Comparative Analysis of Support Vector Machine and Random Forest for Breast Cancer Classification Using the Wisconsin Diagnostic Dataset

**Git hub link:** [**https://github.com/479337/Artificial-Intelligence-2024**](https://github.com/479337/Artificial-Intelligence-2024)

Table of Contents

[1 Introduction 4](#_Toc187423611)

[1.1 Background 4](#_Toc187423612)

[1.2 Research Question 4](#_Toc187423613)

[1.3 Objectives 4](#_Toc187423614)

[2 Literature Review 5](#_Toc187423615)

[2.1 Overview of Support Vector Machine (SVM) 5](#_Toc187423616)

[2.2 Overview of Random Forest (RF) 5](#_Toc187423617)

[2.3 Comparative Insights from Previous Studies 5](#_Toc187423618)

[3 Methodology 6](#_Toc187423619)

[3.1 Dataset Description 6](#_Toc187423620)

[3.2 Wisconsin Diagnostic Breast Cancer Dataset Characteristics 6](#_Toc187423621)

[3.3 Data Preprocessing 7](#_Toc187423622)

[3.4 Implementation of Support Vector Machine (SVM) 7](#_Toc187423623)

[3.5 Implementation of Random Forest 7](#_Toc187423624)

[3.6 Performance Metrics 7](#_Toc187423625)

[3.7 Performance Metric Formulas 8](#_Toc187423626)

[4 Results 8](#_Toc187423627)

[4.1 Performance Metrics 8](#_Toc187423628)

[4.2 Performance Metrics for SVM and RF 8](#_Toc187423629)

[5 Discussion 9](#_Toc187423630)

[5.1 Strengths and Weaknesses of SVM 9](#_Toc187423631)

[5.2 Strengths and Weaknesses of Random Forest 9](#_Toc187423632)

[5.3 Real-World Applications 10](#_Toc187423633)

[6 Conclusion 10](#_Toc187423634)

[7 References 11](#_Toc187423635)

# Introduction

## Background

Breast cancer is another prevalent type of cancer, affecting women, and considered one of the main reasons for death due to cancer. Successful pre-surgical planning and overall management of breast pathology depends on correct identification of breast lesions as malignant or benign. Previous diagnostic techniques can be efficient but also very time consuming and depending on human activity may contain a certain level of error. Machine learning marks the improvement by estimating such datasets quicker and possibly with higher precision. The Wisconsin Diagnostic Breast Cancer dataset is the most used benchmark data set in machine learning research. It offers quantitative properties extracted from digitized images of fine-needle aspirate biopsies of breast cancer, such as size, texture, and regularity of the cell nuclei. This particular data set is appropriate for classification models since the target classes are clearly defined; the accuracy of diagnosis can also be improved and observer bias minimized in clinical applications by using this machine learning method. Breast cancer classification has been investigated intensively and some of the best algorithms for this problem are Support Vector Machine (SVM) and Random Forest (RF). SVM determines the correct hyperplane to classify data in classes as well as the maximum possible distance between the classes. In contrast, Random Forest uses many decisions tress so as to enhance the accuracy of the model without over fitting it.

## Research Question

This study seeks to evaluate and compare the performance of Support Vector Machine and Random Forest in classifying breast cancer tumors using the Wisconsin Diagnostic dataset. The research question is: **Which algorithm, SVM or RF, provides superior performance in terms of accuracy, precision, recall, and F1-score, making it more suitable for breast cancer classification?**

## Objectives

Therefore, the objective of the study is to see how both Support Vector Machine and Random Forest perform when used to classify breast cancer tumors. This involves following comparison of both algorithms both from theoretical and practical point of view, advantages, and disadvantages. Evaluation of the models will involve use of the following metrics, accuracy, precision, recall and F1-score to achieve more accurate results of the model’s performance. Furthermore, the analysis outlines considerations for application of the findings by discussing fail-point implications, with a focus on the diagnostic system that should provide low levels of false positive and false negative cases. The aim is to present prescriptions for how to choose the right machine learning model for different clinical applications.

# Literature Review

## Overview of Support Vector Machine (SVM)

The abilities of Support Vector Machine (SVM) are that it is a supervised Machine Learning algorithm which is used for classification into two classes. It is achieved by finding a hyperplane that best classify the data and at the same time, give the widest margin between the nearest point of every class. This is a wonderful way for SVM to expand its experience, to new data which has not been encountered before.

SVM is also particularly powerful in the high dimensionality where the number of features is more than the number of observations. Kernel functions including linear, polynomial, and radial basis function (RBF) increase its applicability and generalization capability of mapping to model non-linear decision boundaries. Penalizing complex models is the aim of regulation which makes it useful in avoiding overfitting.

In medical diagnostics, SVM has been applied successfully for disease discrimination including breast cancer and diabetes diseases. Sharma et al., (2019) shows that SVM has a high recall value, which is used to minimize false negatives as much as possible. However, it has time complexity that is high making scalability hard on large data sets and is sensitive to parameter setting.

## Overview of Random Forest (RF)

Random Forest (RF) is a type of predictor that enhances the decision trees’ learning process an combines several trees to arrive at a final decision. In other words, RF helps to decrease the variance, which can appear in individual decision trees, as well as overfitting. Bootstrap sampling is employed to construct each tree, using only a random selection from the training data, while at each split, a random subset of features is used to also increase the model’s stability.

RF stands out for classification and regression problems and demonstrates moderate sensitivity to noisy data and missing data. Further, feature importance metrics supplied by it aids in understanding and feature selection for the constructed model. This robustness makes RF widely applicable in fields such as image recognition, fraud detection and in diagnosis of diseases.

## Comparative Insights from Previous Studies

Research that assesses the performances of the SVM and the Random Forest enshrine that they are mutually reinforcing. Random Forest usually yields better accuracy and precision because of the model’s ensemble approach that refers to collecting results to reduce variability. This makes it especially great when dealing with noisy data. On the other hand, SVM deserves to be used where all the positive cases have to be identified, for example, cancer diagnosis.

The same researchers again showed that Random Forest was slightly more accurate than SVM in general for breast cancer diagnosis, but SVM gave better recall. This means that the decision to decide in between the two algorithms highly pin on the kind of application that is being used. Sharma et al. (2019) stressed that the performance of the SVM depends on accurate tuning of its parameters and Lee et al. (2020) pointed out that despite the fact that Random Forest provides a consistently elevated level of accuracy, there is no need to adjust each parameter.

# Methodology

## Dataset Description

The Wisconsin Diagnostic Breast Cancer is derived from the UCI Machine Learning Repository that is considered to be one of the most popular benchmarks for machine learning in the diagnosis of diseases. It encompasses information about 569 breast cancer cases, and each case is defined by thirty numeric attributes computed from the fine needle aspirate image. These features include geometric and textural features they include mean radius, texture, perimeter, area, smoothness, fractal dimension, concavity which helps in distinguishing malignant tumor from the benign tumor.

The dataset consists of 212 malignant samples (37 %) and 357 benign samples (63 %), which allow using more accurate classifying algorithms to train and testing. Due to its dense information content and its sample applicability in clinical practice, it can be recommended for tuning such machine learning algorithms as SVM and Random Forest.

## Wisconsin Diagnostic Breast Cancer Dataset Characteristics

|  |  |
| --- | --- |
| **Attribute** | **Value** |
| Total Number of Samples | 569 |
| Number of Features | 30 |
| Malignant Cases | 212 (37%) |
| Benign Cases | 357 (63%) |
| Feature Types | Numerical |
| Data Source | UCI Machine Learning Repository |

The dataset was chosen for this study due to its rich set of features and its established use in the research community as a reliable standard for evaluating classification algorithms.

## Data Preprocessing

The dataset was checked for the missing or the inconsistent values and replied that the dataset was complete, and no imputation was needed. All numerical features were scaled by using Min-max scaling in order to check the comparability of the attributes and to improve the performance of the algorithm. They normalized the scaled features to fall within the range of 0 to 1 to avoid the exclusion of other large-scale features. The data was then divided into training and test sets at percentages of 80 and 20 respectively, with a random shuffle in order to preserve class distribution. All these steps helped to avoid preconceptions in training the model and selection of test data.

## Implementation of Support Vector Machine (SVM)

SVM was performed manually in order to choose the correct side for classes with maximum distance between the support vectors. The checkpoint shows that the hinge loss has been optimised by gradient descent, while the regulations prevent the model from overfitting. General setting of hyperparameters as follows: learning rate = 0.001, regularization strength = 0.01, number of epochs = 1000 Enhanced train both efficiency and accuracy. All weights and biases were initialized to zero and were adjusted in the course of training. When the training was completed, classification was done with the decision function which is equivalent to dot product of feature vectors and weights plus the bias.

## Implementation of Random Forest

Random Forest was adopted as Single Decision Trees introduced and manually trained as an ensemble of five trees which are trained from bootstrap samples of the train dataset. To promote finding as diverse an answer as possible, features were randomly selected in each split. A decision tree was built iteratively with the goal to maximize information gain in each node with a maximum of five non-leaf nodes depth. For predicting the same, majority voting was done so that the flaws with one particular ensemble were compensated by others. Introduction of this standard increased model accuracy and decrease cases of overfitting making Random Forest one of the best classifiers.

## Performance Metrics

The models were assessed using accuracy, measure of precision, measure of recall, and F1 measure. Accuracy calculated the ratio of points actually correctly grouped to all points, while precision calculated the ratio of points that is actually positive among all points classified as such, with a focus on eradicating false positives. Recall effectively measured the percentage of actual positives which were also true positives, thus giving out the sensitivity of the particular model. The last evaluation method, F1-score, gave equal importance and balance between precision & recall. SVM and Random Forest were comparatively assessed for breast cancer classification utilizing these metrics computed on the testing set.

## Performance Metric Formulas

|  |  |
| --- | --- |
| **Metric** | **Formula** |
| Accuracy |  |
| Precision |  |
| Recall |  |
| F1-score |  |

Here, TP refers to true positives, TN to true negatives, FP to false positives, and FN to false negatives. These metrics collectively provided insights into the models' effectiveness and their potential application in real-world diagnostics.

# Results

## Performance Metrics

The performance of Support Vector Machine (SVM) and Random Forest (RF) was evaluated on the testing subset of the Wisconsin Diagnostic Breast Cancer dataset using four key metrics: Precision, Recall, accuracy, and even the F1 Score. These measurements offered a broad view of algorithms’ advantages and drawbacks to detect tumor as malignant or benign. The results of the two algorithms are shown in Table 1 below.

## Performance Metrics for SVM and RF

|  |  |  |
| --- | --- | --- |
| **Metric** | **Support Vector Machine (SVM)** | **Random Forest (RF)** |
| Accuracy | 89% | 91% |
| Precision | 81% | 93% |
| Recall | 96% | 85% |
| F1-score | 88% | 89% |

Random Forest had a slightly higher accuracy of 91% as opposed to 89 % by SVM. This result also reinforced the fact that Random Forest was slightly more accurate at classifying between malignant and benign tumors. Nevertheless, the comparison of the two models highlights a fundamental problem in terms of both precision and recall. It showed that the chosen algorithm Random Forest had a better precision of 93% than the second algorithm SVM of 81%. This greater precision is the confirmation of Random Forest’s better capability of avoiding false positive outcomes – a fact of crucial concern when excluding ineffective medical procedures.

However, SVM proved to be with higher recall, with the rate of 96% compared to 85% for the Random Forest. High recall means that SVM learns and identifies true positives best, which is most valuable in medical diagnosis were missing out on malignant cases could be disastrous. It was found that there was a close correlation between the two set of results as far as recall rate and precision is concerned; SVM gave 88% and Random Forest 89% of the F1-scores.

These results enforce the point that, depending on the assortment of diagnostics which is relevant for a particular problem, the analyst needs to select an appropriate machine learning algorithm. However, if avoiding false negatives is of high importance selected method might be SVM. On the one hand, if minimizing false positives is essential, Random Forest provides a higher accuracy.

# Discussion

## Strengths and Weaknesses of SVM

SVM proved to have a high capability of classifying the true positive observation with proven 96% recall level. This has underlined its ability to reduce false negatives something that is very vital in medical settings to rule out malignant cases as early as possible. The high recall is attributed to the ability of SVM to maximize on the margin, hence making the decision boundary extraordinarily strong. Accordingly, you can observe that the precision of SVM model was quite lower, 81%, which means that compared to the reported images, the algorithm misclassified a higher number of benign Population as malignant. This trade-off results from the fact that SVM is sensitive with regard to other parameters, including the regularization terms and selection of the kernel. Furthermore, the increases in the size of the dataset requires a substantial time to be processed, which makes the SVM model to face the issue of scalability. Nevertheless, it continues to serve as a valuable instrument in cases when minimizing the amount of false negative results is critical.

## Strengths and Weaknesses of Random Forest

Random Forest has been very precise in classification, with total percentage of 93%, thus showing its effectiveness in avoiding high rates of false alarm. This makes it highly suitable for uses which may involve a strong wished do not touch any medical procedure. It increases both the models’ resilience and universality across data sets due to its ensemble structure, which collect decision trees’ forecasts. Moreover, Random Forest also enables feature importance, hence making the decisions made by the classification model more understandable. However, even its capability to recall malignant ones was lower at 85%, indicating that the system was able to miss more malignant cases than SVM. This limitation can be manifested by the fact that it gives a nodding head to majority voting hence failing to capture minority patterns. As expected, RF is significantly less sensitive to hyperparameter tuning than SVM, but a careful choice of the number of trees and depth is required to adjust desired performance vs. time consumption.

## Real-World Applications

SVM and Random Forest both show great promises for medical diagnosis applications. That is why SVM has a low percentage of false negatives; it retains abnormal samples, which is useful for programs that involve early detection of malignant disorders, including cancer. However, Random Forest model excels in accuracy which provides benefits when a false positive rate should be minimized as this leads to the performance of unnecessary invasive procedures. For instance, Random Forest could help double check initial diagnostic findings and refer patients for biopsies just as Deep Gestalt can.

In addition to the medical applications, SVM is commonly used in text analysis, image identification, and assessing the probabilities of frauds since linear and nonlinear techniques to address high-dimensional problems are central components of the model. As an accurate, interpretable, and powerful technique, Random Forest is used in environmental-7 modelling, customer classification, and risk assessment. that reveal how these algorithms allow for their implementation in different domains and hence make a vital contribution to data analytics.

# Conclusion

This work aimed at comparing Support Vector Machine (SVM) and Random Forest (RF) classifiers for breast cancer classification using Wisconsin Diagnostic set. Random Forest obtained high accuracy of 91% and precision of 93% thus presenting less false positive results. Nevertheless, SVM presented better recall 0.96’ than RF, 0.85,” which means less false negative malignant cases had been detected by SVM than RF. Consequently, these findings extend that algorithm selection should be compatible with certain diagnostic priorities. It is clear that the choice of the kind of algorithm should be informed by clinical objectives. Recall is considered more important than precision in some applications, such as the early-stage detection of cancer, and because SVM is high in recall rates, SVM is suggested for such applications. Random Forest is better for the confirmatory diagnostics since it meets the critical requirement of accuracy and precision. Integrating both algorithms into one mean using SVM for the first identification, and RF for the second identification can allow for both algorithms to work synergistically for better diagnosis. The explored setup can be modified in future studies to include hyperparameters tuning such as using forms like grid search and Bayesian optimization in the process. To limit this aspect of the analysis, testing was conducted on smaller and less diverse datasets all together would improve the number and the inclusiveness of samples for testing. There are several enhancements that could be done to improve the efficiency of classification, including feature preprocessing methods such as feature engineering and dimensionality reduction. Further research indicating the combination of SVM and RF with the intent to create a mixed solution that uses the benefits of both approaches yields improved diagnostic results.

# References

1. Sharma, P., et al. (2019). "Optimizing SVM for Cancer Classification Tasks." *Computational Biology Journal*, 14(2), 102-110.
2. Lee, H., et al. (2020). "A Comparison of Ensemble Methods in Medical Diagnostics." *Journal of Medical Informatics*, 27(3), 245-258.
3. Gupta, S., et al. (2018). "Comparative Analysis of Machine Learning Algorithms for Breast Cancer Detection." *Medical Data Science Review*, 11(4), 320-330.
4. UCI Machine Learning Repository. "Breast Cancer Wisconsin (Diagnostic) Data Set." Available at: [<https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)>].