Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.5, October 2025, pp. 2179~2186
e-ISSN: 2548-6861

Comparing Different KNN Parameters Based on Woman Risk Factors to Predict the Cervical Cancer

Maria Claudia Saletia [1]*, Mochammad Anshori [2]*, M Syauqi Haris [3]*
* Informatika, Institut Teknologi, Sains, dan Kesehatan RS.DR. Soepraoen Kesdam V/BRW
saletiamaria8@gmail.com [1], moanshori@itsk-soepraoen.ac.id [2], haris@itsk-soepraoen.ac.id [3]

Article Info

Article history:

Received 2025-08-11
Revised 2025-09-20
Accepted 2025-09-25

Keyword:

Healthinformatics,
KNN,
Cervical Cancer,
Minkowski,
Classification.

ABSTRACT

Cervical cancer remains a major cause of mortality among women, particularly in low-resource regions where access to conventional screening is limited. Early detection through predictive modeling offers a low-cost and non-invasive alternative to clinical diagnostics. This study aims to evaluate the effectiveness of the k-Nearest Neighbors algorithm for predicting cervical cancer risk using behavioral and psychosocial attributes. The research utilized the publicly available Sobar cervical cancer behavioral dataset comprising 72 instances with 18 input features and a binary target label. Data preprocessing included removal of incomplete records, encoding of categorical variables, and normalization. The algorithm was tested across varying numbers of neighbors and distance metrics, with performance evaluated using 10-fold cross-validation and multiple classification metrics. The optimal configuration was achieved with three neighbors and the Manhattan distance metric, yielding an accuracy of 93.06%, sensitivity of 93.10%, specificity of 85.90%, precision of 93.10%, F1-score of 92.90%, and an area under the curve of 0.8952. This performance surpassed the reported baseline of a probabilistic classifier and demonstrated the algorithm's capability to capture complex behavioral patterns associated with cervical cancer risk. These findings confirm the feasibility of applying optimized instance-based learning to behavioral data for early cancer risk assessment. The approach offers potential for integration into community health programs to support early detection and prevention strategies.

I. INTRODUCTION

Cervical cancer, also referred to as "ca cancer" in certain medical contexts, is recognized by the World Health Organization (WHO) as a malignancy developing in the cervix, the anatomical gateway from the vagina to the uterus within the female reproductive system. This disease originates when healthy cervical cells undergo malignant transformation, most often due to infection with the human papillomavirus (HPV), a pathogen primarily transmitted through sexual contact [1]. Among the early warning signs is abnormal vaginal bleeding, which can be indicative of underlying pathological changes [2]. On a global scale, cervical cancer remains a significant public health concern, with approximately 527,624 women affected and an estimated 265,672 deaths annually. By 2018, it was ranked among the most malignant tumors worldwide, second only to breast cancer in terms of prevalence among women [3].

The situation in Indonesia mirrors this global challenge. Approximately 79.14 million Indonesian women aged 15 years and older are considered at risk of developing cervical cancer. National health data indicate that 13,762 women are diagnosed with the disease each year, and 7,943 succumb to it [4]. This translates to a mortality rate exceeding 50% among diagnosed patients, underscoring the severe limitations of late-stage detection in terms of survival outcomes. Existing medical literature emphasizes that cervical cancer, despite its high mortality, can be effectively treated when detected in its early stages [5]. However, early detection remains challenging due to the difficulty of identifying clear symptoms during the initial phases of the disease [6]. These epidemiological and clinical realities justify the urgency of

implementing effective early prediction strategies to reduce mortality rates and improve patient prognosis.

In recent years, a variety of approaches have been developed and proposed for early cervical cancer detection, with machine learning (ML) emerging as a particularly promising paradigm [7]. As a specialized branch of artificial intelligence within computer science, ML leverages computational algorithms to learn from data patterns, offering considerable advantages in disease prediction. In healthcare applications, ML-based predictive models can be both cost-effective and operationally efficient [8]. Within ML methodologies, supervised learning—where algorithms are trained on labeled datasets—has become the dominant approach for medical diagnosis and risk assessment.

Against this backdrop, the present study proposes a machine learning–driven framework for cervical cancer prediction. Some researchers have focused on the same dataset employed in this study, which originates from the UCI ML Repository's "Cervical Cancer Behavioral Risk" dataset (https://archive.ics.uci.edu/ml/datasets/Cervical+Cancer+Behavior+Risk), comprising 19 features and 72 instances. Previous research efforts have explored a range of ML algorithms for this purpose, yielding varying levels of accuracy. For example, naïve Bayes classifiers have achieved an accuracy rate of 91.67%, and logistic regression accuracy = 87.5%[9], decision stump achieved accuracy = 80%, and C4.5 algorithm achieved accuracy = 78% [10], while multilayer perceptron neural networks reported accuracy of 82.93%, decision tree accuracy = 77.97%, support vector machine = 79.25%[3]. The methodological focus of this study is the K-Nearest Neighbor (KNN) algorithm, a supervised learning method that remains underexplored for cervical cancer prediction in comparison with the aforementioned approaches. Although KNN is relatively simple in its computational design, it possesses a significant advantage: it does not require assumptions about the underlying data distribution [11]. KNN has a benefit over other fundamental techniques like SVM and logistic regression in that it operates non-parametrically, meaning it ignores the linearity of the data and does not make assumptions about it [12]. This property makes KNN especially adaptable in medical datasets, which may not conform to strict statistical distributions. Furthermore, the KNN algorithm operates by classifying new data points based on the majority class among their closest neighbors in the feature space, where "closeness" is typically quantified using a distance metric such as Euclidean or Manhattan distance. This straightforward yet flexible mechanism allows for effective classification in a variety of domains, including healthcare diagnostics.

Despite the promising characteristics of KNN, literature applying it specifically to cervical cancer prediction—especially on behavioral risk datasets—remains limited. Studies that have incorporated KNN into medical classification tasks often emphasize its potential when paired with appropriate feature selection or hyperparameter tuning. In medical imaging, for example, KNN has been used for tumor detection with considerable success when parameters such as the number of neighbors (k) and the choice of distance metric are optimized. Similarly, in non-imaging health datasets, KNN's performance can rival or exceed more complex algorithms if the parameters are carefully calibrated to the data characteristics. This body of evidence suggests that a systematic investigation of KNN's application to cervical cancer behavioral datasets could yield competitive performance benchmarks and expand the algorithm's documented utility in preventive healthcare analytics.

An integrated review of prior studies also reveals a crucial gap in the literature. While high-performing models such as decision trees, random forests, softmax classifiers, and ensemble methods have been extensively tested on the UCI cervical cancer dataset, these models often involve more computational complexity and require substantial tuning to achieve optimal performance. In contrast, KNN's operational simplicity, combined with its distribution-free nature, positions it as an accessible yet powerful alternative—provided its hyperparameters are appropriately optimized. This gap underscores the importance of systematically examining KNN's capabilities and limitations in the specific context of cervical cancer prediction using behavioral and risk factor data.

The primary objective of this study is to address this research gap by conducting a rigorous evaluation of KNN for early cervical cancer prediction. Specifically, the study aims to determine the optimal configuration of KNN parameters—most notably, the number of neighbors (k) and the type of distance metric—to maximize predictive performance on the UCI behavioral risk dataset. The novelty of this research lies in its targeted application of KNN to a domain where it has received minimal attention, coupled with an emphasis on parameter optimization for practical healthcare deployment. By comparing the optimized KNN model to the best-performing baseline reported in previous studies—namely, the naïve Bayes method with an accuracy of 91.67%—this study seeks to establish whether a relatively simple model can match or surpass the performance of established algorithms. The scope of the research is thus both methodological and applied: it contributes to machine learning methodology by demonstrating best practices in KNN tuning, and it serves applied medical informatics by presenting a cost-effective, interpretable predictive model for early-stage cervical cancer risk assessment.

II. METHODOLOGY

This study employed a quantitative experimental approach to evaluate the performance of the k-Nearest Neighbors (KNN) algorithm for predicting cervical cancer risk using behavioral and psychosocial attributes. The methodology was designed to ensure reproducibility, validity, and rigorous evaluation of the proposed model. Referering to Figure 1, the research process comprised six main stages: data acquisition, data preprocessing, algorithm selection, hyperparameter configuration, model training and validation, and evaluation metrics.

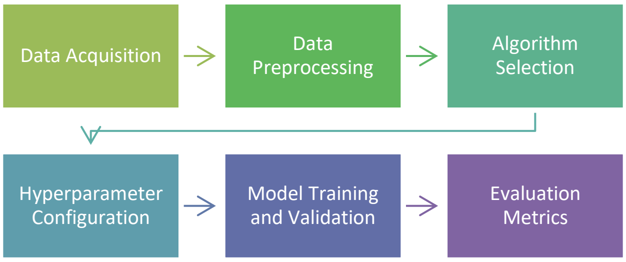

Figure 1. Research Methodology

A. *Data Acquisition*

The dataset used in this study was the publicly available "Sobar" cervical cancer behavioral dataset, retrieved from the University of California Irvine (UCI) Machine Learning Repository [9]. The dataset contains 72 instances with 18 input features describing behavioral and psychosocial risk factors, and a single binary target attribute (Ca_cervix), where "yes" indicates a diagnosis of cervical cancer and "no" indicates absence of the disease. The features include variables related to personal health behavior, reproductive history, and lifestyle patterns. The class distribution is imbalanced, consisting of 51 "no" cases and 21 "yes" cases.

B. *Data Preprocessing*

To ensure data integrity and compatibility with the KNN algorithm, several preprocessing steps were undertaken. First is handling missing data. All instances containing missing or incomplete attribute values were removed from the dataset, resulting in a fully complete dataset for analysis. Second is attribute encoding [13]. Since KNN relies on distance calculations, categorical attributes were numerically encoded, while all numeric attributes were kept in their original form the third is feature scaling. The dataset was normalized to ensure that all features contributed equally to the distance metric and to prevent attributes with larger numeric ranges from dominating the classification process [14].

C. *Algorithm Selection and Justification*

The k-Nearest Neighbors (KNN) algorithm was selected for this study due to its simplicity, non-parametric nature, and adaptability to complex decision boundaries without requiring explicit training models [15], [16]. KNN operates by identifying the k nearest instances in the training dataset to a given query instanc, based on a chosen distance metric, and assigning the most frequent class among these neighbors to the query instance [17]. KNN also called as lazy learner because it does not do generalizations and learning phase do while classification time [18].

Unlike parametric models such as Naïve Bayes or Logistic Regression, KNN can capture nonlinear relationships between features and class labels, which is advantageous when working with behavioral datasets that may exhibit complex interactions among variables [19].

D. *Hyperparameter Configuration*

Two primary hyperparameters were systematically varied to optimize model performance:

1) *Number of Neighbors(k)*: Number of Neighbors (k): Values of k ranging from 1 to 9 were tested to determine the optimal neighborhood size for classification accuracy and robustness.

2) *Distance Metrics*: Four different similarity measures were evaluated:

- Euclidean Distance (L2 norm):
$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$
- Manhattan Distance (L1 norm):
$$d(x,y) = \sum_{i=1}^{n}|x_i - y_i| \quad (2)$$
- Minkowski Distance (generalized Lp norm):
$$d(x,y) = \left(\sum_{i=1}^{n}|x_i - y_i|^r\right)^{1/r} \quad (3)$$
- Chebysev Distance (L∞ norm):
$$d(x,y) = \max_{i=1}^{n}|x_i - y_i| \quad (4)$$

The choice of multiple distance metrics was motivated by prior studies showing that metric selection can significantly influence KNN performance, particularly in high-dimensional or sparse datasets [20].

E. *Model Training and Validation*

fold1	fold2	fold3	fold4	fold5	fold6	fold7	fold8	fold9	fold10
fold1	fold2	fold3	fold4	fold5	fold6	fold7	fold8	fold9	fold10
fold1	fold2	fold3	fold4	fold5	fold6	fold7	fold8	fold9	fold10
fold1	fold2	fold3	fold4	fold5	fold6	fold7	fold8	fold9	fold10
fold1	fold2	fold3	fold4	fold5	fold6	fold7	fold8	fold9	fold10
fold1	fold2	fold3	fold4	fold5	fold6	fold7	fold8	fold9	fold10
fold1	fold2	fold3	fold4	fold5	fold6	fold7	fold8	fold9	fold10
fold1	fold2	fold3	fold4	fold5	fold6	fold7	fold8	fold9	fold10
fold1	fold2	fold3	fold4	fold5	fold6	fold7	fold8	fold9	fold10
fold1	fold2	fold3	fold4	fold5	fold6	fold7	fold8	fold9	fold10

Figure 2. 10-fold cross validation illustration

10-fold cross-validation strategy was adopted to ensure an unbiased and statistically reliable performance estimation. In this approach, the dataset was partitioned into 10 approximately equal folds accroding to Figure 2. For each fold, the model was trained on 9 folds and tested on the remaining fold, and the process was repeated 10 times such that each fold served as the test set exactly once. The average performance across all folds was then computed.

Cross-validation was selected to mitigate overfitting risks associated with small datasets and to provide a more stable estimate of the model's generalization capability compared to a single train-test split [16], [21].

F. Evaluation Metrics

Model performance was assessed using multiple standard classification metrics to capture different aspects of predictive capability:

- Accuracy (ACC): Proportion of correctly classified instances among all instances
$$accuracy = \frac{TP+TN}{total\ data} \quad (5)$$
- Sensitivity (true positive rate): Proportion of correctly predicted positive cases out of all actual positive cases
$$sensitivity = \frac{TP}{TP+FN} \quad (6)$$
- Specificity (true negative rate): Proportion of correctly predicted negative cases out of all actual negative cases
$$specificity = \frac{TN}{TN+FP} \quad (7)$$
- Precision: Proportion of correctly predicted positive cases out of all predicted positive cases
$$precision = \frac{TP}{TP+FP} \quad (8)$$
- F1-Score: Harmonic mean of precision and sensitivity, providing a balance between the two
$$F - measure = 2 \cdot \frac{precision \cdot sensitivity}{precision+sensitivity} \quad (9)$$
- Area Under the Receiver Operating Characteristic Curve (AUC): Measures the ability of the model to discriminate between classes across all possible classification thresholds.

Where TP is true positive, TN is true negative, FP is false positive, and FN is false negative. These metrics were chosen because accuracy alone can be misleading when evaluating models on imbalanced datasets, whereas sensitivity, specificity, and AUC provide deeper insights into classification performance [22].

III. RESULTS AND DISCUSSION

This section presents the findings of the experimental evaluation of the k-Nearest Neighbors (KNN) algorithm on the Sobar cervical cancer behavioral dataset and discusses the implications of these results in the context of prior literature. The analysis follows the sequence of identifying the best-performing model configuration, presenting detailed performance metrics, and interpreting the results relative to existing research.

A. Dataset Characteristics

Based on the Table 1, the ca cervix dataset contains 18 numerical features that measure psychosocial and behavioral aspects, along with one categorical feature (ca_cervix) as a label. The varied range of values for the numerical features (1...5 to 3...15) indicates that the data comes from different measurement scales, rather than from standard clinical data. Therefore, it can be concluded that this study takes a holistic approach, integrating social psychology with health to understand the non-medical factors influencing the risk of cervical cancer.

TABLE 1
DETAILS OF DATASET

Feature	data type	range
behavior_eating	numerical	2 … 10
behavior_personalHygine	numerical	3 … 15
intention_aggregation	numerical	3 … 15
intention_commitment	numerical	2 … 10
attitude_consistency	numerical	6 … 15
attitude_spontaneity	numerical	4 … 10
norm_significantPerson	numerical	1 … 5
norm_fulfillment	numerical	3 … 15
perception_vulnerability	numerical	3 … 15
perception_severity	numerical	2 … 10
motivation_strength	numerical	3 … 15
motivation_willingness	numerical	3 … 15
socialSupport_emotionality	numerical	3 … 15
socialSupport_appreciation	numerical	2 … 10
socialSupport_instrumental	numerical	3 … 15
empowerment_knowledge	numerical	3 … 15
empowerment_abilities	numerical	3 … 15
empowerment_desires	numerical	3 … 15
ca_cervix	categorical	yes, no

TABLE 2
SIGNIFICANCE DATASET

Variable	P value	Sign
behavior_sexualRisk	2,89426E-61	< 0,05
behavior_eating	8,60999E-55	< 0,05
behavior_personalHygine	9,92502E-41	< 0,05
intention_aggregation	5,20182E-34	< 0,05
intention_commitment	1,64845E-53	< 0,05
attitude_consistency	2,63302E-49	< 0,05
attitude_spontaneity	1,12745E-54	< 0,05
norm_significantPerson	8,46737E-19	< 0,05
norm_fulfillment	1,00939E-21	< 0,05
perception_vulnerability	7,50744E-25	< 0,05
perception_severity	9,19068E-19	< 0,05
motivation_strength	8,70752E-43	< 0,05
motivation_willingness	5,71466E-29	< 0,05
socialSupport_emotionality	7,30993E-24	< 0,05
socialSupport_appreciation	6,79656E-26	< 0,05
socialSupport_instrumental	1,90343E-30	< 0,05
empowerment_knowledge	1,05827E-29	< 0,05
empowerment_abilities	2,29683E-27	< 0,05
empowerment_desires	1,85674E-28	< 0,05

Based on the data in the Table 2, it can be concluded that all tested variables have a very low p-value and are consistently less than 0.05. This indicates that every variable, from behavior_sexualRisk to empowerment_desires, has a statistically significant relationship with the factors influencing *Ca Cervix*. Inductively, this result strengthens the argument that various aspects, including behavior, intention, social norms, perception, motivation, social support, and empowerment, collectively play a crucial and interconnected role in explaining vulnerability and risk for cervical cancer. Therefore, a comprehensive approach involving interventions across all these variables is essential for risk mitigation.

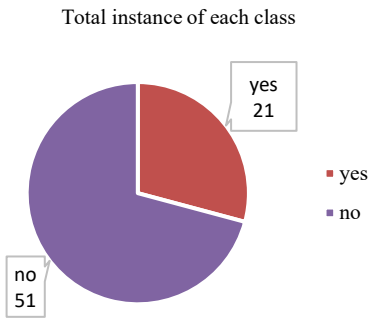

Figure 3. Total instances of eacch class of dataset

Figure 3 above shows the total number of each target class. Based on the figure there are 51 instances with no class and 21 instances with a yes class. It shows if the dataset is imbalanced because the sum no class is more than the yes class about 71% of total data. Since the dataset is not uniform, cross-validation is essential to producing a generic and robust model.

B. Determination of Optimal KNN Configuration

The experiments evaluated multiple combinations of the number of neighbors (k) and distance metrics to determine the optimal KNN configuration. The tested values of k ranged from 1 to 9, and the four-distance metrics considered were Euclidean, Manhattan, Minkowski, and Chebyshev. Performance was assessed using 10-fold cross-validation across all configurations. The tool used in this study is WEKA [23].

Figure 4 above shows the graphic based on the k parameter and distance measure to do a comparison. Based on the figure above, chebyshev distance didn't give a better result. The accuracy result is the lowest than the other distance measure. For chebyshev distance, the optimum $k = 2$. The minkowski distance give accuracy better than chebyshev distance. The graphic of minkowski distance is increasing as the k value but not at $k = 4$. While the $k = 4$ in minkowski, it is the local optimum because the next k the accuracy decreasingly and raising again when $k = 8$ and the higher accuracy is $k = 9$. The euclidean distance give better result than chebyshev and minkowski. But the accuracy between each k value so fluctuates. Best k for this distance measure are at 3 and 5 with the highest accuracy for this measure. The manhattan distance give the best performance than the other measures. With the low k values, the accuracy is as low too. Since k increases, the accuracy result is also raised. Till the $k = 3$ and $k = 6$ the accuracy reach the highest result. But when $k >= 7$, the accuracy is reduced rapidly. It shows if the higher k does not always give the better performance. The high of k will increase the computational times too. In this experiment, we determined the better accuracy based on the k value of each distance measure. The details are shown in Table 3.

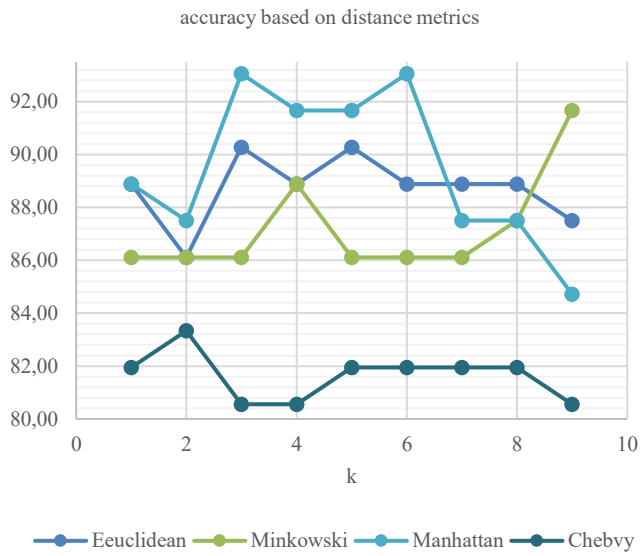

Figure 4. Graphic of comparison accuracy of each iterative k parameter and distance measure

TABLE 3
DETERMINED K OF EACH DISTANCE MEASURE

Distance Measure	k
Euclidean	3
Minkowski	9
Manhattan	3
Chebysev	2

From Table 3 above, next experiment use $k = 3$ for euclidean distance, $k = 9$ for minkowski distance, $k = 3$ for manhattan distance and $k = 2$ for chebyshev distance. Then a comparison is needed to know the best model between the distance measure based on accuracy, specificity, sensitivity, precision, and F-measure percentage. The result shown in Figure 5 below.

C. Performance Metrics of the Best Model

Figure. 5 shows the evaluation between distance measure. Visibly that chebyshev with $k = 2$ get the lowest result. The next lower evaluation results are euclidean with $k = 3$ and minkowski distances with $k = 9$. Between the two measure, euclidean get a higher specificity than minkowski. But the overall result of minkowski has a better result if compare with euclidean distance. Manhattan reaches the best evaluation using $k = 3$. The on accuracy, specificity, sensitivity, precision, and F-measure value are the highest. The accuracy = 93.06, sensitivity = 93.1, specificity = 85.9, precision = 93.1 and f-measure = 92.9. This experiment proves that distance measure and k parameter have an impact to build the KNN

model.

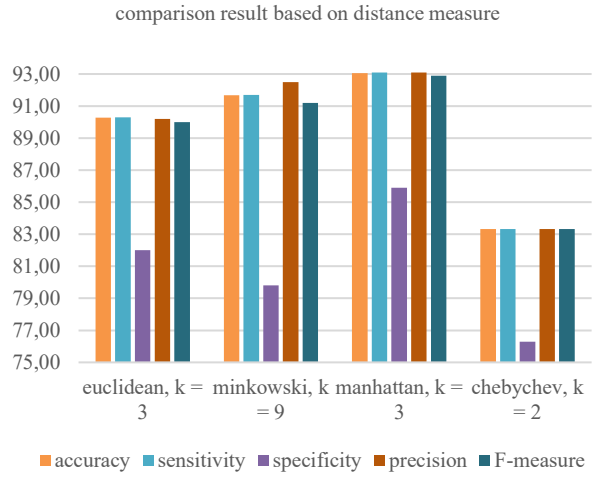

Figure 5. Graphical comparisan of accuracy, sensitivity, specificity, precision, and F-measure based on determined *k* of distance measure.

Referring to Figure 5, the high accuracy demonstrates the model's overall effectiveness in correctly classifying both positive and negative cases. The sensitivity value indicates that the model correctly identified more than 93% of actual positive cases, which is essential in a medical screening context where missing a positive case can have serious consequences. Specificity, while slightly lower than sensitivity, remains strong at 85.9%, suggesting that false positives are relatively limited. The precision and F1-score values confirm that the model maintains a balanced trade-off between capturing true positives and minimizing false positives.

Based on the sensitivity value, the receiver operating characteristic (ROC) curve can be formed. ROC curve based on the x and y-axis. Sensitivity on the vertical axis and 1 – sensitivity on the horizontal axis. ROC curve is a way to visualize the performance of the classifier [24]. ROC curve represents in Figure 6.

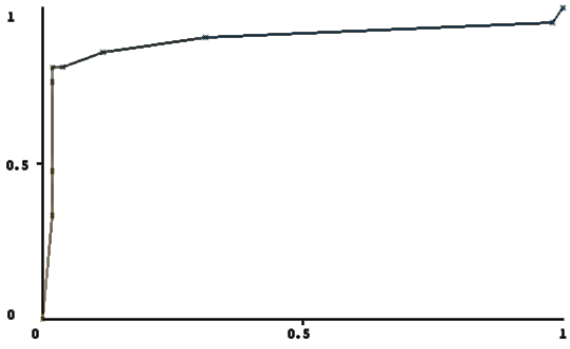

Figure 6. ROC curve with AUC = 0.8952

ROC curve is a graph with value between 0 to 1 [25]. Figure 6 show better classifier model with area under the curve (AUC) value close to 1 and bad model if AUC less than 0.5. AUC of the ROC calculated from averaged test data to compare the discriminant power [26]. Based on Figure. 5 above, the value of AUC is 0.8952.

D. Comparison with Prior Work

TABLE 4
COMPARISON RESULTS BETWEEN THIS STUDI AND PRIOR RESEARCH

Method	Accuracy
Naïve Bayes	91.67%
Logistic Regression	87.5%
C4.5	78%
Decision Stump	80%
Multilayer Perceptron	82.93%
Decision Tree	77.97%
Support Vector Machine	79.25%
Proposed KNN (k=3; distance=Manhattan)	93.06%

Based on the accuracy comparison presented in the Table 4, it can be concluded that the Proposed KNN (with k=3 and Manhattan distance metric) demonstrates superior performance compared to other algorithms used in previous studies. While other algorithms like Naive Bayes and Logistic Regression achieved a reasonably high accuracy (91.67% and 87.5%) [9], the Proposed KNN managed to surpass both with the highest accuracy of 93.06%. This higher accuracy indicates that the Proposed KNN model is more effective in classifying data than other conventional algorithms. This advantage shows that the proposed approach, a modification of KNN, is a more optimal and accurate solution for the classification problem under study.

The obtained accuracy of 93.06% represents an improvement over the Naïve Bayes baseline reported by [9], which achieved 91.67% on the same dataset. While the improvement of 1.39 percentage points may appear modest, it demonstrates that KNN, when tuned appropriately, can outperform simpler probabilistic classifiers on behavioral datasets. The proposed KNN configuration delivers competitive performance while maintaining algorithmic simplicity. This balance between accuracy and computational efficiency makes KNN a viable option for deployment in low-resource settings, where computational infrastructure may be limited.

E. Influence of Hyperparameters on Performance

The provided text argues that a K-Nearest Neighbors (KNN) model, when properly optimized, is a highly effective and practical tool for predicting cervical cancer risk based on behavioral and psychosocial data. The author's central claim is that careful selection of hyperparameters, specifically the value of k and the distance metric, is crucial for the model's success. The text provides a compelling case for the use of KNN in this context by highlighting its high sensitivity, which is vital for medical screening, and its non-invasive, cost-effective nature.

The author systematically supports this claim by explaining the specific roles of the hyperparameters. They argue that the optimal value of *k=3* balances the bias-variance

tradeoff, while the superiority of the Manhattan distance over the Euclidean distance suggests that individual feature differences are more informative for this dataset. This is influenced by the way it works, which is useful when dealing with data that are not normally distributed or when outliers are present [27]. This evidence is presented to justify the model's performance. The text concludes by asserting that the model's simplicity and strong performance metrics—accuracy, sensitivity, and precision—make it well-suited for implementation in resource-constrained environments, positioning KNN as a competitive and practical solution for preliminary cervical cancer screening.

IV. CONCLUSION

This study demonstrated that the k-Nearest Neighbors (KNN) algorithm, when carefully tuned, can serve as an effective predictive model for cervical cancer risk assessment based solely on behavioral and psychosocial attributes. Using the Sobar dataset and evaluating multiple configurations of k values and distance metrics, the optimal performance was achieved with $k = 3$ and the Manhattan distance metric, yielding an accuracy of 93.06%, sensitivity of 93.10%, and an AUC of 0.8952. This result represents an improvement over the Naïve Bayes baseline previously reported for the same dataset and confirms the importance of hyperparameter selection in enhancing model performance. The findings carry practical implications for early cervical cancer screening, particularly in low-resource settings where non-invasive, low-cost, and computationally simple tools are essential. By relying on easily collected behavioral data, the proposed approach can complement existing screening programs and help identify individuals at higher risk for further clinical evaluation. This work contributes to the existing body of knowledge by empirically validating the effectiveness of KNN for behavioral-based medical prediction and highlighting the role of distance metrics in classification accuracy. Future research should focus on validating the model with larger, more diverse datasets and exploring hybrid approaches that integrate clinical variables for improved robustness and generalizability.

REFERENCES

[1] N. Razali, S. A. Mostafa, A. Mustapha, and M. Helmy, "Risk Factors of Cervical Cancer using Classification in Data Mining," *IJECTS 2019*, 2020, doi: 10.1088/1742-6596/1529/2/022102.

[2] P. R. Garg *et al.*, "Women's Knowledge on Cervical Cancer Risk Factors and Symptoms: A Cross Sectional Study from Urban India," *Asian Pacific J. Cancer Prev.*, vol. 23, no. 3, pp. 1083–1090, 2022, doi: 10.31557/APJCP.2022.23.3.1083.

[3] J. Lu, E. Song, A. Ghoneim, and M. Alrashoud, "Machine learning for assisting cervical cancer diagnosis: An ensemble approach," in *Future Generation Computer Systems*, 2020, vol. 106, pp. 199–205, doi: 10.1016/j.future.2019.12.033.

[4] S. Setiawati and Y. Hapsari, "Clinical Manifestations, Diagnosis, Management and Prevention of Cervical Cancer," *J. Biol. Trop.*, vol. 23, no. 4, pp. 382–390, 2023, doi: 10.29303/jbt.v23i4.5594.

[5] S. Gupta and M. K. Gupta, "Computational Prediction of Cervical Cancer Diagnosis Using Ensemble-Based Classification Algorithm," *Comput. J.*, vol. 65, no. 6, pp. 1527–1539, 2022, doi: 10.1093/comjnl/bxaa198.

[6] S. Zhang, H. Xu, L. Zhang, and Y. Qiao, "Cervical cancer: Epidemiology, risk factors and screening," *Chinese J. Cancer Res.*, vol. 32, no. 6, pp. 720–728, 2020, doi: 10.21147/j.issn.1000-9604.2020.06.05.

[7] M. Musthofa and M. Anshori, "Comparing Discriminant Analysis Function for Early Prediction of Smartphone Addiction," *J. Enhanc. Stud. Informatics Comput. Appl.*, vol. 2, no. 1, pp. 1–7, 2025, doi: https://doi.org/10.47794/jesica.v2i1.12.

[8] H. G. Ahmad and M. J. Shah, "Prediction of Cardiovascular Diseases (CVDs) Using Machine Learning Techniques in Health," vol. 4, no. 2, pp. 267–279, 2021.

[9] Sobar, R. Machmud, and A. Wijaya, "Behavior determinant based cervical cancer early detection with machine learning algorithm," *Adv. Sci. Lett.*, vol. 22, no. 10, pp. 3120–3123, 2016, doi: 10.1166/asl.2016.7980.

[10] S. I. Journal, "Cervical Cancer Cell Prediction using Machine Learning Classification Algorithms Cervical Cancer Cell Prediction using," *Eng. Sci. Int. J.*, vol. 8, no. 1, pp. 25–29, 2021, doi: 10.30726/esij/v8.i1.2021.81006.

[11] A. F. Gündüz and A. Karcı, "Heart Sound Classification for Murmur Abnormality Detection Using an Ensemble Approach Based on Traditional Classifiers and Feature Sets," *Anatol. J. Comput. Sci.*, vol. 5, no. 1, pp. 1–13, 2020.

[12] E. Najwaini, Thomas Edyson Tarigan, Fajri Profesio Putra, and Sulistyowati, "Application of the K-Nearest Neighbors (KNN) Algorithm on the Brain Tumor Dataset," *Int. J. Artif. Intell. Med. Issues*, vol. 1, no. 1, pp. 18–26, 2023, doi: 10.56705/ijaimi.v1i1.85.

[13] F. N. Yahya, M. Anshori, and A. N. Khudori, "Evaluasi Performa XGBoost dengan Oversampling dan Hyperparameter Tuning untuk Prediksi Alzheimer," *Techno.Com*, vol. 24, no. 1, pp. 1–12, 2025, doi: 10.62411/tc.v24i1.12057.

[14] Y. Dimas Pratama and A. Salam, "Comparison of Data Normalization Techniques on KNN Classification Performance for Pima Indians Diabetes Dataset," *J. Appl. Informatics Comput.*, vol. 9, no. 3, p. 693, 2025, doi: 10.30871/jaic.v9i3.9353.

[15] R. Katarya and S. Jain, "Comparison of different machine learning models for diabetes detection," *Proc. 2020 IEEE Int. Conf. Adv. Dev. Electr. Electron. Eng. ICADEE 2020*, no. Icadee, pp. 0–4, 2020, doi: 10.1109/ICADEE51157.2020.9368899.

[16] M. Anshori, F. Mar'i, and F. A. Bachtiar, "Comparison of Machine Learning Methods for Android Malicious Software Classification based on System Call," *Proc. 2019 4th Int. Conf. Sustain. Inf. Eng. Technol. SIET 2019*, pp. 343–348, 2019, doi: 10.1109/SIET48054.2019.8985998.

[17] R. Katarya and S. Maan, "Predicting mental health disorders using machine learning for employees in technical and non-technical companies," *Proc. 2020 IEEE Int. Conf. Adv. Dev. Electr. Electron. Eng. ICADEE 2020*, no. Icadee, 2020, doi: 10.1109/ICADEE51157.2020.9368923.

[18] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning," *Decis. Anal. J.*, vol. 3, no. May, p. 100071, 2022, doi: 10.1016/j.dajour.2022.100071.

[19] N. M. Mathkunti and S. Rangaswamy, "Machine Learning Techniques to Identify Dementia," *SN Comput. Sci.*, vol. 1, no. 3, pp. 1–6, 2020, doi: 10.1007/s42979-020-0099-4.

[20] N. Wulandari, Y. Cahyana, and H. Hikmayanti Handayani, "Sentiment Analysis on the Relocation of the National Capital (IKN) on Social Media X Using Naive Bayes and K-Nearest Neighbor (KNN) Methods," *J. Appl. Informatics Comput.*, vol. 9, no. 3, pp. 724–731, 2025, doi: 10.30871/jaic.v9i3.9552.

[21] R. Choirunnisa, M. Anshori, and W. T. Kusuma, "Improving Random Forest Evaluation in Mental Health Disorder Identification with Cross Validation," *J. Artif. Intell. Digit. Bus.*, vol. 4, no. 2, pp. 3526–3534, 2025.

[22] M. Anshori and M. S. Haris, "Predicting Heart Disease using Logistic Regression," *Knowl. Eng. Data Sci.*, vol. 5, no. 2, p. 188,

2022, doi: 10.17977/um018v5i22022p188-196.
[23] E. Frank, M. A. Hall, and I. H. Witten, "The WEKA workbench," *Data Min.*, pp. 553–571, 2017, doi: 10.1016/b978-0-12-804291-5.00024-6.
[24] B. P. Doppala, D. Bhattacharyya, M. Janarthanan, and N. Baik, "A Reliable Machine Intelligence Model for Accurate Identification of Cardiovascular Diseases Using Ensemble Techniques," *J. Healthc. Eng.*, vol. 2022, no. 1, 2022, doi: 10.1155/2022/2585235.
[25] W. Andriyani *et al.*, *Matematika Pada Kecerdasan Buatan*, Pertama., vol. 7, no. 2. Makassar: CV Tohar Media, 2024.
[26] J. R. Khan, S. Chowdhury, H. Islam, and E. Raheem, "Machine Learning Algorithms To Predict The Childhood Anemia In Bangladesh," *J. Data Sci.*, vol. 17, no. 1, pp. 195–218, 2021, doi: 10.6339/jds.201901_17(1).0009.
[27] Z. Shapcott, *An Investigation into Distance Measures in Cluster Analysis*, no. April. 2024.