Accredited SINTA 2 Ranking
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online at: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 8 No. 2 (2024) 189 - 198 e-ISSN: 2580-0760

Comparison of Genetic Algorithm and Recursive Feature Elimination on High Dimensional Data

Yoga Pristyanto[1], Dipa Wirantanu[2]
[1,2]Department of Information System, Faculty of Computer Science, Universitas Amikom, Yogyakarta, Indonesia
[1]yoga.pristyanto@amikom.ac.id, [2]dipaw@students.amikom.ac.id

Abstract

The use of big data in companies is currently used in file processing. With large capacity files, it can affect the performance in terms of time in the company so to overcome the problem of high-dimensional data, feature selection is used in selecting the number of features. On the wdbc dataset with 30 features and 569 data, feature selection is performed using the Recusive Feature Elimination (RFE) and Genetic Algorithm (GA) models. Then a comparison of evaluation values is made to determine which feature selection is best for solving the problem. From the 14 tables of evaluation results and discussion in tables 1 to 14, it is found that in the evaluation of accuracy, and the use of weighted macros on precision, recall and f1 score, using GA selection features has slightly higher results than RFE so it is concluded that GA selection features are better at solving problems in high-dimensional data.

Keywords: feature selection; high-dimensional data; recursive feature elimination; genetic algorithm

How to Cite: Y. Pristyanto and D. Wirantanu, "Comparison of Genetic Algorithm and Recursive Feature Elimination on High Dimensional Data", J. RESTI (Rekayasa Sist. Teknol. Inf.), vol. 8, no. 2, pp. 189 - 199, Mar. 2024.
DOI: https://doi.org/10.29207/resti.v8i2.5375

1. Introduction

The development of the scope of work in the data archive section of the company that manages a large amount of data focuses on new situations and tasks so that it needs to adapt to current developments. In the process of development, new problems and changes arise, increasing the response and responsibility as well as job security risks in prevention and control to promote the level of file processing specialization to accelerate the development of corporate archives to contribute positively and effectively [1]. File processing in today's big data era involves combining information from multiple sources with different representations because the diversity and quality of data vary greatly from one source to another, even within the same field [2]. So in other words, the object of research discussed in this study is the use of feature selection on data that has many data features. This research is considered important because, with fewer features tested, it can improve accuracy results [3].

Research [3] examines feature selection using Heuristic Search Wrappers with optimization using Autoencoder and Model-based elimination, on the Wisconsin Breast Center dataset using 20% testing data obtained an accuracy of 96. 78% with the AMBER method, then the unique difference with this research is that two selection features are used between genetic algorithms and recursive feature elimination with 10%, 20%, and 30% testing data, for the algorithm methods used there are as many as six namely Logistic Regression, Support Vector Machines, Naïve Bayes, K-Nearest Neighbors, Random Forest and AdaBoost, which are expected to improve accuracy in previous researchers. Then the discussion will be detailed in the evaluation of accuracy, precision, recall and f1 score.

Research [4] examines using CCEA (Cooperative Co-Evolutionary Algorithm) selection features and machine learning algorithms that are the same as this research are Naïve Bayes, SVM, KNN, RF, and LR tested on WDBC datasets. However, for split data in research [4] there is no further information so even so, later in the discussion section researchers will compare with 10% testing data in each evaluation. The accuracy, precision, recall and f1 scores for the NB algorithm were 93.15%, 93.10%, 93.10%, and 93.10% respectively, as well as for the SVM algorithm 92.79%, 93%, 92.80%, and 92.70% respectively, for the KNN

Received: 21-08-2023 | Accepted: 07-01-2024 | Published Online: 29-03-2025

algorithm 92. 44%, 92.40%, 92.40%, and 92.40%, for the RF algorithm they are 93.32%, 93.30%, 93.30%, and 93.30%, respectively, for the last algorithm LR they are 92.44%, 92.40%, 92.40%, and 92.40%, respectively.

Research [5] examines the use of VIM (Variable Importance Measure) feature selection and the use of HCRF (Hierarchical Clustering Random Forest) for classification. The algorithms used are Decision Tree, AdaBoost, Random Forest and HCRF. The training data used is 30%. On the WDBC dataset, the consecutive accuracy results based on the above algorithms are 91.46%, 93.33%, 96.37%, and 97.05% and the precision results are 87.67%, 91.15%, 95.97%, and 97.32% respectively. The unique difference in this study is to compare the accuracy and precision results on 30% testing data based on the AdaBoost and Random Forest algorithm models to find out whether the VIM selection feature is more effective than the GA (Genetic Algorithm) and RFE (Recursive Feature Elimination) selection features or not.

Research [6] examines the use of PCA selection features with classification using PNN (Probabilistic NN), LDA (Linear DA) and KNN algorithms. The testing data used is 40%, 25%, and 10%. Because the 10% testing data used is the same as this study the results are only displayed in the 10% testing data section, the following are the results of the research [6], for the KNN algorithm on the WDBC dataset, the accuracy is 97.77%, sensitivity or recall is 98.08%, and precision or positive predictive value is 95.78%. The difference that wants to be carried out in this study is to compare the results of the KNN algorithm on 10% testing data and evaluation, especially on accuracy, recall and precision.

The purpose of this study is to determine the best feature selection between Genetic Algorithm (GA) and Recursive Feature Elimination (RFE) by displaying the results as well as discussing the results of the evaluation of accuracy, precision, recall and f1 score on the six algorithms used, the algorithms are Logistic Regression, Support Vector Machines, Naïve Bayes, K-Nearest Neighbors, Random Forest and AdaBoost based on 10%, 20%, and 30% testing data that have high dimensions in the data.

The method used is part of Supervised Classification with an evaluation in the form of an average class label, and an average evaluation value in the form of accuracy, precision, recall, F-measure, and so on [7]. Then the method used is Wrapper feature selection, namely Recursive Feature Elimination (RFE) with Genetic Algorithm (GA). The logic of the wrapper method is the first process by creating a learning model using a subset of features and by repeatedly (backward or foreward) training a prediction model using these features. Based on the results of the model, the irrelevant ones are removed [8]. Wrapper methods are used to generate more model-based solutions [9]. One of the models is the RFE method [10]. RFE itself is used to improve classification performance by optimizing a subset of features [11] by giving each feature a weight to determine the ranking based on the importance of the feature [12] and of course also based on some specific machine learning method chosen [11]. Since its first implementation in 1989, the Genetic Algorithm (GA) has served as a search algorithm. Then this algorithm in the last three years made fantastic developments in research on Feature Selection (FS), with global search capabilities, able to optimize population-based metaheuristics [13].

In this paper, the methods used are Recursive Feature Elimination (RFE) and Genetic Algorithm (GA) as feature selection and Logistic Regression (LR), Support Vector Machines (SVM), Naïve Bayes, K-Nearest Neighbor (KNN), AdaBoost, and Random Forest (RF) as classification. In RFE features, X and y are initialized as well as converted from string to float on y, where X is a total of 30 features and y is the class, then feature selection is performed using the RFE library with parameters (SVC estimator with linear kernel, total features selected, and step of 1). After that, split data is performed with X new RFE features and y. Then classification is carried out using the six algorithms above using weighted average and macros on precision, recall and f1, but for the six algorithms without any parameters. In the GA feature, first initialize X and y, where X is a total of 30 features and y is the class dataset, then perform feature selection with the GAFeatureSelectionCV library with parameters (estimator using SVC with gamma auto, generations 80, and cv as many as 3) then split data with X new GA features and y. Then do split data with X new features and y. Then classification is carried out using the six algorithms above using average weighted and macros on precision, recall and f1, but for the six algorithms without any parameters.

It is hoped that this research can be useful for further researchers in terms of comparing the use of GA and RFE selection features with the algorithms used are Logistic Regression, Support Vector Machines, Naïve Bayes, K-Nearest Neighbors, Random Forest and AdaBoost in terms of accuracy, precision, recall and f1 score results based on 10%, 20%, and 30% testing data on high-dimensional data.

The main contributions of this research are: (1) Logistic Regression, Support Vector Machines, Naïve Bayes, K-Nearest Neighbors, Random Forest and AdaBoost algorithms can be proposed to detect breast cancer, (2) Can find out which selection feature is the best between GA and RFE by minimizing the number of features that are to be small so that it can improve accuracy results.

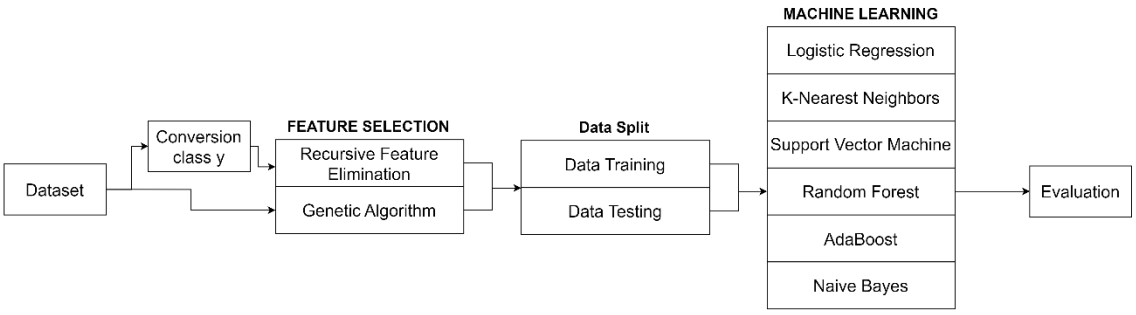

Figure 1. High-dimensional Data Classification Evaluation System

2. Research Methods

Figure 1 shows the research method for evaluating the use of feature selection for high-dimensional data classification based on the average of all classes in the dataset.

The initial stage is the dataset, the dataset used is wdbc, and this dataset comes from the UC Irvine Machine Learning Repository. The Wdbc dataset stands for Wisconsin Diagnostic Breast Center so it discusses breast diagnosis in Wisconsin. The wdbc dataset has a total of 32 columns, including 30 columns of features to be tested, one column is a class consisting of Benign (B) and Malignant (M) and the last column is the id not used.

At the Conversion class y stage, the X (all features of the wdbc dataset) and y (class of the wdbc dataset) data are first determined, which are then converted only to the RFE selection feature model because the class y of the WDBC dataset in the form of B (benign) and M (malignant) is a string so it is converted to B: 0, and M: 1.

The two features selection (FS) used are Recursive Feature Elimination (RFE) and Genetic Algorithm (GA). Both FSs are used on the wdbc dataset and then compared to produce an average evaluation of the use of the algorithm model from all classes or labels. Calculation of evaluation results on algorithms with the average of all classes or labels. Then the way the FS works in this research, after X and y data are initialized feature selection is carried out which will produce True and False in GA, where True is the selected feature, and False is the discarded or removed feature. Meanwhile, the RFE feature selection produces a number ranking, where the 1st order is the selected feature. For clarity, here is a description of the RFE and GA selection features.

The RFE selection feature used in previous research applies random forest with cross-validation to maintain the height of accuracy of large data [14]. Then in research [15] examined several novels with random by measuring using mutual information and RFE with Support Vector Machine (SVM) as the estimator. Results with 99% accuracy were obtained using 316 genes out of 3571. In researcher [16] examined using MGRFE for cancer classification, it was concluded that the MGRFE method is feasible to generalize to problems in high-dimensional data with large p small n paradigm characteristics and applied in several fields. Researchers [17] used the SVM-RFE and GA models to diagnose Parkinson's disease, this is used as a review because it uses the same method to compare the accuracy of the results that will be obtained later. Obtained accuracy results using Genetic Algorithm and SVM-RFE of 88.71% increased by 1.02% and sensivity increased from 52.08% to 70.83%. In this study, RFE without cross-validation is used to select the most optimal features with six different algorithms that aim to calculate the final evaluation results on accuracy, precision, recall and f score based on weighted and macro. The way RFE works in this study is by determining X and y, where X is the number of all features or labels, and y is the class. Then determine the estimator, the estimator used is SVR with a linear kernel. Then determine the selector with the RFE library with estimator parameters, the number of features to be tested follows the total number of GA selection features, and step. Then the selector is ranked to proceed to the next process.

GA selection features in previous studies were used to classify bank marketing data, but the selection features used were still unable to improve the accuracy of the algorithms used in bank marketing data classification [18]. Researchers [19] on the classification of positive and negative Go-Jek reviews using SVM, obtained GA-SVM results with an accuracy of 0.895, while using SVM only 0.621. In researcher [20] using the feature selection method based on isolation weight, it was found that the weight of the feature on the f-value can signification increase the efficiency of using the selection feature and reduce the error rate which is similar to relieving the pressure on classification calculations. In this study, GA was used for feature selection with 80 genes to minimize overfitting in other data, then combined with six different methods to calculate the evaluation results of accuracy, precision, recall and f score based on weighted and macro. The workflow of GA feature selection is almost the same as RFE, the difference is that the estimator part uses a Support Vector Machine (SVM) with gamma auto, and the selector library used is GAFeatureSelectionCV with estimator parameters, genes of 80, and cross-validation

of 3. Then display True or False output to determine whether or not to use it for the next process.

After the features have been selected, the next step is to replace X data with new features that have been selected previously with a split testing data ratio of 10%, 20%, and 30%. The way this stage works is by determining the specific X data according to the columns or features that have been selected in the GA and RFE selection features.

The calculation uses machine learning algorithm methods, the details of the algorithms used are Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forest, AdaBoost, and Naïve Bayes. This stage uses weighted averages and macros to calculate the final evaluation based on the average of each label with outputs such as accuracy, precision, recall and F score. Weighted is used if the classes have a balanced distribution of samples that can provide an indicator of the overall performance of the model. Macros are used if the classes have an unbalanced number of samples to provide an overview of the performance of the minority class or the smallest sample. The following is an explanation of the above algorithm.

Logistic regression is a form of regression analysis used when the response variable is binary [21]. variable is binary [21]. This method is used to calculate only the four evaluations of accuracy, precision, recall and f score. Formula 1 is used for classification regression analysis if a small sample is used on high-dimensional data depending on the data.

$$2 \log L(D_0 \left| M_j \right.) = 2 \log L\ (D_0 | M_0) + t \quad (1)$$

t is a random variable sampled from a c2 distribution with df degrees of freedom. Then $L(D_0 \mid M_j)$ is the maximum likelihood under a single logistic regression, and M_j and D_0 are surrogate data.

K-Nearest Neighbors (KNN) is a theory that has strong properties. As the data size increases the algorithm is guaranteed to produce an upper bound error rate no more than twice the optimal achievable [21]. Here is a strong theory of the property.

P (x, y) = 0 if x=y
P (x, y) = p (y, x)
P (x, y) ≤ p (x, z) + p(z,y) for z is a set of X.

Support Vector Machines (SVM) is one of the standard algorithms currently used for machine learning [21]. Formula 2 is used to classify high-dimensional data.

$$S = \left((x_1, y_1), \ldots, (x_\lambda, y_\lambda) \right) \subseteq (X\ x\ Y)^\lambda \quad (2)$$

X and Y denote the input space and output domain. Then the classification binary Y ∈ {-1,1} and the classification class Y = {1,2,…m}. Then finally for is the number of examples and x_i, i = 1,2,... λ is an n-dimensional column vector (to generate row vectors).

Random forest has the advantage of processing the original training data which makes this method one step ahead of other methods. One step ahead is in the random selection of features performed on each node of the tree. The way it works is that a subset of features is randomly selected from the entire feature set, and the best features will be sought to divide certain nodes to improve predictions in the final evaluation results in this study [21]. Formula 3 is the result of a random forest, namely out-of-bag (OOB) where the label $C_{bag}(xi)$ is compared with the actual label yi and the unbiased error estimate for OOB can be described by the formula.

$$N^{-1} \sum_{i=1}^{N} I\ (C_{bag}(x_i) \neq y_i \quad (3)$$

I is an indicator function whose value is 1, otherwise, if $C_{bag}\ (x_i) \neq y_i$ is true and 0.

The way the AdaBoost algorithm works is that the hypothesis is performed with a large enough set of training examples that are unlikely to be wrong, in other words, the approximately correct hypothesis is the one that has an error probability bounded by a small positive constant ε. Formula 4 is the final formula to get a prediction by combining Y_m with X_{test}.

$$Y_m(X_{test}) = sign(\sum_{m=1}^{M} a_m(2y_m(X_{test}) - 1)\ ) \quad (4)$$

It is assumed that the classifier can cope with the weighted case. In other words, cases are selected as possible with weights until the data set is as large as the original training set.

Naïve Bayes is a feature that assumes independent features allowing it to reduce the problem of multiple variations of dimension D to protect against overfitting. Then in terms of classification, when calculating the accuracy of naïve Bayes classification, it is recommended to use selection features to enable the accuracy to be improved [22]. Formula 5 is a feature that is assumed to be conditionally independent in the calculation of naïve Bayes.

$$P\left(C_j \left| A \right.\right)\ \propto \mathrm{P}\left(C_j\right) \prod_{i=1}^{D} P(A_i | C_j) \quad (5)$$

The feature vector A consists of individual features A_i, i = 1,..., D. Then assuming independent features reduces the multi-dimensional variation D in the problem of (estimating $P(A_1,..., A_D|C_j)$) to a univariate problem with (estimating $P(A_1|C_j),...., P(A_D|C_j)$).

The last stage is evaluation, after the evaluation of the model on the wdbc dataset is obtained, then an evaluation is carried out or provides an explanation related to the relationships of accuracy, precision, recall and F-score results based on the average of each weighted table. The following is an explanation of the evaluation.

Accuracy is the number of correct predictions (TP and TN) divided by the number of all samples. Formula 6 is the formula for calculating accuracy.

$$Accuracy = \frac{\mathrm{TP+TN}}{TP+TN+FP+FN} \quad (6)$$

TP is true positive or true positive, TN is true negative or true negative, FP is false positive or false positive, and FN is false negative or false negative. In the evaluation if an algorithm has more true positives and true negatives while fewer false positives and false negatives then it gives more reasonable and better results than other algorithms.

Precision is used to measure how many samples predicted to be positive turn out to be positive (true positive). Formula 7 is a precision measurement to determine TP (true positive or true positive).

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

TP is true positive, and FP is false positive. This precision is specifically used to limit the number of false positives. In other words, the limit should not be more than it should be, so it is expected to have a high precision value.

Recall is used to measure how many positive samples are captured by positive predictions. Formula 8 is a measurement of recall.

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

TP is true positive, FN is false negative. Recall is used specifically to avoid false negatives.

F-score is the combined harmonic mean between precision and recall. In evaluation, it is better to look at the F score rather than comparing recall and precision because the two evaluations are inversely proportional, if precision is higher then recall will be lower and vice versa if recall is higher then precision will be lower. Formula 9 is the f-score calculation formula.

$$F = 2 \, . \, \frac{precision - recall}{precision + recall} \quad (9)$$

From the above formulas, recall remains a better evaluation measure than accuracy on unbalanced classification datasets. However, the f-score is more difficult to interpret and explain than accuracy.

3. Results and Discussions

The dataset used is only one, namely WDBC. The WDBC (Wisconsin Diagnostic Breast Center) dataset was created by William Wolberg, Olvi Mangasarian, Nick Street, and W. Street. This dataset comes from the UC Irvine Machine Learning Repository with a total of 569 data and a total of 30 features. This dataset discusses the diagnosis of Wisconsin women with breast cancer. Features were calculated from digitized images of fine needle aspirations (FNA) of breast masses. Ten real-valued features were calculated for each cell nucleus: Radius, texture, perimeter, area, smoothness, compactness, concavity, concave point, symmetry, and fractal dimension.

3.1 Results

Before comparison, accuracy does not need to use the average parameter, so there is no need to use weighted and macro. This is because accuracy itself can calculate the overall truth prediction whereas the accuracy library already includes the overall average on true positives and negatives.

Table 1. Accuracy Results Using RFE Feature Selection

Split Ratio		Accuracy Value					
Training	Testing	Logistic Regression	KNN	SVM	Random Forest	AdaBoost	Naive Bayes
90	10	98.25%	94.74%	98.25%	94.74%	98.25%	91.23%
80	20	96.49%	92.98%	93.86%	96.49%	92.98%	92.11%
70	30	95.91%	91.81%	91.81%	92.98%	95.91%	92.40%

Table 2. Accuracy Results Using GA Feature Selection

Split Ratio		Accuracy Value					
Training	Testing	Logistic Regression	KNN	SVM	Random Forest	AdaBoost	Naive Bayes
90	10	98.25%	94.74%	98.25%	98.25%	98.25%	92.98%
80	20	96.49%	92.98%	93.86%	92.98%	92.98%	91.23%
70	30	95.91%	95.32%	92.98%	93.57%	95.32%	91.81%

Comparison of the results of using feature selection is carried out by describing the results of accuracy, precision, and recall to f-score. Beginning in Tables 1, and 2 is a table that describes the evaluation results of the accuracy value of the RFE and GA selection features. It was found that 10% of testing data on RFE with the highest accuracy value was obtained by the Logistic Regression, SVM, and AdaBoost model with an average value of 98.25%. Then the data testing is 20% the highest accuracy value on the Logistic Regression model, and the Random forest is 96.49%. Then 30% testing data, the highest accuracy value is obtained by the Logistic Regression and AdaBoost models of 95.91%.

For the results of the GA selection feature. It was found that the 10% testing data with the highest accuracy value was held by the Logistic Regression, SVM, Random Forest, and AdaBoost models with a value of 98.25%, then in the 20% testing data the highest accuracy value was obtained at 96.49% in the Logistic Regression model, then in the 30% testing data, the highest accuracy value was 95.91% in the Logistic Regression model.

Before comparison, this precision is used to determine the positive prediction value is correct, in precision false positive has more impact than false negative because what is predicted in false positive may be part of the negative.

Table 3. Weighted Precision Results Using RFE Feature Selection

Split Ratio		Precision Value					
Training	Testing	Logistic Regression	KNN	SVM	Random Forest	AdaBoost	Naive Bayes
90	10	98.36%	94.69%	98.29%	94.91%	98.36%	92.26%
80	20	96.49%	93.05%	93.88%	95.66%	93.37%	92.68%
70	30	95.90%	91.81%	91.85%	92.98%	95.90%	92.54%

Table 4. Precision Macro Results Using RFE Selection Features

Split Ratio		Precision Value					
Training	Testing	Logistic Regression	KNN	SVM	Random Forest	AdaBoost	Naive Bayes
90	10	96.88%	94.10%	98.84%	92.53%	96.88%	87.61%
80	20	96.10%	93.35%	94.02%	96.10%	91.61%	90.62%
70	30	95.64%	90.94%	91.99%	92.24%	95.64%	91.25%

Table 5. Weighted Precision Results Using GA Feature Selection

Split Ratio		Precision Value					
Training	Testing	Logistic Regression	KNN	SVM	Random Forest	AdaBoost	Naive Bayes
90	10	98.36%	94.69%	98.29%	98.36%	98.36%	93.51%
80	20	96.49%	93.05%	93.88%	93.37%	93.37%	91.65%
70	30	95.93%	95.31%	93.39%	93.60%	95.32%	91.90%

Table 6. Precision Macro Results Using GA Feature Selection

Split Ratio		Precision Value					
Training	Testing	Logistic Regression	KNN	SVM	Random Forest	AdaBoost	Naive Bayes
90	10	96.88%	94.10%	98.84%	96.88%	96.88%	89.93%
80	20	96.10%	93.35%	94.02%	91.61%	91.61%	89.75%
70	30	95.32%	95.18%	94.47%	92.24%	94.82%	90.72%

Tables 3, 4, 5, and 6 are tables that explain the evaluation results of the weighted and macro precision values on the RFE and GA selection features. The highest precision results of the RFE selection feature were obtained on 10% testing data with a weighted average of 98.36% in the LR and AdaBoost models, while the macro was 98.84% in the SVM model, then 20% testing data with a weighted average of 96.49% in the LR model, while the macro was 96.10% in the LR and RF models, then on 30% testing data with a weighted average of 95.90% in the LR and AdaBoost models, while the macro was 95.64% in the LR and AdaBoost models.

Furthermore, the highest precision results of GA selection features were obtained in 10% testing data with a weighted average of 98.36% in LR, RF and AdaBoost models, while macros were 98.84% in the SVM model, then 20% testing data with a weighted average of 96.49% in the LR model, while macros were 96.10% in the LR model, then in 30% testing data with a weighted average of 95.93% in the LR model, while macros were 95.32% in the LR model.

Before comparison, Recall/sensitivity is used to determine the true positive prediction value, even though it is predicted false negative it could be a true positive.

Tables 7, 8, 9 and 10 are tables that explain the evaluation results of the recall/sensitivity values of weighted and macro on the RFE and GA selection features.

Table 7. Weighted Recall Results Using RFE Feature Selection

Split Ratio		Recall Value					
Training	Testing	Logistic Regression	KNN	SVM	Random Forest	AdaBoost	Naive Bayes
90	10	98.25%	94.74%	98.25%	94.74%	98.25%	91.23%
80	20	96.49%	92.98%	93.86%	95.61%	92.98%	92.11%
70	30	95.91%	91.81%	91.81%	92.98%	95.91%	92.40%

Table 8. Macro Recall Results Using RFE Selection Features

Split Ratio		Recall Value					
Training	Testing	Logistic Regression	KNN	SVM	Random Forest	AdaBoost	Naive Bayes
90	10	98.81%	92.14%	96.67%	94.29%	98.81%	91.90%
80	20	96.10%	87.03%	87.74%	92.82%	91.49%	92.15%
70	30	95.27%	90.14%	91.08%	93.53%	93.93%	91.79%

Table 9. Weighted Recall Results Using GA Feature Selection

Split Ratio		Recall Value					
Training	Testing	Logistic Regression	KNN	SVM	Random Forest	AdaBoost	Naive Bayes
90	10	98.25%	94.74%	98.25%	98.25%	98.25%	92.98%
80	20	96.49%	92.98%	93.86%	92.98%	92.98%	91.23%
70	30	95.91%	95.32%	92.98%	93.57%	95.32%	91.81%

Table 10. Macro Recall Results Using GA Feature Selection

Split Ratio		Recall Value					
Training	Testing	Logistic Regression	KNN	SVM	Random Forest	AdaBoost	Naive Bayes
90	10	98.81%	92.14%	96.67%	98.81%	98.81%	93.10%
80	20	96.10%	90.97%	92.26%	93.44%	93.44%	91.49%
70	30	95.67%	94.42%	90.23%	92.24%	94.82%	91.34%

The highest recall/sensitivity results of RFE selection features were obtained on 10% of testing data with a weighted average of 98.25% on LR, SVM, and AdaBoost models, while macros were 98.81% on LR, AdaBoost models, then 20% testing data with a weighted average of 96.49% on LR models, while macros were 96.10% on LR models, then on 30% testing data with a weighted average of 95.91% on LR and AdaBoost models, while macros were 95.27% on LR models.

Furthermore, the highest recall/sensitivity results of GA selection features are obtained in 10% testing data with a weighted average of 98.25% in LR, SVM, RF and AdaBoost models, while macros are 98.81% in LR, RF, AdaBoost models, then 20% testing data with a weighted average of 96.49% in the LR model, while macros are 96.10% in the LR model, then in 30% testing data with a weighted average of 95.91% in the LR model, while macros are 95.67% in the LR model.

Before comparison, the F1 score is used to measure the balance between precision and recall or sensitivity.

Table 11. Weighted F-Score Results Using RFE Feature Selection

Split Ratio		F-Score Value					
Training	Testing	Logistic Regression	KNN	SVM	Random Forest	AdaBoost	Naive Bayes
90	10	98.26%	94.68%	98.23%	94.79%	98.26%	91.46%
80	20	96.49%	92.88%	93.80%	95.63%	93.06%	92.21%
70	30	95.90%	95.90%	95.90%	95.90%	95.90%	92.44%

Table 12. Macro F-Score Results Using RFE Selection Features

Split Ratio		F-Score Value					
Training	Testing	Logistic Regression	KNN	SVM	Random Forest	AdaBoost	Naive Bayes
90	10	97.78%	93.06%	97.69%	93.35%	97.78%	89.34%
80	20	96.10%	92.00%	93.05%	96.10%	92.38%	91.47%
70	30	95.45%	90.94%	90.71%	92.24%	95.45%	91.69%

Table 13. Weighted F-Score Results Using GA Feature Selection

Split Ratio		F-Score Value					
Training	Testing	Logistic Regression	KNN	SVM	Random Forest	AdaBoost	Naive Bayes
90	10	98.26%	94.68%	98.23%	98.26%	98.26%	93.12%
80	20	96.49%	92.88%	93.80%	93.06%	93.06%	91.32%
70	30	95.91%	95.30%	92.81%	93.58%	95.32%	91.84%

Table 14. Macro F-Score Results Using GA Feature Selection

Split Ratio		F-Score Value					
Training	Testing	Logistic Regression	KNN	SVM	Random Forest	AdaBoost	Naive Bayes
90	10	97.78%	93.06%	97.69%	97.78%	97.78%	91.31%
80	20	96.10%	92.00%	93.05%	92.38%	92.38%	90.48%
70	30	95.49%	94.78%	91.88%	92.24%	94.82%	91.01%

Tables 11, 12, 13, and 14 are tables that explain the evaluation results of the F1 score weighted and macro values on the RFE and GA selection features. The highest F1 score results from the RFE selection feature were obtained in 10% of testing data with a weighted average of 98.26% in the LR model, and AdaBoost, while the macro was 97.78% in the LR model, AdaBoost, then 20% testing data with a weighted average of 96.49% in the LR model, while the macro was 96.10% in the LR and RF models, then in 30% testing data with a weighted average of 95.90% in the LR, KNN, SVM, RF and AdaBoost models, while the macro was 95.45% in the LR and AdaBoost models.

Furthermore, the highest F1 score results from GA selection features are obtained in 10% testing data with a weighted average of 98.26% in LR, RF and AdaBoost models, while macros are 97.78% in LR, RF, AdaBoost models, then 20% testing data with a weighted average of 96.49% in the LR model, while macros are 96.10% in the LR model, then in 30% testing data with a weighted average of 95.91% in the LR model, while macros are 95.49% in the LR model.

So the topic of this research is a comparison of the use of selection features between RFE and GA using six classification algorithms on the WDBC dataset which discusses women in the Wisconsin region who are diagnosed with breast cancer. We will further discuss the results of accuracy, precision, recall/sensitivity, and f1 score related to breast cancer diagnosis in the discussion section.

3.2 Discussion

Before comparison with previous research, there is some validity information that needs to be considered, namely in the wdbc dataset in this study the number of classes B: M is 357: 212. Then in the comparison later what is used is weighted because weighted can provide indicators of overall model performance between benign and malignant classes, while macros are used to provide indicators of model performance in the minority class or smallest number of classes, where the smallest in this dataset is Malignant. Then the six algorithms used in this study use default parameters according to their respective rules because the purpose of this study is to present an overall picture with the use of 2 selection features along with which algorithm is the best or highest.

Comparison of 20% testing data results. In research [3] using the HSW (Heuristic Search Wrappers) selection feature with autoencoder optimization and MBE (Model Based Elimination) with the AMBER algorithm obtained an accuracy of 96.78%. While in this study using RFE and GA selection features has the same highest accuracy of 96.49% on the LR (Logistic Regression) model. From these results, research [3] is better by 0.29%. Regarding the same wdbc dataset used, then because previous researchers have higher accuracy, it can affect the diagnosis test including benign or malignant breast cancer.

In research [4] obtained results using the CCEA selection feature with the NB, SVM, KNN, RF, and LR algorithms. Of all the research algorithms [4] which has the highest accuracy is RF at 93.32%, while in this study the RFE and GA selection features have the highest accuracy comparable to 98.25% where the GA algorithm model is LR, SVM, RF and AdaBoost, while RFE without any RF algorithm. From this comparison, this research is better by 4.93%. Then for precision, recall, and the best f1 score in research [4] is RF at 93.30%, while in this study RFE and GA have the highest precision comparable to 96.49% with the LR algorithm model, then the recall value of RFE and GA is comparable to 98.25% where the GA algorithm model is LR, SVM, RF, and AdaBoost, while RFE is without RF, then the f1 score value of RFE and GA is comparable to 98.26% where the GA algorithm model is LR, RF and AdaBoost, while RFE is LR and AdaBoost. From the comparison of precision, recall and f1 score, this research is better by 3.19%, 4.95%, and 4.96% respectively. So the conclusion when compared to research [4] is that this research is better at evaluating the accuracy, precision, recall and f1 score by 4.93%, 3.19%, 4.95%, and 4.96% respectively. Then about datasets that both use WDBC, because in terms of overall evaluation this research has a higher percentage than in terms of diagnosis tests, determining predictions whether the whole is positive including in the false positive section, if there is a negative then it is discarded, then determining positive predictions on false negative (fn), it could be that the fn is a true positive (tp), and finally in managing the balance between recall and precision.

In research [5] using VIM selection features with AdaBoost and RF to compare with this research. The training data used is 30%. The accuracy of AdaBoost and RF was 93.33%, and 96.37%, respectively. while in this study, it was obtained that the AdaBoost model had better RFE selection features with an accuracy of 95.91%, than for the RF model, the GA selection feature

was better with an accuracy of 93.57%. From the comparison of these accuracy results in the AdaBoost model, this research is better with a distance of 2.58% with these results the RFE selection feature is better than the VIM selection feature, then conversely in the RF model, the VIM selection feature is better than the GA with a distance of 2.8%. In research [5] the precision results of the AdaBoost and RF algorithm models are 91.15% and 95.97% respectively, while in this study in the AdaBoost model, the RFE selection feature is better with a precision of 95.90%, then in the RF model the GA selection feature is better with a precision of 93.60%. From the comparison of the precision results, the AdaBoost model is better than this research with a distance of 4.75% so in terms of diagnosis test, the RFE-AdaBoost selection feature is better. Then on the other hand, for the RF algorithm, the VIM selection feature is better than GA with a distance of 2.37% so in terms of accuracy when in the false positive section (fp) all positives are better VIM-RF selection features.

In research [6] PCA selection features using the KNN algorithm on 10% testing data to compare with this research. In research [6] obtained an accuracy of 97.77%, precision of 95.78%, and recall of 98.08%, while in this study RFE and GA have the same highest results, namely the accuracy of 94.74%, precision of 94.69%, recall of 94.74%. From these results, the distance of the accuracy value is 3.03%, the distance of the precision value is 1.09%, and the recall distance is 3.34%, the three evaluations are better when using the PCA selection feature, in other words, in terms of diagnosis testing, determining the accuracy of the false positive (FP) section whether all of them are positive, and determining whether in the false negative section, whether all of them are negative, it is all better for researchers [6] using the PCA-KNN selection feature.

4. Conclusion

In this paper, from the results of using RFE and GA selection features in selecting features with support using machine learning algorithms including Logistic Regression, Naïve Bayes, KNN, Support Vector Machines, AdaBoost, and Random Forest tested on the wdbc dataset, overall the GA selection feature is better from accuracy to average Weighted and macro on precision, recall and f1 score for selecting on high-dimensional data.

It is expected that in the future there needs to be in-depth literacy related to the use of parameters in the GA and RFE selection features as well as understanding the use of parameters in the algorithm that you want to use, with that hopefully evaluation results such as accuracy, precision, recall and f1 score have results that are close to 99%.

Acknowledgements

Thanks Department of Research and Community Service, Universitas Amikom Yogyakarta, for funding, supporting this research and striving to finish this research.

References

[1] F. Peng, H. Wang, L. Zhuang, M. Wang, and C. Yang, "Methods of enterprise electronic file content information mining under big data environment," in *2020 International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*, Oct. 2020, pp. 5–8. doi: 10.1109/ICBASE51474.2020.00008.

[2] M. Lamrini and M. Y. Chkouri, "Decomposition and Visualization of High-Dimensional Data in a Two-Dimensional Interface," in *2019 1st International Conference on Smart Systems and Data Science (ICSSD)*, Oct. 2019, pp. 1–5. doi: 10.1109/ICSSD47982.2019.9002846.

[3] S. Ramjee and A. El Gamal, "Efficient Wrapper Feature Selection using Autoencoder and Model Based Elimination," May 2019, [Online]. Available: http://arxiv.org/abs/1905.11592

[4] A. N. M. B. Rashid, M. Ahmed, L. F. Sikos, and P. Haskell-Dowland, "A Novel Penalty-Based Wrapper Objective Function for Feature Selection in Big Data Using Cooperative Co-Evolution," *IEEE Access*, vol. 8, pp. 150113–150129, 2020, doi: 10.1109/ACCESS.2020.3016679.

[5] Z. Huang and D. Chen, "A Breast Cancer Diagnosis Method Based on VIM Feature Selection and Hierarchical Clustering Random Forest Algorithm," *IEEE Access*, vol. 10, pp. 3284–3293, 2022, doi: 10.1109/ACCESS.2021.3139595.

[6] F. S. Fogliatto, M. J. Anzanello, F. Soares, and P. G. Brust-Renck, "Decision Support for Breast Cancer Detection: Classification Improvement Through Feature Selection," *Cancer Control*, vol. 26, no. 1, p. 107327481987659, Jan. 2019, doi: 10.1177/1073274819876598.

[7] I. Cholissodin and A. A. Soebroto, "AI , MACHINE LEARNING & DEEP LEARNING (Teori & Implementasi)," no. December, 2021.

[8] T. Almutiri and F. Saeed, "Chi Square and Support Vector Machine with Recursive Feature Elimination for Gene Expression Data Classification," in *2019 First International Conference of Intelligent Computing and Engineering (ICOICE)*, Dec. 2019, pp. 1–6. doi: 10.1109/ICOICE48418.2019.9035165.

[9] N. Jamshidpour, A. Safari, and S. Homayouni, "Multiview Active Learning Optimization Based on Genetic Algorithm and Gaussian Mixture Models for Hyperspectral Data," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 1, pp. 172–176, 2020, doi: 10.1109/LGRS.2019.2914858.

[10] X. Ding, F. Yang, and F. Ma, "An efficient model selection for linear discriminant function-based recursive feature elimination," *J. Biomed. Inform.*, vol. 129, p. 104070, May 2022, doi: 10.1016/j.jbi.2022.104070.

[11] L. Li, W.-K. Ching, and Z.-P. Liu, "Robust biomarker screening from gene expression data by stable machine learning-recursive feature elimination methods," *Comput. Biol. Chem.*, vol. 100, p. 107747, Oct. 2022, doi: 10.1016/j.compbiolchem.2022.107747.

[12] P. R. Kannari, N. S. Chowdary, and R. Laxmikanth Biradar, "An anomaly-based intrusion detection system using recursive feature elimination technique for improved attack detection," *Theor. Comput. Sci.*, vol. 931, pp. 56–64, Sep. 2022, doi: 10.1016/j.tcs.2022.07.030.

[13] W. Liu and J. Wang, "Recursive elimination current algorithms and a distributed computing scheme to accelerate wrapper feature selection," *Inf. Sci. (Ny).*, vol. 589, pp. 636–654, Apr. 2022, doi: 10.1016/j.ins.2021.12.086.

[14] U. Das, A. Y. Srizon, M. Al Mehedi Hasan, J. Rahman, and M. K. Ben Islam, "Effective Data Dimensionality Reduction Workflow for High-Dimensional Gene Expression Datasets," in *2020 IEEE Region 10 Symposium (TENSYMP)*, 2020, pp.

182–185. doi: 10.1109/TENSYMP50017.2020.9230847.

[15] N. Koul and S. S. Manvi, "Ensemble Feature Selection from Cancer Gene Expression Data using Mutual Information and Recursive Feature Elimination," in *2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAECC)*, Dec. 2020, pp. 1–6. doi: 10.1109/ICAECC50550.2020.9339518.

[16] C. Peng, X. Wu, W. Yuan, X. Zhang, Y. Zhang, and Y. Li, "MGRFE: Multilayer Recursive Feature Elimination Based on an Embedded Genetic Algorithm for Cancer Classification," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 18, no. 2, pp. 621–632, Mar. 2021, doi: 10.1109/TCBB.2019.2921961.

[17] J. Goyal, P. Khandnor, and T. C. Aseri, "Analysis of Parkinson's disease diagnosis using a combination of Genetic Algorithm and Recursive Feature Elimination," in *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*, Jul. 2020, pp. 268–272. doi: 10.1109/WorldS450073.2020.9210415.

[18] Yoga Religia, Agung Nugroho, and Wahyu Hadikristanto, "Klasifikasi Analisis Perbandingan Algoritma Optimasi pada Random Forest untuk Klasifikasi Data Bank Marketing," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 1, pp. 187–192, 2021, doi: 10.29207/resti.v5i1.2813.

[19] W. M. P.D. and Haryoko, "Optimization Of Parameter Support Vector Machine (SVM) using Genetic Algorithm to Review Go-Jek's Services," in *2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, Nov. 2019, pp. 301–304. doi: 10.1109/ICITISEE48480.2019.9003894.

[20] L. Qiang, X. Zhijie, and Z. Zhisheng, "A Feature Selection Method Based on Variable Weight in Fault Isolation," in *2021 International Conference on Computer, Control and Robotics (ICCCR)*, Jan. 2021, pp. 256–261. doi: 10.1109/ICCCR49711.2021.9349378.

[21] O. Okun, *Feature Selection and Ensemble Methods for Bioinformatics*. IGI Global snippet, 2011. doi: 10.4018/978-1-60960-557-5.

[22] D. Hand, "The top ten algorithms in data mining," Chapman & Hall/CRC Press, 2009, pp. 163–177. doi: 10.1201/9781420089653.ch9.

