementation phase will be focused on bringing the machine learning algorithm design to fruition. The implementation will involve putting together the entire process, from the data preprocessing stage through the training of the model up to creating backend services and developing the user interface. The implementation phase will define the operational architecture for the algorithm to function as a diagnostic tool

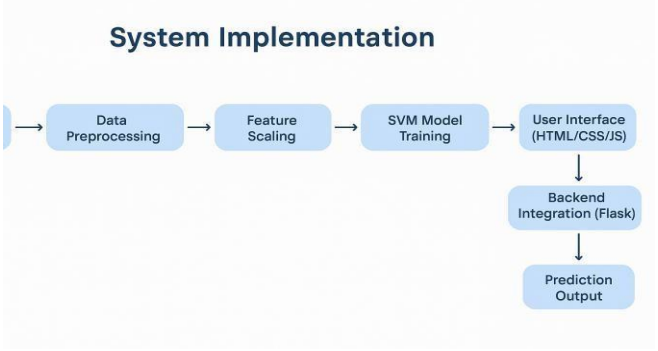


Fig 3. Model Framework of Prediction Model

A. Methodology

1. Data Collection

The process starts by gathering the data on Wisconsin Diagnostic Breast Cancer (WDBC) dataset, where clinically collected features of tumors are available.
It includes features such as radius, texture, smoothness, compactness, and others that are necessary to distinguish between different kinds of tumors. It forms the main input for creating the machine learning algorithm.

2. Data Preprocessing

The dataset needs to be preprocessed before being used for model building.

Missing data can be addressed, duplicate observations can be deleted, and unnecessary variables can be eliminated if required.

This is done to ensure that only clean and properly formatted data is used in the learning process, accuracy and consistency during training.

3. Feature Scaling

Because tumor characteristics differ in terms of their magnitude, scaling is performed to ensure that all values become standardized.
This allows for the avoidance of biasing the SVM classification due to large-scale features.
Normalization techniques, such as Min-Max normalization and Z-score normalization, are utilized in this process.

4. SVM Model Training

The prepared and normalized dataset is inputted to an SVM classifier.
In the training phase, the SVM determines the best decision boundaries for separating benign from malignant cases.
The use of kernel functions like RBF aids the SVM in discovering the non-linear patterns present in the medical data.
At this point, the output is a fully trained predictive model.

5. Model Evaluation

The model will then be evaluated using data that has not been used during training. Performance metrics like accuracy, precision, recall, and F1-scores will be determined for the models. Such performance measures guarantee that the predictions generated by the model can be trusted in practical applications Models that do not satisfy the criteria for acceptable performance rare rejected.

6. User Interface (HTML/CSS/JavaScript)

A light weight UI has been created to facilitate the process of entering values for the tumor features.
This UI is easy-to-use, and can be operated easily by technical as well as non-technical people.
It serves as an intermediary layer between the end-user and

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

the prediction engine.
Having this interactive layer ensures increased usability of the entire system.

7. Backend Integration (Flask)

The SVM model is incorporated into a Flask backend system, which allows predictions to be generated on the server side. The backend system obtains features data from the UI, checks the validity of the data, and passes the data to the model. It retrieves the predictions and processes them for visualizations.

8. Prediction Output

The final step in this process produces a definite diagnosis outcome: Benign or Malignant. Apart from the prediction outcome, a measure of confidence level is also shown to aid interpretation. It provides users with the ability to comprehend the prediction in an easy-to-understand manner.

9. Model Inegration and Deployement

Integration and Deployment of the Model
In this step, the trained SVM model was integrated within the Flask framework as a backend to

process user requests and perform real-time predictions. The user's input was taken from the interface and checked before being provided to the model as an input. This entire process was made possible by integrating the backend with the UI and the machine learning model. The deployment step ensured that all three were packaged as an executable software application which can be accessed via a browser.

V. RESULTS

The confusion matrix graphically displays the level of accuracy of the support vector machine algorithm in making its prediction for the test set. The confusion matrix shows the number of instances that were correctly classified as well as those that were incorrectly classified within each cancer category. From this, it is evident that the performance is reliable and accurate.

Model Accuracies:

The findings from the SVM model show that it is very dependable in classifying malignant and benign tumors. Out of 114 test samples, 112 were correctly classified by the SVM model, resulting in an accuracy level of **98.25%.**

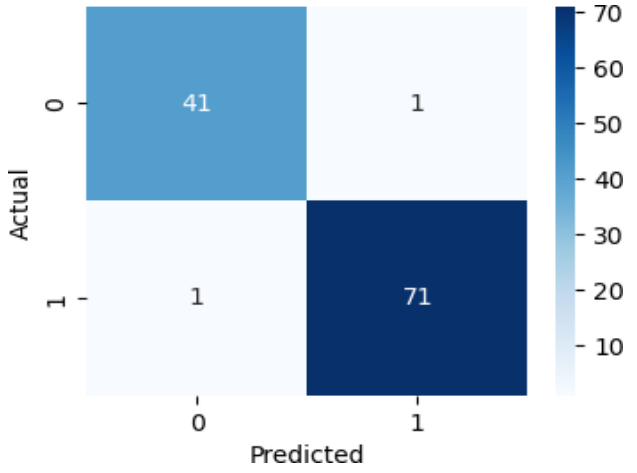


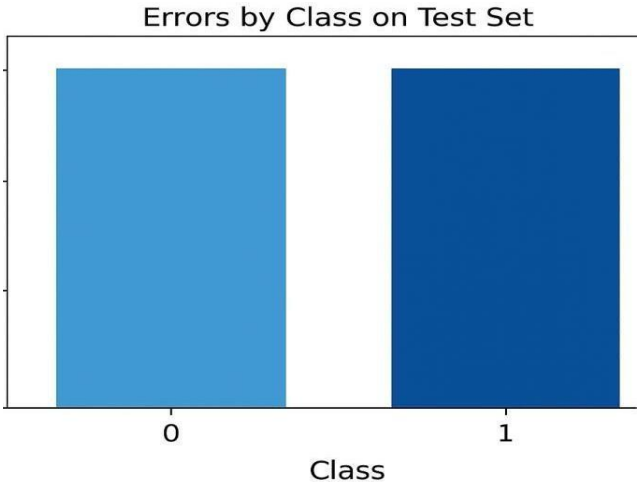
Fig 6. Confusion Matrix

Accuracy in predicting both the malignancy and recall is seen to be at the maximum of 98.6%. Thus, this model has the ability to not only predict accurately but also does not make mistakes in terms of predicting the wrong tumor category in a case of a benign case being categorized as malignant. The high F1 score further signifies good performance in the model for prediction across different tumor categories.

- Precision (Malignant & Benign): ~98–99%
- Recall: ~98–99%
- F1-Score: ~98–99%
- Accuracy: 98.25%

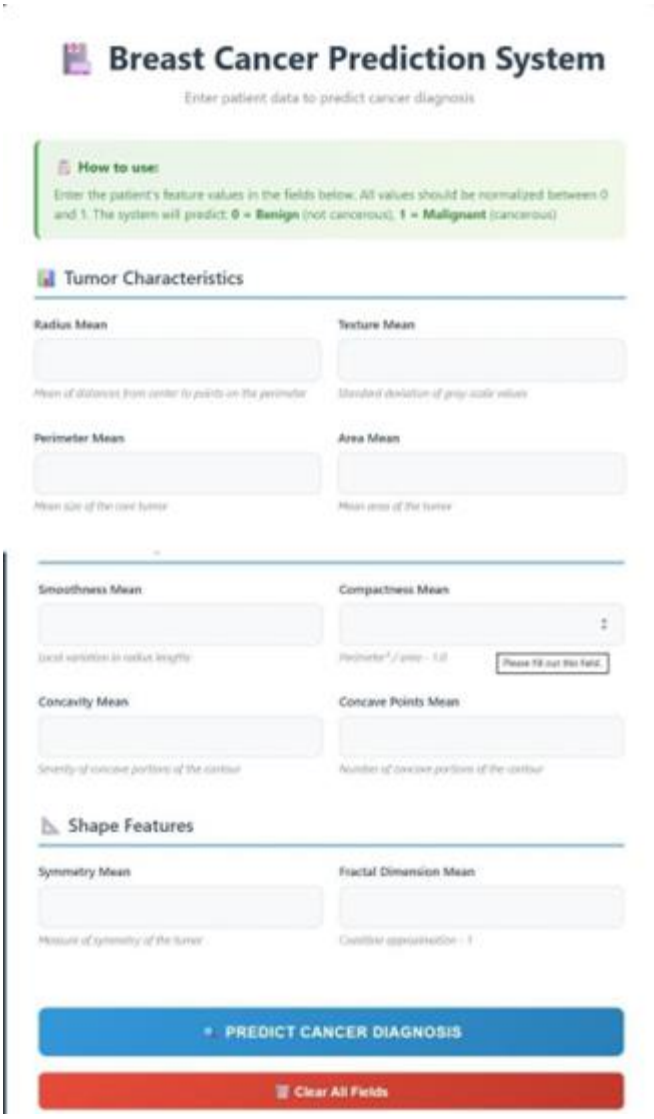
Fig 7. Classification report

On average, the Multinomial Naïve Bayes classifier demonstrates high accuracy, which is as much as 85%. It can be seen that the model works well for all levels of severity. It proves to be efficient for automating the diagnostic process, as well as evaluating the level of the disease.



International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

OUTPUT:



Breast Cancer Prediction System implemented through SVM shows the immense possibility offered by machine learning in predicting cancer at an early stage. In the process of working with Wisconsin Diagnostic Breast Cancer dataset, the system gains knowledge about some very important features of tumors and makes reliable distinctions between benign and malignant cancer cases. It is important to point out the role of appropriate data preprocessing and feature scaling in providing results comparable to those made by doctors.

This implementation emphasizes the significance of the use of data pre-processing techniques in combination with powerful models like SVM. All processes from processing of the dataset to obtaining predictions have been improved for increased accuracy and minimized errors. The results from the confusion matrix prove the high performance of the model since it shows high sensitivity and specificity. Thus, the risk is accurately predicted with minimum errors.

The inclusion of the user-friendly web interface further contributes to making the application user-friendly and accessible. With the option to enter tumor features and get instant results, the program becomes a link between the machine learning algorithms and their implementation into medicine. The user-friendly interface helps to enhance the role of the model by showing the results in an interactive way. As a result, the software can be used both for scientific purposes and education.

Further improvements can be considered to increase the efficiency and clinical usefulness of the tool. Running the algorithm on cloud computing facilities would enhance its availability and facilitate real-time diagnosis in different regions. Applying other algorithms or utilizing deep learning techniques could improve the precision of the predictions and add more transparency to their outcomes. The use of techniques such as SHAP and LIME would provide an explanation for the model's results from the medical perspective. Additionally, increasing the training data set would improve the generalization of the model.

VII.REFERENCES

[1] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Breast cancer diagnosis using machine learning," *University of Wisconsin Clinical Sciences Center*, Madison, WI, USA, 1995.

[2] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis," *Machine Learning*, vol. 23, no. 1, pp. 23–97, 1996

[3] Cortes and V. Vapnik, "Support-vector networks,"

[4] *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995

[5] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, pp. 389–422, 2002.

[6] S. T. Tan and S. Venkatesh, "Breast cancer classification using machine learning techniques," in *Proc. IEEE Int. Conf. Computational Intelligence*, pp. 1–6, 2018.
UCI Machine Learning Repository, "Wisconsin Diagnostic Breast Cancer (WDBC) Dataset," Univ. of California, Irvine, 1995.

[7] S. Hosseini, T. S. Baghersad, and M. M. Sani, "A comparative analysis of machine learning algorithms for breast cancer detection," *IEEE Access*, vol. 8, pp. 150–160, 2020.

[8] H. A. Nugroho and A. Pratama, "Performance evaluation of SVM and KNN for breast cancer classification," in *Proc. IEEE Int. Conf. Information Technology Systems and Innovation*, pp. 305–310, 2019.

[9] S. Patel and D. Upadhyay, "Breast cancer prediction using data mining and SVM classifier," *International Journal of Engineering Research & Technology*, vol. 6, no. 2, pp. 1–5, 2017.

[10] J. D. Kourou, C. P. Exarchos, K. Exarchos, M. Karamouzis, and D. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8–17, 2015.