Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 1, February 2026, Page. 361-372
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5422

Optimized KNN Performance with PCA and K-Fold Cross-Validation for Colorectal Cancer Survival Prediction

Yuke Manza[1], Rika Rosnelly[*2], Mhd Furqan[3], Bob Subhan Riza[4]

[1,2,4]Computer Science, Universitas Potensi Utama, Indonesia
[3]Universitas Islam Negeri Sumatera Utara, Indonesia

Email: [1]rika@potensi-utama.ac.id

Received : Nov 3, 2025; Revised : Dec 14, 2025; Accepted : Dec 15, 2025; Published : Feb 15, 2025

Abstract

Colorectal cancer remains a leading cause of global mortality, necessitating effective predictive tools for patient survival. While Machine Learning algorithms like K-Nearest Neighbors (KNN) utilize patient data for prediction, standard KNN implementations often suffer from the curse of dimensionality and overfitting, leading to unreliable performance on complex medical datasets. This study aims to evaluate and optimize the performance of the KNN algorithm by integrating Principal Component Analysis (PCA) for dimensionality reduction and K-Fold Cross-Validation (KFCV) to enhance model stability. The research utilized a quantitative approach on a global colorectal cancer dataset, processing demographic and clinical features through a rigorous pipeline of imputation, encoding, and normalization. Three model configurations were systematically compared: Standard KNN, KNN combined with PCA, and an optimized KNN model utilizing both PCA and KFCV across various neighbor values. The results demonstrate a distinct trade-off between predictive sensitivity and model stability. While the Standard KNN and PCA-enhanced models achieved higher recall, indicating a strong ability to identify survivors in a single data split, the fully optimized KNN+PCA+KFCV model provided the most stable and generalized accuracy with minimal deviation. These findings indicate that while PCA effectively reduces computational complexity without information loss, the integration of cross-validation is crucial for obtaining an honest assessment of model performance. This research contributes to clinical informatics by highlighting the necessity of prioritization between high sensitivity and generalization stability when developing survival prediction models for complex, inseparable medical data.

Keywords: *Colorectal cancer prediction, K-fold cross validation, K-Nearest Neighbors, Machine learning, Principal Component Analysis, Survival prediction.*

1. INTRODUCTION

Colorectal cancer (CRC), a malignancy that grows in the colon or rectum [1], stands as one of the most prevalent cancers and the second leading cause of cancer-related mortality worldwide [1-7]. A multitude of risk factors contribute to its development, including age, diet, physical inactivity, and a family history of the disease [2, 4, 5, 8, 9]. The cancer's slowly progressive nature presents a critical window for early detection through established screening methods like colonoscopy or fecal tests [5, 8]. Early detection is paramount as it significantly increases the chances of successful treatment and recovery. Despite this, public awareness regarding the importance of screening remains insufficient, often compounded by limited access to healthcare services [5, 8, 9].

In this context, machine learning (ML) offers a promising and powerful solution by providing a quantitative and data-driven approach to analyzing and predicting colorectal cancer [10]. Machine learning, a prominent branch of Artificial Intelligence (AI) [11, 12], excels at identifying complex patterns within large datasets [13], processing big data efficiently [14], and accelerating complex decision-making processes [12]. Its operational principle is analogous to human learning, where it

Vol. 7, No. 1, February 2026, Page. 361-372
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5422

improves its performance based on data [15, 16]. Among the various algorithms, K-Nearest Neighbors (KNN) is a well-established method [17, 18] proven to be effective for datasets of both small and large dimensions [13, 19]. Its utility has been demonstrated with high accuracy in diverse predictive tasks, including diagnosing heart disease [20, 21], predicting skin diseases [19], predicting lung cancer [22], optimizing biodiesel production [23], detecting cybercrime [24], classifying images [25-27], and analyzing sentiment on social media [28, 29].

However, despite being a powerful tool, the standard KNN algorithm possesses several inherent weaknesses that limit its performance on complex medical data. These include relatively low accuracy without optimization, high sensitivity to the choice of the k-value, slow computation times on large and high-dimensional datasets (the curse of dimensionality), susceptibility to irrelevant features, and the risk of overfitting [25].

Various studies have explored KNN optimization. For instance, some studies [20, 21] focused on feature selection or Grid Search for heart disease diagnosis. Another study [22] compared KNN with methods like XGBoost for lung cancer prediction but often did not deeply address model stability. While techniques like K-Fold Cross-Validation (KFCV) [30-33] and Principal Component Analysis (PCA) [34-40] are well-known, and some recent 2025 studies have even compared KNN with PCA for CRC classification [41], a significant research gap persists. These studies, such as Bahrambanan et al. [41] which noted PCA's ability to improve Random Forest accuracy more significantly than KNN accuracy, primarily focus on classification accuracy alone. The specific combined application of both PCA and KFCV to systematically analyze the critical performance trade-off—balancing stable, generalized accuracy from KFCV against the high clinical sensitivity (recall) of standard models—remains an underexplored area, particularly in the context of patient survival prediction rather than just diagnosis.

Therefore, this study aims to fill this gap by developing and evaluating a systematically optimized KNN model. The primary objective of this research is to analyze the performance and metric trade-offs of a KNN model optimized using Principal Component Analysis (PCA) for dimensionality reduction and K-Fold Cross-Validation (KFCV) for validation and hyperparameter tuning. This research will compare three model configurations (Standard KNN, KNN+PCA, and KNN+PCA+KFCV) to determine which optimization pipeline yields the most stable accuracy versus the highest recall for CRC patient survival prediction.

2. METHOD

This research method uses two approaches:

a. A qualitative approach for the collection and pre-processing of colorectal cancer data.
b. A quantitative approach using the K-Nearest Neighbors method, optimized with K-Fold Cross Validation and Principal Component Analysis, for colorectal cancer prediction.

This research was conducted through a systematic quantitative approach, encompassing several key stages from data preparation and model development to performance evaluation. The methodology is designed to build and optimize a predictive model for colorectal cancer using the K-Nearest Neighbors algorithm (Figure 1).

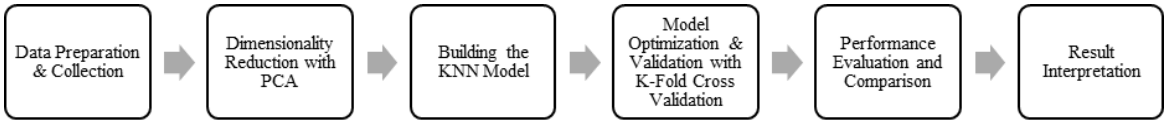

Figure 1. Research Flow

2.1. Data Preparation and Collection

Vol. 7, No. 1, February 2026, Page. 361-372
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5422

The initial phase of this study involved collecting a dataset suitable for colorectal cancer analysis from a trusted public repository (https://www.kaggle.com/datasets/ankushpanday2/colorectal-cancer-global-dataset-and-predictions). This dataset includes various features (variables) that are relevant and crucial for prediction, including demographic data such as age and gender, as well as clinical and lifestyle factors such as family history and smoking history. The dataset consists of 167,497 rows and 28 columns.

To ensure the quality and suitability of the data for modeling, a rigorous pre-processing pipeline was applied. This process included three main steps: (1) Handling missing values was performed using median imputation for numerical features to preserve the data distribution; (2) Categorical data (e.g., gender) were encoded using one-hot encoding; and (3) Numerical features were normalized to a 0 to 1 scale to prevent variables with larger ranges from disproportionately influencing the model. All data processing and model implementation were conducted using the Python (version 3.x) programming language with the scikit-learn, pandas, and numpy libraries.

2.2. Dimensionality Reduction with Principal Component Analysis (PCA)

To address the potential issue of high dimensionality, reduce computational complexity, and mitigate multicollinearity, Principal Component Analysis (PCA) was employed as a feature reduction technique, shown in Equation (1). The process began with the extraction of all numerical features from the pre-processed dataset. Subsequently, a covariance matrix was calculated to map the inter-variable relationships. Through eigenvalue decomposition of this matrix, a set of orthogonal principal components was generated. Only the components that collectively explained more than 95% of the total variance in the data were selected for the final model. Finally, the original dataset was transformed into a new, lower-dimensional feature space defined by these selected components

$$\Sigma w = \lambda w. \tag{1}$$

To interpret the contribution of original features to the principal components, an inverse PCA transformation can be applied, as shown in Equation (2):

$$X_{\text{reconstructed}} = Z \cdot W^T \tag{2}$$

Where Z is the matrix of principal components and W^{T} is the transpose of the component loadings matrix.

2.3. Building the KNN Model

The transformed dataset was partitioned into a training set and a testing set, using an 80-20 split. Eighty percent of the data was allocated for model training, while the remaining twenty percent was reserved for an unbiased evaluation of its performance. A baseline K-Nearest Neighbors (KNN) model was first constructed. The initial configuration involved determining key hyperparameters, such as the number of neighbors (k) and the distance metric (Euclidean distance), shown in Equation (3). The model was trained using the training data. To establish a performance benchmark, this unoptimized model was evaluated on the test set. Its performance was measured using standard classification metrics, including accuracy, precision, recall, and F1-score.

$$d(p, q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \tag{3}$$

2.4. Model Optimization and Validation with K-Fold Cross-Validation

To enhance the model's robustness, prevent overfitting, and obtain a more reliable estimate of its performance, a 10-Fold Cross-Validation technique was implemented on the training data, shown in

Vol. 7, No. 1, February 2026, Page. 361-372
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5422

Equation (4). In this procedure, the training set was divided into ten equal-sized folds. The model was trained and validated ten times, with each iteration using a different fold as the validation set and the remaining nine folds as the training set. The model's hyperparameters (such as the optimal value of *k*) were tuned based on the average performance across all folds. This ensures the resulting model is stable and generalizes well to unseen data.

$$CV_{Score} = \frac{1}{k}\sum_{i=1}^{k} Score_i \quad (4)$$

2.5. Performance Evaluation and Comparison

The final, optimized model's performance was comprehensively evaluated on the reserved test set. The evaluation was based on several key metrics, shown in Equation (5, 6, 7, 8):

a. Accuracy: The proportion of correctly classified instances.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

b. Precision: The model's ability to correctly identify positive cases from all predicted positive cases.

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

c. Recall (Sensitivity): The model's ability to identify all actual positive cases.

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

d. F1-Score: The harmonic mean of precision and recall, providing a balanced measure.

$$F1 - Score = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (8)$$

e. Confusion Matrix: A table used to visualize the performance of the classification model, detailing true positives, true negatives, false positives, and false negatives.

A central part of this research was the comparison of the model's performance before and after optimization with PCA and K-Fold Cross-Validation. This comparison quantitatively demonstrates the effectiveness of the proposed optimization techniques.

2.6. Result Interpretation

The final stage of the methodology involves the interpretation of the model's results. An important feature analysis will be conducted by applying an inverse PCA transform to identify which of the original clinical or demographic features contributed most significantly to the principal components used for prediction. Furthermore, the findings will be discussed in the context of their clinical implications, exploring how the developed model could potentially be used as a tool to support clinicians in the early diagnosis or risk stratification of colorectal cancer patients.

2.7. Data Ethics Statement

This research utilized a publicly available and anonymized dataset from the Kaggle repository. All data used was de-identified, ensuring patient privacy and adhering to the data usage terms provided by the platform.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 1, February 2026, Page. 361-372
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5422

3. RESULT

This section presents the empirical findings of the study, beginning with an exploratory data analysis (EDA) to understand the characteristics of the patient dataset, followed by a comparative performance evaluation of the developed K-Nearest Neighbors (KNN) models.

3.1. Exploratory Data Analysis (EDA)

Exploratory data analysis was conducted to identify patterns and distributions within the colorectal cancer patient data.

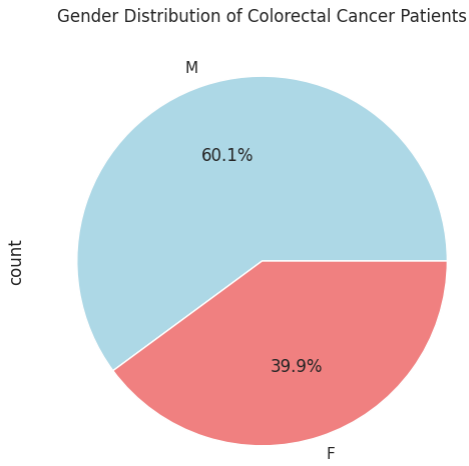

Figure 2. Gender Distribution of Colorectal Cancer Patients

Figure 2 illustrates the gender distribution within the dataset. The data shows a predominance of male (M) patients, accounting for 60.1% of the total cases, while female (F) patients comprise the remaining 39.9%.

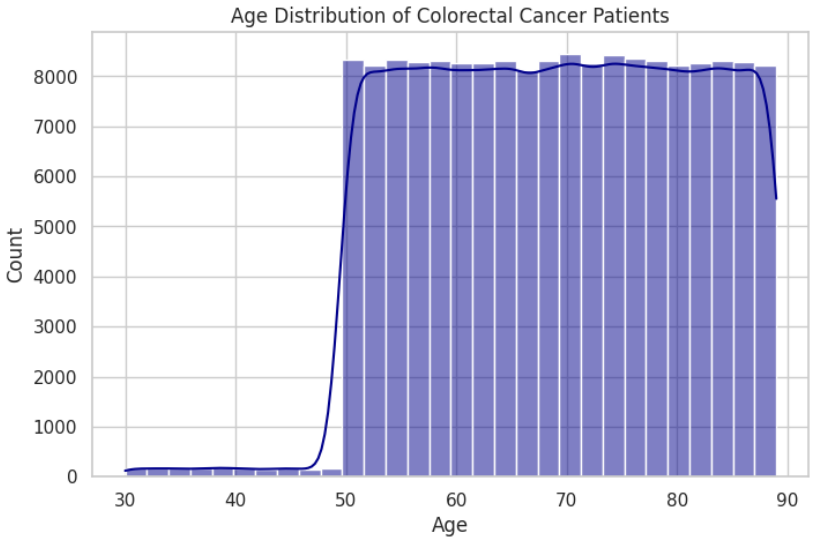

Figure 3. Age Distribution of Colorectal Cancer Patients

The age distribution of patients is presented in Figure 3. A sharp increase in the number of patients is observed starting around age 50. This distribution remains relatively high and stable for

Vol. 7, No. 1, February 2026, Page. 361-372
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5422

patients in the 50 to 90 age range, indicating that the majority of patients in this dataset are over 50 years old.

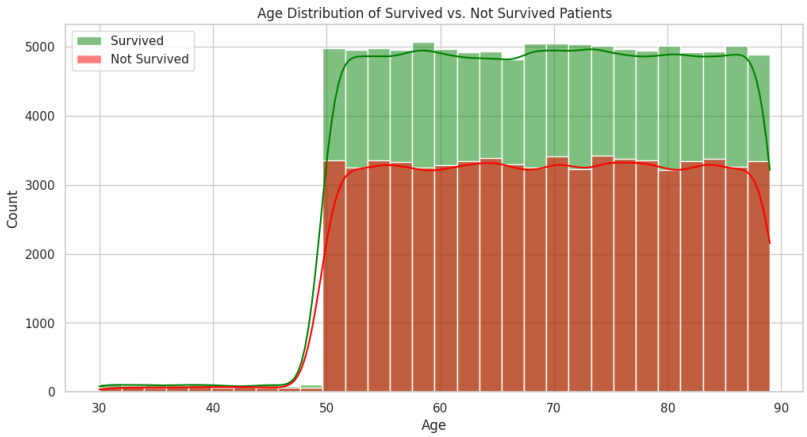

Figure 4. Age Distribution of Survived vs. Not Survived Patients

Figure 4 provides a more detailed view by comparing the age distribution between 'Survived' (green) and 'Not Survived' (red) patients. Both groups exhibit a similar trend with a spike in cases after age 50. Across the entire 50-90 age range, the count of 'Survived' patients consistently appears higher than the count of 'Not Survived' patients.

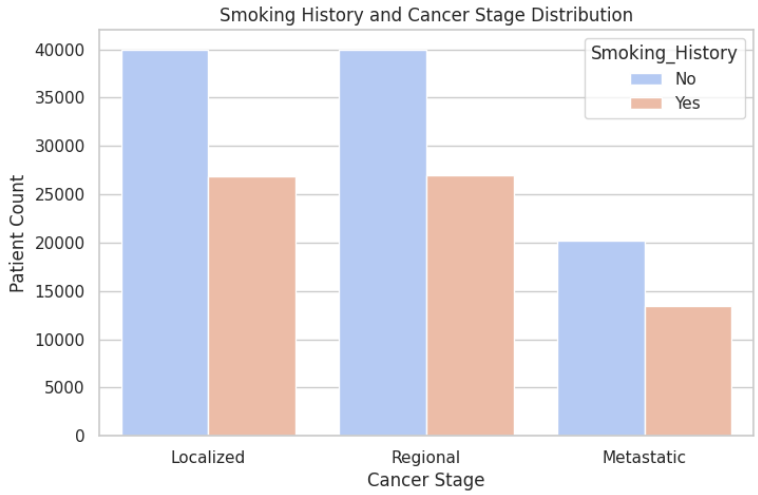

Figure 5. Smoking History and Cancer Stage Distribution

Figure 5 presents a comparative analysis of smoking history across three cancer stages (Localized, Regional, and Metastatic). In all three stages, the number of patients who do not smoke ('No') is substantially higher than the number of patients with a smoking history ('Yes').

3.2. Principal Component Analysis Variance

PCA was applied to the numerical features of the dataset. The analysis identified that 16 principal components were sufficient to explain 90% of the total variance in the data. This reduction from 23 features to 16 components significantly reduced computational complexity while retaining the vast majority of the data's informational content. Subsequent models were trained using this transformed dataset.

3.3. Model Performance Evaluation

Vol. 7, No. 1, February 2026, Page. 361-372
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5422

The quantitative assessment of the proposed methods is presented in Table 1. This evaluation compares three distinct configurations: standard K-Nearest Neighbors (KNN), KNN with Principal Component Analysis (KNN + PCA), and the optimized KNN model incorporating both PCA and K-Fold Cross-Validation (KNN + PCA + KFCV). The models were tested across varying K-values (K=3, 5, 7, 9) to analyze the impact of neighborhood size on classification performance.

Table 1. Comparative Classification Report of KNN Models

K Value	Classification Report	KNN	KNN + PCA	KNN + PCA + KFCV (k=10)
3	Accuracy	52.80%	52.87%	52.85% (Std: 0.26%)
	Precision	59.82%	59.87%	51.90%
	Recall	64.82%	64.96%	52.85%
	F1-Score	62.22%	62.31%	52.25%
5	Accuracy	53.32%	53.46%	53.59% (Std: 0.48%)
	Precision	59.72%	59.78%	51.93%
	Recall	68.05%	68.40%	53.59%
	F1-Score	63.61%	63.80%	52.40%
7	Accuracy	53.82%	53.99%	54.10% (Std: 0.25%)
	Precision	59.73%	59.84%	51.91%
	Recall	70.55%	70.72%	54.10%
	F1-Score	64.69%	64.83%	52.38%
9	Accuracy	54.39%	54.54%	54.71% (Std: 0.33%)
	Precision	59.81%	59.90%	52.08%
	Recall	72.97%	73.18%	54.71%
	F1-Score	65.74%	65.88%	52.47%

KNN: K-Nearest Neighbors
PCA: Principal Component Analysis
KFCV: K-Fold Cross Validation

The data reveals a consistent trend where increasing the K-value yields marginal improvements in accuracy across all configurations. The highest accuracy was achieved at K=9, with the KNN + PCA + KFCV model reaching 54.71%.

Interestingly, while the KNN + PCA + KFCV configuration provided the highest accuracy and stability (indicated by low standard deviations, e.g., $\pm 0.33\%$ at K=9), the standard KNN and KNN + PCA models demonstrated significantly higher values for Precision, Recall, and F1-Score. Specifically, at K=9, the KNN + PCA model achieved a Recall of 73.18% and an F1-Score of 65.88%, compared to the KFCV model which recorded a Recall of 54.71% and an F1-Score of 52.47%.

The integration of PCA (comparing Standard KNN vs. KNN + PCA) resulted in slight performance uplifts across all metrics. For instance, at K=9, PCA improved the F1-Score from 65.74% to 65.88%, suggesting that dimensionality reduction successfully retained the essential variance required for classification while slightly mitigating noise.

4. DISCUSSIONS

This study aimed to optimize the K-Nearest Neighbors algorithm for colorectal cancer survival prediction using dimensionality reduction (PCA) and robust validation techniques (KFCV). The results present a complex picture of the trade-offs between model stability, generalization, and predictive capability on this specific dataset.

4.1. Performance and Data Complexity

Vol. 7, No. 1, February 2026, Page. 361-372
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5422

The overall accuracy across all models hovers between 52% and 55%. While the KNN + PCA + KFCV model achieved the highest accuracy of 54.71% (K=9), these results suggest that the classification task is challenging. The overlap between the 'Survived' and 'Not Survived' classes in the feature space appears significant, making it difficult for a linear or distance-based classifier like KNN to draw a distinct decision boundary. However, the upward trend in performance as K increases (from 3 to 9) indicates that a broader neighborhood helps smooth out local noise, leading to slightly better generalization.

4.2. The Role of PCA and Optimization

The comparison between standard KNN and KNN + PCA validates the utility of dimensionality reduction. The KNN + PCA model consistently outperformed the standard KNN, albeit by narrow margins (e.g., +0.15% Accuracy at K=9). This confirms that PCA successfully reduced the computational burden and filtered out some variance attributed to noise without losing critical information.

4.3. Stability vs. Sensitivity (KFCV Analysis)

A critical observation is the discrepancy between the single-split models (KNN and KNN + PCA) and the cross-validated model (KNN + PCA + KFCV). The single-split models achieved relatively high Recall (~73%) and F1-Scores (~65%), whereas the KFCV model showed lower, flattened metrics (F1-Score ~52%). This disparity likely indicates that the single train-test split used in the first two columns may have contained a favorable distribution of data, artificially inflating the sensitivity (Recall). In contrast, K-Fold Cross-Validation provides a more rigorous and honest estimate of the model's performance. The low standard deviation in the KFCV results (ranging from 0.25% to 0.48%) demonstrates that while the accuracy is modest, the model is highly stable and robust to variations in the training data. The KFCV results reveal the true difficulty of generalizing on this dataset, stripping away the bias of a "lucky" data split.

4.4. Clinical Implications and Trade-offs

From a clinical perspective, the choice of model depends on the metric of priority. If the goal is to maximize the identification of survivors (Recall), the KNN + PCA configuration appears superior with a Recall of 73.18%. However, the low Precision (~59%) implies a considerable number of false positives. Conversely, the KNN + PCA + KFCV model offers a more conservative and stable prediction baseline.

4.5. Limitations

The modest accuracy highlights the limitations of using KNN on this dataset. The features derived from the clinical and demographic data may lack sufficient separability in their current form. Future work should focus on feature engineering, non-linear transformations, or the application of more complex ensemble methods (such as Random Forest or Gradient Boosting) to better capture the underlying patterns of colorectal cancer survival.

5. CONCLUSION

This research successfully demonstrates the application and rigorous evaluation of the K-Nearest Neighbors algorithm for predicting colorectal cancer patient survival. Through the systematic integration of Principal Component Analysis and K-Fold Cross-Validation, the study reveals that the optimization strategy fundamentally alters the model's performance profile. The application of PCA proved effective in reducing the dimensionality of the clinical dataset, streamlining computational efficiency while maintaining the informational integrity required for classification. However, the core

Vol. 7, No. 1, February 2026, Page. 361-372
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5422

finding of this investigation lies in the significant performance divergence observed between single-split testing and cross-validation.

The analysis concludes that relying solely on standard train-test splits can yield optimistically biased sensitivity metrics, whereas the K-Fold Cross-Validation approach provides a more conservative but highly stable and reliable estimation of the model's true capabilities on unseen data. Although the overall accuracy indicates that the linear separation of survival classes remains a complex challenge, the low standard deviation achieved by the optimized model confirms its robustness against data variability. Consequently, this study suggests that while KNN serves as a functional baseline, future development of clinical decision support systems for this specific domain should prioritize model stability over inflated recall scores and consider exploring non-linear ensemble methods to overcome the inherent separability issues within the feature space.

CONFLICT OF INTEREST

The authors declare no conflict of interest regarding the research, authorship, or publication of this article.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the Directorate General of Higher Education, Research, and Technology (Ditjen Diktiristek), Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia, for the financial support provided through the Master's Thesis Research Grant scheme for the 2025 fiscal year, which made this research possible.

Sincere appreciation is also extended to Prof. Dr. Rika Rosnelly, S.Kom., M.Kom. as thesis supervisor, for their invaluable guidance, direction, and constructive discussions throughout this research and manuscript preparation.

The authors are also grateful to the provider of the "Colorectal Cancer Global Dataset and Predictions" dataset, which was made publicly available on the Kaggle platform and served as the primary data object for this study.

REFERENCES

[1] A. Lewandowska, G. Rudzki, T. Lewandowski, A. Stryjkowska-Góra, and S. Rudzki, "Risk factors for the diagnosis of colorectal cancer," *Cancer Control*, vol. 29, p. 10732748211056692, 2022. [Online]. Available: https://journals.sagepub.com/doi/full/10.1177/10732748211056692

[2] Y. Xi and P. Xu, "Global colorectal cancer burden in 2020 and projections to 2040," *Transl. Oncol.*, vol. 14, no. 10, p. 101174, Oct. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1936523321001662

[3] S. Shinji *et al.*, "Recent advances in the treatment of colorectal cancer: a review," *J. Nippon Med. Sch.*, vol. 89, no. 3, pp. 246–254, 2022. [Online]. Available: https://www.jstage.jst.go.jp/article/jnms/89/3/89_JNMS.2022_89-310/_article/-char/ja/

[4] J. C. Sedlak, Ö. H. Yilmaz, and J. Roper, "Metabolism and colorectal cancer," *Annu. Rev. Pathol. Mech. Dis.*, vol. 18, no. 1, pp. 467–492, 2023. [Online]. Available: https://www.annualreviews.org/content/journals/10.1146/annurev-pathmechdis-031521-041113

[5] S. M. Alzahrani, H. A. Al Doghaither, and A. B. Al-Ghafari, "General insight into cancer: An overview of colorectal cancer," *Mol. Clin. Oncol.*, vol. 15, no. 6, p. 271, Dec. 2021. [Online]. Available: https://www.spandidos-publications.com/10.3892/mco.2021.2433

[6] M. Bretthauer *et al.*, "Effect of colonoscopy screening on risks of colorectal cancer and related death," *N. Engl. J. Med.*, vol. 387, no. 17, pp. 1547–1556, Oct. 2022. [Online]. Available: https://www.nejm.org/doi/full/10.1056/NEJMoa2208375

[7] D. C. Chung *et al.*, "A cell-free DNA blood-based test for colorectal cancer screening," *N. Engl. J. Med.*, vol. 390, no. 11, pp. 973–983, Mar. 2024. [Online]. Available:

Vol. 7, No. 1, February 2026, Page. 361-372
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5422

https://www.nejm.org/doi/full/10.1056/NEJMoa2304714
[8] T. Sawicki, M. Ruszkowska, A. Danielewicz, E. Niedźwiedzka, T. Arłukowicz, and K. E. Przybyłowicz, "A review of colorectal cancer in terms of epidemiology, risk factors, development, symptoms and diagnosis," *Cancers (Basel)*, vol. 13, no. 9, p. 2025, May 2021. [Online]. Available: https://www.mdpi.com/2072-6694/13/9/2025
[9] V. A. Ionescu, G. Gheorghe, N. Bacalbasa, A. L. Chiotoroiu, and C. Diaconu, "Colorectal cancer: from risk factors to oncogenesis," *Medicina (Kaunas)*, vol. 59, no. 9, p. 1646, Sep. 2023. [Online]. Available: https://www.mdpi.com/1648-9144/59/9/1646
[10] N. H. Minh, T. Q. Quy, N. D. Tam, T. M. Tuan, and L. H. Son, "A practical approach for colorectal cancer diagnosis based on machine learning," *PLoS One*, vol. 20, no. 4, p. e0321009, Apr. 2025. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC12040227/
[11] M. R. Wayahdi, F. Ruziq, and S. H. Ginting, "AI approach to predict student performance (Case study: Battuta University)," *J. Sci. Soc. Res.*, vol. 7, no. 4, pp. 1800–1807, Nov. 2024. [Online]. Available: https://www.jurnal.goretanpena.com/index.php/JSSR/article/view/2332
[12] V. Galaz *et al.*, "Artificial intelligence, systemic risks, and sustainability," *Technol. Soc.*, vol. 67, p. 101741, Nov. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0160791X21002165
[13] R. K. Halder, M. N. Uddin, M. A. Uddin, S. Aryal, and A. Khraisat, "Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications," *J. Big Data*, vol. 11, no. 1, p. 113, Aug. 2024. [Online]. Available: https://link.springer.com/article/10.1186/s40537-024-00973-y
[14] M. R. Wayahdi and M. Zaki, "The Role of AI in Diagnosing Student Learning Needs: Solutions for More Inclusive Education," *Int. J. Educ. Insights Innov.*, vol. 2, no. 1, pp. 1–7, Mar. 2025. [Online]. Available: https://ijedins.technolabs.co.id/index.php/ijedins/article/view/6
[15] M. M. Taye, "Understanding of machine learning with deep learning: architectures, workflow, applications and future directions," *Computers*, vol. 12, no. 5, p. 91, Apr. 2023. [Online]. Available: https://www.mdpi.com/2073-431X/12/5/91
[16] K. Sharifani and M. Amini, "Machine learning and deep learning: A review of methods and applications," *World Inf. Technol. Eng. J.*, vol. 10, no. 07, pp. 3897–3904, 2023. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4458723
[17] S. Ramadhani and M. R. Wayahdi, "K-Nearest Neighbor and Random Forest Algorithms in Loan Approval Prediction," *J. Minfo Polgan*, vol. 13, no. 1, pp. 1307–1313, Dec. 2024. [Online]. Available: https://jurnal.polgan.ac.id/index.php/jmp/article/view/14345
[18] M. R. Wayahdi and F. Ruziq, "Predicting Smartphone Addiction Levels with K-Nearest Neighbors Using User Behavior Patterns", *J. Tek. Inform. (JUTIF)*, vol. 6, no. 5, pp. 3379–3391, Oct. 2025. [Online]. Available: https://jutif.if.unsoed.ac.id/index.php/jurnal/article/view/4905
[19] M. A. Araaf, K. Nugroho, and D. R. Setiadi, "Comprehensive analysis and classification of skin diseases based on image texture features using K-nearest neighbors algorithm," *J. Comput. Theor. Appl.*, vol. 1, no. 1, pp. 31–40, Sep. 2023. [Online]. Available: https://dl.futuretechsci.org/id/eprint/85/
[20] M. A. Khan *et al.*, "Optimal feature selection for heart disease prediction using modified Artificial Bee colony (M-ABC) and K-nearest neighbors (KNN)," *Sci. Rep.*, vol. 14, no. 1, p. 26241, Oct. 2024. [Online]. Available: https://www.nature.com/articles/s41598-024-78021-1
[21] S. Hidayat, H. M. Ramadhan, and E. Y. Puspaningrum, "Comparison of K-nearest neighbor and decision tree methods using principal component analysis technique in heart disease classification," *Indones. J. Data Sci.*, vol. 4, no. 2, pp. 87–96, Jul. 2023. [Online]. Available: https://www.jurnal.yoctobrain.org/index.php/ijodas/article/view/70
[22] M. R. Wayahdi and F. Ruziq, "KNN and XGBoost Algorithms for Lung Cancer Prediction," *J. Sci. Technol. (JoSTec)*, vol. 4, no. 1, Jan. 2022. [Online]. Available: https://ejournal.ipinternasional.com/index.php/jostec/article/view/251
[23] A. Sumayli, "Development of advanced machine learning models for optimization of methyl ester biofuel production from papaya oil: Gaussian process regression (GPR), multilayer perceptron (MLP), and K-nearest neighbor (KNN) regression models," *Arab. J. Chem.*, vol. 16, no. 7, p. 104833, Jul. 2023. [Online]. Available:

Vol. 7, No. 1, February 2026, Page. 361-372
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5422

https://www.sciencedirect.com/science/article/pii/S1878535223002952

[24] D. M. Cao *et al.*, "Advanced cybercrime detection: A comprehensive study on supervised and unsupervised machine learning approaches using real-world datasets," *J. Comput. Sci. Technol. Stud.*, vol. 6, no. 1, pp. 40–48, Jan. 2024. [Online]. Available: https://www.neliti.com/publications/589855/advanced-cybercrime-detection-a-comprehensive-study-on-supervised-and-unsupervis

[25] M. R. Wayahdi, D. Syahputra, and S. H. Ginting, "Evaluation of the K-Nearest Neighbor Model With K-Fold Cross Validation on Image Classification," *Infokum*, vol. 9, no. 1, pp. 1–6, Dec. 2020. [Online]. Available: http://seaninstitute.org/infor/index.php/infokum/article/view/72

[26] M. Jagdish, A. M. Guzman, G. F. Sancho, and A. Guerrero-Luzuriaga, "Detection and classification of caterpillar using image processing with K-nearest neighbor classification technique," *Turk. J. Comput. Math. Educ.*, vol. 12, no. 5, pp. 719–728, 2021. [Online]. Available: https://www.proquest.com/openview/5fdb289afbaaf45f991102c89e259cf2/1?cbl=2045096&pq-origsite=gscholar

[27] S. Anraeni, D. Indra, D. Adirahmadi, S. Pomalingo, and S. H. Mansyur, "Strawberry ripeness identification using feature extraction of RGB and K-nearest neighbor," in *2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, 2021, pp. 395–398. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9431854

[28] M. I. Hutapea and A. P. Silalahi, "Moderna's Vaccine Using the K-Nearest Neighbor (KNN) Method: An Analysis of Community Sentiment on Twitter," *J. Penelit. Pendidik. IPA*, vol. 9, no. 5, pp. 3808–3814, May 2023. [Online]. Available: https://jppipa.unram.ac.id/index.php/jppipa/article/view/3203

[29] S. Masturoh, R. L. Pratiwi, M. R. Saelan, and U. Radiyah, "Application of the k-nearest neighbor (KNN) algorithm in sentiment analysis of the Ovo e-wallet application," *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, vol. 8, no. 2, pp. 84–89, Jan. 2023. [Online]. Available: https://ejournal.nusamandiri.ac.id/index.php/jitk/article/view/3997

[30] Z. R. Tembusai, H. Mawengkang, and M. Zarlis, "K-nearest neighbor with k-fold cross validation and analytic hierarchy process on data classification," *Int. J. Adv. Data Inf. Syst.*, vol. 2, no. 1, pp. 1–8, 2021. [Online]. Available: https://www.neliti.com/publications/396954/k-nearest-neighbor-with-k-fold-cross-validation-and-analytic-hierarchy-process-o

[31] Z. A. Sejuti and M. S. Islam, "A hybrid CNN–KNN approach for identification of COVID-19 with 5-fold cross validation," *Sens. Int.*, vol. 4, p. 100229, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666351123000037

[32] A. M. P. Chacón, I. S. Ramírez, and F. P. G. Márquez, "K-nearest neighbour and K-fold cross-validation used in wind turbines for false alarm detection," *Sustain. Futures*, vol. 6, p. 100132, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S266618882300028X

[33] A. Yacob, N. E. Ghazali, and F. M. Hassan, "Sentiment Analysis of ChatGPT Using the KNN Algorithm and K-Fold Cross-Validation Optimization of the K Value," *J. Int. J. Inform. Comput.*, vol. 1, no. 2, pp. 48–55, 2024. [Online]. Available: https://www.researchgate.net/publication/389515999_Sentiment_Analysis_of_ChatGPT_Using_the_KNN_Algorithm_and_K-Fold_Cross-Validation_Optimization_of_the_K_Value

[34] S. Hidayat, H. M. T. Ramadhan, and E. Y. Puspaningrum, "Comparison of K-nearest neighbor and decision tree methods using principal component analysis technique in heart disease classification," *Indones. J. Data Sci.*, vol. 4, no. 2, pp. 87–96, 2023. [Online]. Available: https://www.jurnal.yoctobrain.org/index.php/ijodas/article/view/70

[35] A. Razzaque and A. Badholia, "PCA based feature extraction and MPSO based feature selection for gene expression microarray medical data classification," *Meas. Sens.*, vol. 31, p. 100945, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2665917423002817

[36] R. S. Rao, S. Dewangan, A. Mishra, and M. Gupta, "A study of dealing class imbalance problem with machine learning methods for code smell severity detection using PCA-based feature selection technique," *Sci. Rep.*, vol. 13, no. 1, p. 16245, 2023. [Online]. Available: https://www.nature.com/articles/s41598-023-43380-8

Vol. 7, No. 1, February 2026, Page. 361-372
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5422

[37] J. Barth, D. Katumullage, C. Yang, and J. Cao, "Classification of wines using principal component analysis," *J. Wine Econ.*, vol. 16, no. 1, pp. 56–67, 2021. [Online]. Available: https://www.cambridge.org/core/journals/journal-of-wine-economics/article/abs/classification-of-wines-using-principal-component-analysis/447CE06A9FA61D6950E3163FCF655ADF

[38] X. Yan *et al.*, "Classification of plastics using laser-induced breakdown spectroscopy combined with principal component analysis and K nearest neighbor algorithm," *Results Opt.*, vol. 4, p. 100093, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666950121000419

[39] T. Aljrees, "Improving prediction of cervical cancer using KNN imputer and multi-model ensemble learning," *PLoS ONE*, vol. 19, no. 1, p. e0295632, 2024. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0295632

[40] D. Cheng, D. Zhao, J. Zhang, C. Wei, and D. Tian, "PCA-based denoising algorithm for outdoor lidar point cloud data," *Sensors (Basel)*, vol. 21, no. 11, p. 3703, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/11/3703

[41] F. Bahrambanan, M. Alizamir, K. Moradveisi, S. Heddam, S. Kim, S. Kim, M. Soleimani, S. Afshar, and A. Taherkhani, "The development of an efficient artificial intelligence-based classification approach for colorectal cancer response to radiochemotherapy: deep learning vs. machine learning," *Sci. Rep.*, vol. 15, no. 62, Jan. 2025. [Online]. Available: https://www.nature.com/articles/s41598-024-84023-w