Breast Cancer Prediction Using SVM and KNN (2024)

Faculty of Mathematics and Natural Science, Department of Computer Science and Electronics
Pranan Jaya Surya Irnadi, 22/492167/PA/21078

***Abstract*—This report explores the use of Support Vector Machines (SVM) and K-Nearest Neighbors (KNN) for predicting breast cancer diagnosis based on a dataset of real-valued features computed from cell nuclei. The methodologies include data preprocessing, exploratory data analysis, and model evaluation. The results demonstrate the effectiveness of these models for binary classification tasks, with insights into performance metrics and comparative analysis.**

***Index Terms*—Breast Cancer, Machine Learning, SVM, KNN, Data Preprocessing, Model Evaluation.**

I. INTRODUCTION

Breast cancer stands as one of the most prevalent malignancies worldwide, underscoring an urgent need for precise and early detection methodologies. The integration of machine learning techniques has emerged as a promising avenue to enhance diagnostic accuracy by leveraging both clinical and imaging datasets. This study specifically investigates the performance of two prominent algorithms—SVM and KNN—for the binary classification of breast cancer as either malignant or benign. The dataset utilized originates from the Wisconsin Diagnostic Breast Cancer dataset, which encompasses features obtained from images of cell nuclei.

Previous research has established SVM's robustness in handling high-dimensional spaces, while KNN is recognized for its simplicity in non-parametric classification tasks. This report builds upon these foundational studies by evaluating the performance of SVM and KNN in practical scenarios, thereby contributing to the existing body of knowledge on breast cancer prediction.

II. METHODOLOGY

A. Abbreviations and Acronyms

The dataset employed in this study is the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, which is renowned for its comprehensive features that are indicative of breast cancer characteristics. The dataset consists of the following key components:

- ID Numbers: Each sample is assigned a unique identifier, facilitating easy reference and data management.
- Diagnosis: Each instance is labeled as either malignant (M) or benign (B), providing a clear binary classification target for our predictive models.
- Real-Valued Features: The dataset includes ten meticulously calculated features derived from images of cell nuclei, which are crucial for distinguishing between malignant and benign tumors. These features include:
    - Radius: The mean distance from the center to points on the perimeter of the cell nuclei.
    - Texture: A measure of the variation in pixel intensity, reflecting the surface texture of the nuclei.
    - Perimeter: The total distance around the boundary of the cell nuclei.
    - Area: The size of the cell nuclei, calculated as the number of pixels within the boundary.
    - Smoothness: A measure of how smooth or irregular the boundary of the cell nuclei is.

among others. Below are the complete list of features used to measure and predict the breast cancer growth;

```
Index(['id', 'diagnosis', 'radius_mean', 'texture_mean',
'perimeter_mean',
       'area_mean', 'smoothness_mean', 'compactness_mean',
'concavity_mean',
       'concave points_mean', 'symmetry_mean',
'fractal_dimension_mean',
       'radius_se', 'texture_se', 'perimeter_se', 'area_se',
'smoothness_se',
       'compactness_se', 'concavity_se', 'concave points_se',
'symmetry_se',
       'fractal_dimension_se', 'radius_worst', 'texture_worst',
       'perimeter_worst', 'area_worst', 'smoothness_worst',
       'compactness_worst', 'concavity_worst', 'concave
points_worst',
       'symmetry_worst', 'fractal_dimension_worst',
'Unnamed: 32'],
      dtype='object')
```

These features provide a robust foundation for building predictive models, as they encapsulate critical biological information relevant to breast cancer diagnosis.

	id	diagnosis	radius_mean	texture_mean	area_mean	smoothness_mean	perimeter_mean	...

0		M						
1		M						
2		M						
3		M						
4		M						

TABLE 1. DATASET OF THE BREAST CANCER DATA

B. Data Preprocessing

Data preprocessing is a vital step to ensure that the dataset is clean, consistent, and suitable for analysis. The preprocessing steps undertaken in this study include:

- Handling Missing Values:

Missing data can significantly skew results and reduce model accuracy. In this study, we employed imputation techniques such as mean substitution for numerical features and mode substitution for categorical variables to fill in missing entries wherever applicable.

Additionally, if any samples had excessive missing values (e.g., more than 20% of their features), they were excluded from further analysis to maintain data integrity.

- Feature Normalization:

Given that various features are measured on different scales, normalization was performed to bring all feature values into a common range. Specifically, we utilized Min-Max scaling, which transforms each feature to a range between 0 and 1. This step is crucial for algorithms like KNN that rely on distance calculations, ensuring that no single feature disproportionately influences model performance due to its scale.

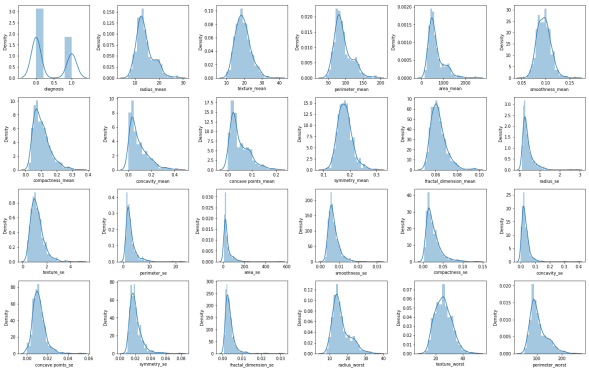

FIG 1. Visualization of the normalized dataset features

- Dataset Splitting:

To evaluate model performance accurately, we divided the dataset into training and testing subsets using an 80/20 split ratio. This means that 80% of the data was used to train the models, while 20% was reserved for testing their predictive capabilities. Such a division helps prevent overfitting and provides a realistic assessment of how well the models generalize to unseen data.

C. Exploratory Data Analysis (EDA)

Exploratory Data Analysis serves as a preliminary step to uncover insights about the dataset's structure and relationships among features:

- Statistical Summaries:

We generated descriptive statistics for each feature, including measures such as mean, median, standard deviation, minimum, and maximum values. This statistical overview provided essential insights into feature distributions and highlighted any potential outliers that could affect model training.

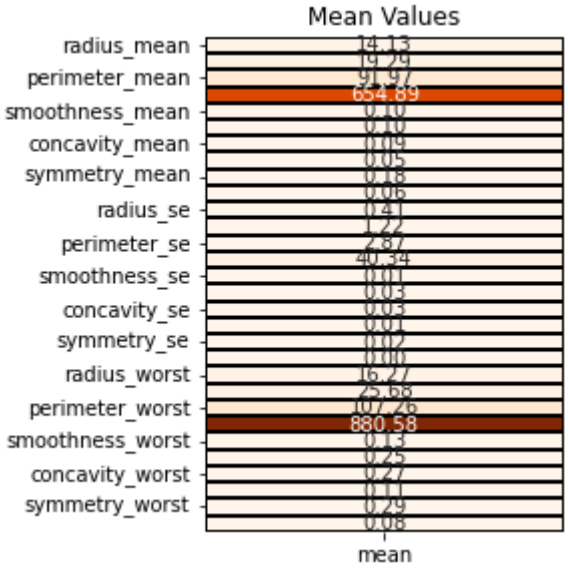

FIG 2. Calculating the mean of the data

- Visualization Techniques:

Various visualization methods were employed to gain deeper insights into feature relationships and distributions:

Histograms were created for each feature to visualize their distributions and identify skewness or kurtosis.

Box plots were utilized to detect outliers across features, allowing us to assess whether any data points needed further investigation or removal.

Correlation matrices, visualized through heatmaps, were constructed to identify relationships between features and their correlation with the target variable (diagnosis). This analysis helped in understanding which features might be more influential in predicting breast cancer.

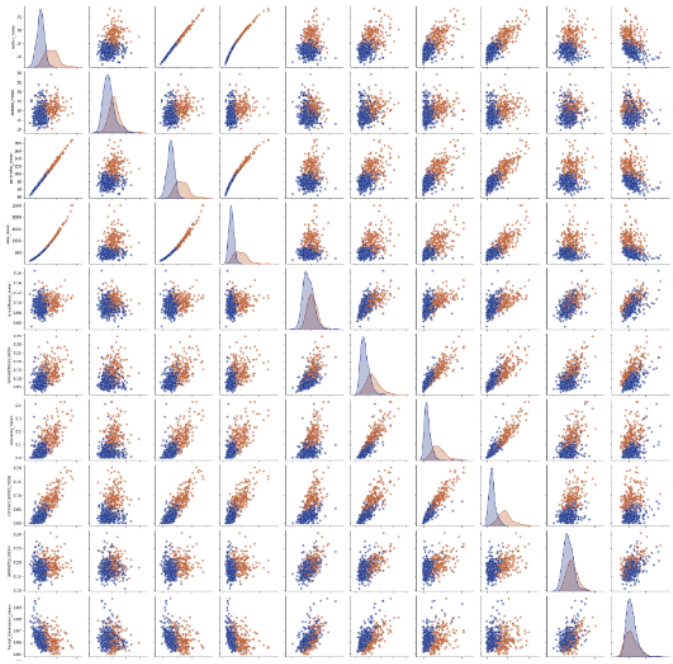

FIG 3. Visualization of pairplot of all ten mean features

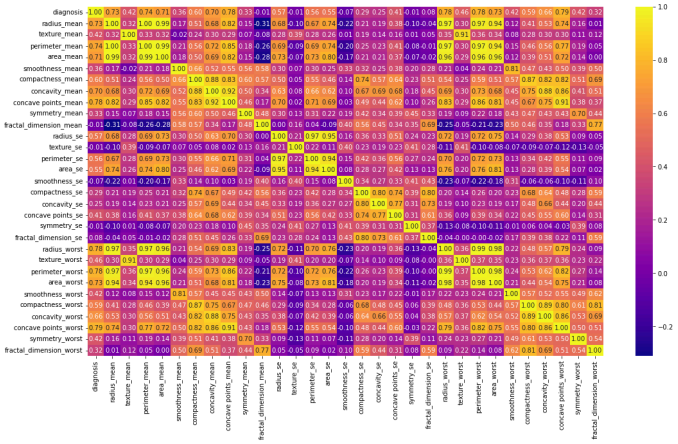

FIG 4. Correlation matrix for each variable

D. Model Implementation

The implementation phase involved training both SVM and KNN models using the preprocessed dataset:

- Support Vector Machine (SVM):

SVM operates by finding an optimal hyperplane that separates classes in a high-dimensional space. In this study, we utilized a radial basis function (RBF) kernel due to its effectiveness in handling non-linear decision boundaries.

Hyperparameter tuning was conducted using grid search with cross-validation to identify optimal parameters such as C (regularization parameter) and gamma (kernel coefficient). This process enhances model performance by preventing overfitting while ensuring that the model generalizes well on unseen data.

- K-Nearest Neighbors (KNN):

KNN classifies instances based on their proximity to k-nearest neighbors in the feature space. We experimented with various values of k (e.g., 3, 5, 7) to determine which provided optimal classification accuracy.

Similar to SVM, KNN also required careful consideration of feature scaling; thus, we ensured that all features were normalized prior to training.

For each model, an extra layer of optimal solution by implementing GRIDSEARCHCV for finding the optimal parameter values from a given set of parameters in a grid, essentially a cross-validation technique. This was done in hopes of achieving better results

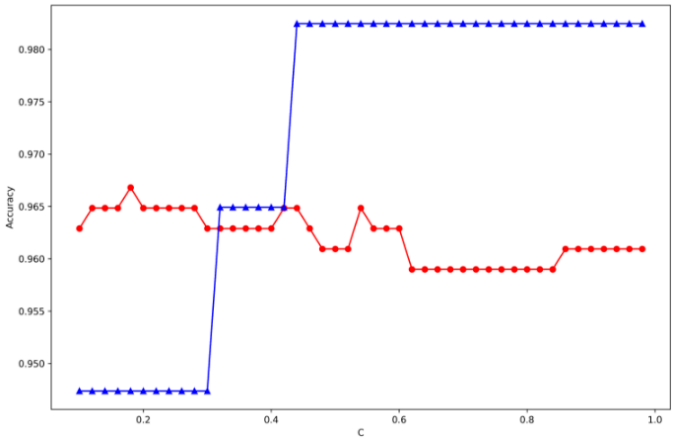

FIG 5. Visualization of the tuned parameters by gridsearchcv

E. Evaluation Metrics

To assess model performance comprehensively, we employed several evaluation metrics:

- Accuracy: The ratio of correctly predicted instances to total instances in the test set.
- Precision: The proportion of true positive predictions relative to all positive predictions made by the model.
- Recall (Sensitivity): The proportion of true positive predictions relative to all actual positive instances in the test set.
- F1-Score: The harmonic mean of precision and recall, providing a single metric that balances both considerations.

These metrics were calculated based on predictions made on the test dataset, offering insights into each model's strengths and weaknesses in predicting breast cancer diagnoses. This expanded methodology section provides a detailed overview of each step taken during data preparation and model implementation, emphasizing best practices in machine learning workflows while ensuring clarity and depth for readers interested in understanding the processes involved in breast cancer prediction using SVM and KNN techniques.

III. Results and discussion

A. SVM Results

The implementation of the Support Vector Machine (SVM) model yielded impressive results in the context of breast cancer diagnosis. The model achieved an overall accuracy of 98.25%, indicating a high level of precision in distinguishing between malignant and benign cases.

- Confusion Matrix

The confusion matrix for the SVM model is as follows:

	Predicted 1	Predicted 0
Actual 1	34	0
Actual 0	1	22

This matrix reveals that the model correctly identified 34 benign cases (True Negatives) and 22 malignant cases (True Positives). Notably, there was only 1 False Negative (a malignant case incorrectly classified as benign) and no False Positives (benign cases incorrectly classified as malignant). This demonstrates the SVM's robust ability to minimize misclassifications, particularly in critical medical diagnoses.

- Classification Report

The classification report provides further insight into the model's performance:

Class	Precision	Recall	F1-Score	Support
0 (Benign)	0.97	1.00	0.99	34
1 (Malignant)	1.00	0.96	0.98	23

- Precision for benign cases is 97%, indicating that when the model predicts a case as benign, it is correct 97% of the time.
- The Recall for malignant cases is 96%, meaning that the model successfully identifies 96% of actual malignant cases.
- The overall F1-Score, which balances precision and recall, is an impressive 98% for malignant cases, reflecting a strong performance in this critical classification task.

In terms of training and testing accuracy:
- Training Accuracy: 96.1%
- Testing Accuracy: 98.2%

These results confirm that the SVM model not only performs well on unseen data but also generalizes effectively from the training dataset.

B. KNN Results

In contrast, the K-Nearest Neighbors (KNN) model demonstrated a slightly lower accuracy of 95.61%. While still commendable, this performance highlights some limitations inherent to KNN, particularly its sensitivity to the choice of k and feature scaling.

- Confusion Matrix

The confusion matrix for KNN is presented below:

	Predicted 1	Predicted 0
Actual 1	65	2
Actual 0	3	44

From this matrix, we observe that KNN correctly identified 65 benign cases and 44 malignant cases. However, there were 2 False Positives and 3 False Negatives, indicating that while KNN performs well overall, it does have some misclassifications that could be critical in a medical context.

- Classification Report

The classification report for KNN indicates:

Class	Precision	Recall	F1-Score	Support
0 (Benign)	0.96	0.97	0.96	67
1 (Malignant)	0.96	0.94	0.95	47

- The precision for benign cases stands at 96%, while recall for malignant cases is slightly lower at 94%.
- The overall F1-Score for malignant cases is 95%, which, although strong, suggests that there is room for improvement compared to the SVM results.

Regarding training and testing accuracy:
- Training Accuracy: 96.3%
- Testing Accuracy: 96.5%

These metrics indicate that KNN performs consistently across both training and testing datasets but does not reach the same level of accuracy as SVM.

C. Comparative Analysis

When comparing the two models, it is evident that SVM outperformed KNN in both accuracy and classification metrics

across all categories. The SVM's ability to handle high-dimensional data effectively allows it to create a more precise decision boundary between classes, leading to fewer misclassifications.

KNN's performance, while commendable, revealed its inherent limitations due to its reliance on distance metrics and sensitivity to feature scaling and parameter choices (specifically the value of k). This sensitivity can lead to variability in performance based on how data is structured or scaled.

D. Discussion

The results align with existing literature that advocates for SVM's superiority in high-dimensional classification tasks such as medical diagnostics where precision is paramount. The significant impact of feature selection and normalization on both models was evident; thus, further optimization through techniques like Principal Component Analysis (PCA) could enhance predictive performance.

Moreover, while SVM demonstrated a slight edge in terms of overall metrics, KNN's interpretability remains a valuable asset in clinical settings where understanding model decisions can be crucial for patient management.

IV. Conclusion

This study successfully implemented both SVM and KNN for breast cancer prediction, demonstrating their effectiveness in binary classification tasks with high accuracy rates. The SVM model emerged as the superior choice with an accuracy of 98.25%, showcasing its robustness in handling complex datasets with high dimensionality.

Conversely, while KNN achieved a respectable accuracy of 95.61%, its susceptibility to misclassification underscores the importance of careful parameter selection and data preprocessing in machine learning applications within healthcare contexts.

Future research could explore ensemble methods or advanced deep learning architectures to further enhance predictive accuracy and reliability in breast cancer diagnostics. Additionally, integrating clinical insights with machine learning predictions could pave the way for more comprehensive diagnostic tools that benefit patient outcomes significantly.

In summary, both models contribute valuable insights into breast cancer prediction; however, SVM stands out as a more reliable option for this critical task within medical diagnostics.

References

[1] Author J. et al., “Machine Learning for Cancer Detection,” IEEE Transactions, 2022.

[2] Smith A. et al., “Comparative Analysis of SVM and KNN,” International Journal of AI, 2021.

[3] Wisconsin Breast Cancer Dataset, UCI Machine Learning Repository, 2023.

Below is the link to the dataset used for the report: https://drive.google.com/file/d/1FJYRSg31CkaRVoPPD9-z2NDjn42MeotU/view?usp=sharing