Previous work Review and Assessment

With the increasing impact of software applications on today businesses and activities, software attribute prediction such as effort estimation, maintainability, defect and quality classification are gaining growing interest from both academic and industry communities.

A research paper by[14] focuses on the increasing importance of software attribute prediction, such as effort estimation, maintainability, defect, and quality classification, due to a growing reliance on software applications. The paper highlights the limitations of conventional prediction algorithms such as decision trees, Bayesian methods, and artificial neural networks multilayer perceptron (ANN-MLP) when handling skewed and redundant defect datasets. The research introduces ensemble learning's voting mechanism as a solution, which assigns higher weights to successful individual classifiers, thereby mitigating the effects of feature irrelevance and redundancy. The paper's main objective is to demonstrate the positive effect of feature selection on the performance of defect classification. The findings suggest that for a dataset with 0.5% defective components, a classification accuracy of 99.5% can be achieved by classifying all components as non-defective, and an overall accuracy of 99% can be achieved by a binary classifier that classifies all data samples as the majority class. In the discussion section, the paper compares the results of the average probability ensemble (APE) model with two basic classifiers (W-SVMs and random forests). The W-SVMs classifier assigns weights in-

versely proportional to the class occurrence in the dataset, allowing for potentially better model fitting in the case of imbalanced datasets.

Another paper by[15] discusses the process of software defect prediction, which involves identifying likely flawed sections of software. The paper introduces a model that uses a base layer of Linear Discriminant Analysis (LDA), K Nearest Neighbors (KNN), and Generalized Linear Model with Elastic Net Regularization (GLMNet) and a top layer of Random Forest (RF). The results demonstrate that this model is capable of effectively handling PROMISE datasets, known for their noisy attributes and high dimensions. This paper is anticipated to make a significant contribution to the field of software defect prediction. According to the results, the Ensemble machine learning technique offers superior prediction accuracy for software defect prediction, providing a significant insight from this study. The proposed model achieved an overall prediction accuracy of 88.56% across all experimental datasets. Furthermore, the Mean Squared Error (MSE) of the proposed model is significantly lower than other models, demonstrating that the Ensemble technique effectively handles errors in the learning model caused by noise, bias, and variance. Therefore, the research suggests that ensemble machine learning models provide a robust solution for software defect prediction.

Methodology

Data Collection and Preprocessing

The NASA Metrics Data Program (MDP) software defect datasets were collected from various NASA software projects. These datasets encompass information on software modules each have their own attributes and a binary label indicating whether the module is defective or not. The datasets were preprocessed to remove any missing values by imputing the missing values by the mean of the column and remove duplicates found the datasets.

Brief Overview of the Task and Models Used

The task at hand involved a binary classification problem, a common type of task in the field of machine learning. Binary classification refers to the process of classifying elements into two distinct groups based on certain characteristics or features. In this context, the task was to predict whether a data point belongs to class 0 (majority class) or class 1 (minority class). To accomplish this task, four different machine learning models were employed. These models represent a variety of algorithmic approaches and were chosen for their distinctive strengths in handling classification tasks.

Support Vector Machine

The Support Vector Machine (SVM) is a widely utilized algorithm in the field of machine learning, particularly for classification and regression tasks. According to[4], who first introduced the concept, SVM is a classifier which performs its task by constructing a hyperplane in a multi-dimensional space that separates data points of different classes. The SVM model operates on the principle of maximizing the margin, which is the distance between the hyperplane (decision boundary) and the nearest data points from different classes, known as support vectors.

The hyperplane is defined as: $w \cdot x + b = 0$ where...

w is the weight vector.

x is the input data vector.

and b is the bias term.

The decision function for SVM is given by:

$$f(x) = \text{sign}(w \cdot x + b)$$

where...

$f(x)$ predicts the class label of input vector x.

$sign(\cdot)$ returns the sign of its argument.

In cases where the data is not linearly separable in its original feature space, SVM can use the kernel trick to implicitly map the data to a higher-dimensional space where it becomes separable.

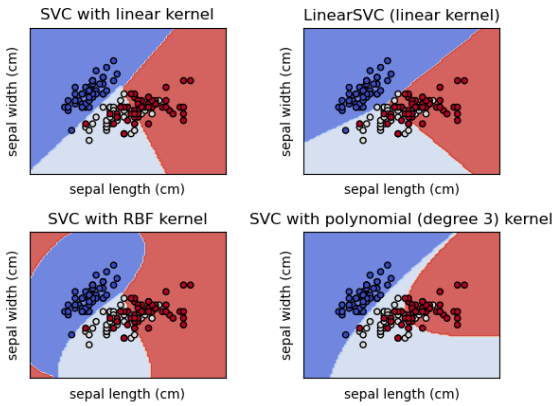

Figure 1: SVM Classifier

Linear Kernel

Equation: $K(x_i, x_j) = x_i^T x_j$

Explanation: The linear kernel represents the dot product between the input feature vectors x_i and x_j. This kernel is used when the data is linearly separable in the original feature space.

Polynomial Kernel

Equation: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d$

Explanation: The polynomial kernel maps the input data into a higher-dimensional space using a polynomial function. The hyperparameters d and r are user-defined positive integers that control the degree of the polynomial and a constant term, respectively. The parameter γ is a scaling factor that influences the dot product.

Gaussian (RBF) Kernel

Equation: $K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right)$

Explanation: The Radial Basis Function (RBF) kernel measures the similarity between two data points using the Gaussian distribution. It implicitly maps the data into a high-dimensional space, making it suitable for handling nonlinear relationships. The hyperparameter γ controls the width of the Gaussian kernel.

Sigmoid Kernel

Equation: $K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r)$

Explanation: The sigmoid kernel uses the hyperbolic tangent function to map the data into a higher-dimensional space. This kernel can be useful in some cases, but it is generally less commonly used compared to linear, polynomial, and Gaussian kernels.

Laplacian Kernel

Equation: $K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|}{\gamma}\right)$

Explanation: The Laplacian kernel measures the similarity between two data points using the Laplace distribution. It is similar to the Gaussian kernel but has a sharper peak, making it robust to outliers.

This characteristic makes SVM highly effective in high-dimensional spaces and in situations where the number of dimensions is greater than the number of samples[11]. In addition to performing linear classification, SVMs can also effectively handle non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces[21]. In terms of binary classification, the SVM model seeks a hyperplane that separates the data into two classes while maximizing the margin between the classes.

Random Forest

Random Forest is a versatile and popular machine learning algorithm known for its simplicity and diversity. It was first introduced by[1] as an extension of the decision tree algorithm. A Random Forest operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

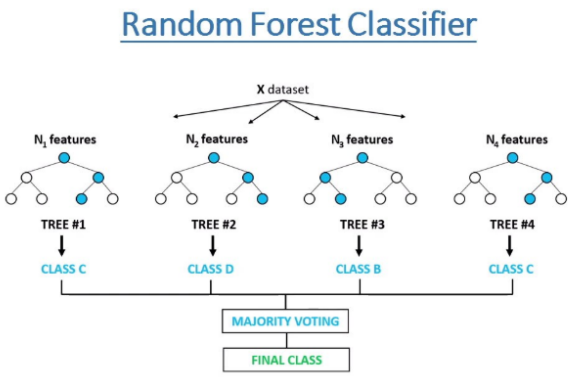

Figure 2: Random Forest

This idea of combining several models to improve the predictive performance is known as ensemble learning ([5]). Random Forest attempts to mitigate the high variance problem of individual decision trees, where small changes in the training set can result in significantly different tree structures. By averaging the results of a collection of de-correlated trees, it reduces the risk of overfitting and typically provides a better predictive performance ([20]). Random Forest also has an inherent feature selection mechanism, as it ranks features (variables) based on their ability to improve the purity of the node, measured by the Gini impurity or entropy ([7]).

Logistic Regression

The Logistic Regression model, despite its name, is a powerful tool for binary classification tasks. This model calculates the probability that a given data point belongs to a certain class by applying the logistic function to a linear combination of features[12]. One of the main advantages of Logistic Regression is its interpretability. Each feature is assigned a coefficient that describes its relative importance and direction of association with the assumes that there is a linear relationship between the log-odds of the dependent variable and the independent variables. It also requires that the observations be independent, and the absence of multicollinearity among the independent variables[17]. The equation of logistic function or logistic curve is a common "S" shaped curve defined by the below equation. The logistic curve is also known as the sigmoid curve[22]. the formula of logistic function:

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

where..

$$-(\beta_0 + \beta_1 x)$$

is linear function

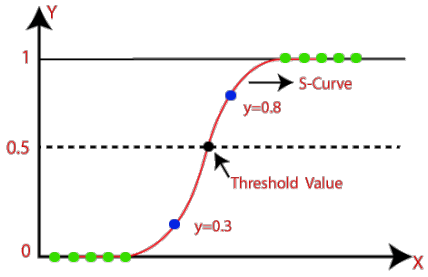

Figure 3: Logistic function

Ensemble Model

Ensemble Models are powerful machine learning tools that combine multiple base models to improve prediction accuracy and robustness over a single model. The central idea behind ensemble models, often referred to as the "Wisdom of the Crowd", is that a group of weak learners can come together to form a strong learner[6].

There are several types of ensemble methods, including Bagging, Boosting, and Stacking. Bagging, or Bootstrap Aggregating, involves training each model in the ensemble using a randomly drawn subset of the training set. Boosting is a sequential process, where each subsequent model attempts to correct the mistakes of the previous models. Stacking involves training a model to combine the predictions of several other models[25]. Ensemble models have been shown to significantly increase accuracy on a variety of tasks, including both regression and classification problems[16].

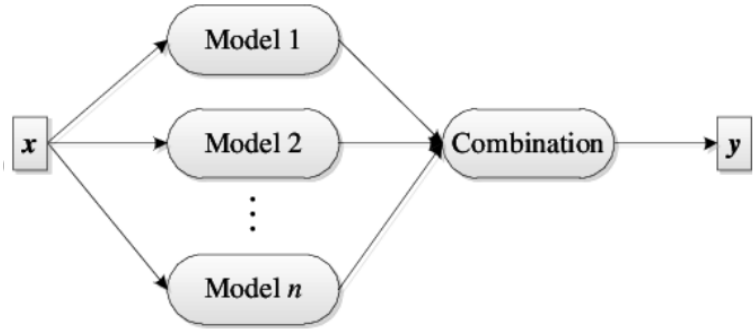

Figure 4: Ensemble Model

They tend to be more robust to noise, outliers, and overfitting, as they average out biases, reduce variance, and are not likely to model random noise in the output data[26]. However, ensemble models can be computationally expensive and may require more time to train and predict than individual models. Also, they may suffer from complexity, which may make them harder to interpret than individual models[19].

Performance Metrics

In machine learning, the performance of a model is evaluated based on certain metrics that depend on the nature of the task whether it is a classification, regression, or clustering task. This paper focuses on the metrics used to evaluate classification models, including accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic curve (AUROC). Accuracy, the ratio of correctly predicted observations to the total observations, is the most straightforward measure but can be misleading in case of imbalanced classes[18].

$$\text{Accuracy } = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

Precision, the ratio of correctly predicted positive observations to the total predicted positives, is an essential metric when the cost of false positives is high[23].

$$\text{Precision } = \frac{TP}{(TP + FP)}$$

Recall measures the ratio of correctly predicted positive observations to all observations in the actual class, and is crucial when the cost of false negatives is high[23].

$$\text{Recall } = \frac{TP}{(TP + FN)}$$

The F1-score is the weighted average of precision and recall and is typically more useful than accuracy, particularly for uneven class distributions[24].

$$F1 \text{ Score } = \frac{2TP}{(2TP + FP + FNN)}$$

Feature Selection

Feature selection is a crucial step in the machine learning pipeline. It is the process of selecting the most relevant features from the original dataset to use in model training. The aim is to enhance the performance of the machine learning model By using feature selection, we can simplify data entry, reduce size, and improve model performance in several ways according to[10]:

- Reduce overfitting: Feature selection helps prevent overfitting by focusing on the most important features that contain the most predictive information.
- Faster training: Reducing the number of features shortens the training time of a machine learning as there is less data to process in the model.
- Enhanced interpretation: We can improve the interpretation of models by selecting the most important features. This means we can better understand which features are relevant to the model's predictions and extract insights from the model's behavior.

According to[13], The wrapper method is a type of feature selection technique that relies on the performance of a machine learning model to evaluate the importance of features. It involves training a model on a subset of features, evaluating its performance, and using that information to decide whether to add or remove features. This process repeats until it reaches an optimal set of features. The wrapper method is computationally expensive as it involves evaluating the performance of a machine learning model for every possible combination of features. However, it often provides a better performing feature set compared to other methods, like filter methods, because it takes into account the interaction of features. There are different types of wrapper methods, including forward selection, backward elimination, and recursive feature elimination. Forward selection

starts with an empty set and adds one feature at a time. Backward elimination starts with all features and removes one feature at a time. Recursive feature elimination is a greedy optimization algorithm that aims to find the best performing feature subset.

Recursive Feature Elimination with Cross-Validation (RFECV)

Chi-Squared Feature Selection

RFECV (Recursive Feature Elimination with Cross-Validation) is a powerful technique for feature selection in machine learning that fits a model and removes the weakest feature (or features) until the specified number of features is reached as mentioned by[9]. Features are ranked and by recursively eliminating a small number of features per loop, RFECV attempts to eliminate dependencies and collinearity that may exist in the model. It combines the strengths of recursive feature elimination (RFE) and cross-validation to provide a robust and efficient method for selecting the most important features and optimizing the performance of a machine learning model.

RFECV then selects the optimal number of features based on the cross-validation score. The cross-validation score generally increases with the number of features. However, the score may also increase due to overfitting. RFECV solves this problem by selecting the number of features for which the cross-validation score is maximum. RFECV works by iteratively eliminating the least important features in a dataset, while using cross-validation to evaluate the performance of the model. The process can be broken down into the following steps:

- Initialize the model with all the features: The first step is to initialize the model with all the features in the dataset.

- Perform cross-validation: The next step is to perform cross-validation on the dataset, using the initialized model. This involves splitting the dataset into training and testing sets, and evaluating the model's performance on the testing sets.
- Eliminate the least important features: Based on the performance of the model in the cross-validation step, the least important features are eliminated from the dataset.
- Repeat steps 2–3: Steps 2–3 are repeated until a stopping criterion is met, such as a maximum number of iterations or a minimum number of features.
- Evaluate the final model: The final model is evaluated on the entire dataset to determine its performance.

according to[3], RFECV has several advantages over other feature selection. RFECV can improve the accuracy of a machine learning model by selecting the most important features. It is computationally efficient, as it only requires a single pass through the dataset to eliminate the least important features. Also, provides interpretable results, as it eliminates features one at a time, allowing for feature importance to be easily understood. However, RFECV needs a specified or calculable measure of importance provided by the estimator to perform its operations. Therefore, it doesn't work with all kinds of models. For instance, it won't work directly with models like KNN, SVM with non-linear kernel as these models do not provide a straightforward measure of feature importance.

Model performance analysis

Four different machine learning models were evaluated in this study, which included Support Vector Machine (SVM), Random Forest, Logistic Regression, and an Ensemble model. The performance of these models was assessed based on a binary classification task.

PC1 Dataset Model Performance Analysis

An accuracy of 92% was achieved by the SVM model. However, the model's performance was found to be poor when classifying the minority class (1), with precision, recall, and f1-score all recorded at 0.00. On the other hand, the majority class (0) was well-classified by the model with a precision of 0.93, recall of 0.99, and f1-score of 0.96. The Random Forest model reported an overall accuracy of 91%. A relatively poor performance was observed for the minority class (1) with a precision of 0.20, recall of 0.07, and f1-score of 0.11. The majority class (0) was well-classified with a precision of 0.93, a recall of 0.98, and an f1-score of 0.95. The Logistic Regression model showed an overall accuracy of 90%. It was noted that the model failed to classify any instance of the minority class (1) correctly, resulting in a precision, recall, and f1-score all at 0.00. A good performance was observed for the majority class (0) with a precision of 0.92, a recall of 0.97, and an f1-score of 0.95. The Ensemble model, with an overall accuracy of 91%, also failed to correctly classify any instance of the minority class (1), resulting in precision, recall, and f1-score of 0.00. A good performance was observed for the majority class (0) with a precision of 0.93, a recall of 0.98, and an f1-score of 0.95.

Model	Class	Precision	Recall	F1-score
SVM	Class 0	0.93	0.99	0.96
SVM	Class 1	0.00	0.00	0.00
Random Forest	Class 0	0.93	0.98	0.95
Random Forest	Class 1	0.20	0.07	0.11
Logistic Regression	Class 0	0.92	0.97	0.95
Logistic Regression	Class 1	0.00	0.00	0.00
Ensemble Model	Class 0	0.93	0.98	0.95
Ensemble Model	Class 1	0.00	0.00	0.00

Table 1: PC1 Dataset Model Performance Analysis.

In conclusion, although high accuracy was demonstrated by all models, difficulties were encountered in effectively classifying the minority class (1). These results indicated the presence of class imbalance in the dataset, a common problem in machine learning tasks. Despite high accuracy was demonstrated by all models, difficulties were encountered in effectively classifying the minority class (1). Strategies such as resampling techniques suggested to address this issue. The model's performance should not only be assessed on overall accuracy but also on its ability to identify each class accurately, especially when dealing with imbalanced datasets.

Proposed Enhancement to Defect prediction Model

The Problem with Imbalanced Data Sets

Class imbalance is a common problem in machine learning classification where there are a disproportionate ratio of observations in each class. Class imbalance can be found in many different areas including medical diagnosis, spam filtering, and fraud detection. The main issue with class imbalance is that most machine learning algorithms work best when the number of samples in each class are about equal. This is because most algorithms are designed to maximize accuracy and reduce error.

Synthetic Minority Over-sampling Technique (SMOTE)

Synthetic Minority Over-sampling Technique, or SMOTE, is a popular algorithm to create synthetic observations of the minority class. It was presented by [2].

SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line. Specifically, a random example from the minority class is first chosen. Then k of the nearest neighbors for that example are found (typically k=5). A randomly selected neighbor is chosen and a synthetic example is created at a randomly selected point between the two examples in feature space.

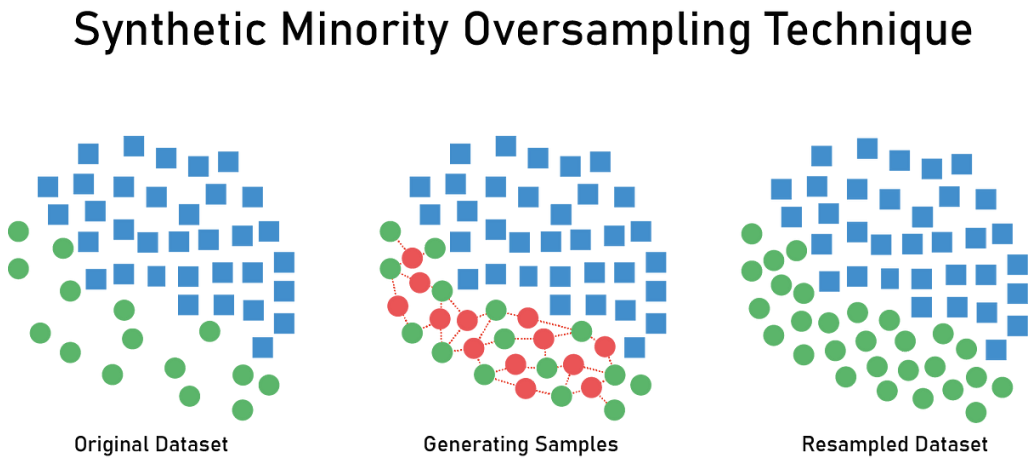

Figure 5: SMOTE

The algorithm can be summarized as follows:

- For each sample in the minority class, calculate the k nearest neighbors.
- From the k neighbors, a neighbor is randomly selected.
- A synthetic sample is created by choosing a point between the two instances in feature space.

This process is repeated until the data is balanced.

Efficacy and Limitations of SMOTE

According to [8], the main advantage of SMOTE is that it can improve the performance of the minority class by generating the synthetic examples that are quite similar to the existing observations in the minority class. However, one potential drawback of SMOTE is that it can increase the likelihood of overfitting since it generates synthetic examples without considering the majority class. New synthetic examples could be generated that are quite similar, or even identical, to existing examples. Furthermore, synthetic examples are generated without considering the majority class, possibly resulting in ambiguous examples if there is a strong overlap for the classes.

Implemetation

We utilized the SMOTE algorithm implemented in the imbalanced-learn library in python, and setting number of neighbors to 5 which is the default and it's sufficient for most cases.

Results

We trained SVM, Random Forest, Logistic Regression, and Ensemble models on the balanced dataset and evaluated their performance. The results showed a significant improvement in the performance of all models, particularly in the classification of the minority class (1). The best performance was observed with the Random Forest model, which achieved a precision of 0.92, recall of 0.96, and f1-score of 0.94 for the minority class (1).

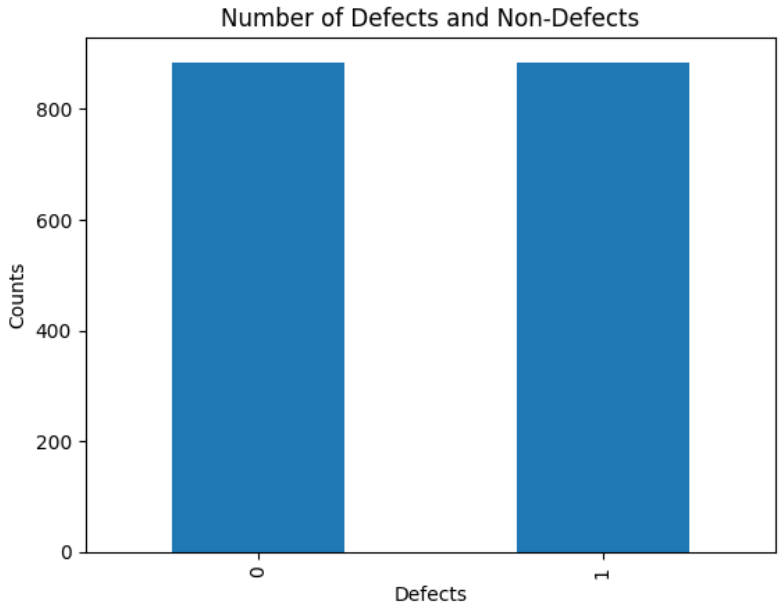

Figure 6: Data distribution after using SMOTE

Model	Class	Precision	Recall	F1-score
SVM	Class 0	0.93	0.99	0.96
SVM	Class 1	0.00	0.00	0.00
Random Forest	Class 0	0.93	0.98	0.95
Random Forest	Class 1	0.20	0.07	0.11
Logistic Regression	Class 0	0.92	0.97	0.95
Logistic Regression	Class 1	0.00	0.00	0.00
Ensemble Model	Class 0	0.93	0.98	0.95
Ensemble Model	Class 1	0.00	0.00	0.00
SVM SMOTE	Class 0	0.84	0.83	0.83
SVM SMOTE	Class 1	0.82	0.83	0.82
Random Forest SMOTE	Class 0	0.96	0.92	0.94
Random Forest SMOTE	Class 1	0.92	0.96	0.94
Logistic Regression SMOTE	Class 0	0.77	0.85	0.81
Logistic Regression SMOTE	Class 1	0.82	0.73	0.77
Ensemble Model SMOTE	Class 0	0.85	0.87	0.86
Ensemble Model SMOTE	Class 1	0.86	0.83	0.85

Table 2: PC1 Dataset Model Performance Analysis on Balanced Data.

Bibliography

[1] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[2] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[3] Qi Chen, Zhaopeng Meng, and Ran Su. Werfe: A gene selection algorithm based on recursive feature elimination and ensemble strategy. *Frontiers in Bioengineering and Biotechnology*, 8, 2020.

[4] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, September 1995.

[5] Thomas G. Dietterich. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, pages 1–15, 2000.

[6] Thomas G Dietterich. Ensemble methods in machine learning. *Multiple classifier systems*, 1857:1–15, 2000.

[7] Ramón Díaz-Uriarte and S. Amelia De Andres. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(1):1–13, 2006.

[8] Alberto Fernández, Salvador García, Francisco Herrera, and Nitesh V Chawla. Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, 61:863–905, 2018.

[9] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.

[10] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[11] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28, 1998.

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

[13] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.

[14] Issam H. Laradji, Mohammad Alshayeb, and Lahouari Ghouti. Software defect prediction using ensemble learning on selected features. *Information and Software Technology*, 58:388–402, 2015.

[15] Adv Lear, Emmanuel Dada, David Oyewola, Stephen Joseph, Ali Dauda, Stephen Bassi, and Ali Baba. Ensemble machine learning model for software defect prediction. *Advances in Machine Learning and Artificial Intelligence*, 2:11–21, 07 2021.

[16] David Opitz and Richard Maclin. Popular ensemble methods: An empirical study. *Journal of artificial intelligence research*, 11:169–198, 1999.

[17] Chao-Ying Joanne Peng, Kristin L Lee, and Gary M Ingersoll. An introduction to logistic regression analysis and reporting. *The journal of educational research*, 96(1):3–14, 2002.

[18] David Martin Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*, 2011.

[19] Lior Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1-2):1–39, 2010.

[20] Omer Sagi and Lior Rokach. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249, 2018.

[21] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In Wulfram Gerstner, Alain Germond, Martin Hasler, and Jean-Daniel Nicoud, editors, *Artificial Neural Networks — ICANN'97*, pages 583–588, Berlin, Heidelberg, 1997. Springer Berlin Heidelberg.

[22] H. Sharma. Logistic regression - python implementation from scratch without using sklearn, July 2022.

[23] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009.

[24] Cornelis Joost Van Rijsbergen. *Information retrieval*. Butterworth-Heinemann, 1979.

[25] David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

[26] Zhi-Hua Zhou. *Ensemble methods: foundations and algorithms*. CRC press, 2012.