International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Improving Breast Cancer Classification Using Support Vector Classifier with Grid search CV for Optimal Hyperparameter Tuning

Anusha Engle [1], Madaboni Chandrika [2], B. Sanjay Kumar [3], Mrs. M.Reddi Durgasree[4]

[1,2,3] *UG Scholars,* [4]*Assistant Professor*
[1,2,3,4] *Department of CSE[Artificial Intelligence & Machine Learning],*
[1,2,3,4] *Guru Nanak Institutions Technical Campus, Hyderabad, Telangana, India*

Abstract - Breast cancer continues to be a major global health challenge, particularly for women, where early and reliable detection significantly improves patient outcomes. This research introduces a refined machine learning framework utilizing a Support Vector Classifier (SVC) in conjunction with GridSearchCV to optimize its hyperparameters for binary classification of breast tumors. The proposed approach systematically evaluates combinations of parameters such as kernel type, regularization constant (C), and gamma value, using 5-fold cross-validation. The optimized model was applied to the Wisconsin Diagnostic Breast Cancer dataset and achieved notable improvements across various evaluation metrics, including accuracy, recall, precision, and F1-score. Results demonstrate the efficacy of hyperparameter tuning in enhancing predictive performance, validating the method's potential as a reliable diagnostic support tool for early cancer screening.

1. INTRODUCTION

Image translation is a crucial task in modern computer vision Among all forms of cancer affecting women, breast cancer remains one of the most commonly diagnosed and deadliest. Detecting malignancies at an early stage plays a pivotal role in improving survival rates and guiding effective treatment plans. Traditional diagnostic procedures often fall short in capturing complex, nonlinear patterns in medical datasets. To address this, machine learning (ML) methods have gained popularity due to their ability to model intricate relationships and deliver high accuracy in classification tasks.

Support Vector Machines (SVM), particularly the Support Vector Classifier (SVC), have proven effective for binary classification tasks, such as distinguishing between malignant and benign breast tumors.[1] However, the performance of an SVC is highly influenced by its hyperparameters, including the kernel function, penalty parameter (C), and gamma. Poorly selected parameters can lead to misclassification or limited generalization on new data. GridSearchCV, an exhaustive parameter optimization method, facilitates systematic exploration of hyperparameter combinations using cross-validation, thereby enhancing model robustness.

This study aims to develop a high-performing SVC model tailored for breast cancer classification by integrating GridSearchCV for optimal parameter selection.[2] Using the WDBC dataset, the paper evaluates the tuned model's ability to accurately predict tumor classes and discusses the implications of tuning in medical diagnostic contexts.

2. LITERATURE SURVEY

The integration of hyperparameter optimization with SVC has been widely studied in the context of breast cancer diagnosis. Aggarwal and Sharma (2024) demonstrated that applying GridSearchCV significantly boosts SVC performance across various breast cancer datasets, showcasing improvements in predictive accuracy. Deshwal and Sharma (2019) conducted comparative analysis before and after parameter tuning using GridSearchCV, and observed considerable gains in classification accuracy and overall model reliability.Additional research by Mat Radzi et al. (2021) compared traditional grid search methods with automated machine learning tools like TPOT. Their findings highlighted that while automated tools can streamline model development, GridSearchCV remains competitive, especially for support vector-based models. Chaudhury et al.

(2020) explored the combined impact of Principal Component Analysis (PCA) and hyperparameter tuning. Their approach led to a significant increase in classification accuracy, reinforcing the value of pre-processing and optimization techniques when working with high-dimensional medical data.

Other studies have proposed ensemble-based solutions, but these often introduce added complexity and computational costs. By contrast, a well-tuned SVC provides a simpler yet highly effective alternative with strong generalization capabilities.

3. PROBLEM STATEMENT

Detecting breast cancer at an early stage is vital for increasing survival rates and improving the effectiveness of treatment. However, traditional medical screening methods often lack consistency and may struggle with complex patterns in diagnostic data. While machine learning algorithms—especially Support Vector Classifiers (SVCs)—have shown promise in automating tumor classification, their accuracy depends heavily on choosing the right combination of model parameters.[3]

If the values for key hyperparameters such as the kernel type, regularization strength (C), or kernel coefficient (gamma) are selected poorly, the model may perform well during training but fail to generalize to new, unseen cases. Manually tuning these parameters is inefficient and error-prone, which makes the process unreliable, especially for medical applications where accuracy is critical.

This research addresses that challenge by introducing an automated method for selecting optimal parameters using GridSearchCV. By combining this technique with the SVC model, the aim is to build a classification system that can reliably distinguish between benign and malignant breast tumors.[4] The ultimate goal is to support clinical decision-making by improving both the accuracy and dependability of cancer detection.

4. PROPOSED METHODOLOGY

The methodology adopted in this study follows a structured approach designed to maximize the predictive performance of a Support Vector Classifier for breast cancer classification. The steps involved are detailed as follows:

4.1 MODULES

1. Dataset Preparation

The Wisconsin Diagnostic Breast Cancer dataset is used as the basis for training the model. Before modeling, the dataset is cleaned and standardized. All input features are scaled to ensure uniform influence, and the data is split into training and testing sets to allow for fair performance evaluation.

2. Model Configuration

A Support Vector Classifier is initialized, allowing flexibility in choosing different kernels. The primary goal at this stage is to configure the base model and define the search space for tuning, including values for the kernel function, regularization constant, and gamma coefficient.

3. Hyperparameter Tuning

GridSearchCV is used to automate the selection of the best hyperparameter values. This technique systematically tests combinations of parameters through a process called cross-validation. Each configuration is evaluated across multiple folds of the training data to prevent bias and overfitting.

4. Model Training

After selecting the best-performing set of hyperparameters, the model is retrained on the full training data. This ensures that the classifier benefits from all available learning data before final evaluation.

5. Evaluation Metrics

The model's effectiveness is measured using key classification metrics: accuracy, recall, precision, and the F1 score. These metrics help determine how well the model differentiates between malignant and benign tumors. A confusion matrix is also used to provide a detailed view of the classification results.

6. Baseline Comparison

For reference, the tuned model's performance is compared against a default SVC with no tuning. This highlights the improvement made possible through GridSearchCV and validates the use of hyperparameter optimization in sensitive diagnostic applications.

4.1.2 SYSTEM ARCHITECTURE :

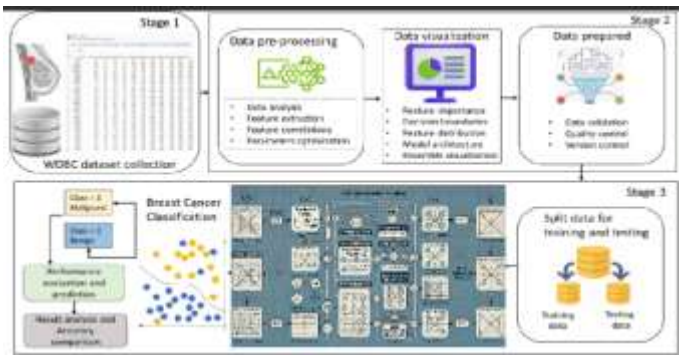

Figure 1: Represents the architecture of breast cancer.

The proposed architecture for breast cancer classification is designed as a modular machine learning pipeline that systematically progresses through data preprocessing, model training, hyperparameter tuning, and evaluation.[5] It combines the interpretability of traditional machine learning with the precision enabled by modern optimization techniques.

1. Data Acquisition and Preprocessing Layer

This component is responsible for loading the diagnostic dataset—such as the Wisconsin Breast Cancer Diagnostic (WBCD) dataset—and performing essential data transformations. Key operations include:

Handling missing values

Label encoding of categorical features

Feature scaling using techniques like StandardScaler to normalize data for SVC, which is sensitive to feature magnitude

2. Model Construction Layer

This stage initializes the Support Vector Classifier (SVC), specifying a range of hyperparameters to be optimized:

Kernel types: linear, polynomial, RBF

Regularization parameter (C)

Gamma (for non-linear kernels)

The SVC acts as the core predictive engine of the architecture.

3. Hyperparameter Tuning Layer

This module integrates GridSearchCV, a cross-validation-based exhaustive search mechanism. It systematically explores the defined parameter grid and evaluates model configurations using k-fold cross-validation (commonly k=5), ensuring optimal generalization.[6]

4. Evaluation and Performance Analysis Layer

Once the best parameters are identified, the final model is tested on unseen data using:

Accuracy

Precision, Recall, F1-score

Confusion Matrix

ROC-AUC Curve

These metrics provide insights into both the correctness and clinical relevance of the model's predictions.

5. Deployment Layer

The optimized SVC model can be serialized using tools like joblib or pickle for deployment.[7]

This model may be integrated into a web-based diagnostic tool using frameworks such as Flask or Streamlit for real-time tumor prediction and clinical decision support.

3.1.3. Algorithm

The algorithm designed for this project follows a structured approach to classify breast cancer tumors as either benign or malignant using a machine learning technique called Support Vector Classifier (SVC).

STEP 1:

The process begins by importing a reliable dataset known as the Wisconsin Diagnostic Breast Cancer dataset. This dataset includes various features measured from cell samples taken from breast tissue.

STEP 2:

Once the data is loaded, it goes through a preparation phase where all features are normalized. Normalization ensures that all input values fall within a similar range, helping the model perform

better. After preprocessing, the data is divided into two parts—one for training the model and the other for testing its accuracy.[8]

STEP 3:

Next, an SVC model is initialized. SVC is a type of classifier that works by finding a boundary (called a hyperplane) that best separates the two categories of data. However, the performance of this model depends on certain parameters like the type of kernel used, the regularization strength (C), and the gamma value.[9] To find the best values for these parameters, a technique called GridSearchCV is used. This method tests different combinations of parameters using cross-validation and selects the one that gives the best results.

STEP 4:

After tuning the model, the best version of the SVC is trained on the training data. Finally, the trained model is tested on the test data, and its performance is measured using metrics like accuracy, precision, recall, and F1-score.[10] These metrics show how well the model can identify cancer correctly and avoid wrong predictions. The result is a fine-tuned model that can assist in early breast cancer detection.[11]

4.1.4 RESULT

Figure 2: Representation of malignant cancer.

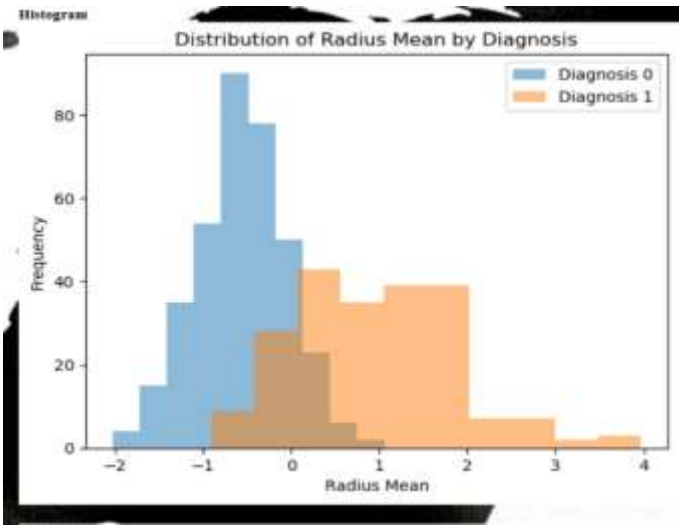

Figure 3 :Distribution of Radius Mean

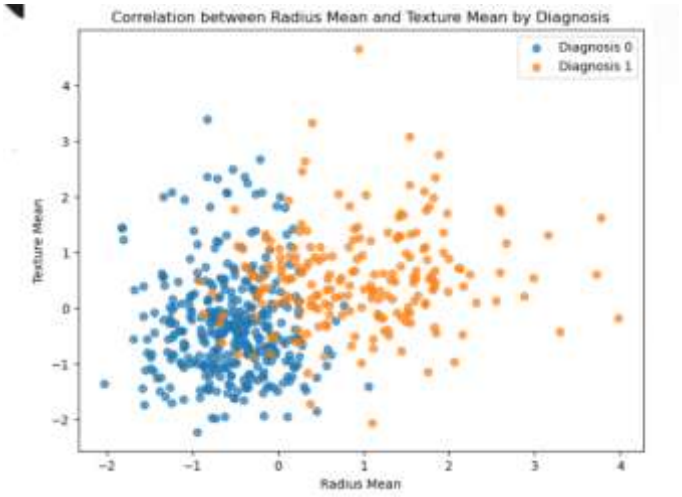

Figure 4 :Correlation between Radius Mean and Texture Mean

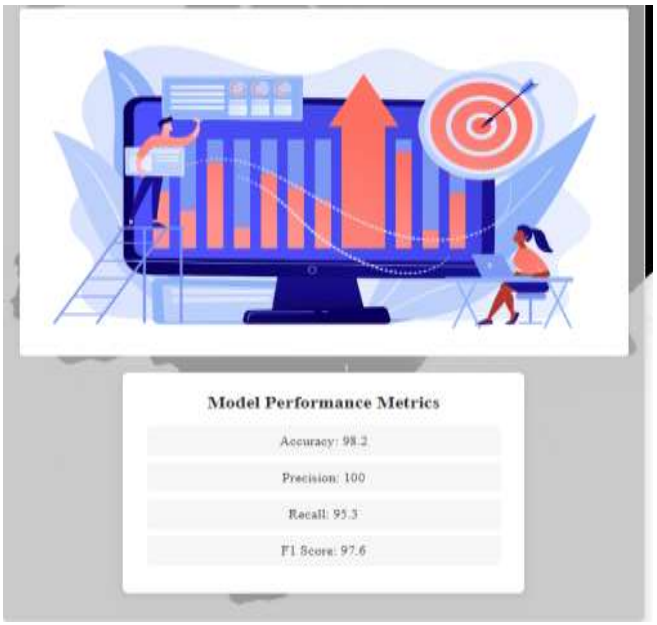

Figure 5 : Model Performance Metrics

4.2. TECHNIQUE USED OR ALGORITHM USED

4.2.1. SUPPORT VECTOR CLASSIFIER:

The proposed approach leverages a Support Vector Classifier (SVC) in combination with GridSearchCV to improve the accuracy and robustness of breast cancer diagnosis.[12] SVC is a supervised learning algorithm specifically designed for classification tasks, known for its ability to handle complex, high-dimensional data. In medical datasets such as the Wisconsin Breast Cancer Diagnostic dataset, where class boundaries may be non-linear, SVC offers a dependable solution when configured properly.

A key challenge in using SVC lies in the selection of its hyperparameters—such as the penalty parameter C, kernel function type, and the kernel coefficient gamma. Improper tuning of these values may result in suboptimal performance.[13] To address this, the proposed method employs GridSearchCV, a technique that systematically explores a predefined range of hyperparameter combinations. By performing cross-validation at each step, GridSearchCV identifies the configuration that yields the best performance, thereby minimizing risks of overfitting or underfitting.

The overall methodology includes the following stages:

Data Preparation:

The input dataset is cleaned and normalized using standard scaling techniques to ensure consistency across features.

Model Construction:

An SVC model is initialized, with kernel options (e.g., linear, RBF) and parameter grids defined for optimization.

Hyperparameter Tuning: GridSearchCV conducts an exhaustive search across the parameter grid, applying k-fold cross-validation to evaluate each configuration.

Model Evaluation:

The optimized SVC is tested on unseen data, and its performance is measured using accuracy, recall, precision, F1-score, and confusion matrix analysis.

This framework enhances diagnostic accuracy by using a structured and data-driven approach to model tuning. By combining the strength of SVC with the exhaustive search capabilities of GridSearchCV, the system achieves a high degree of precision in classifying breast tumors, making it a practical tool for aiding clinical decisions.

5. CONCLUSION

This work introduces a refined machine learning approach for breast cancer detection using Support Vector Classifiers (SVC) combined with GridSearchCV for hyperparameter optimization. By systematically tuning parameters such as the kernel type, regularization strength, and gamma value, the model achieves improved accuracy and robustness on clinical data like the Wisconsin Breast Cancer Diagnostic dataset. The process includes thorough data preprocessing, model training, and evaluation to ensure dependable classification of malignant and benign cases. Its modular structure allows for smooth integration into medical diagnostic systems and real-time applications. The study emphasizes the importance of fine-tuning in building effective clinical AI tools and suggests future directions, including advanced feature selection, personalized diagnostics through clinical and genetic inputs, and the adoption of explainable and ensemble-based models.

6. REFERENCES

[1] C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.

[2] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene selection for cancer classification using support vector machines,” Machine Learning, vol. 46, no. 1–3, pp. 389–422, 2002.

[3] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[4] V. Deshwal and M. Sharma, “Breast Cancer Detection using SVM Classifier with Grid Search

Technique," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 5, no. 3, 2019.

[5] S. F. M. Radzi et al., "Hyperparameter Tuning and Pipeline Optimization via Grid Search Method and Tree-Based AutoML in Breast Cancer Prediction," 2021.

[6] A. Aggarwal and A. Sharma, "Breast Cancer Prediction via Grid Search Hyperparameter Optimization," 2024.

[7] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," Department of Computer Science, National Taiwan University, Taipei, 2003.

[8] I. H. Witten, E. Frank, and M. A. Hall, "Data Mining: Practical Machine Learning Tools and Techniques," 3rd ed., Morgan Kaufmann, 2011.

[9] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," Artificial Intelligence in Medicine, vol. 34, no. 2, pp. 113–127, 2005.

[10] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Information Processing & Management, vol. 45, no. 4, pp. 427–437, 2009.

[11] R. Detrano et al., "International application of a new probability algorithm for the diagnosis of coronary artery disease," The American Journal of Cardiology, vol. 64, no. 5, pp. 304–310, 1989.

[12] W. H. Wolberg and O. L. Mangasarian, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology," Proceedings of the National Academy of Sciences, vol. 87, no. 23, pp. 9193–9196, 1990.

[13] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory (COLT), 1992, pp. 144–152.

[14] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[15] A. Srinivasu et al., "Classification of skin disease using deep learning and SVM," *Computers, Materials & Continua*, vol. 70, no. 2, pp. 2151–2167, 2022.