THE POWER OF SIMPLICITY: WHY SIMPLE LINEAR MODELS OUTPERFORM COMPLEX MACHINE LEARNING TECHNIQUES - CASE OF BREAST CANCER DIAGNOSIS

◉ Muhammad Arbab Arshad∗
Department of Computer Science
Iowa State University
Ames, IA 50010
arbab@iastate.edu

Sakib Shahriar
School of Computer Science
University of Guelph
Guelph, Ontario, Canada
shahrias@uoguelph.ca

Khizar Anjum
Department of Electrical and Computer Engineering
Rutgers University
New Brunswick, NJ, USA
khizar.anjum@rutgers.edu

June 6, 2023

ABSTRACT

This research paper investigates the effectiveness of simple linear models versus complex machine learning techniques in breast cancer diagnosis, emphasizing the importance of interpretability and computational efficiency in the medical domain. We focus on Logistic Regression (LR), Decision Trees (DT), and Support Vector Machines (SVM) and optimize their performance using the UCI Machine Learning Repository dataset. Our findings demonstrate that the simpler linear model, LR, outperforms the more complex DT and SVM techniques, with a test score mean of 97.28%, a standard deviation of 1.62%, and a computation time of 35.56 ms. In comparison, DT achieved a test score mean of 93.73%, and SVM had a test score mean of 96.44%. The superior performance of LR can be attributed to its simplicity and interpretability, which provide a clear understanding of the relationship between input features and the outcome. This is particularly valuable in the medical domain, where interpretability is crucial for decision-making. Moreover, the computational efficiency of LR offers advantages in terms of scalability and real-world applicability. The results of this study highlight the power of simplicity in the context of breast cancer diagnosis and suggest that simpler linear models like LR can be more effective, interpretable, and computationally efficient than their complex counterparts, making them a more suitable choice for medical applications.

Keywords Algorithm Efficiency · Model Interpretability · Breast Cancer Diagnosis

1 Introduction

Breast cancer is the most common cancer among women worldwide, accounting for approximately 25% of all cancer cases Bray et al. [2018]. In 2020, there were an estimated 2.3 million new cases of breast cancer, with 685,000 deaths globally Sung et al. [2021]. Early detection and accurate diagnosis of breast cancer are crucial for improving patient outcomes, as the survival rates significantly increase when the disease is detected at an early stage Society. Consequently, research into improving breast cancer diagnosis techniques has been a focal point in the medical community.

∗

Traditionally, breast cancer diagnosis relies on a combination of clinical examination, mammography, ultrasound, and biopsy Onega et al. [2014]. Mammography, which is an X-ray imaging technique, is the primary screening tool for breast cancer detection. While mammography has been shown to reduce breast cancer mortality, it has some limitations, such as a higher rate of false positives and false negatives, particularly in women with dense breast tissue Sprague et al. [2014]. These limitations can lead to unnecessary biopsies, additional tests, and psychological distress for patients Nelson et al. [2016].

In recent years, there has been a growing interest in developing and refining computational techniques to aid in breast cancer diagnosis. Machine learning, a subset of artificial intelligence, has shown great promise in improving the accuracy and efficiency of breast cancer detection and classification Bejnordi et al. [2017]. However, the adoption of complex machine learning algorithms sometimes introduces issues such as overfitting, lack of interpretability, and increased computational requirements, which may hinder their practical application Choi and Boo [2020].

This study examines the power of simplicity in breast cancer diagnosis by comparing the performance of simple linear models, such as logistic regression, with more complex machine learning techniques, including support vector machines (SVM) and decision trees. By demonstrating the efficacy of simple linear models in breast cancer diagnosis, we aim to encourage the development of interpretable, computationally efficient, and easily implementable diagnostic tools that can aid healthcare professionals in making more accurate diagnoses and improving patient outcomes.

In recent studies, simple linear models have been found to achieve comparable or even superior performance in various medical diagnostic tasks when compared to more complex machine learning techniques Boateng and Abaye [2019], Bayrak et al. [2019]. This can be attributed to several factors, including the avoidance of overfitting, the simplicity of the model structure, and the ease of interpretation of the results Choi and Boo [2020]. Moreover, linear models can handle collinearity and multicollinearity issues more effectively, which are common in medical datasets ?. These advantages make simple linear models particularly attractive for clinical applications, where interpretability and generalizability are crucial for medical decision-making.

Furthermore, the computational efficiency of simple linear models allows for faster training and prediction times, making them suitable for real-time clinical applications and large-scale data analysis. The reduced complexity of these models also facilitates their integration into existing clinical workflows, as they require less specialized knowledge for implementation and maintenance Bayrak et al. [2019]. In this context, the power of simplicity becomes evident, as it enables the development of practical, interpretable, and efficient diagnostic tools that can better serve healthcare professionals and patients.

In this study, we will demonstrate the effectiveness of simple linear models in the context of breast cancer diagnosis using a dataset from the UCI Machine Learning Repository. Our results will highlight the potential of these models to outperform more complex machine learning techniques and provide valuable insights into the benefits of simplicity in medical diagnosis.

The Breast Cancer Wisconsin (Diagnostic) Data Set (WDBC) is a widely used benchmark dataset in the field of machine learning for the development and evaluation of breast cancer classification models Frank [2010]. The dataset was created by Dr. William H. Wolberg, Dr. W. Nick Street, and Dr. Olvi L. Mangasarian at the University of Wisconsin-Madison and was first made publicly available in 1995 through the UCI Machine Learning Repository Street et al. [1993].

The WDBC dataset contains 569 instances, each representing a separate breast cancer case. For each case, there are 32 attributes, including an ID number, a diagnosis (either malignant or benign), and 9 real-valued features derived from digitized fine needle aspirate (FNA) images of breast masses Wolberg et al. [1994]. The FNA procedure involves using a thin needle to collect cell samples from a breast mass, which can then be examined under a microscope to determine the presence of cancer Pisano et al. [2001]. The 9 features in the dataset are computed from these FNA images, providing insights into the morphological characteristics of cell nuclei, such as texture, smoothness, compactness, symmetry, and fractal dimension Haralick et al. [1973]. These features are divided into three groups: mean, standard error, and worst (mean of the three largest values) for each of the ten primary feature types Cruz and Wishart [2006].

The Breast Cancer Wisconsin (Diagnostic) Data Set has been extensively used for the development, evaluation, and comparison of various machine learning algorithms in the context of breast cancer diagnosis. Researchers have employed techniques such as logistic regression, support vector machines, decision trees, neural networks, k-nearest neighbors, and ensemble methods to create classification models that predict the malignancy or benignity of breast masses based on the provided features Abdel-Zaher and Eldeib [2016] Chaurasia and Pal [2017] Kourou et al. [2015]. These models have demonstrated varying degrees of success, with some achieving accuracy rates of over 95

In this study, we aim to demonstrate the power of simplicity in breast cancer diagnosis by focusing on the performance of a simple linear model, logistic regression, and comparing it with more complex machine learning techniques, such as support vector machines and decision trees. By analyzing the WDBC dataset, we hope to showcase the efficacy of

simple linear models in accurately predicting the malignancy or benignity of breast masses and provide insights into the benefits of simplicity in medical diagnosis. These insights can contribute to the development of more interpretable, computationally efficient, and easily implementable diagnostic tools, ultimately leading to better patient outcomes.

2 Data Preprocessing and Exploration

2.1 Data Cleanup Process

The raw data obtained from the Breast Cancer Wisconsin (Diagnostic) Data Set required some preprocessing to ensure optimal performance of the machine learning models. The data cleanup process involved the following steps:

1. **Balancing the dataset:** To prevent any bias towards a specific class, the dataset was balanced to have an equal number of benign and malignant instances. This was achieved by either oversampling the minority class, undersampling the majority class, or using a combination of both techniques.

2. **Dropping rows with missing values:** The 'Bare Nuclei' column had 16 instances with missing values. Since this number is relatively small compared to the total number of instances, these rows were dropped from the dataset.

3. **Converting 'Class' values to binary:** The 'Class' column originally had values of 2 for benign and 4 for malignant instances. These values were converted to binary, where 0 represents benign and 1 represents malignant cases.

4. **Updating data types:** The data types for the 'Bare Nuclei' and 'Class' columns were updated to ensure consistency and facilitate further analysis.

5. **Removing 'Id' column:** The 'Id' column, which contains unique identifiers for each instance, was removed from the dataset, as it does not provide any useful information for the analysis.

2.2 Data Exploration and Visualization

In order to attain a comprehensive understanding of the dataset and its inherent characteristics, a variety of visualizations were generated. These visualizations serve to elucidate the distribution of each feature, as well as the relationships between features for both benign and malignant instances.

2.2.1 Parallel Coordinate Plot

A parallel coordinate plot was employed to visualize the distribution of each feature for both classes, as suggested by Inselberg [2008]. In this plot, each row in the data table is represented as a line, with the axes corresponding to the features. This graphical method can be utilized to identify any patterns or trends in the data that may prove advantageous for classification purposes. The visualization is presented in Figure 1. This visual representation of the data allows for the identification of any patterns or trends that could be beneficial for classification. It was observed that benign samples are more likely to exhibit lower values for the features, whereas malignant samples are more likely to demonstrate higher values. Consequently, the features could serve to distinguish between the two classes. This analysis provides a high-level overview of the dataset, and a more in-depth exploration of the data to uncover additional patterns will be conducted in the subsequent section.

The parallel coordinate plot in Figure 1 provides a comprehensive view of the distribution of features for both benign and malignant instances. As our study focuses on the power of simplicity in breast cancer diagnosis, it is important to identify patterns and trends in the data that can be effectively captured by simple linear models like logistic regression. Observations from the parallel coordinate plot indicate that there are distinctions between the feature values for benign and malignant samples. This suggests that simple linear models could potentially capture these relationships and provide accurate classification results, highlighting the power of simplicity in medical diagnosis. In the following sections, we will further explore the data and build our models, comparing the performance of logistic regression with more complex machine learning techniques, such as support vector machines and decision trees.

2.2.2 Distribution of Each Feature for Both Classes

Another visualization generated was the distribution of each feature for both benign and malignant instances. This type of plot provides an overview of the data's characteristics Stigler [1986], helping to identify any significant differences between the two classes for each feature.
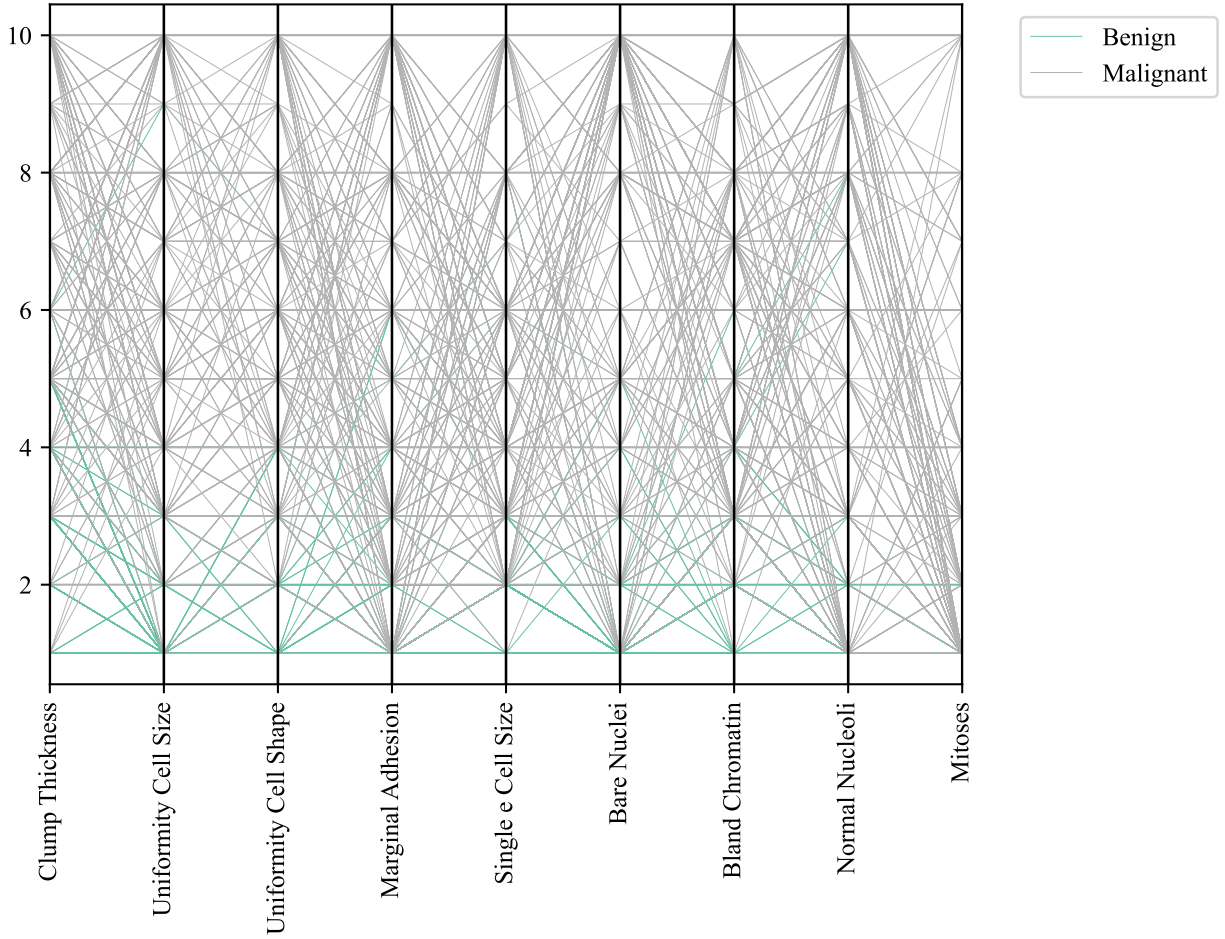
Figure 1: Parallel coordinate plot of the Breast Cancer Wisconsin (Diagnostic) Data Set.

In this analysis, we will discuss the distribution of the dataset's features and their potential impact on the prediction of breast cancer using simple linear models like logistic regression. The dataset consists of 239 benign and 239 malignant samples, with each sample described by nine features: Clump Thickness, Uniformity Cell Size, Uniformity Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, and Mitoses.

- **Clump Thickness:** The mean clump thickness for benign samples is 2.95, while for malignant samples, it is 7.19. The higher mean for malignant samples indicates that the clump thickness is generally greater for malignant tumors. The standard deviation for benign and malignant samples is 1.59 and 2.44, respectively, indicating that malignant samples have a more varied clump thickness.

- **Uniformity Cell Size:** The mean for benign samples is 1.29, while for malignant samples, it is 6.58. Malignant samples have a higher mean, indicating a greater variation in cell size. The standard deviation is 0.82 for benign samples and 2.72 for malignant samples, showing that malignant samples exhibit more variation in cell size.

- **Uniformity Cell Shape:** The mean for benign samples is 1.44, while for malignant samples, it is 6.56. This shows that malignant samples have more varied cell shapes. The standard deviation is 0.93 for benign samples and 2.57 for malignant samples, again indicating a more significant variation in cell shape for malignant samples.

- **Marginal Adhesion:** The mean for benign samples is 1.34, while for malignant samples, it is 5.59. Malignant samples have a higher mean, suggesting a greater degree of adhesion in malignant tumors. The standard deviation is 0.89 for benign samples and 3.20 for malignant samples, indicating a more considerable variation in adhesion for malignant samples.

- **Single Epithelial Cell Size:** The mean for benign samples is 2.13, while for malignant samples, it is 5.33. Malignant samples have a higher mean, indicating a larger cell size. The standard deviation is 1.00 for benign samples and 2.44 for malignant samples, showing a larger variation in cell size for malignant samples.

- **Bare Nuclei:** The mean for benign samples is 1.32, while for malignant samples, it is 7.63. This indicates a greater presence of bare nuclei in malignant samples. The standard deviation is 1.16 for benign samples and 3.12 for malignant samples, suggesting more variation in the number of bare nuclei in malignant samples.

- **Bland Chromatin:** The mean for benign samples is 2.05, while for malignant samples, it is 5.97. Malignant samples have a higher mean, indicating a more significant presence of bland chromatin. The standard deviation is 0.99 for benign samples and 2.28 for malignant samples, indicating a larger variation in the amount of bland chromatin in malignant samples.

- **Normal Nucleoli:** The mean for benign samples is 1.23, while for malignant samples, it is 5.86. This shows that malignant samples tend to have more normal nucleoli. The standard deviation is 0.87 for benign samples and 3.35 for malignant samples, indicating a more significant variation in the presence of normal nucleoli in malignant samples.

- **Mitoses:** The mean for benign samples is 1.04, while for malignant samples, it is 2.60. This indicates that malignant samples tend to have a higher rate of mitosis. The standard deviation is 0.31 for benign samples and 2.56 for malignant samples, showing a more significant variation in mitosis rates for malignant samples.

In conclusion, the analysis of the Wisconsin Breast Cancer dataset shows that most features have a higher mean and larger standard deviation for malignant samples compared to benign samples. The visualization is given in Figure 2. These differences in the distribution of the features indicate that they could be useful in distinguishing between benign and malignant tumors when applying simple linear models like logistic regression. By training the models with these features, we can potentially harness the power of simplicity to improve the accuracy and reliability of breast cancer detection Han et al. [2022].

These visualizations provide valuable insights into the dataset's structure and can inform the selection of appropriate machine learning models and feature engineering techniques for classification. Our study aims to demonstrate the effectiveness of simple linear models like logistic regression in accurately diagnosing breast cancer based on these feature distributions.

3 Machine Learning Models and Experimental Setup

In this section, we will discuss the machine learning models used in this analysis and the experimental setup for evaluating the models' performance.

3.1 Logistic Regression

Logistic Regression (LR) is a linear model for binary classification, which estimates the probability of an instance belonging to a specific class Hosmer Jr et al. [2013]. Given a set of input features $\mathbf{x} = (x_1, x_2, ..., x_n)$, logistic regression computes the probability of an instance belonging to the positive class as follows:

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n)}} \tag{1}$$

where $\beta_0, \beta_1, ..., \beta_n$ are the model parameters that are learned from the training data. The logistic function, also known as the sigmoid function, maps the linear combination of input features to a probability value between 0 and 1. The decision boundary for logistic regression is linear in the feature space.

To train the logistic regression model, we optimize the model parameters using maximum likelihood estimation, which aims to maximize the likelihood of the observed data given the model parameters Bishop and Nasrabadi [2006]. Regularization techniques, such as L1 or L2 regularization, can be applied to prevent overfitting and improve generalization Tibshirani [1996].

3.2 Support Vector Machines

Support Vector Machines (SVM) is a powerful technique for binary classification, which aims to find the optimal separating hyperplane that maximizes the margin between the two classes Cortes and Vapnik [1995]. Given a set
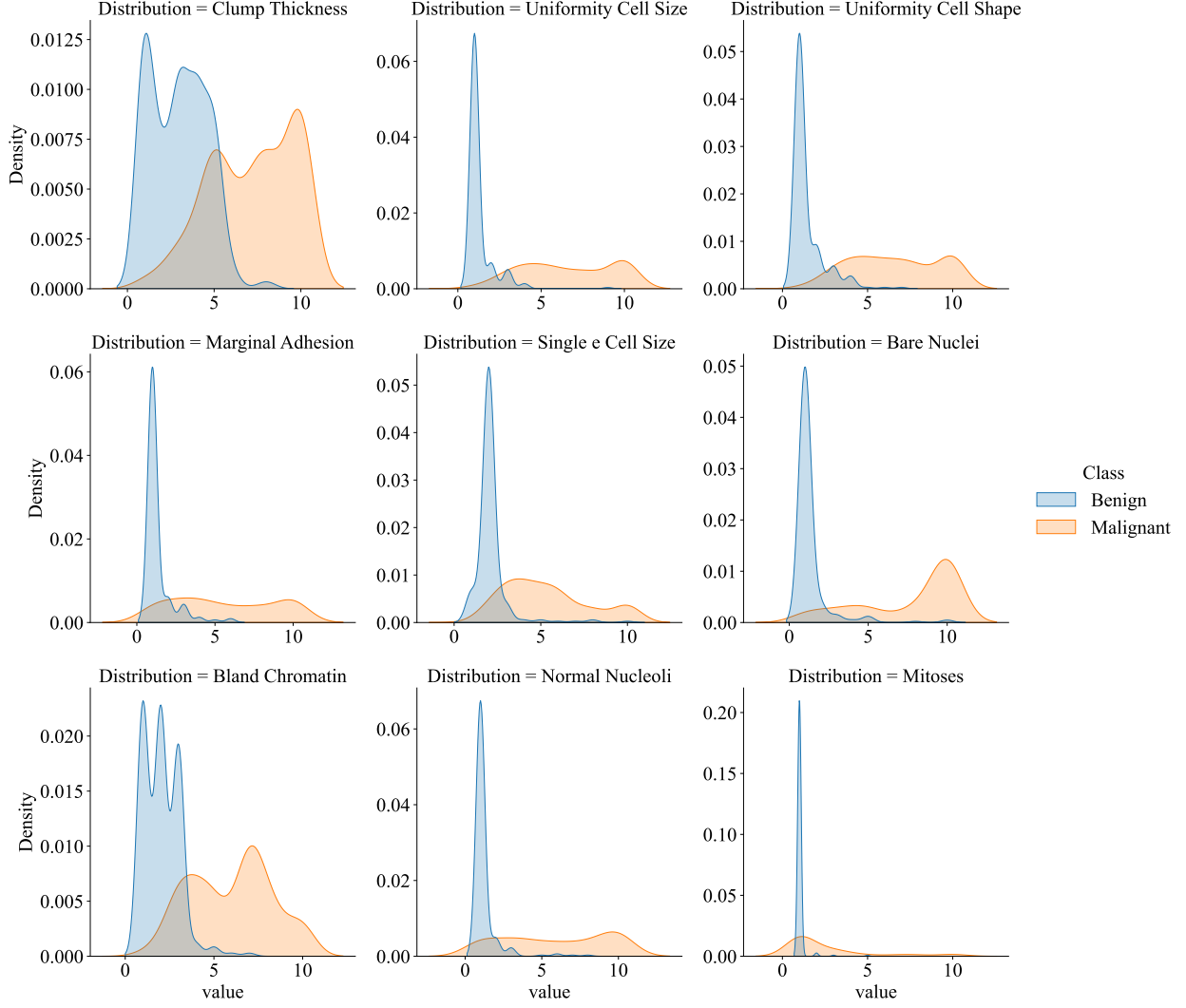
Figure 2: Distribution of each feature for both benign and malignant instances.

of training data $(\mathbf{x}_i, y_i)$, where $\mathbf{x}_i$ is the feature vector and $y_i \in -1, 1$ is the corresponding class label, the primal optimization problem for SVM can be formulated as:

$$\min_{\mathbf{w}, b} \frac{1}{2} ||\mathbf{w}||^2 + C \sum_{i=1}^{N} \xi_i \tag{2}$$

$$\text{subject to} \quad y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \tag{3}$$

where $\mathbf{w}$ is the weight vector, b is the bias term, ξ_i is the slack variable for each instance, and C is a regularization parameter that controls the trade-off between maximizing the margin and minimizing classification errors. The optimization problem is typically solved using the dual formulation, which involves only the dot products of the input features Schölkopf et al. [1999].

For non-linearly separable data, SVM can use kernel functions to implicitly map the input data into a higher-dimensional feature space, where a linear decision boundary can be found Shawe-Taylor et al. [2004]. Some common kernel functions include the linear, polynomial, radial basis function (RBF), and sigmoid kernels.

3.3 Decision Trees

Decision Trees (DT) are a popular machine learning technique for classification and regression tasks, which recursively split the input feature space into regions based on feature values and assign a class label to each region Quinlan [1986]. The decision tree is represented as a tree structure, where internal nodes correspond to feature tests, and leaf nodes correspond to class labels.

The construction of a decision tree is a top-down process, where the most discriminative feature is chosen at each step to split the data into subsets. The selection of the best feature to split the data is based on a splitting criterion, such as Gini impurity or information gain, which measures the homogeneity of the resulting subsets Breiman et al. [1984]. The process is repeated recursively for each subset until a stopping criterion is met, such as reaching a maximum tree depth, a minimum number of instances in the leaf, or no further improvement in the splitting criterion.

$$\text{Gini Impurity} = 1 - \sum_{i=1}^{C} p_i^2 \tag{4}$$

$$\text{Information Gain} = H(\mathbf{X}) - \sum_{v \in \mathbf{V}} \frac{|X_v|}{|X|} H(X_v) \tag{5}$$

where C is the number of classes, p_i is the proportion of instances of class i in a subset, $H(\mathbf{X})$ is the entropy of the data set $\mathbf{X}$, $\mathbf{V}$ is the set of possible values for the feature, and X_v is the subset of instances with the feature value v.

Decision trees can be prone to overfitting, especially when the tree is deep and complex. To mitigate overfitting, pruning techniques can be applied, which involve removing branches of the tree that provide little improvement in the splitting criterion Chen et al. [2009]. Ensemble methods, such as Random Forests and Gradient Boosting, can also be used to improve the performance and robustness of decision trees by combining multiple trees into a single model Friedman [2001] Breiman [2001]

4   Methodology

The methodology used in this study comprises several steps, including hyperparameter identification, grid search, and cross-validation. This process was applied to each of the three machine learning models - Logistic Regression, Support Vector Machines, and Decision Trees - to optimize their performance and enable a fair comparison.

4.1   Hyperparameter Identification

The first step in the methodology was to identify the key hyperparameters for each machine learning model. Hyperparameters are external configurations that cannot be learned by the model during training, and they significantly influence the performance of the model. Identifying the most relevant hyperparameters for each model is crucial for successful optimization.

4.2   Grid Search

Grid search is a technique used to perform an exhaustive search of the hyperparameter space to find the best combination of hyperparameter values for a given model. The process involves defining a range of values for each hyperparameter and evaluating the model's performance with each combination. The best set of hyperparameters is the one that results in the highest performance metric, such as accuracy or F1 score.

4.2.1   Support Vector Machines (SVM)

For the SVM model, the following hyperparameters were identified:

- **C:** The regularization parameter, which controls the trade-off between maximizing the margin and minimizing classification errors. The range of values considered for C was [0.1, 0.5, 1, 3, 9, 100].
- **gamma:** The kernel coefficient for non-linear kernels, which controls the shape of the decision boundary. The range of values considered for gamma was [0.1, 1, 10].
- **kernel:** The function used to transform the input data into a higher-dimensional space. The possible kernels considered were ['linear', 'poly', 'rbf', 'sigmoid'].
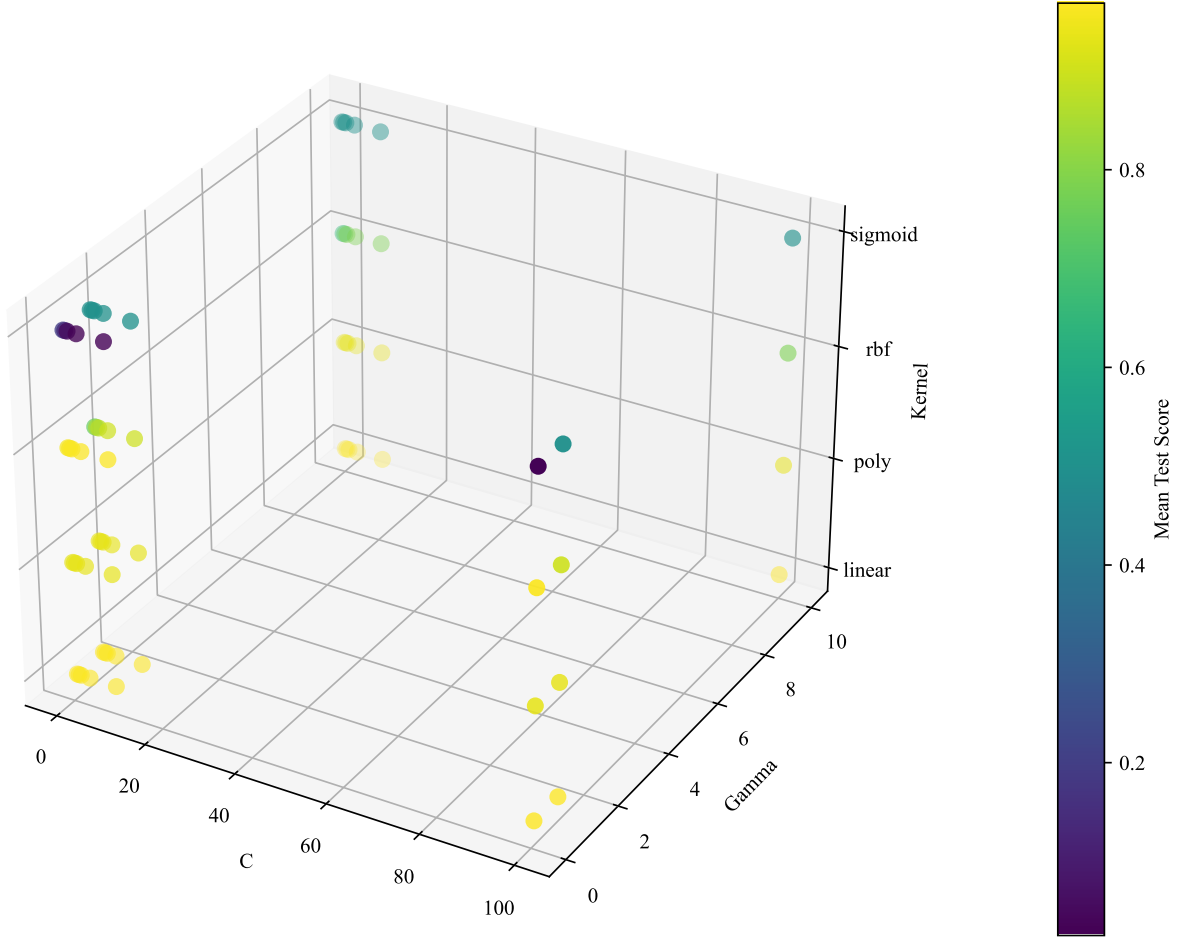
Figure 3: Grid search results for SVM

The grid search identified the best hyperparameters for the SVM model as:

- C: 9
- gamma: 0.1
- kernel: linear

4.2.2 Decision Trees

For the Decision Trees model, the following hyperparameters were identified:

- **Max Depth:** The maximum depth of the tree, which controls the complexity of the model and prevents overfitting. The range of values considered for Max Depth was [1, 2, 5, 10, 15, 20, 30, 50, 100].
- **Min Samples Split:** The minimum number of samples required to split an internal node. The range of values considered for Min Samples Split was [2, 5, 10, 15, 20, 30, 50].
- **Min Samples Leaf:** The minimum number of samples required to be at a leaf node. The range of values considered for Min Samples Leaf was [1, 2, 5, 10, 15, 20,30, 50].

The grid search identified the best hyperparameters for the Decision Trees model as:
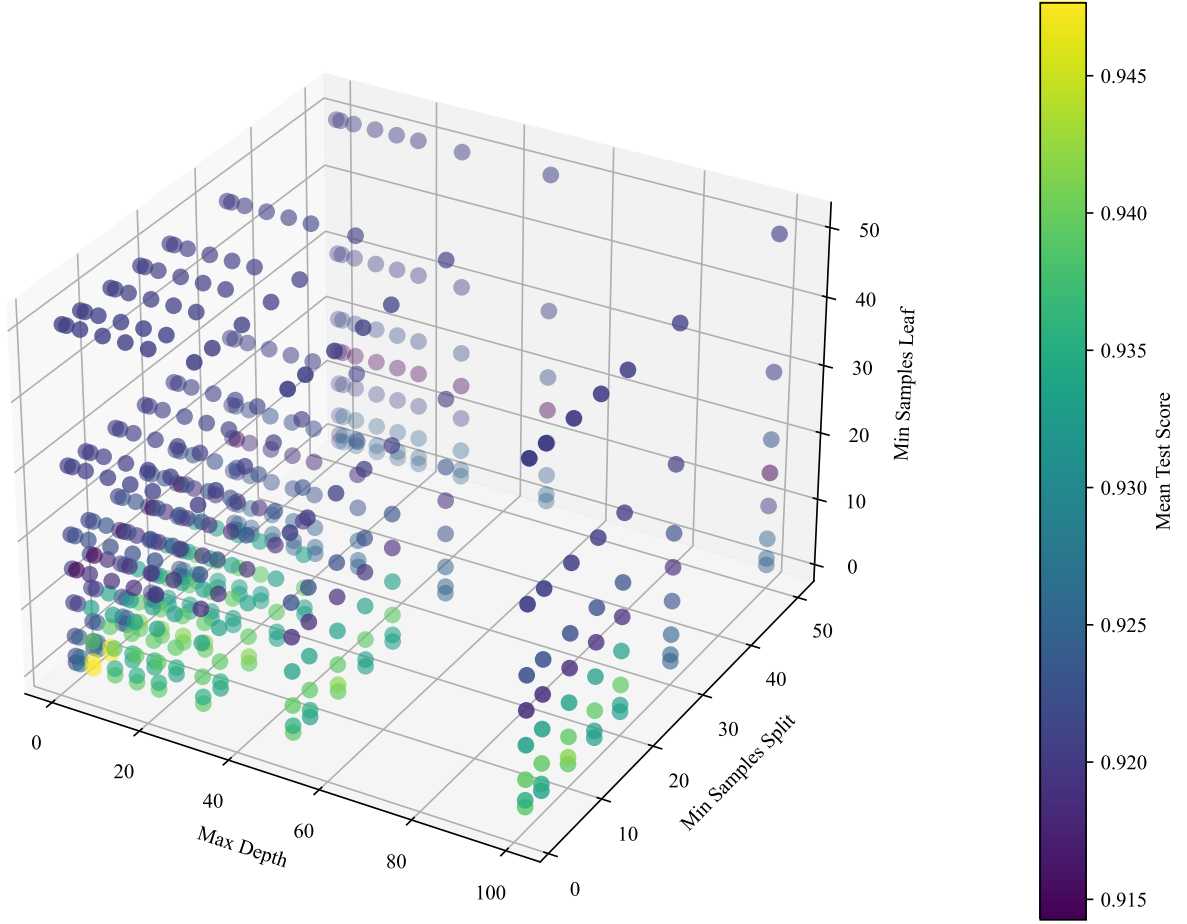
- Max Depth: 5

Figure 4: Grid search results for Decision Trees

- Min Samples Split: 2

- Min Samples Leaf: 2

4.2.3 Logistic Regression

For the Logistic Regression model, the following hyperparameters were identified:

- **C:** The inverse of regularization strength, with smaller values specifying stronger regularization. The range of values considered for C was [0 ... 10], incremented by 0.5.

- **Solver:** The algorithm used for optimization. The possible solvers considered were ['lbfgs', 'newton-cg', 'newton-cholesky', 'sag', 'saga'].

- **Penalty:** The norm used in the penalization. The possible penalties considered were ['none', 'l1', 'l2'].

The grid search identified the best hyperparameters for the Logistic Regression model as:

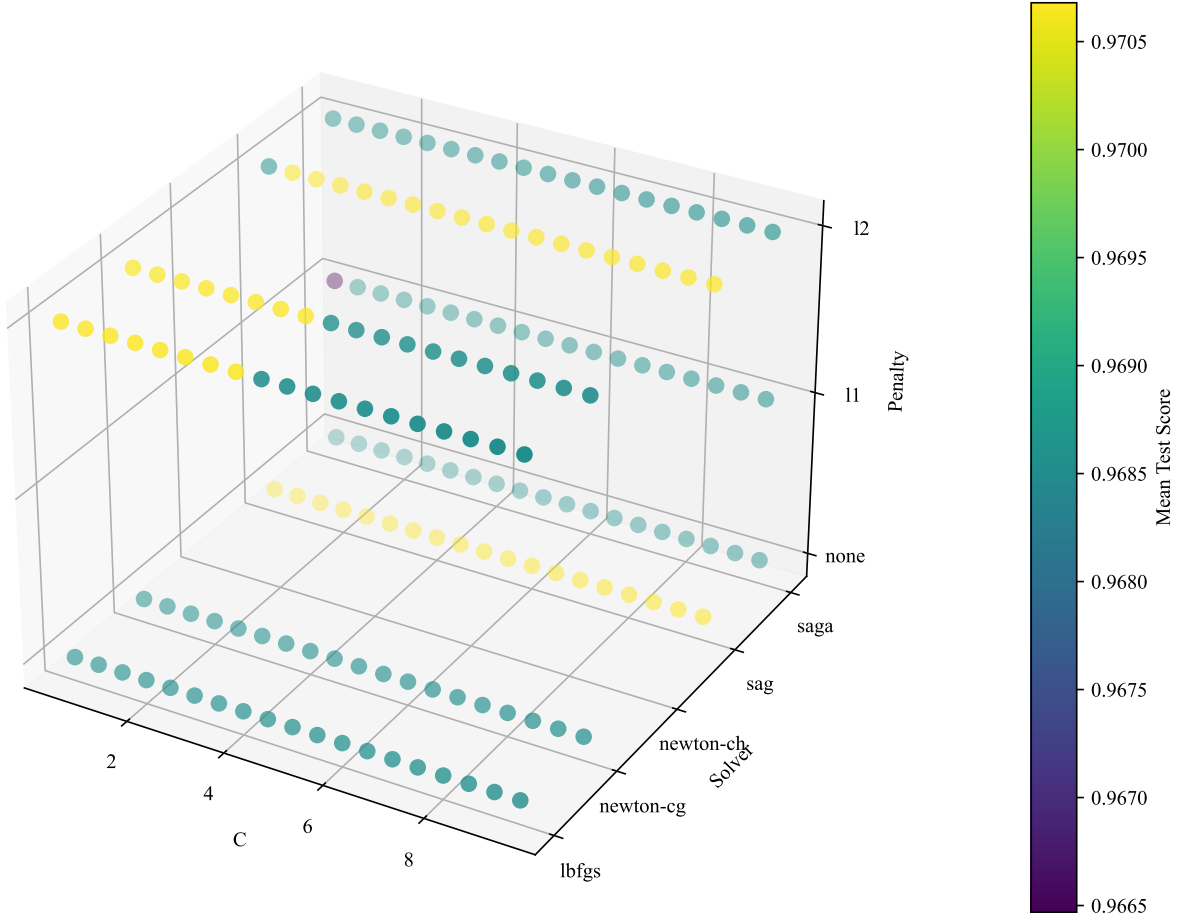- C: 0.5

- Solver: sag

- Penalty: none

Figure 5: Grid search results for Logistic Regression

4.3   Cross-Validation

To assess the performance of the optimized models, we employed 10-fold cross-validation. This technique involves dividing the dataset into ten equal parts or folds, training the model on nine of these folds, and testing the model on the remaining fold. This process is repeated ten times, with each fold used as the test set once. The mean and standard deviation of the performance metric across the ten iterations are used to evaluate the model's performance. Cross-validation helps to reduce overfitting and provides a more reliable estimate of the model's performance on unseen data.

5   Model Comparison and Analysis

The performance of the three models, Logistic Regression (LR), Decision Trees (DT), and Support Vector Machines (SVM), is illustrated in Figure 6. The figure presents the train and test scores, standard deviations, and computation time for each model. Based on these metrics, we can compare the advantages and disadvantages of each model and analyze why Logistic Regression performed the best. A summary table is given in Table 1.

5.1   Performance Comparison

- **Logistic Regression (LR):** LR demonstrates the highest performance, with a test score mean of 97.28% and the lowest test score standard deviation of 1.62%. The computation time for LR is 35.56 ms. With superior test score mean and standard deviation, LR is especially suitable for binary classification problems. Although

its computation time is moderate compared to other models, LR has a less complex hyperparameter tuning process than Support Vector Machines.

- **Decision Trees (DT):** Despite having the highest train score mean (98.19%), DT's test score mean is the lowest among the three models (93.73%). The test score standard deviation is 3.08%, and the computation time for DT is the least expensive at 8.97 ms. DT exhibits a lower test score mean and a tendency for overfitting, as indicated by the discrepancy between train and test scores. Nonetheless, DT models are suitable for both numerical and categorical data and require the least computation time among the three models. The complexity of tuning Decision Trees is moderate.

- **Support Vector Machines (SVM):** SVM exhibits consistent performance, with a test score mean of 96.44% and a test score standard deviation of 1.63%. However, it has a longer computation time of 82.33 ms compared to LR and DT. SVM models are effective in handling non-linear relationships and managing high-dimensional spaces with appropriate kernel functions. Nevertheless, they require the slowest computation time among the three models and have a more complex hyperparameter tuning process, necessitating careful parameter selection.
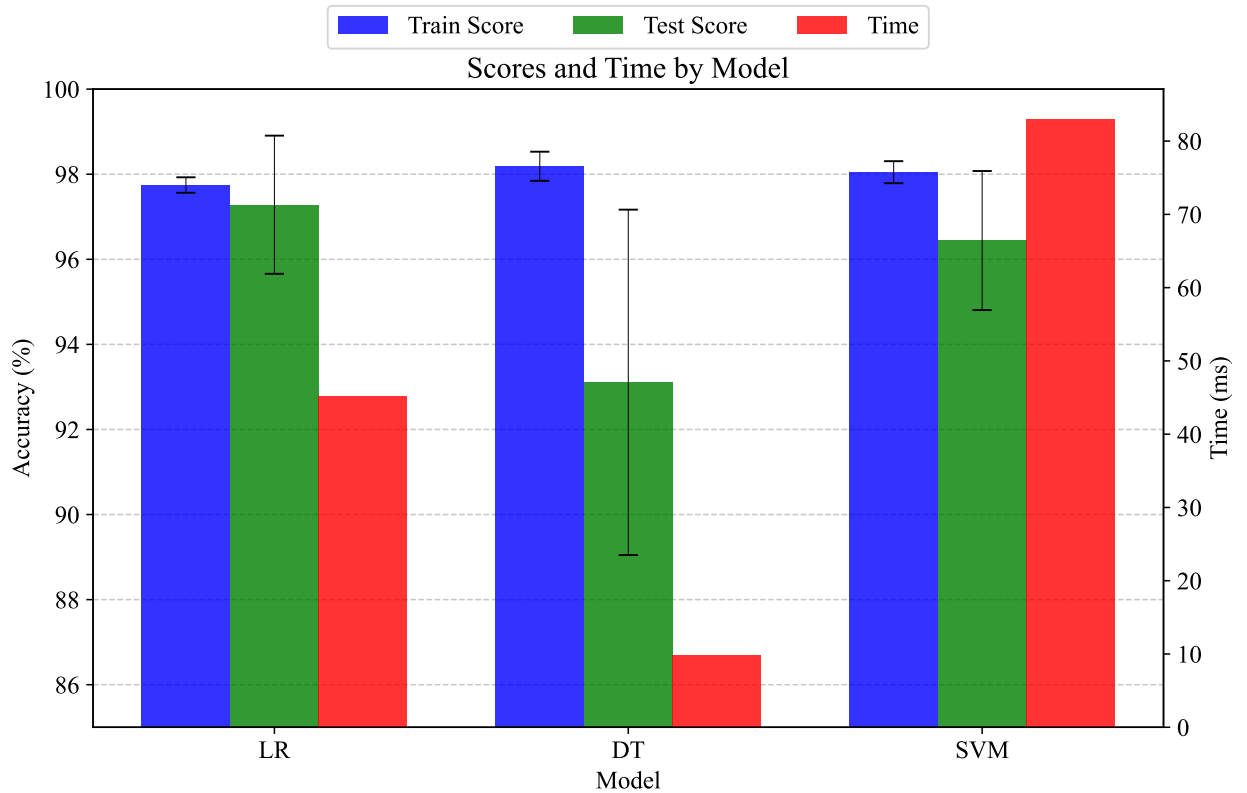

Figure 6: Model performance comparison

6 The Power of Simplicity: Linear Models Outperforming Complex Techniques

Our analysis above demonstrates that Logistic Regression (LR) outperforms both Decision Trees (DT) and Support Vector Machines (SVM) in the context of breast cancer diagnosis. This section delves deeper into the reasons behind LR's superior performance and the implications of using simpler models for classification tasks.

11

6.1 Interpretability and Complexity

One key advantage of LR over DT and SVM is its simplicity and interpretability. LR models are easy to understand, as the coefficients of the model directly indicate the effect of each feature on the outcome. This allows for better clinical decision-making, as the model's predictions can be directly traced back to the input features.

In contrast, DT and SVM models involve more complex decision boundaries and are harder to interpret. The complexity of these models can lead to overfitting, as seen in DT's discrepancy between train and test scores. The higher complexity of SVM also results in a longer computation time and more challenging hyperparameter tuning.

6.2 Generalization and Robustness

LR's linear decision boundary allows for better generalization, as demonstrated by its higher test score mean and lower standard deviation. By avoiding overfitting, LR models are more robust to variations in the data and can better handle new, unseen samples. This is crucial in the context of medical diagnosis, where models must be able to handle a wide range of patients and conditions.

In contrast, the more complex decision boundaries of DT and SVM may lead to overfitting and reduced generalization capabilities. The discrepancy between train and test scores for DT indicates the model's inability to generalize well to new data. SVM, while having consistent performance, still falls short of LR in terms of test score mean and has a longer computation time.

6.3 Scalability and Computational Efficiency

The computational efficiency of LR is another advantage over more complex models like SVM. Although DT has the least computation time among the three models, it suffers from lower test score mean and overfitting issues. LR, on the other hand, balances computational efficiency and performance, making it an ideal choice for large-scale applications and real-world settings.

SVM models, while effective at handling non-linear relationships and high-dimensional spaces, have the slowest computation time among the three models. This can be a significant drawback, particularly when dealing with large datasets or when rapid diagnosis is critical.

6.4 Tradeoffs and Model Selection

On the other hand, Decision Trees and Support Vector Machines showed certain advantages, such as faster computation times and the ability to handle non-linear relationships. However, these benefits were offset by other disadvantages, such as overfitting in the case of Decision Trees and slower computation times for Support Vector Machines. These factors demonstrate the importance of carefully considering the trade-offs between model performance and complexity when approaching classification tasks. In particular, the use of simpler models like Logistic Regression can be beneficial in real-world settings, where interpretability, generalization, and computational efficiency are critical. But in cases where non-linear relationships are present, more complex models like Support Vector Machines may be more appropriate. In any case, it is essential to consider the specific requirements of the task at hand and select the model that best fits these requirements.

Table 1: Comparison of Logistic Regression (LR), Decision Trees (DT), Support Vector Machines (SVM) Models Based on Specific Aspects

Model	Performance	Classification Capability	Computation Time	Hyperparameter Tuning
LR	Superior	Suitable for binary	Moderate	Less complex
DT	Lower test score mean Tendency for overfitting	Applicable to both numerical and categorical	Least expensive	Moderate complexity
SVM	Consistent	Effective in non-linear relationships	Slowest	Greater complexity

7 Conclusion

In this study, we have presented a comprehensive comparison of three machine learning techniques, namely Logistic Regression (LR), Decision Trees (DT), and Support Vector Machines (SVM), for the diagnosis of breast cancer using the UCI Machine Learning Repository dataset. Our results demonstrate that the simpler linear model, LR, outperforms the more complex DT and SVM techniques in terms of test score mean, standard deviation, and computational efficiency.

The superior performance of LR can be attributed to several factors. First, its simplicity and interpretability provide a clear understanding of the relationship between input features and the outcome, which is particularly valuable in the context of medical diagnosis. Second, LR's linear decision boundary allows for better generalization and robustness, ensuring reliable performance on new, unseen data. Lastly, the computational efficiency of LR offers advantages in terms of scalability and real-world applicability.

However, it is important to note that Decision Trees and Support Vector Machines showed certain advantages as well. DT models demonstrated faster computation times compared to LR, making them suitable for applications where efficiency is a priority. Additionally, SVM models excel in handling non-linear relationships and high-dimensional spaces with appropriate kernel functions, providing a valuable tool in scenarios where complex decision boundaries are present.

Nevertheless, these benefits of DT and SVM were offset by their respective disadvantages. DT models exhibited a tendency for overfitting, as indicated by the discrepancy between train and test scores. This limitation raises concerns about their generalization capabilities, particularly in the context of medical diagnosis, where reliable predictions on new data are crucial. On the other hand, SVM models required longer computation times compared to LR, which can be a significant drawback in large-scale applications or time-sensitive scenarios.

These considerations highlight the importance of carefully weighing the trade-offs between model performance and complexity when approaching classification tasks. While simpler linear models like LR offer advantages in terms of interpretability, generalization, and computational efficiency, they may not capture the intricacies of non-linear relationships as effectively as more complex techniques like SVM. Therefore, the choice of the model should be guided by the specific problem, the underlying data characteristics, and the desired trade-offs between performance, interpretability, and computational efficiency. In the context of privacy-preservation, simpler models such as logistic regression can often offer advantages over more complex models. One primary reason is their inherent transparency. Logistic regression models are easier to interpret, as they provide clear coefficients for each feature to understand the contribution of that feature to the model's decisions. This interpretability can help identify and mitigate privacy risks, such as the inadvertent inclusion of a sensitive or personally identifiable information (PII) feature. In contrast, complex models like deep neural networks often act as 'black boxes' where the intricate interactions between layers can obscure the contribution of individual features Peake and Wang [2018]. This opacity can make it difficult in ensuring that these models aren't learning or leaking sensitive information. Moreover, simpler models can perform similar to complex models for many tasks while requiring less data Hastie et al. [2009], potentially offering a better privacy-utility trade-off.

Future work could explore the following directions:

1. **Feature engineering and selection:** Investigate the impact of feature engineering and selection techniques on the performance of simpler linear models. This could include the use of domain knowledge to create new features, as well as the application of statistical and machine learning methods for feature selection.

2. **Ensemble methods:** Explore the potential benefits of combining simpler models, such as LR, with more complex techniques, like DT and SVM, through ensemble methods. This could lead to improved performance and robustness by leveraging the strengths of multiple models.

3. **Model interpretability:** Develop methods to improve the interpretability of more complex models, such as DT and SVM, while maintaining their performance. This could involve the use of visualization techniques, surrogate models, or local explanations to provide better insights into the decision-making process.

4. **Real-world deployment:** Conduct further studies on the deployment of simpler models like LR in real-world clinical settings, focusing on aspects such as integration with existing workflows, user acceptance, and ethical considerations.

In conclusion, this study emphasizes the power of simplicity in machine learning models for medical diagnosis. By demonstrating the advantages of simpler linear models like Logistic Regression over more complex techniques. We hope that this work will inspire further research in this direction and encourage the use of simpler models in real-world applications.

References

Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018.

Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

American Cancer Society. Cancer facts & figures 2021. URL https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2021/cancer-facts-and-figures-2021.pdf.

Tracy Onega, Elisabeth F Beaber, Brian L Sprague, William E Barlow, Jennifer S Haas, Anna NA Tosteson, Mitchell D. Schnall, Katrina Armstrong, Marilyn M Schapira, Berta Geller, et al. Breast cancer screening in an era of personalized regimens: A conceptual model and national cancer institute initiative for risk-based and preference-based approaches at a population level. *Cancer*, 120(19):2955–2964, 2014.

Brian L Sprague, Ronald E Gangnon, Veronica Burt, Amy Trentham-Dietz, John M Hampton, Robert D Wellman, Karla Kerlikowske, and Diana L Miglioretti. Prevalence of mammographically dense breasts in the united states. *JNCI: Journal of the National Cancer Institute*, 106(10), 2014.

Heidi D Nelson, Miranda Pappas, Amy Cantor, Jessica Griffin, Monica Daeges, and Linda Humphrey. Harms of breast cancer screening: systematic review to update the 2009 us preventive services task force recommendation. *Annals of internal medicine*, 164(4):256–267, 2016.

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017.

YoungJin Choi and YooKyung Boo. Comparing logistic regression models with alternative machine learning methods to predict the risk of drug intoxication mortality. *International journal of environmental research and public health*, 17(3):897, 2020.

Ernest Yeboah Boateng and Daniel A Abaye. A review of the logistic regression model with emphasis on medical research. *Journal of data analysis and information processing*, 7(4):190–207, 2019.

Ebru Aydındag Bayrak, Pınar Kırcı, and Tolga Ensari. Comparison of machine learning methods for breast cancer diagnosis. In *2019 Scientific meeting on electrical-electronics & biomedical engineering and computer science (EBBT)*, pages 1–3. IEEE, 2019.

A Frank. Uci machine learning repository. irvine, ca: University of california, school of information and computer science. *http://archive. ics. uci. edu/ml*, 2010.

W Nick Street, William H Wolberg, and Olvi L Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *Biomedical image processing and biomedical visualization*, volume 1905, pages 861–870. SPIE, 1993.

William H Wolberg, W Nick Street, and Olvi L Mangasarian. Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates. *Cancer letters*, 77(2-3):163–171, 1994.

Etta D Pisano, Laurie L Fajardo, Daryl J Caudry, Nour Sneige, William J Frable, Wendie A Berg, Irena Tocino, Stuart J Schnitt, James L Connolly, Constantine A Gatsonis, et al. Fine-needle aspiration biopsy of nonpalpable breast lesions in a multicenter clinical trial: results from the radiologic diagnostic oncology group v. *Radiology*, 219(3):785–792, 2001.

Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621, 1973.

Joseph A Cruz and David S Wishart. Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, 2:117693510600200030, 2006.

Ahmed M Abdel-Zaher and Ayman M Eldeib. Breast cancer classification using deep belief networks. *Expert Systems with Applications*, 46:139–144, 2016.

Vikas Chaurasia and Saurabh Pal. A novel approach for breast cancer detection using data mining techniques. *International journal of innovative research in computer and communication engineering (An ISO 3297: 2007 Certified Organization) Vol*, 2, 2017.

Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8–17, 2015.

A Inselberg. Parallel coordinates: Visual multidimensional geometry and its applications. 233 spring street, new york, ny 10013, 2008.

Stephen M Stigler. *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press, 1986.

Jiawei Han, Jian Pei, and Hanghang Tong. *Data mining: concepts and techniques*. Morgan kaufmann, 2022.

David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.

Bernhard Schölkopf, Christopher JC Burges, Alexander J Smola, et al. *Advances in kernel methods: support vector learning*. MIT press, 1999.

John Shawe-Taylor, Nello Cristianini, et al. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1:81–106, 1986.

Leo Breiman, JH Friedman, RA Olshen, and CJ Stone. Classification and regression trees. monterey, ca: Wadsworth & brooks, 1984.

Jie Chen, Xizhao Wang, and Junhai Zhai. Pruning decision tree using genetic algorithms. In *2009 International Conference on Artificial Intelligence and Computational Intelligence*, volume 3, pages 244–248. IEEE, 2009.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

Georgina Peake and Jun Wang. Explanation mining: Post hoc interpretability of latent factor models for recommendation systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2060–2069, 2018.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.