Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 4, August 2025, Page. 1925-1940
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5017

Comparison of ANOVA and Chi-Square Feature Selection Methods to Improve Machine Learning Performance in Anemia Classification

Tiko Nur Annisa*[1], Jasmir Jasmir[2], Nurhadi Nurhadi[3]

[1]Magister of Information Systems Student, Universitas Dinamika Bangsa, Indonesia
[2,3]Magister of Information Systems, Universitas Dinamika Bangsa, Indonesia

Email: [1]tikonurannisa10@gmail.com

Received : Jul 2, 2025; Revised : Aug 10, 2025; Accepted : Aug 12, 2025; Published : Aug 18, 2025

Abstract

Anemia is a prevalent hematological condition marked by decreased hemoglobin concentration in the blood, which can lead to serious health complications if undetected. Although machine learning has shown potential in supporting early diagnosis, its effectiveness is often hindered by irrelevant or excessive features. This study investigates the impact of ANOVA and Chi-Square feature selection methods in improving the effectiveness of three distinct machine learning models algorithms, Naive Bayes, K-Nearest Neighbor (KNN), and Support Vector Machine (SVM) for anemia classification. Using a Kaggle dataset consisting of 15,300 instances and 25 features, the evaluation of each model was conducted with reference to its accuracy, precision, recall, and F1-score, both before and after applying feature selection. Experimental results show a substantial improvement in classification performance after feature selection, with the SVM + ANOVA combination achieving the highest accuracy of 94.61%. In contrast, models without feature selection performed below 90%, highlighting the need for appropriate feature reduction techniques. This study contributes a comparative analysis framework for medical data classification, emphasizing the role of statistical feature selection in optimizing model accuracy. Its novelty lies in demonstrating consistent performance improvement across algorithms using real-world anemia data and providing evidence that ANOVA and Chi-Square can significantly enhance model generalization in medical diagnostic contexts.

Keywords : *Anemia, Classification, Improvement, Machine Learning, Performance*

This work is an open access article and licensed under a Creative Commons Attribution-Non Commercial 4.0 International License

1. INTRODUCTION

Anemia is a common medical condition marked by a reduced oxygen-carrying capacity of the blood, typically due to a low red blood cell count or abnormal hemoglobin [1]. Erythrocytes play a vital role in delivering oxygen to body tissues and facilitating the removal of carbon dioxide [2]. Clinically, anemia is diagnosed when in women a hemoglobin concentration that is considered low is less than 11 g/dL, while in men it is less than 12 g/dL [3].

According to WHO, around 1.62 billion people suffer from anemia globally, including 43% of children under five and 300,000 infants [4][5]. In 2019, the global prevalence reached 22.8% [6]. In Indonesia, anemia affects 27.2% of girls and 20.3% of boys aged 15–24, making it a significant public health issue, especially among adolescent girls[7].

Anemia is commonly diagnosed through complete blood count, serum ferritin, and hemoglobin electrophoresis tests [8]. However, these manual methods face challenges such as limited resources, data interpretation errors, and delayed diagnosis [9][10].

The advancement of information technology, especially machine learning, supports the use of intelligent systems to assist in disease diagnosis more quickly, accurately, and efficiently, particularly in areas with limited medical resources [11][12][13]. Machine learning has been successfully applied to disease classification with promising results [14]. Model performance is strongly influenced by data

Vol. 6, No. 4, August 2025, Page. 1925-1940
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5017

quality and feature relevance [15]. Irrelevant features may reduce computational efficiency and prediction accuracy [16][17], making feature selection essential for optimizing model performance [18].

Research on machine learning for disease classification has been widely conducted, but studies comparing feature selection methods to optimize anemia classification are still limited [19]. A study by Tuba et al. (2021) used the Naïve Bayes algorithm to classify anemia with 74% accuracy but did not incorporate a feature selection stage, resulting in a higher computational burden [20]. Another study by Manasvi et al. (2023) applied the K-Nearest Neighbor algorithm to anemia classification and achieved only 63.15% accuracy. This study also compared several classification models but did not include feature selection, thereby missing an opportunity for performance optimization [21]. Another study by Justice et al. (2024) employed Support Vector Machine for anemia prediction with an accuracy of 89.45%, yet also excluded any feature selection method, which limited the model's potential [22]. Algoritma Naïve Bayes, K-Nearest Neighbor, and Support Vector Machine are commonly used in medical diagnosis tasks due to their interpretability, simplicity, and proven performance in classification problems involving health data [23].

Although machine learning has been widely applied in disease classification, comparative studies evaluating ANOVA and Chi-Square feature selection methods for anemia classification remain scarce. Most prior research either omits feature selection or employs a single method, limiting model optimization opportunities [20][21][22]. Additionally, clinical datasets often include mixed numeric and categorical variables, necessitating tailored selection strategies that are rarely addressed [24]. Improper selection of features in mixed-type data can lead to biased model learning and decreased generalization ability, especially in small-to-medium healthcare datasets [25]. Shanthi (2024) showed that feature selection improves prediction accuracy and computational efficiency in high-dimensional health data [26]. Therefore, this study fills the gap by comparing ANOVA and Chi-Square feature selection across Naive Bayes, K-Nearest Neighbor, and Support Vector Machine to determine the optimal approach for anemia classification [27][28].

ANOVA is generally used to analyze the statistical impact of numerical features by analyzing variance among classes, while Chi-Square evaluates categorical features based on their independence from class labels [29][30]. Both methods are computationally efficient and suitable for handling mixed-type datasets. The data source applied in this study, sourced from Kaggle, includes both numerical and categorical attributes, making it essential to apply the appropriate feature selection strategies to reduce information redundancy and boost the effectiveness of the model. The dataset consists of approximately 15.300 instances and 25 features, including attributes such as age, hemoglobin levels, mean corpuscular volume, and presence of clinical symptoms.

This study aims to make a significant contribution to the field of medical informatics by providing a comprehensive comparison of ANOVA and Chi-Square feature selection methods across different machine learning classifiers. The results indicate that expected to enhance the performance of anemia prediction models and serve as a practical reference for scholars and healthcare professionals in the development of accurate, efficient, and interpretable diagnostic tools.

2. METHOD

In this study, several stages were carried out such as the anemia dataset underwent data cleaning, then divided into a training set and a testing set. Two approaches were tested, one without feature selection and one using ANOVA and Chi-Square to select relevant features. Each approach was evaluated using Naive Bayes, K-Nearest Neighbor (KNN), and Support Vector Machine (SVM). Model performance was measured using a confusion matrix with accuracy, precision, recall, and F1 score to compare results and assess the impact of feature selection. The overall process is presented in Figure 1.

Vol. 6, No. 4, August 2025, Page. 1925-1940
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5017

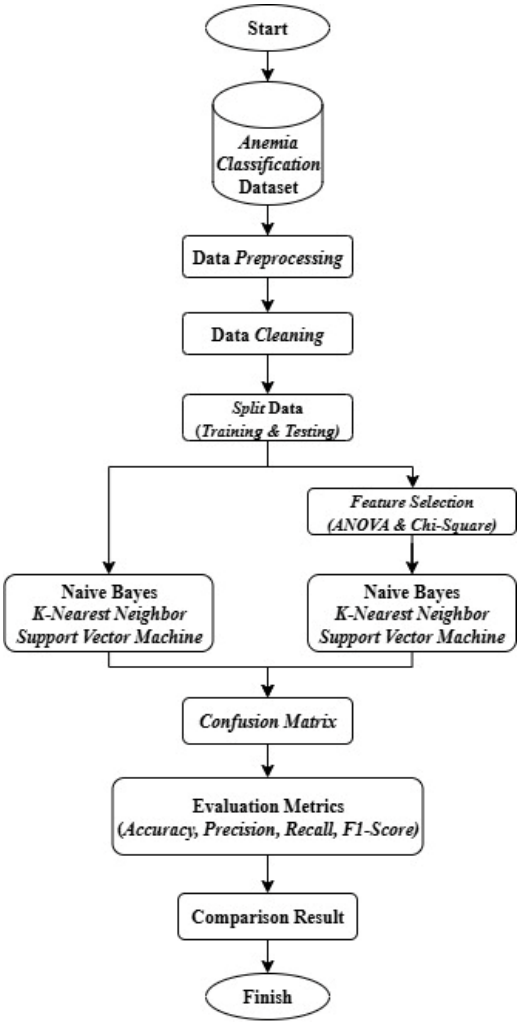

Figure 1. Research Stages

2.1. Dataset

The dataset used in this study was obtained from the Kaggle.com platform, comprising 15,300 records and 25 attributes. This dataset served as the basis for the training and testing processes of the classification models in this research

2.2. Data Preprocessing

The preprocessing of data is a fundamental phase in the machine learning workflow, as the standard of the resulting model directly depends on this step. It involves cleaning, integrating, transforming, and enriching data for managing missing values, outliers, and inconsistent scales, thereby ensuring high-quality data for subsequent processes [31].

2.2.1 Data Cleaning

Data cleaning is an essential initial step in machine learning methods employed to confirm that the dataset is free from errors. This process can be performed manually or automatically using specialized tools. Proper data cleaning prepares the dataset used for precise and reliable analysis [32].

2.2.2 Data Split

The dataset is partitioned into 80% training data and 20% testing data ensure the model receives sufficient data for learning while providing sufficient test data for performance evaluation [33].

2.3 Application of Machine Learning Algorithms

This study employs machine learning algorithms to develop an anemia prediction model based on data patterns, utilizing Naive Bayes, K-Nearest Neighbor, and Support Vector Machine with different classification approaches.

2.3.1 Naïve Bayes Algorithm

Naive Bayes is a classification method grounded on Bayes' Theorem that utilizes probabilistic principles, assuming that each feature is independent of others with respect to the target class. Despite this independence assumption rarely being fully met in real-world data, Naive Bayes remains efficient, accurate, and widely applied in various classification problems [34]. The Bayes theorem formula used in the Naïve Bayes Classifier is as follows:

$$P(H|X) = \frac{P(H|X).P(H)}{P(X)} \quad (1)$$

Description :
X = Unknown data class.
H = Hypothesis of data X is a specific class
P(X|H) = Probability of hypothesis H based on condition X (posteriori prob)
P(H) = Probability of hypothesis H (prior prob)
P(X|H) = Probability of X based on the condition
P(X) = Probability of X

2.3.2 K-Nearest Neighbor Algorithm

The K-Nearest Neighbor (KNN) algorithm is a classification method which categorizes a new data instance according to the predominant class among the K most proximate neighbors in the feature domain. The classification process includes computing the distance betwee the new instance and all data in the training set, selecting the K nearest data points as references [35]. The formula used in K-Nearest Neighbor is as follows:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(xi - yi)^2} \quad (2)$$

Description:
d(x,y) = Distance between test data and training data
n = Number of training data
x = Training data
y = Test data

2.3.3 Support Vector Machine Algorithm

Support Vector Machine (SVM) is a classification method that maps data into a higher-dimensional space and constructs a hyperplane with the maximum margin to separate classes, using kernel functions for nonlinear data [36]. The formula used in Support Vector Machine (SVM) is as follows :

$$Xi.w + b \geq 1 \; untuk \; Yi = 1 \quad (3)$$

$$Xi.w + b \leq 1 \; untuk \; Yi = -1 \quad (4)$$

Vol. 6, No. 4, August 2025, Page. 1925-1940
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5017

Description:
Xi = i-th data
w = support vector weight value perpendicular to the hyperplane
b = bias value
Yi = i-th data class

2.4 Feature Selection Using ANOVA and Chi-Square

Feature selection improves model performance by eliminating less relevant features, enhancing computational efficiency without compromising accuracy. In this study, ANOVA and Chi-Square methods were employed.

2.4.1 Analysis of Variance (ANOVA)

ANOVA represents a statistical tool used to measure the role of the mean value differences between groups. In classification, ANOVA plays a role in selecting the most relevant features to the target variable, thereby improving model performance [37]. The ANOVA formula used is generally expressed as:

$$F = \frac{variance\ between\ groups}{variance\ within\ groups} \quad (5)$$

$$variance\ between\ groups = \frac{\sum_i^n ni\ (Yi-Y)^2}{(k-1)} \quad (6)$$

$$variance\ within\ groups = \frac{\sum_{i=1}^{k}\sum_{j=1}^{ni}(Yij-Yi)^2}{(n-k)} \quad (7)$$

Description:
ni = number of data in the i-th group
Yi = average of the i-th group
n = total number of data
k = number of groups

2.4.2 Chi-Square

Chi-square is a statistical method applied to quantify the correlation between a feature (attribute) and a target variable in classification. This test is performed by comparing recorded and predicted frequencies, with a high Chi-square value indicating strong feature relevance to the target. In feature selection, this method helps eliminate less significant attributes, thereby improving model accuracy and efficiency [38]. The Chi-square formula is expressed as:

$$Ea = \frac{(a+b)(a+c)}{t} \quad (8)$$

$$x^2 = \sum_{i=1}^{n} \frac{(Oi-Ei)^2}{Ei} \quad (9)$$

$$x^2 = \frac{(a-Ea)^2}{Ea} + \frac{(b-Eb)^2}{Eb} + \frac{(c-Ec)^2}{Ec} + \frac{(d-Ed)^2}{Ed} \quad (10)$$

Description:
a+b = number of documents that have the feature
a+c = number of documents from a certain class
t = total of all documents
Oi = actual value (observation)

Vol. 6, No. 4, August 2025, Page. 1925-1940
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5017

Ei = expected value (expectation)
a,b,c,d = number of documents in each category (feature present/absent and positive/negative class)
Ea,Eb,Ec,Ed = expected value for each category

2.5 Model Evaluation

Classification model performance evaluation was performed using a Confusion Matrix, which provides information on the quantity of right and wrong predictions. Based on the Confusion Matrix, model performance is measured using four main metrics accuracy, precision, recall, and F1-score [39].

$$Accuracy = \frac{TP+TN}{TP+TP+FP+FN} \times 100\% \quad (11)$$

$$Precision = \frac{TP}{TP+FP} \times 100\% \quad (12)$$

$$Recall = \frac{TP}{TP+FN} \times 100\% \quad (13)$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \times 100\% \quad (14)$$

Description:
True Positives (TP): The number of positive cases correctly classified by the model.
False Positives (FP): The number of negative cases mistakenly classified as positive.
True Negatives (TN): The number of negative cases correctly identified by the model.
False Negatives (FN): The number of positive cases incorrectly classified as negative.

3. RESULT

3.1. Data Representation

Data representation aims to describe the structure and content of the dataset systematically. In this study, data from Kaggle related to anemia disease is utilized. The dataset contains various relevant attributes, such as demographic information, clinical indicators, and laboratory test results, which are essential for supporting the analysis and prediction process. These attributes serve as the foundation for building models to detect and predict anemia cases more accurately.

Table 1. Research Dataset

No.	Gender	WBC	NE	LY	MO	EO	BA	RBC	HGB	…	Class
1	1	10.63	6.31	2.79	0.91	0.56	0.06	4.31	12.7	…	1
2	1	5.08	2.50	1.87	0.43	0.26	0.02	4.34	12.8	…	1
3	1	13.68	9.4	2.69	1.55	0.03	0.01	3.18	9.4	…	1
4	1	5.6	3.94	0.83	0.54	0.26	0.03	3.35	10.5	…	1
5	1	3.57	2.03	1.25	0.1	0.18	0.01	1.31	5.1	…	1
…	…	…	…	…	…	…	…	…	…	…	…
15296	0	11.99	10.21	1.14	0.62	0.00	0.02	3.58	12.0	…	0
15297	0	9.14	5.38	2.74	0.79	0.18	0.05	4.77	13.1	…	0
15298	0	5.91	2.96	2.09	0.57	0.23	0.06	4.95	13.7	…	0
15299	0	7.65	3.77	3.26	0.45	0.15	0.02	4.48	12.1	…	0
15300	0	8.18	6.32	1.26	0.54	0.04	0.02	4.78	13.8	…	0

Table 1 shows the raw data of laboratory tests from a number of individuals. Each row represents a single data sample, while the columns contain the numerical values of the measurements of various

Vol. 6, No. 4, August 2025, Page. 1925-1940
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5017

blood parameters. This data is used as input in the training process of a machine learning model to perform classification based on the patterns contained therein.

3.2 Data Split

The initial dataset comprised 15.300 rows, which after undergoing thorough cleaning and validation to remove incomplete, inconsistent, and duplicate entries, resulted in 15.212 high-quality records. The refined dataset was then divided using an 80:20 ratio, allocating 12.169 rows for the training phase and 3.043 rows for testing. This division strive to ensure the model obtains sufficient and diverse data for effective learning whilst ensuring a representative test set for an objective evaluation of its generalization capability. The visualization of this data split is shown in Figure 2.

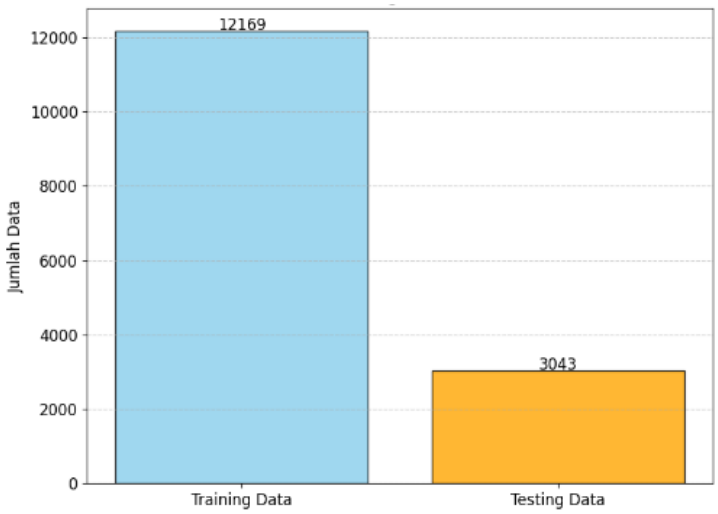

Figure 2. Split Data

Figure 2 shows the results of dividing the dataset into 12.169 rows of training data used to build the model, and 3043 rows of testing data used to measure the model's performance on previously unseen data. Model performance evaluation was performed using accuracy, precision, recall, and confusion matrix metrics. This dataset consists of two classes, namely “yes” and “no”, with the label distribution shown in Figure 3.

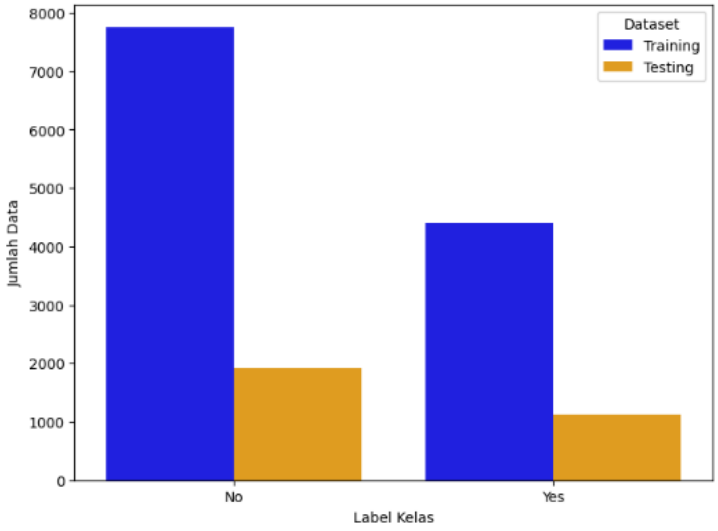

Figure 3. Sample of Each Label

Vol. 6, No. 4, August 2025, Page. 1925-1940
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5017

Based on Figure 3, the results of the data division show that the label distribution in the dataset is relatively balanced, both in the training data and the testing data. Of the total 12.169 training data, there are 7.758 data with the label "No" and 4,411 data with the label "Yes". Meanwhile, of the 3.043 testing data, 1.914 are labeled "No" and 1,129 are labeled "Yes". This balanced distribution is important to ensure that the model is not biased towards one class.

3.3 Model Evaluation

At this stage, an evaluation of the performance of the anemia prediction model built using three classification algorithms, namely Naive Bayes, K-Nearest Neighbor (KNN), and Support Vector Machine (SVM) is carried out. The evaluation aims to measure the ability of each model to classify data accurately. The evaluation metrics used include accuracy, precision, recall, and F1-score. By comparing the results of the three algorithms, an overview of the most effective and efficient model in making predictions can be obtained.

3.3.1 Evaluation Without Features

At this stage, an initial evaluation of the anemia disease prediction model was carried out using the Naive Bayes, K-Nearest Neighbor, and Support Vector Machine algorithms without going through the feature selection process. All attributes available in the dataset are used directly to build the model, in order to determine the basic performance of each algorithm before optimization is carried out through the feature selection process. The comparison results are presented in Figure 4.

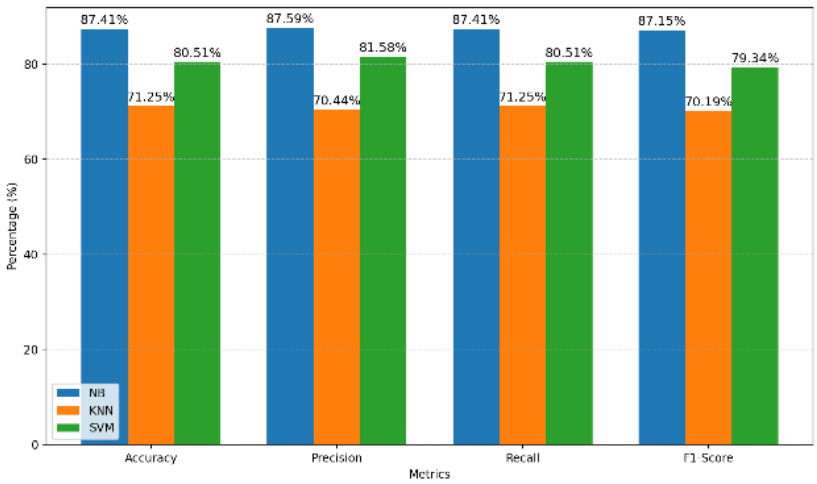

Figure 4. Evaluation Results Without Features

Figure 4 shows a comparison of the accuracy of the three classification algorithms used in building an anemia prediction model, namely Naive Bayes, K-Nearest Neighbor, and Support Vector Machine without going through the feature selection process. According to the evaluation findings, the Naive Bayes algorithm obtained the highest accuracy of 87.41%, followed by Support Vector Machine with an accuracy of 80.51%, and K-Nearest Neighbor with the lowest accuracy of 71.25%. These results indicate that in the early stages without feature optimization, Naive Bayes has the most optimal performance in classifying anemia data.

3.3.2 Evaluation Using ANOVA Features

Evaluation using the Analysis of Variance (ANOVA) feature aims to test the significance of the average difference between data groups based on one or more independent variables. ANOVA is used to determine the extent to which independent variables have a significant effect on dependent

Vol. 6, No. 4, August 2025, Page. 1925-1940
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5017

variables, by measuring and comparing variances between groups statistically. The comparison results are presented in Figure 5.

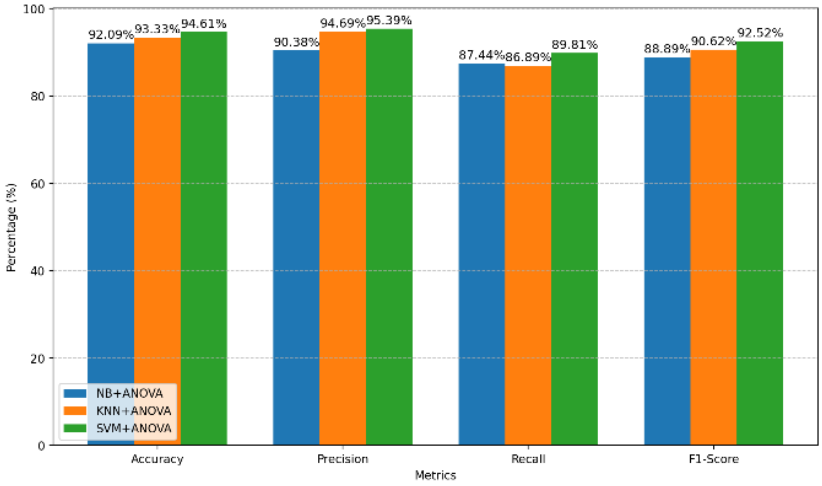

Figure 5. Evaluation results using the ANOVA Features

Figure 5 presents the results of the evaluation of the accuracy of the anemia prediction model after feature selection using the Analysis of Variance (ANOVA) method on three different classification algorithms. Based on these results, the Support Vector Machine (SVM) algorithm showed the highest accuracy of 94.61 percent, followed by the K-Nearest Neighbor (KNN) algorithm with an accuracy of 93.33 percent, and Naive Bayes (NB) with an accuracy of 92.09 percent. The increase in accuracy shown by these three models indicates that the application of feature selection using the ANOVA method is able to significantly improve the performance of the classification model.

3.3.3 Evaluation Using Chi-Square Features

Chi-Square feature selection is a common method in classification modeling used to assess the association between features and the target variable. It helps retain only the most relevant features, making model training more efficient and improving prediction accuracy. This technique is especially effective for categorical data due to its simplicity, speed, and ease of interpretation. Comparison results are shown in Figure 6.

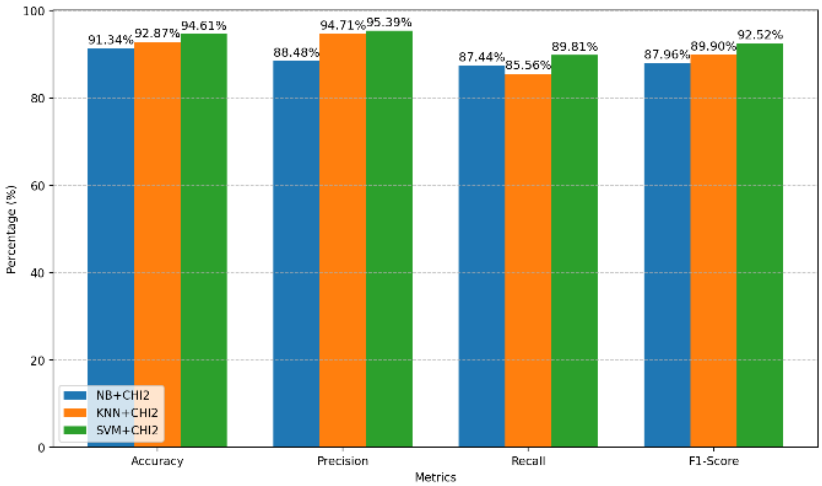

Figure 6. Evaluation results using the Chi-Square Features

Figure 6 presents the results of the evaluation of the accuracy of the anemia prediction classification model after applying feature selection using the Chi-Square method. Based on the

Vol. 6, No. 4, August 2025, Page. 1925-1940
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5017

evaluation results, the Support Vector Machine algorithm obtained the highest accuracy of 94.61%, followed by K-Nearest Neighbor with an accuracy of 92.87%, and Naive Bayes at 91.34%. These results indicate that the application of Chi-Square feature selection can significantly improve model performance in the classification process.

3.3.4 Comparison of Evaluation Results

The evaluation results show that models using feature selection methods, such as ANOVA and Chi-Square, perform significantly better than those without. ANOVA and Chi-Square have proven effective in identifying relevant features, thereby improving model efficiency and accuracy. The comparison is illustrated in Figure 7.

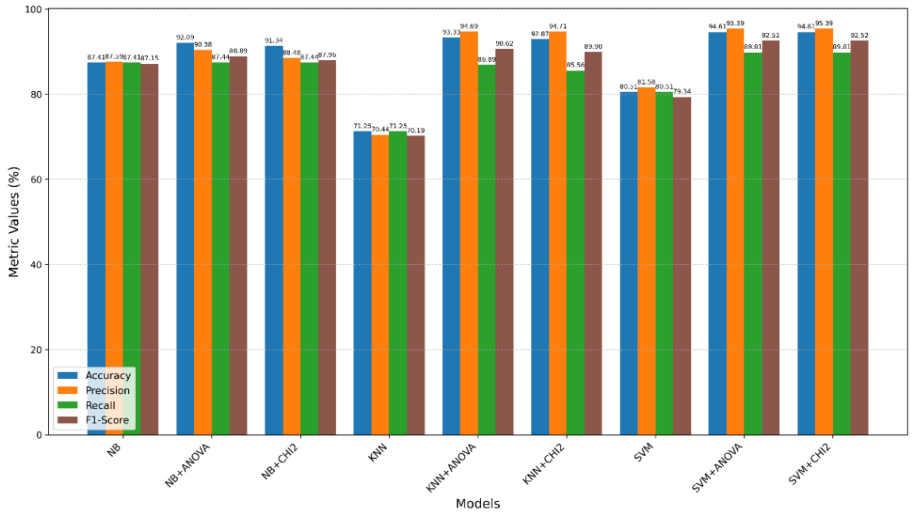

Figure 7. Comparison of Evaluation Results

Figure 7 presents a comparison of the accuracy of nine classification models, both without and with feature selection using ANOVA and Chi-Square (Chi2). The results show that the SVM+Chi2 and SVM+ANOVA models have the highest accuracy of 94.61%, followed by KNN+ANOVA with an accuracy of 93.33% and KNN+Chi2 of 92.87%. Meanwhile, the NB+Chi2 and NB+ANOVA models recorded accuracies of 91.34% and 92.09%, respectively. On the other hand, the accuracy of the model without feature selection is lower, namely NB at 87.41%, SVM at 80.51%, and KNN with the lowest at 71.25%. These findings indicate that the application of feature selection, especially with the Chi-Square and ANOVA methods, can significantly improve prediction accuracy.

3.3.5 Validation Process on Model

The model training process uses the k-fold cross-validation technique with 5-fold and 10-fold schemes to analyze the extent of generalization towards new data. In 5-fold, the the data is partitioned into five balanced subsets, where four subsets are used for training and one for testing alternately. In 10-fold, the division and testing are carried out with similar principles on ten subsets. This method effectively reduces bias in model evaluation.

1) Model Training with 5 Fold Cross-Validation

Model training using 5-fold cross-validation was performed to evaluate performance more reliably and reduce the chance of overfitting occurring. The training data was divided into five subsets of approximately the same size. The training and validation processes were performed five times, where in each iteration four subsets were used for training and one subset for validation. The results of the validation process are presented in the image 8.

Vol. 6, No. 4, August 2025, Page. 1925-1940
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5017

Figure 8 shows a comparison of the accuracy of three classification algorithms, namely Naive Bayes, K-Nearest Neighbor, and Support Vector Machine, both without and with feature selection using ANOVA and Chi-Square. The best results were obtained in the SVM + ANOVA model with an accuracy of 93.98%, followed by KNN + Chi2 at 93.16%, and SVM + Chi2 at 93.44%. The Naive Bayes model also increased from 84.43% to 92.20% with ANOVA and 91.05% with Chi2. K-Nearest Neighbor without feature selection recorded the lowest accuracy, which was 71.15%, but increased significantly after feature selection. These findings indicate that the use of feature selection methods can significantly improve model accuracy, especially in models with low initial performance.

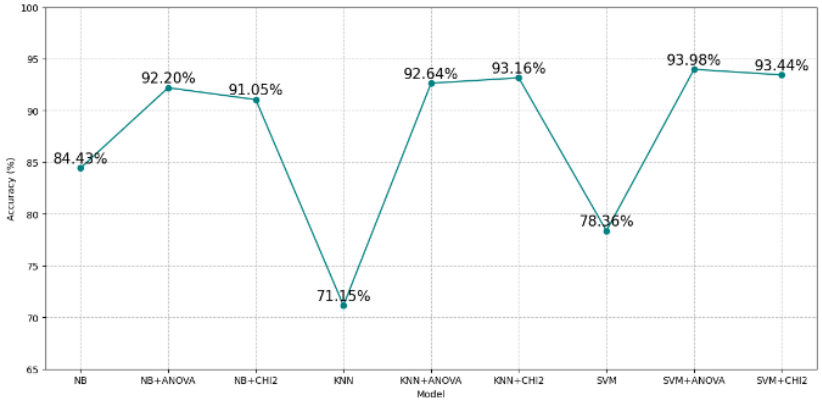

Figure 8. 5 Fold Cross-Validation Test Results

2) Model Training with 10 Fold Cross-Validation

Model training using the 10-fold cross-validation method is done by dividing the dataset into 10 balanced subsets. In each iteration, 9 subsets are used for training and 1 subset for testing, and this process is repeated 10 times until each subset becomes the test data once. The evaluation is obtained from the average accuracy of all iterations, thus providing more stable and reliable results than a single data division. This approach is also effective in reducing the risk of overfitting, especially on small to medium-sized datasets. The validation results are shown in Figure 9.

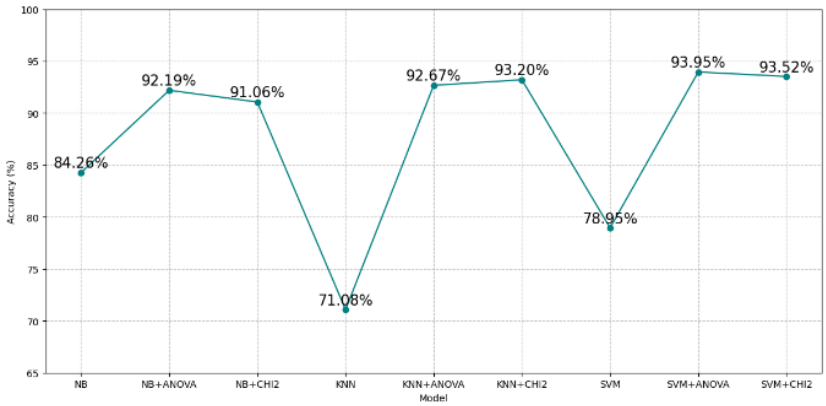

Figure 9. 10 Fold Cross-Validation Test Results

Figure 9 shows a comparison of the accuracy of the Naive Bayes, K-Nearest Neighbor, and Support Vector Machine algorithms, both without and with feature selection using ANOVA and Chi-Square. The best results were achieved by SVM + ANOVA with an accuracy of 93.95%, followed by SVM + Chi2 at 93.52% and KNN + Chi2 at 93.20%. The Naive Bayes model also increased from 84.26% to 92.19% with the ANOVA feature and 91.06% with Chi2. Meanwhile, K-Nearest Neighbor

Vol. 6, No. 4, August 2025, Page. 1925-1940
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5017

without feature selection produced the lowest accuracy of 71.08%, but increased significantly after feature selection. Overall, the application of the feature selection technique has been proven successful in improving the accuracy of the classification model.

3.6 Model Validation

To evaluate the model performance comprehensively, a validation process was carried out using three approaches, namely full training, 5-fold cross-validation, and 10-fold cross-validation. Full training validation was carried out without data sharing, thus providing an overview of the model's accuracy against all available data. Meanwhile, the 5-fold and 10-fold cross-validation approaches were used to test the stability and generalization capabilities of the model by alternating data sharing between training and testing. The results of the model performance comparison of the three validation methods are shown in Figure 10, which shows that the application of the feature selection method consistently increases accuracy, both in the full training and cross-validation schemes.

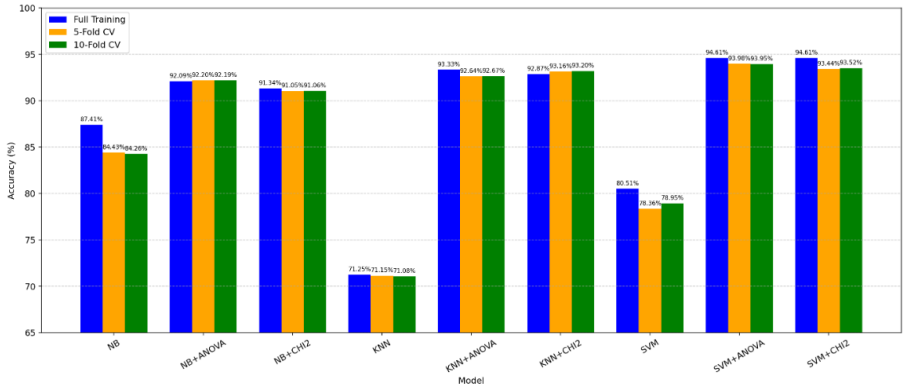

Figure 10. Model Validation Comparisoasd

Figure 10 shows the results of the accuracy evaluation of nine combinations of classification models. To measure the performance and stability of the model as a whole, a validation process was carried out using three approaches, namely full training, 5-fold cross-validation, and 10-fold cross-validation. The validation results are shown in Figure 9 and show that the combination of models with feature selection produces higher and more consistent accuracy than models without feature selection. The SVM + ANOVA model showed the best performance with an accuracy of 93.95%, 93.75%, and 93.54% for each validation method. Followed by SVM + Chi2 at 93.52%, 93.31%, and 93.10%, and KNN + Chi2 at 93.20%, 92.97%, and 92.78%. Meanwhile, KNN + ANOVA recorded an accuracy of 92.67%, 92.45%, and 92.24%.

The Naive Bayes model also showed improvement after applying feature selection. The combination of NB + ANOVA recorded accuracies of 92.19%, 91.98%, and 91.77%, while NB + Chi2 produced accuracies of 91.06%, 90.85%, and 90.62%. In contrast, the model without feature selection produced lower accuracies, namely NB at 84.26%, 83.95%, and 83.71%, KNN at 71.08%, 70.89%, and 70.67%, and SVM at 78.95%, 78.76%, and 78.53%. These findings indicate that the feature selection method not only improves accuracy but also strengthens the model's generalization ability to unseen data, as reflected by the consistency of the results across validation methods.

4. DISCUSSIONS

This discussion provides a deeper interpretation of the results obtained. The Support Vector Machine (SVM) consistently outperformed other algorithms after feature selection due to its ability to find an optimal hyperplane in high-dimensional space. The feature selection process using ANOVA or Chi-Square removes irrelevant features, reducing noise and allowing SVM to focus on the most informative attributes. The K-Nearest Neighbor (KNN) model showed the most significant improvement after feature selection (from 71.25% to above 92%). This is because KNN is highly

sensitive to irrelevant features due to the curse of dimensionality, and removing such features improves its distance calculations.

The similarity in results between ANOVA and Chi-Square indicates that both methods successfully identified nearly the same set of predictive features, despite differences in statistical approach. Comparing with previous studies, the achieved accuracy of 94.61% using SVM+ANOVA surpasses the accuracy reported by Tuba et al. (2021) with 74% using Naive Bayes without feature selection, and Justice et al. (2024) with 89.45% using SVM without feature selection. This confirms the essential contribution of feature selection to optimizing model performance.

From a practical perspective, a model with 94.61% accuracy can serve as a reliable tool to support early anemia detection in healthcare settings, potentially assisting medical practitioners in accelerating diagnosis and improving patient outcomes. Nevertheless, this study has a limitation in that the dataset was obtained from a single source (Kaggle). Therefore, future research should validate the model with clinical datasets collected from diverse populations and varied medical environments to ensure its robustness and applicability.

To evaluate the model performance comprehensively, a validation process was carried out using three approaches, namely full training, 5-fold cross-validation, and 10-fold cross-validation. Full training validation was carried out without data sharing, thus providing an overview of the model's accuracy against all available data. Meanwhile, the 5-fold and 10-fold cross-validation approaches were used to test the stability and generalization capabilities of the model by alternating data sharing between training and testing. The results of the model performance comparison of the three validation methods are shown in Figure 10, which shows that the application of the feature selection method consistently increases accuracy, both in the full training and cross-validation schemes.

This study demonstrates that statistical methods for selecting features such as ANOVA and Chi-Square can significantly improve the accuracy, efficiency, and interpretability of machine learning models. These findings are applicable to various datasets with high feature dimensionality problems, contributing to advancements in data mining, predictive analytics, and decision support systems.

5. CONCLUSION

This study demonstrated that applying ANOVA and Chi-Square feature selection significantly improves the accuracy of Naive Bayes, KNN, and SVM algorithms for anemia classification. The SVM+ANOVA combination achieved the highest accuracy of 94.61%, outperforming models without feature selection. Cross-validation confirmed the stability of these results. Contribution to computer science lies in providing empirical evidence that filter-based feature selection is a crucial preprocessing step for medical datasets with mixed numerical attributes, especially for boundary-based classifiers like SVM. Future work should explore other algorithms such as ensemble methods or deep learning, apply advanced techniques for selecting features (such as Recursive Feature Elimination, LASSO), and validate models on real clinical datasets. Developing a prototype decision support system implementing the best model is also recommended.

ACKNOWLEDGEMENT

This research is supported and funded by a research grant from the Ministry of Higher Education, Science and Technology. [Master Grant Contract Number: 131/C3/DT.05.00/PL/2025]

REFERENCES

[1] T. Qadah and A. Munshi, “Synthesis and Prediction of Anemia from Multi-Data Attribute Co-Existence,” *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3510636.

[2] Muljono, S. A. Wulandari, H. Al Azies, M. Naufal, W. A. Prasetyanto, and F. A. Zahra, “Breaking Boundaries in Diagnosis: Non-Invasive Anemia Detection Empowered by AI,” *IEEE Access*, vol. 12, pp. 9292–9307, 2024, doi: 10.1109/ACCESS.2024.3353788.

[3] T. Qadah and A. Munshi, “Synthesis and Prediction of Anemia from Multi-Data Attribute Co-Existence,” *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3510636.

[4] A. Al-Jawaldeh *et al.*, "Are countries of the eastern mediterranean region on track towards meeting the world health assembly target for anemia? A review of evidence," *Int J Environ Res Public Health*, vol. 18, no. 5, p. 2449, 2021, Accessed: Aug. 06, 2025. [Online]. Available: https://doi.org/10.3390/ijerph18052449

[5] *Development of a Semaphore of Anemia: Screening Method Based on Photographic Images of the Ungueal Bed Using a Digital Camera*. 2020. doi: 10.0/Linux-x86_64.

[6] W. Gardner and N. Kassebaum, "Global, Regional, and National Prevalence of Anemia and Its Causes in 204 Countries and Territories, 1990–2019," *Curr Dev Nutr*, vol. 4, p. nzaa053_035, 2020, doi: https://doi.org/10.1093/cdn/nzaa053_035.

[7] K. Aprilianti Cia *et al.*, "Asupan Zat Besi dan Prevalensi Anemia pada Remaja Usia 16-18 Tahun." doi: https://doi.org/10.31004/jerkin.v3i4.628.

[8] D. Chaerul Ekty Saputra, K. Sunat, and T. Ratnaningsih, "SMOTE-MRS: A Novel SMOTE-Multiresolution Sampling Technique for Imbalanced Distribution to Improve Prediction of Anemia," *IEEE Access*, vol. 12, pp. 154675–154699, 2024, doi: 10.1109/ACCESS.2024.3482968.

[9] M. M. Ali, M. S. Islam, M. N. Uddin, and M. A. Uddin, "A conceptual IoT framework based on Anova-F feature selection for chronic kidney disease detection using deep learning approach," *Intell Based Med*, vol. 10, Jan. 2024, doi: 10.1016/j.ibmed.2024.100170.

[10] B. Heinrichs and S. B. Eickhoff, "Your evidence? Machine learning algorithms for medical diagnosis and prediction," *Hum Brain Mapp*, vol. 41, no. 6, pp. 1435–1444, Apr. 2020, doi: 10.1002/hbm.24886.

[11] J. W. Asare, P. Appiahene, and E. T. Donkoh, "Detection of anaemia using medical images: A comparative study of machine learning algorithms – A systematic literature review," Jan. 01, 2023, *Elsevier Ltd*. doi: 10.1016/j.imu.2023.101283.

[12] T. Karagül Yıldız, N. Yurtay, and B. Öneç, "Classifying anemia types using artificial learning methods," *Engineering Science and Technology, an International Journal*, vol. 24, no. 1, pp. 50–70, Feb. 2021, doi: 10.1016/j.jestch.2020.12.003.

[13] J. W. Asare, W. L. Brown-Acquaye, M. M. Ujakpa, E. Freeman, and P. Appiahene, "Application of machine learning approach for iron deficiency anaemia detection in children using conjunctiva images," *Inform Med Unlocked*, vol. 45, Jan. 2024, doi: 10.1016/j.imu.2024.101451.

[14] O. Peretz, M. Koren, and O. Koren, "Naive Bayes classifier – An ensemble procedure for recall and precision enrichment," *Eng Appl Artif Intell*, vol. 136, Oct. 2024, doi: 10.1016/j.engappai.2024.108972.

[15] Z. Khan *et al.*, "A Framework for Segmentation and Classification of Blood Cells Using Generative Adversarial Networks," *IEEE Access*, vol. 12, pp. 51995–52015, 2024, doi: 10.1109/ACCESS.2024.3378575.

[16] A. Kumar, N. Gaur, and A. Nanthaamornphong, "Machine learning RNNs, SVM and NN Algorithm for Massive-MIMO-OTFS 6G Waveform with Rician and Rayleigh channel," *Egyptian Informatics Journal*, vol. 27, Sep. 2024, doi: 10.1016/j.eij.2024.100531.

[17] T. Wahyuningsih, D. Manongga, I. Sembiring, and S. Wijono, "Comparison of Effectiveness of Logistic Regression, Naive Bayes, and Random Forest Algorithms in Predicting Student Arguments," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 349–356. doi: 10.1016/j.procs.2024.03.014.

[18] *Development of a Semaphore of Anemia: Screening Method Based on Photographic Images of the Ungueal Bed Using a Digital Camera*. 2020. doi: 10.0/Linux-x86_64.

[19] H. Sabir *et al.*, "Fingertip Video Dataset for Non-Invasive Diagnosis of Anemia Using ResNet-18 Classifier," *IEEE Access*, vol. 12, pp. 68880–68892, 2024, doi: 10.1109/ACCESS.2024.3398353.

[20] T. Karagül Yıldız, N. Yurtay, and B. Öneç, "Classifying anemia types using artificial learning methods," *Engineering Science and Technology, an International Journal*, vol. 24, no. 1, pp. 50–70, Feb. 2021, doi: 10.1016/j.jestch.2020.12.003.

[21] M. J. Maasthi, H. L. Gururaj, V. Ravi, B. D, M. Almeshari, and Y. Alzamil, "Decision-making Support System for Predicting and Eliminating Malnutrition and Anemia," *Open Bioinforma J*, vol. 16, no. 1, Nov. 2023, doi: 10.2174/0118750362246898230921054021.

[22] J. W. Asare, W. L. Brown-Acquaye, M. M. Ujakpa, E. Freeman, and P. Appiahene, "Application of machine learning approach for iron deficiency anaemia detection in children using conjunctiva images," *Inform Med Unlocked*, vol. 45, Jan. 2024, doi: 10.1016/j.imu.2024.101451.

[23] P. Iacobescu, V. Marina, C. Anghel, and A. D. Anghele, "Evaluating Binary Classifiers for Cardiovascular Disease Prediction: Enhancing Early Diagnostic Capabilities," Dec. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/jcdd11120396.

[24] D. Zhao *et al.*, "Ant colony optimization with horizontal and vertical crossover search: Fundamental visions for multi-threshold image segmentation," *Expert Syst Appl*, vol. 167, p. 114122, 2021, doi: 10.1016/j.eswa.2020.114122.

[25] F. Hasanzadeh, C. B. Josephson, G. Waters, D. Adedinsewo, Z. Azizi, and J. A. White, "Bias recognition and mitigation strategies in artificial intelligence healthcare applications," Dec. 01, 2025, *Nature Research*. doi: 10.1038/s41746-025-01503-7.

[26] M. A. Shanthi, "Optimizing predictive accuracy: A comparative study of feature selection strategies in the healthcare domain," *The Scientific Temper*, vol. 15, no. spl-1, pp. 217–229, Oct. 2024, doi: 10.58414/scientifictemper.2024.15.spl.26.

[27] N. K. Naik, M. V. Subbarao, P. K. Sethy, S. K. Behera, and G. R. Panigrahi, "Machine learning with analysis-of-variance-based method for identifying rice varieties," *J Agric Food Res*, vol. 18, Dec. 2024, doi: 10.1016/j.jafr.2024.101397.

[28] N. Nasution, F. Nasution, E. Erlin, and M. Hasan, "Evaluation Study of the Chi-Square Method for Feature Selection in Stroke Prediction with Random Forest Regression," European Alliance for Innovation n.o., May 2024. doi: 10.4108/eai.30-10-2023.2343096.

[29] M. M. Ali, M. S. Islam, M. N. Uddin, and M. A. Uddin, "A conceptual IoT framework based on Anova-F feature selection for chronic kidney disease detection using deep learning approach," *Intell Based Med*, vol. 10, Jan. 2024, doi: 10.1016/j.ibmed.2024.100170.

[30] S. Bani Hani and M. Ahmad, "Effective Prediction of Mortality by Heart Disease Among Women in Jordan Using the Chi-Squared Automatic Interaction Detection Model: Retrospective Validation Study," *JMIR Cardio*, vol. 7, 2023, doi: https://doi.org/10.2196/48795.

[31] I. Moura, A. Teles, D. Viana, J. Marques, L. Coutinho, and F. Silva, "Digital Phenotyping of Mental Health using multimodal sensing of multiple situations of interest: A Systematic Literature Review," Feb. 01, 2023, *Academic Press Inc.* doi: 10.1016/j.jbi.2022.104278.

[32] G. Y. Lee, L. Alzamil, and B. Doskenov, "A Survey on Data Cleaning Methods for Improved Machine Learning Model Performance", doi: 10.48550/arXiv.2109.07127.

[33] S. Shahrabadi, T. Adão, E. Peres, R. Morais, L. G. Magalhães, and V. Alves, "Automatic Optimization of Deep Learning Training through Feature-Aware-Based Dataset Splitting," *Algorithms*, vol. 17, no. 3, Mar. 2024, doi: 10.3390/a17030106.

[34] R. Blanquero, E. Carrizosa, P. Ramírez-Cobo, and M. R. Sillero-Denamiel, "Variable selection for Naïve Bayes classification," *Comput Oper Res*, vol. 135, Nov. 2021, doi: 10.1016/j.cor.2021.105456.

[35] X. Ou *et al.*, "Hyperspectral Image Target Detection via Weighted Joint K-Nearest Neighbor and Multitask Learning Sparse Representation," *IEEE Access*, vol. 8, pp. 11503–11511, 2020, doi: 10.1109/ACCESS.2019.2962875.

[36] A. Venkataramana, K. Suresh Kumar, N. Suganthi, and R. Rajeswari, "Prediction of Brinjal Plant Disease Using Support Vector Machine and Convolutional Neural Network Algorithm Based on Deep Learning," *Journal of Mobile Multimedia*, vol. 18, no. 3, pp. 771–788, 2022, doi: 10.13052/jmm1550-4646.18315.

[37] C. Wu *et al.*, "SEMG Measurement Position and Feature Optimization Strategy for Gesture Recognition Based on ANOVA and Neural Networks," *IEEE Access*, vol. 8, pp. 56290–56299, 2020, doi: 10.1109/ACCESS.2020.2982405.

[38] N. Su, X. An, C. Yan, and S. Ji, "Incremental attribute reduction method based on chi-square statistics and information entropy," *IEEE Access*, vol. 8, pp. 98234–98243, 2020, doi: 10.1109/ACCESS.2020.2997013.

[39] I. Markoulidakis and G. Markoulidakis, "Probabilistic Confusion Matrix: A Novel Method for Machine Learning Algorithm Generalized Performance Analysis," *Technologies (Basel)*, vol. 12, no. 7, Jul. 2024, doi: 10.3390/technologies12070113.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 4, August 2025, Page. 1925-1940
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5017