

IMPROVING BREAST CANCER CLASSIFICATION USING SUPPORT VECTOR CLASSIFIER WITH GRIDSEARCHCV FOR OPTIMAL HYPERPARAMETER TUNING

Sumayya Banu¹, Taleya Rahman², Uzma Begum³, Ms. Nishat Taleha⁴

^{1,2,3}UG Scholar, Dept of CSE, Shadan Women's College of Engineering and Technology, Hyderabad,

⁴Asst. Professor, Dept of CSE, Shadan Women's College of Engineering and Technology

ABSTRACT

Breast cancer is one of the leading causes of cancer-related deaths among women, making early and precise detection a critical challenge. Traditional methods for breast cancer classification often rely on static models that may not fully capture the complexities of the data. In this study, we propose an enhanced approach for breast cancer detection using a Support Vector Classifier (SVC), coupled with a GridSearchCV procedure for hyperparameter optimization. The goal of this model is to achieve better classification performance by systematically exploring different parameter configurations, including the regularization parameter (C), kernel type, and gamma. The model was trained and tested on the Wisconsin Breast Cancer Diagnostic (WBCD) dataset, with GridSearchCV employed to select the optimal hyperparameters through a 5-fold cross-validation process. The resulting SVC model achieved significant improvements in accuracy, precision, recall, F1 score, and AUC when compared to other standard machine learning models. This approach demonstrates the potential of SVC with grid search as a robust solution for breast cancer classification, offering better diagnostic capabilities and aiding in more accurate predictions of malignant and benign tumors. Our findings highlight the importance of hyperparameter optimization in machine learning for healthcare applications and suggest that the proposed model could play a vital role in the early diagnosis and treatment of breast cancer.

INTRODUCTION

Breast cancer is one of the leading causes of death among women worldwide, with early detection playing a critical role in improving survival rates. The timely identification and accurate classification of breast tumors as malignant or benign significantly impact clinical decision-making and treatment strategies. As the complexity of medical data increases, traditional diagnostic methods often fall short in providing consistent and reliable predictions. This challenge has led to the growing use of machine learning (ML) algorithms, which are capable of learning complex patterns from data and making predictions with higher accuracy. In the realm of breast cancer diagnosis, machine learning has shown great potential in improving classification performance, reducing the reliance on manual interpretation, and providing more efficient and scalable solutions.

Among the various machine learning algorithms, Support Vector Machines (SVM), specifically the Support Vector Classifier (SVC), have been widely adopted for classification tasks due to their ability to handle high-dimensional data and provide clear decision boundaries. However, the performance of SVC models is highly dependent on the selection of appropriate hyperparameters, such as the regularization parameter (C), the kernel type (linear or non-linear), and the gamma parameter. Improper tuning of these parameters can lead to underfitting or overfitting, which negatively affects the model's predictive accuracy. To address this issue, hyperparameter optimization becomes crucial. One of the most effective methods for hyperparameter tuning

is GridSearchCV, which systematically evaluates different combinations of hyperparameters and selects the best performing configuration based on cross-validation. By incorporating GridSearchCV into the training process, it is possible to enhance the performance of the SVC model and achieve more reliable results for breast cancer classification.

In this study, we propose an enhanced approach to breast cancer detection by utilizing a Support Vector Classifier (SVC) combined with GridSearchCV for hyperparameter optimization. The Wisconsin Breast Cancer Diagnostic (WBCD) dataset, which includes a variety of features extracted from breast cancer biopsies, serves as the dataset for training and testing our model. The dataset includes both benign and malignant tumor samples, providing a diverse set of cases for classification. The primary objective of this research is to demonstrate the potential of an optimized SVC model for accurate breast cancer classification. Through systematic hyperparameter tuning with GridSearchCV, we aim to improve key performance metrics such as accuracy, precision, recall, and the F1 score, which are critical for assessing the reliability and robustness of the model. Furthermore, by comparing the results of the optimized SVC model with traditional classifiers, we aim to showcase the advantages of this approach in terms of diagnostic accuracy and early detection capabilities. Ultimately, this study aims to contribute to the growing body of research on using machine learning for healthcare applications, specifically in breast cancer detection. The findings from this research could have significant implications for clinical practice, offering a more accurate and

efficient tool for healthcare professionals to make informed decisions regarding breast cancer diagnosis and treatment.

SCOPE OF THE PROJECT

The scope of this project revolves around the development and optimization of a machine learning model for breast cancer classification using a Support Vector Classifier (SVC). The primary focus is to enhance the classification performance through systematic hyperparameter tuning with GridSearchCV. The project will utilize the Wisconsin Breast Cancer Diagnostic (WBCD) dataset, which contains various features extracted from breast cancer biopsies, to train and test the model. The data preprocessing phase will include cleaning the dataset, handling missing values, and scaling the features to prepare them for optimal learning. The project aims to build an SVC model that can accurately classify breast tumors as benign or malignant by fine-tuning its key hyperparameters, such as the regularization parameter (C), kernel type, and gamma. Performance evaluation will focus on key metrics like accuracy, precision, recall, F1 score, and AUC to assess the model's effectiveness. Ultimately, the goal of the project is to create a reliable and efficient machine learning tool for early breast cancer detection, providing insights into the potential of hyperparameter optimization in improving classification accuracy.

OBJECTIVE

The objective of this project is to develop and optimize a machine learning model that accurately classifies breast cancer tumors as benign or malignant using a Support Vector Classifier (SVC). A key focus is on improving the model's performance through hyperparameter optimization via GridSearchCV, where parameters such as the regularization parameter (C), kernel type, and gamma will be fine-tuned to identify the best configuration. The project will also involve comprehensive data preprocessing of the Wisconsin Breast Cancer Diagnostic (WBCD) dataset, including cleaning, handling missing values, and scaling the features to ensure the model receives high-quality input for training. Once the model is trained, it will be evaluated using multiple classification metrics, such as accuracy, precision, recall, F1 score, and AUC, to assess its effectiveness in distinguishing between benign and malignant tumors. The optimized SVC model will be compared to other traditional machine learning classifiers to benchmark its performance and highlight its strengths. The ultimate goal of the project is to create a reliable, high-performing tool for early breast cancer detection, offering healthcare

professionals a more accurate diagnostic tool to improve clinical decision-making and enhance patient outcomes.

EXISTING SYSTEM:

The existing model, Toward Improving Breast Cancer Classification Using an Adaptive Voting Ensemble Learning Algorithm, addresses a significant challenge in the field of medical diagnostics by utilizing an ensemble learning approach to improve the classification of breast cancer tumors. In recent years, the use of machine learning (ML) in healthcare has revolutionized diagnostic methods, particularly in cancer detection, by enabling faster, more accurate analysis of complex datasets. Early detection of breast cancer is crucial for improving patient outcomes, but traditional diagnostic methods can be limited by human error and the complexity of medical data. To overcome these limitations, machine learning models offer the potential to automate and enhance the diagnostic process, offering more precise predictions for malignancy. Ensemble learning is a method where multiple machine learning models are combined to produce a stronger overall model, with the goal of improving classification performance. The fundamental idea behind ensemble learning is that a group of weak learners, or individual models that perform just slightly better than random guessing, can be combined to form a strong learner. This collective decision-making mechanism helps to reduce the variance and bias of individual models, leading to better generalization on unseen data.

In this model, an adaptive voting ensemble learning algorithm is used, which combines four distinct classifiers: Extra Trees Classifier (ETC), Light Gradient Boosting Machine (LGBM), Ridge Classifier (RC), and Linear Discriminant Analysis (LDA). Each classifier brings a unique advantage to the ensemble. The Extra Trees Classifier (ETC) is known for its robustness against overfitting and its ability to handle high-dimensional datasets. The Light Gradient Boosting Machine (LGBM) is particularly effective for large datasets and complex patterns, as it uses gradient boosting to optimize the model iteratively. Ridge Classifier (RC) is effective for linear relationships in data, particularly when dealing with collinearity or high-dimensional features. Finally, Linear Discriminant Analysis (LDA) is a statistical method that seeks to reduce the dimensionality of the dataset while maximizing the separation between different classes, making it particularly suitable for classification tasks where the data has distinct groups. The adaptive voting mechanism plays a central role in this ensemble approach. In the traditional voting ensemble, each classifier votes on a class label, and the majority vote

determines the final prediction. However, adaptive voting adjusts the weight or influence of each classifier based on its performance during the training process. Classifiers that perform well on specific subsets of the data are given more influence, while weaker classifiers are down-weighted. This adaptability allows the ensemble to focus on the most informative models for a given dataset, improving the overall decision-making process and enhancing classification accuracy. This approach was applied to the Wisconsin Breast Cancer Diagnostic (WBCD) dataset, which consists of features derived from breast cancer biopsies, such as radius, texture, smoothness, compactness, and symmetry. These features serve as the basis for distinguishing between malignant and benign tumors. The ensemble model's strength lies in its ability to combine the predictions of these diverse classifiers, capturing different aspects of the data and minimizing the weaknesses of individual models. By integrating multiple algorithms, the ensemble classifier can handle various complexities in the data and make more accurate predictions about the malignancy of a tumor. The proposed adaptive voting ensemble model showcases the power of combining multiple machine learning models to tackle complex problems, particularly in medical diagnostics. By addressing the inherent limitations of individual classifiers, the ensemble method improves the robustness and reliability of predictions. It also demonstrates how ensemble learning can be applied in healthcare for more effective and accurate cancer detection, which is critical for early diagnosis and better clinical outcomes. The adaptability of this approach suggests it can be extended to other medical domains, making it a versatile tool for enhancing diagnostic systems and supporting healthcare professionals in their decision-making processes.

Existing System Disadvantages:

- **Complexity:** The use of multiple classifiers in an ensemble increases model complexity, making it harder to interpret.
- **Computational Cost:** Ensemble methods are computationally expensive and require more resources compared to individual models.
- **Overfitting:** The model may still overfit if the base models are not well-tuned, despite ensemble methods reducing overfitting.
- **Data Imbalance Sensitivity:** Ensemble learning may be sensitive to imbalanced datasets, potentially affecting classification performance.

LITERATURE SURVEY

Breast cancer classification using snapshot ensemble deep learning model and distributed stochastic

neighbor embedding, X N. Sharma, K. P. Sharma, M. Mangla, and R. Rani, 2023.

The paper by Sharma et al. (2023), titled "Breast cancer classification using snapshot ensemble deep learning model and t-distributed stochastic neighbor embedding", presents an advanced method for improving the accuracy of breast cancer detection using machine learning techniques. The study integrates a snapshot ensemble deep learning model, which combines multiple models to improve prediction reliability, with t-distributed stochastic neighbor embedding (t-SNE) for better visualization of the data. In this approach, snapshot ensembles are used to capture different "snapshots" of a deep learning model trained at various stages, which are then combined to enhance classification performance. t-SNE, a dimensionality reduction technique, is employed to visualize high-dimensional breast cancer data in lower dimensions, facilitating better understanding and interpretation of the model's decisions. This methodology allows for effective classification of breast cancer data, aiming to improve both accuracy and the interpretability of results, which is crucial in medical diagnosis. This paper aims to address the challenges in breast cancer classification by using ensemble learning and dimensionality reduction, demonstrating a novel approach that may lead to more accurate and robust models for medical diagnostics.

An ensemble architecture based on deep learning model for click fraud detection in Pay-Per-Click advertisement campaign. A. Batool and Y.-C. Byun, 2022.

The paper "An Ensemble Architecture Based on Deep Learning Model for Click Fraud Detection in Pay-Per-Click Advertisement Campaign" by Amreen Batool and Y.-C. Byun focuses on detecting click fraud in online advertising, a practice that involves fraudulent clicks aimed at wasting an advertiser's budget. The authors propose an innovative ensemble architecture combining deep learning models and machine learning techniques. The model integrates a Convolutional Neural Network (CNN) and a Bidirectional Long Short-Term Memory (BiLSTM) network for feature extraction, followed by a Random Forest (RF) classifier for final prediction. This hybrid approach addresses the challenge of identifying fraudulent clicks by automatically extracting features from click data and then classifying it as fraudulent or legitimate. To further enhance the model's effectiveness, it also incorporates a preprocessing module to handle issues like imbalanced data, which is common in fraud detection. The overall goal of this method is to improve the reliability and efficiency of click fraud detection in pay-per-click advertising.

offering a solution that can safeguard advertiser budgets and maintain the integrity of online advertising campaigns.

A case study of microarray breast cancer classification using machine learning algorithms with grid search cross validation, N. Mohd Ali, R. Besar, and N. A. A. Aziz, 2023.

The article you referenced discusses a study on breast cancer (BC) classification using machine learning techniques. It specifically focuses on the use of three algorithms: Support Vector Machine (SVM), Random Forest (RF), and Extra Trees (ET), for classifying microarray data. The researchers employed GridSearchCV for hyperparameter tuning to optimize the models' performance. The dataset used for the study combined two microarray datasets (GSE45255 and GSE15852), resulting in a total of 182 breast cancer patients and 43 normal cases, with 22,215 genes per sample. The study demonstrates a preprocessing phase that involved normalization and standardization of the gene expression data. Data splitting was performed with an 80-20 training-test ratio, and cross-validation (CV) was applied to assess model performance. Performance metrics such as classification accuracy, F1-score, sensitivity, and specificity were calculated to evaluate the models' effectiveness. In terms of results, the authors found that the classification accuracy and other evaluation metrics were influenced by the choice of model and the hyperparameters selected through GridSearchCV. The study's findings emphasize the importance of proper model tuning and cross-validation in achieving robust classification results for microarray-based breast cancer data.

Performance analysis of averaged perceptron machine learning classifier for breast cancer detection, V. Birchha and B. Nigam, 2023

The paper by Birchha and Nigam evaluates the Averaged Perceptron (AP) classifier's effectiveness for breast cancer detection. It highlights the model's ability to handle noisy data and perform well in multi-class classification. The study also emphasizes the role of parameter optimization and feature selection to improve the classifier's performance. By using breast cancer datasets, they demonstrate that AP can be a reliable tool in early diagnosis, although its accuracy is highly dependent on proper data preprocessing and tuning

Using PBL and agile to teach artificial intelligence to undergraduate computing students. V. A. M. De Barros, H. M. Paiva, and V. T. Hayashi, 2023.

The study titled "Using PBL and Agile to Teach Artificial Intelligence to Undergraduate Computing

Students" by Vitor Augusto Menten De Barros, Henrique Mohallem Paiva, and Victor Takashi Hayashi, focuses on the implementation of Project-Based Learning (PBL) integrated with Agile methodologies, specifically Scrum, to teach Artificial Intelligence (AI) to undergraduate students in computing programs. The case study, conducted with 30 students, highlights how these students successfully developed machine learning models for a real-world breast cancer prediction task. The key contribution of the study is its exploration of how PBL and Agile can enhance student learning by fostering practical skills such as problem-solving, teamwork, critical thinking, and communication, which are crucial in real-world applications of AI. The use of Agile principles like collaboration, iteration, and feedback loops within the learning process allowed students to gain deeper insights into the iterative nature of AI development. The study serves as a valuable resource for future AI teaching methodologies and aims to bridge gaps in the research on the application of Scrum and Agile in computing education.

1.6 PROPOSED SYSTEM

The proposed system centers around optimizing a Support Vector Classifier (SVC) model, chosen for its robustness in handling high-dimensional data and its effectiveness in binary classification tasks. SVC works by identifying a hyperplane that maximizes the margin between different classes, thereby reducing misclassification and improving model accuracy. This approach allows SVC to generalize well, particularly when dealing with complex datasets that might otherwise lead to overfitting in other algorithms. To enhance the model's performance further, the system employs a hyperparameter tuning technique via GridSearchCV, which iterates over a predefined set of hyperparameters to find the optimal combination. Specifically, the model examines variations of the regularization parameter C , kernel types (such as linear and radial basis function (RBF)), and gamma settings (scale and auto). These hyperparameters are crucial, as they control the flexibility and regularization of the decision boundary, directly influencing how well the model balances bias and variance.

GridSearchCV uses cross-validation, generally set to 5 folds, to test each parameter combination rigorously across different subsets of the data. This ensures that the best parameters selected are not only optimal for the training set but also generalize well across unseen data. The iterative search and validation help mitigate issues of overfitting and underfitting, often improving the model's predictive accuracy and stability. Overall, this system is designed to efficiently classify data by

balancing complexity and generalization, with the structured approach of GridSearchCV enhancing the SVC model's capability in finding a stable, high-performing solution for various classification tasks. This setup is especially advantageous in scenarios with limited labeled data or where precision in separating classes is paramount.

Proposed System Advantages:

- **High Accuracy:** SVC is effective in high-dimensional spaces, making it well-suited for tasks with complex and high-dimensional data.
- **Robust to Overfitting:** The SVC model's margin maximization reduces overfitting, providing more reliable generalization on new data.
- **Effective with Small Datasets:** SVC performs well with smaller datasets, as it does not require a large amount of data to find an optimal hyperplane.
- **Hyperparameter Optimization:** GridSearchCV ensures that the best parameter settings are chosen, enhancing model accuracy and consistency.

METHODOLOGY

A machine learning-based classification system to improve the accuracy of breast cancer detection. By analyzing medical datasets, specifically the Wisconsin Breast Cancer Diagnostic (WBCD) dataset, the goal is to create a system that can distinguish between malignant and benign tumors. Accurate and timely diagnosis of breast cancer plays a critical role in improving survival rates, as early detection is linked to better treatment outcomes. Machine learning offers significant advantages over traditional methods by enabling automated, data-driven insights. In particular, using advanced algorithms allows for more consistent and objective classifications, which can assist healthcare professionals in making informed decisions. This approach leverages the power of data mining techniques to identify patterns in medical data, increasing diagnostic efficiency and reducing the potential for human error.

The system is designed to enhance breast cancer classification by using robust algorithms that are fine-tuned for optimal performance. These algorithms learn from historical data to improve their predictions, which is crucial in handling complex medical cases with high dimensionality. Through model optimization techniques, such as hyperparameter tuning and cross-validation, the system aims to achieve high classification accuracy, reliability, and scalability, contributing to the ongoing effort to utilize artificial intelligence (AI) and machine learning in healthcare. Ultimately, the project aspires to provide a valuable tool for

clinicians, improving the diagnostic process by providing a faster, more accurate means of identifying cancerous tumors. By leveraging the power of machine learning, the project helps advance the field of medical diagnostics and paves the way for integrating AI solutions into clinical practices.

MODULES

1) Data Collection and Preprocessing:

The primary dataset used is the Wisconsin Breast Cancer Diagnostic (WBCD) dataset, which includes features of tumors such as size, texture, and shape. Preprocessing steps typically include handling missing values, scaling/normalizing the data to ensure all features contribute equally, and splitting the data into training and testing sets to avoid overfitting.

2) Model Selection:

The project leverages the Support Vector Classifier (SVC) model, a supervised machine learning algorithm particularly effective for classification tasks. SVC creates a decision boundary by maximizing the margin between data points of different classes, providing an optimal separation for accurate classification.

3) Hyperparameter Tuning:

GridSearchCV is used to optimize the model by selecting the best set of hyperparameters through an exhaustive search over a predefined parameter grid. This helps find the most suitable parameters for the SVC, such as the regularization parameter C, kernel type (linear or radial basis function (RBF)), and gamma, which affect the decision boundary's flexibility and generalization ability.

4) Cross-Validation:

To ensure the model's performance generalizes well to unseen data, cross-validation (with 5-folds in your case) is applied. This technique involves partitioning the training data into subsets, training the model on some and validating it on others, which helps mitigate issues of overfitting.

5) Ensemble Methods:

In addition to the SVC model, the existing model uses an ensemble learning approach, combining multiple classifiers—Extra Trees Classifier (ETC), Light Gradient Boosting Machine (LightGBM), Ridge Classifier (RC), and Linear Discriminant Analysis (LDA)—through a voting mechanism. Ensemble methods typically improve model performance by combining the strengths of individual classifiers to reduce bias and variance in the final predictions.

6) Model Evaluation:

The performance of the models is evaluated using various metrics, including accuracy, precision, recall, and F1-score, to ensure they provide a well-rounded

assessment of classification performance. These metrics help determine how well the model distinguishes between benign and malignant tumors.

7) Comparison with Existing Methods:

The results from proposed SVC model are compared against those of the ensemble model, with the goal of assessing whether the fine-tuned SVC can match or exceed the performance of the ensemble approach.

EXISTING TECHNIQUE: -

➤ ADAPTIVE VOTING ENSEMBLE LEARNING

The existing algorithm utilizes an adaptive voting ensemble learning approach that combines four distinct machine learning models: Extra Trees Classifier (ETC), Light Gradient Boosting Machine (LightGBM), Ridge Classifier (RC), and Linear Discriminant Analysis (LDA). These models each offer unique strengths, and by combining them into a single framework, the ensemble approach leverages the diversity of the individual classifiers to produce a more accurate and robust prediction. Extra Trees Classifier (ETC) is an ensemble of decision trees that enhances the decision-making process by introducing additional randomness in the tree-building process, such as selecting random subsets of features at each split. This randomness helps in reducing overfitting and ensures the model generalizes better to unseen data. ETC is particularly useful for high-dimensional datasets and handles both categorical and numerical features well.

Light Gradient Boosting Machine (LightGBM) is a gradient boosting framework that efficiently builds decision trees in a sequential manner. Unlike traditional boosting methods, LightGBM uses a histogram-based technique to split continuous features into bins, significantly reducing memory usage and speeding up training. This makes it particularly effective for large datasets and tasks requiring high prediction accuracy. The boosting nature of LightGBM enables the model to focus on correcting the mistakes of previous trees, making it highly precise. Ridge Classifier (RC) is a linear model that applies ridge regression to classification tasks. It adds regularization to prevent overfitting, especially when the data has a large number of features or is collinearly correlated. Ridge classification is effective when the relationship between the features and the target variable is approximately linear, which allows for quick and stable solutions.

Linear Discriminant Analysis (LDA) is a statistical method that projects data onto a lower-dimensional space to separate classes in a way that maximizes the variance between classes while minimizing the variance within each class. LDA is particularly useful

for classification tasks with fewer training samples but many features, making it suitable for high-dimensional datasets like the one used in breast cancer detection. By combining these four classifiers, the adaptive voting ensemble method aggregates their individual predictions through a majority vote or weighted voting, depending on the specific implementation. This combination reduces the model's overall error rate and improves the robustness and stability of predictions, especially when faced with complex and imbalanced datasets like those used in medical diagnostics.

PROPOSED TECHNIQUE

➤ SUPPORT VECTOR CLASSIFIER (SVC):

The proposed algorithm utilizes the Support Vector Classifier (SVC), a machine learning model widely used for classification tasks, particularly in scenarios where data is not linearly separable. The core idea behind SVC is to find a hyperplane that best separates data points of different classes, maximizing the margin between them. In the context of breast cancer classification, this model aims to differentiate between benign and malignant tumor cases, relying on the features extracted from the dataset. The model's performance can vary significantly based on its hyperparameters, so GridSearchCV is applied to systematically explore different configurations. This involves testing a range of hyperparameters, including the C parameter, which balances the trade-off between margin width and classification error; the kernel, which defines how the data is mapped into higher dimensions (with options like 'linear' and 'rbf'); and gamma, which governs the influence of individual data points in constructing the decision boundary. The grid search method ensures the model selects the best parameters that yield the highest accuracy for the specific dataset.

SYSTEM ARCHITECTURE:

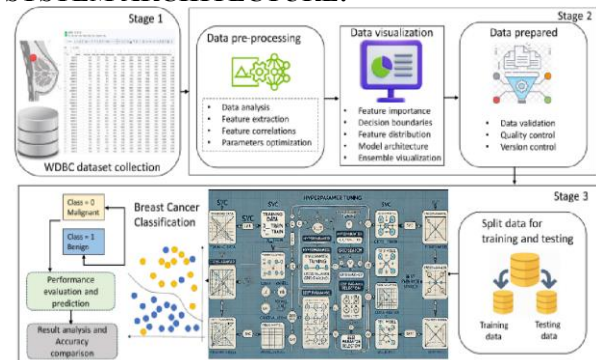


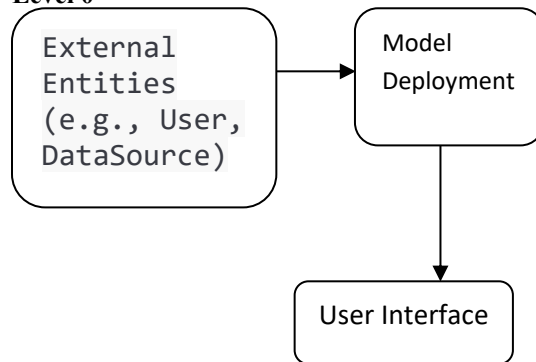
Fig 4.11: System Architecture

To assess the robustness of the model, 5-fold cross-validation is employed. This technique divides the

dataset into five subsets, using each subset once as the test set while training the model on the remaining four. By averaging the results from each fold, the model's generalization ability is improved, and overfitting is minimized. This step is crucial in medical applications like breast cancer detection, where the cost of misclassification is high. Once the optimal parameters are selected, the SVC model is trained on the entire dataset, allowing it to make accurate predictions about new, unseen data. This approach, combining careful hyperparameter tuning with robust cross-validation, enhances the model's ability to generalize and perform well on real-world, unseen breast cancer data. The use of SVC for this task is especially relevant due to its effectiveness in handling complex, high-dimensional spaces, typical in medical diagnostic datasets.

DATA FLOW DIAGRAM

Level 0



Level 1

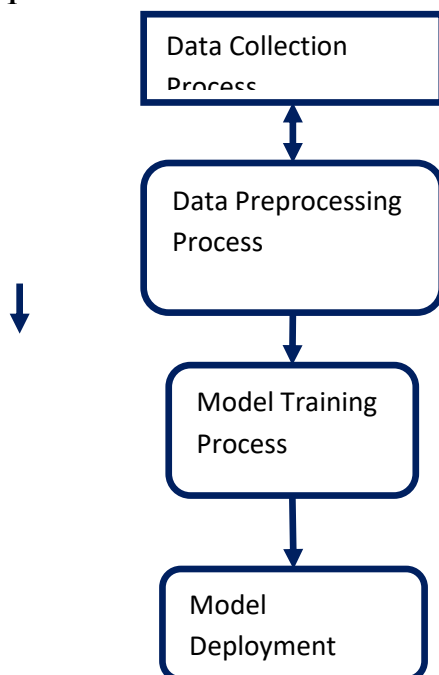


Fig 4.9: Data Flow Diagrams

FUTURE ENHANCEMENT

For future enhancements of the breast cancer classification system, several key improvements can be made to further optimize its performance and adaptability. One potential direction is the integration of deep learning models, such as Convolutional Neural Networks (CNNs), which are highly effective in extracting hierarchical features from complex data, particularly in image or multi-dimensional datasets (LeCun et al., 2015). These models can potentially outperform traditional classifiers like SVC by learning more intricate patterns from the data. Additionally, addressing the common issue of class imbalance in medical datasets can be improved through techniques like Synthetic Minority Over-sampling Technique (SMOTE) or class weighting in the SVC model, ensuring the model does not favor the majority class and improving detection accuracy for the minority class (Chawla et al., 2002). Furthermore, enhancing feature selection and engineering can contribute to better model performance. Methods such as Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) could be applied to reduce dimensionality and retain the most relevant features, which would simplify the model while maintaining its predictive power. Lastly, exploring ensemble learning methods, such as stacking or boosting, could combine the strengths of multiple classifiers to create a more robust model, improving both accuracy and generalization. These future enhancements could significantly boost the model's ability to detect breast cancer more reliably, particularly in complex or unbalanced datasets.

CONCLUSION

In conclusion, the proposed Support Vector Classifier (SVC) algorithm for breast cancer classification offers a robust method for distinguishing between benign and malignant tumors, leveraging powerful machine learning techniques. By employing GridSearchCV for hyperparameter optimization and 5-fold cross-validation for model evaluation, the system ensures high accuracy and generalization ability. This approach can effectively handle complex datasets, providing reliable predictions in medical diagnostics. However, future enhancements such as the integration of deep learning models, techniques to manage class imbalance, and improvements in feature selection and ensemble learning could further boost the model's performance, making it more adaptable and accurate in detecting breast cancer. These enhancements would pave the way for even more effective early diagnosis and decision-making

support in clinical settings, ultimately contributing to better patient outcomes.

REFERENCES

- [1] Breast Cancer Statistics, NBCC, 2022.
- [2] B. He, H. Sun, M. Bao, H. Li, J. He, G. Tian, and B. Wang, "A crosscohort computational framework to trace tumor tissue-of-origin based onRNA sequencing," *Sci. Rep.*, vol. 13, no. 1, p. 15356, Sep. 2023.
- [3] W. Wang, R. Jiang, N. Cui, Q. Li, F. Yuan, and Z. Xiao, "Semi-supervisedvision transformer with adaptive token sampling for breast cancer classification," *Frontiers Pharmacol.*, vol. 13, Jul. 2022, Art. no. 929755.
- [4] S. Chen, Y. Chen, L. Yu, and X. Hu, "Overexpression of SOCS4 inhibits proliferation and migration of cervical cancer cells by regulating JAK1/STAT3 signaling pathway," *Eur. J. Gynaecol. Oncol.*, vol. 42, no. 3, pp. 554–560, 2021.
- [5] I. A. for Research on Cancer, W. H. Organization, Globocan 2012: Cervical Cancer-Estimated Incidence, Mortality and Prevalence Worldwide in 2012, 2018.
- [6] Z.-R. Jiang, L.-H. Yang, L.-Z. Jin, L.-M. Yi, P.-P. Bing, J. Zhou, and J.-S. Yang, "Identification of novel cuproptosis-related lncRNA signatures to predict the prognosis and immune microenvironment of breast cancer patients," *Frontiers Oncol.*, vol. 12, Sep. 2022, Art. no. 988680.
- [7] N. Kumar Sinha, "Developing a web based system for breast cancer prediction using XGboost classifier," *Int. J. Eng. Res.*, vol. V9, no. 6, pp. 852–856, Jun. 2020.
- [8] M. Umer, M. Naveed, F. Alrowais, A. Ishaq, A. A. Hejaili, S. Alsubai, A. A. Eshmawi, A. Mohamed, and I. Ashraf, "Breast cancer detection using convoluted features and ensemble machine learning algorithm," *Cancers*, vol. 14, no. 23, p. 6015, Dec. 2022.
- [9] B. He, Y. Zhang, Z. Zhou, B. Wang, Y. Liang, J. Lang, H. Lin, P. Bing, L. Yu, D. Sun, H. Luo, J. Yang, and G. Tian, "A neural network framework for predicting the tissue-of-origin of 15 common cancer types based on RNA-Seq data," *Frontiers Bioeng. Biotechnol.*, vol. 8, p. 737, Aug. 2020.
- [10] S. A. Yazdan, R. Ahmad, N. Iqbal, A. Rizwan, A. N. Khan, and D.-H. Kim, "An efficient multi-scale convolutional neural network based multi-class brain MRI classification for SaMD," *Tomography*, vol. 8, no. 4, pp. 1905–1927, Jul. 2022.
- [11] Imran, N. Iqbal, S. Ahmad, and D. H. Kim, "Health monitoring system for elderly patients using intelligent task mapping mechanism in closed loop healthcare environment," *Symmetry*, vol. 13, no. 2, p. 357, Feb. 2021.
- [12] J. Zhang, Q. Shen, Y. Ma, L. Liu, W. Jia, L. Chen, and J. Xie, "Calcium homeostasis in Parkinson's disease: From pathology to treatment," *Neurosci. Bull.*, vol. 38, no. 10, pp. 1267–1270, Oct. 2022.
- [13] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, Sep. 2023, Art. no. 100804.
- [14] Y. Li, "Performance evaluation of machine learning methods for breast cancer prediction," *Appl. Comput. Math.*, vol. 7, no. 4, pp. 212–216, 2018.
- [15] B. He, C. Dai, J. Lang, P. Bing, G. Tian, B. Wang, and J. Yang, "A machine learning framework to trace tumor tissue-of-origin of 13 types of cancer based on DNA somatic mutation," *Biochim. et Biophys. Acta (BBA)-Mol. Basis of Disease*, vol. 1866, no. 11, 2020, Art. no. 165916.
- [16] A. Bharat, N. Pooja, and R. A. Reddy, "Using machine learning algorithmsfor breast cancer risk prediction and diagnosis," in *Proc. 3rd Int. Conf. Circuits, Control, Commun. Comput. (I4C)*, Oct. 2018, pp. 1–4.
- [17] K. Jain and M. S. Sharma, "Breast cancer diagnosis using machine earning echniques," *Int. J. Innov. Sci., Eng. Technol.*, vol. 5, no. 5, 2018.
- [18] A. S. Elkorany, M. Marey, K. M. Almustafa, and Z. F. Elsharkawy, "Breast cancer diagnosis using support vector machines optimized by whale optimization and dragonfly algorithms," *IEEE Access*, vol. 10, pp. 69688–69699, 2022.
- [19] L. Liu, "Research on logistic regression algorithm of breast cancer diagnose data by machine learning," in *Proc. Int. Conf. Robots Intell. Syst.(ICRIS)*, May 2018, pp. 157–160.
- [20] V. Amudha, R. G. Babu, K. Arunkumar, and A. Karunakaran, "Machinelearning-based Performance comparison of breast cancer detection using support vector machine," in *Proc. AIP Conf.*, vol. 2519, no. 1, 2022, Art. no. 050011.
- [21] E. Halim, P. P. Halim, and M. Hebrard, "Artificial intelligent models for breast cancer early detection," in *Proc. Int. Conf. Inf. Manage. Technol. (ICIMTech)*, Sep. 2018, pp. 517–521.
- [22] B. S. Abunasser, M. R. J. Al-Hiealy, I. S. Zaqout, and S. S. Abu-Naser, "Breast cancer detection and classification using deep learning Xception algorithm," *Breast Cancer*, vol. 13, no. 7, 2022.
- [23] D. Albashish, "Ensemble of adapted convolutional neural networks (CNN) methods for classifying colon histopathological images," *PeerJ Comput. Sci.*, vol. 8, p. e1031, Jul. 2022.
- [24] M. Dabass, S. Vashisth, and R. Vig, "A convolution neural network with multi-level

convolutional and attention learning for classification of cancer grades and tissue structures in colon histopathological images,” *Comput. Biol. Med.*, vol. 147, Aug. 2022, Art. no. 105680.

[25] P. Goyal, S. Sehgal, S. Ghosh, D. Aggarwal, P. Shukla, A. Kumar, R. Gupta, and S. Singh, “Histopathological correlation of atypical (C3) and suspicious (C4) categories in fine needle aspiration cytology of the breast,” *Int. J. Breast Cancer*, vol. 2013, Sep. 2013, Art. no. 965498

[26] Shashank A. (2025). Metadata-driven data integration framework: Automating enterprise data integration through declarative approaches. *European Modern Studies Journal*, 9(4), 9.

[27] Karim, M. F., & Intakhab, B. Microwave-Based Detection of Early-Stage Renal Cell Carcinoma Using UHF Range Antenna.

[28] Kumar, Anuj. "Battery Thermal Management Systems for Electric Vehicles Using Phase Change Materials." *International Journal of Engineering, Science and Mathematics*, Vol. 10, Issue 3, 2021, pp. 202-211