J-KOMA Journal of Computer Science and Applications

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

Open Access

***E-ISSN***: *2620 - 4872*
***Vol***.*09*, ***No***.*01*
***Doi***: *10.21009/JKOMA.091.04*

***Recivied***: *29 June 2026*
***Accepted***: *07 July 2026*
***Published***: *31 July 2026*

***Keywords:***
*Cervical Cancer;*
*Discriminant Analysis;*
*Healthinformatics;*
*LDA;*
*QDA.*

***Correspondence Email:***
*moanshori@itsk-soepraoen.ac.id*

# Comparing Linear and Quadratic Discriminant Analysis for Cervical Cancer Risk Prediction Using Behavioral Attributes

**Mochammad Anshori[1*], Nindynar Rikatsih[2]**

[1,2] Institut Teknologi, Sains, dan Kesehatan RS.DR. Soepraoen Kesdam V/BRW, Indonesia.

**Abstract**

Cervical cancer remains a critical global health challenge, yet early prediction using non-medical determinants is underexplored. This study aims to compare linear and quadratic discriminant analysis for early cervical cancer risk assessment based on behavioral and psychosocial factors. A quantitative experimental approach was utilized, analyzing a publicly available behavioral dataset of seventy-two instances and eighteen psychosocial features. Following rigorous statistical validation, feature normalization, and stratified data splitting, both discriminant models were trained and evaluated using accuracy and area under the receiver operating characteristic curve metrics. The empirical findings reveal a striking performance disparity between the classifiers. Linear discriminant analysis emerged as the superior model, achieving an exceptional accuracy of 93.33% and an area under the curve of 0.9464, demonstrating robust discriminative capability. Crucially, it attained perfect sensitivity with zero false negatives, eliminating the most dangerous diagnostic errors. Conversely, quadratic discriminant analysis exhibited complete predictive failure, performing no better than random guessing. This extreme divergence indicates that the underlying behavioral data possesses linearly separable characteristics. The additional complexity of estimating class-specific covariance matrices in the quadratic model resulted in severe overfitting and numerical instability given the limited sample size, proving that simpler, assumption-aligned models outperform complex alternatives. Ultimately, linear discriminant analysis provides an accurate, interpretable, and cost-effective decision support tool for identifying cervical cancer vulnerability. These findings validate that integrating social psychology with machine learning offers a robust framework for early disease screening in resource-constrained environments.

## INTRODUCTION

Clinically known as ca cervix, cervical cancer is a serious worldwide health concern that is typified by the malignant transformation of cervical epithelial cells and is mainly caused by a persistent infection with the human papillomavirus (HPV), a pathogen that is primarily spread through sexual contact [1], [2]. The disease manifests through various clinical indicators, with abnormal vaginal bleeding serving as one of the earliest warning signs indicatives of underlying pathological changes [3]. According to recent epidemiological data, cervical cancer affects approximately 527,624 women worldwide annually, resulting in an estimated 265,672 deaths. In terms of incidence rates, this illness was the second most common cancer among women worldwide in 2018, only topped by breast cancer [4], [5]. These statistics underscore

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

the persistent burden of cervical cancer on global public health infrastructure and highlight the urgent need for more effective prevention and early detection strategies.

The Indonesian context mirrors this global crisis with alarming severity. Cervical cancer is thought to be a risk factor for 79.14 million Indonesian women who are 15 years of age or older. According to national health surveillance data, 13,762 women are diagnosed with cervical cancer each year, and 7,943 of them die from the illness [6], [7]. In terms of survival outcomes, this amounts to a mortality rate of more than 50% among diagnosed patients, highlighting the serious limits of current late-stage detection procedures. Despite its high death rate, cervical cancer can be successfully treated if diagnosed in its early clinical stages, according to current medical research [8]. However, early detection is still clinically problematic because it is hard to find distinct symptom indications in the early stages of the illness [9]. the need to implement effective, reliable, and easily available early prediction methods. Together, these epidemiological and clinical realities support the goal of improving patient prognosis and reducing death rates across a range of healthcare settings.

The principal research problem centers on the critical gap between the availability of effective treatment for early-stage cervical cancer and the persistent difficulty in achieving timely diagnosis. This challenge is particularly pronounced in resource-limited settings where access to sophisticated clinical screening infrastructure remains inadequate. Consequently, there is an imperative need for alternative predictive approaches that can identify high-risk individuals using non-invasive, cost-effective, and readily accessible indicators. The use of computational intelligence, particularly machine learning (ML) techniques, is the general answer that is emerging in modern health informatics. It can produce precise risk estimates by analyzing intricate patterns in behavioral, psychological, and clinical data. By using computational algorithms to learn from data patterns, machine learning (ML), a specific area of artificial intelligence within computer science. It offers significant benefits in disease prediction by spotting subtle associations that may be missed by conventional statistical techniques [10]. In healthcare applications, ML-based predictive models demonstrate potential for being both cost-effective and operationally efficient, making them particularly suitable for deployment in diverse clinical and community health settings [11].

In response to this challenge, various supervised machine learning approaches—where algorithms are trained on labeled datasets—have been developed and proposed for early cervical cancer detection, with varying degrees of success. For this goal, prior research has investigated a variety of machine learning techniques, producing accuracy rates that cover a wide range. For example, logistic regression showed an accuracy rate of 87.5%, whereas naïve Bayes classifiers achieved an accuracy rate of 91.67% [12], Tree-based methods have shown more modest performance, with the decision stump algorithm achieving 80% accuracy and the C4.5 algorithm reaching 78% [13], Neural network approaches, specifically multilayer perceptron architectures, reported accuracy of 82.93%, while traditional decision tree implementations achieved 77.97% accuracy [4]. Support vector machine classifiers demonstrated 79.25% accuracy in comparable evaluations. More recently, the K-nearest neighbors (KNN) algorithm with Manhattan distance metric and k=3 parameter configuration has produced accuracy of approximately 93.06% [14]. These prior investigations collectively establish a performance baseline while revealing opportunities for methodological improvement through alternative algorithmic approaches.

In the context of cervical cancer prediction based on behavioral and psychosocial risk factors, discriminant analysis techniques, particularly Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA). These algorithms remain relatively understudied despite this substantial body of work using various classification algorithms. A popular machine learning technique for dimensional reduction that concurrently preserves strong classification abilities for binary and multiclass tasks is discriminant analysis. [15]. LDA has gained increasing prominence for reducing high-dimensional data while maintaining exceptional accuracy on linearly separable datasets [16]. Although QDA shares methodological similarities with LDA, it is specifically designed for classification tasks involving nonlinearly separated data distributions [17]. Both techniques have demonstrated effectiveness in various healthcare applications. But their comparative performance in cervical cancer risk prediction using behavioral determinants has not been systematically investigated. This represents a significant research gap, as discriminant analysis offers distinct advantages in terms of computational efficiency, interpretability, and theoretical rigor that may prove particularly valuable for clinical decision support systems in resource-constrained environments.

In light of this, the current study suggests a machine learning-driven framework that uses discriminant analysis to predict cervical cancer using a dataset that includes 18 behavioral and psychosocial

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

characteristics from 72 cases. [14]. The main goal is to assess and contrast how well LDA and QDA algorithms perform in identifying cervical cancer risk early on using non-medical behavioral characteristics. This study aims to determine whether the performance of well-known sophisticated algorithms can be matched or exceeded by these comparatively simple, assumption-based models. The novelty of this study lies in its systematic comparison of linear versus quadratic discriminant functions within the specific context of behavioral risk factor analysis for cervical cancer, addressing a methodological gap in the existing literature. The research's scope is both methodological and applied; it advances machine learning methodology by showcasing best practices in applying discriminant analysis to unbalanced medical datasets. Another objective is advances applied medical informatics by offering an affordable, comprehensible predictive model for early-stage cervical cancer risk assessment that combines social psychology and health data analytics.

## METHODOLOGY

The performance of Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) algorithms for predicting cervical cancer risk based on behavioral and psychosocial characteristics was assessed and compared in this work using a quantitative experimental research approach. The methodological framework was methodically organized to guarantee rigorous performance evaluation, internal validity, and reproducibility. Data collection, data preprocessing, algorithm selection and mathematical formulation, hyperparameter setup, model training and validation, and thorough performance evaluation utilizing different metrics comprised the six sequential steps of the research process, as shown in FIGURE 1.

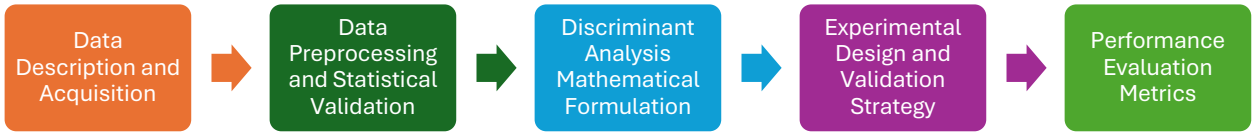


**FIGURE 1.** Research methodology

### Dataset Description and Acquisition

The publicly accessible "Sobar" cervical cancer behavioral dataset, which was acquired from the University of California, Irvine Machine Learning Repository, was the dataset used in this study [12]. In addition to a single binary target attribute where "yes" denotes a cervical cancer diagnosis and "no" denotes the absence of the disease, this dataset has 72 instances with 18 input features that represent behavioral and psychosocial risk factors. The characteristics include those pertaining to lifestyle trends, reproductive history, and individual health behavior. There is an imbalance in the class distribution, with 21 "yes" cases and 51 "no" cases, representing approximately 71% and 29% of the total dataset respectively. This imbalanced distribution necessitates the application of stratified sampling techniques to ensure representative model training and evaluation.

### Data Preprocessing and Statistical Validation

To ensure data integrity, compatibility with discriminant analysis algorithms, and statistical robustness, several systematic preprocessing procedures were implemented. The first stage involved significance testing using p-value analysis to determine the statistical reliability of each variable. Variables with p-values less than 0.05 were deemed statistically significant and retained for subsequent analysis. The second preprocessing stage involved attribute encoding [17], wherein numerical attributes were kept in their original form and categorical attributes were numerically encoded. In order to guarantee that every feature contributed equally to distance metrics and to keep attributes with wider numerical ranges from controlling the classification process, the third stage used feature scaling through min-max normalization [18]. The normalization transformation is mathematically expressed in Equation (1):

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

$$\bar{x} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

This normalization procedure produces scaled values ranging from 0 (minimum) to 1 (maximum), ensuring uniform feature contribution regardless of original measurement scales.

**Discriminant Analysis Mathematical Formulation**

Discriminant analysis represents a preferred machine learning methodology for dimensionality reduction while maintaining robust classification capabilities for both binary and multiclass tasks [15]. Two different discriminant analysis methods were used in this study: Quadratic Discriminant Analysis (QDA) and Linear Discriminant Analysis (LDA). LDA is increasingly employed for reducing high-dimensional data while preserving classification accuracy, particularly excelling with linearly separable data distributions [16]. Although QDA shares methodological similarities with LDA, it is specifically designed for classification tasks involving nonlinearly separated data distributions [16].

The mathematical formulations for LDA and QDA in binary classification scenarios are presented in Equations (2) and (3) respectively. Whereas The classification decision rule for binary cases is determined by Equation (4).

$$\delta(x) \coloneqq 2(\sum^{-1}(\mu_2 - \mu_1))^T x + ((\mu_1 - \mu_2)^T(\sum^{-1}(\mu_1 - \mu_2)) + 2\, ln(\frac{\pi_2}{\pi_2}) \quad (2)$$

$$\delta(x) \coloneqq x^T(\sum_1 - \sum_2)^{-1}x + 2(\sum_2^{-1}\mu_2 - \sum_1^{-1}\mu_1)^T x + (\mu_1^T\sum_1^{-1}\mu_1 - \mu_2^T\sum_2^{-1}\mu_2) + \ln\left(\frac{|\sum_1|}{|\sum_2|}\right) + 2\ln(\frac{\pi_2}{\pi_2}) \quad (3)$$

$$\hat{C}(x) = \begin{cases} 1, & if\delta(x) < 0 \\ 2, & if\delta(x) > 0 \end{cases} \quad (4)$$

Where $\delta_k(x)$ represents the discriminant function for class $k, \mu_k$ denotes the mean vector for class $k, \Sigma$ represents the pooled covariance matrix (LDA) or class-specific covariance matrix $\Sigma_k$ (QDA), and $\pi_k$ indicates the prior probability of class $k$ [19].

The underlying differences between LDA and QDA are found in the covariance structure assumptions they make. LDA produces linear decision boundaries that are represented by straight lines since it assumes identical covariance matrices across classes. Conversely, QDA permits different covariance matrices for each class, generating quadratic decision boundaries that can capture more complex, nonlinear relationships between classes [20]. This theoretical distinction has profound implications for model performance, particularly when dealing with limited sample sizes where QDA's additional parameters may lead to overfitting and numerical instability.

**Experimental Design and Validation Strategy**

To address optimization and ensure robust model evaluation, this study implemented stratified holdout validation with an 80:20 training-to-testing data split ratio. This strategy adheres to the Pareto principle [21], allocating 80% of the data (approximately 58 instances) for model training and 20% (approximately 14 instances) for performance evaluation. The stratification ensures that both training and testing sets maintain proportional representation of the imbalanced class distribution, thereby preventing sampling bias and ensuring generalizable performance estimates. The selection of this holdout strategy is based on earlier research which suggests that strict separation of test data from the start can minimize optimistic bias owing to information leaking which typically occurs in conventional cross-validation [22].

Following discriminant analysis model development, comprehensive experimental scenarios were executed to evaluate model performance under identical conditions. The experimental design facilitated direct comparison between LDA and QDA classifiers, enabling identification of the superior model based on multiple evaluation metrics.

**Performance Evaluation Metrics**

The final methodological stage involved rigorous performance evaluation using multiple complementary metrics to ensure comprehensive assessment of classification effectiveness. The primary evaluation metrics employed were accuracy and Receiver Operating Characteristic - Area Under the Curve

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

(ROC-AUC). Accuracy, defined as the proportion of correctly classified instances among all predictions, is mathematically expressed in Equation (5):

$$Accuracy = \frac{TP + TN}{total\ data} \tag{5}$$

where TP denotes true positives, and TN represents true negatives. The ROC-AUC metric provides a more nuanced evaluation of classification performance, particularly valuable for imbalanced datasets where accuracy alone may prove misleading [23]. The likelihood that a randomly chosen positive case would earn a higher classification score than a randomly chosen negative instance is measured by the AUC. The ROC-AUC calculation is expressed in Equation (6):

$$ROC - AUC = \int_0^1 TPR(FPR^{-1}(X))\, dx \tag{6}$$

where TPR (True Positive Rate) represents sensitivity and FPR (False Positive Rate) equals 1 - specificity. AUC values approaching 1.0 indicate strong class separability and excellent discriminative capability [24].

Additionally, confusion matrix visualization was implemented to quantify misclassification patterns and identify class pairs exhibiting overlapping feature spaces. This diagnostic tool supports detailed interpretation of model behavior, error patterns, and the nature of classification mistakes (Type I versus Type II errors), providing actionable insights for model refinement and clinical application. These metrics were deliberately selected because accuracy alone can be misleading when evaluating models on imbalanced datasets, whereas ROC-AUC provides deeper insights into classification performance across various decision thresholds [23]. The combination of these complementary metrics ensures a comprehensive, multi-faceted evaluation of model performance suitable for clinical decision support applications.

## RESULT AND DISCUSSION

### Descriptive Statistical Analysis

The experimental results are presented in this section through a comprehensive multi-stage analytical framework. Initially, the dataset underwent rigorous descriptive analysis to characterize its structural properties and distributional characteristics. The class distribution, illustrated in FIGURE 2, reveals a binary classification scheme where instances are categorized as either "yes" (cervical cancer present) or "no" (cervical cancer absent). The dataset comprises 51 instances (71%) in the "no" class and 21 instances (29%) in the "yes" class, demonstrating a pronounced class imbalance. This imbalance ratio of approximately necessitates the implementation of stratified sampling techniques during model training and validation to prevent classifier bias toward the majority class and ensure robust generalization across both classes.

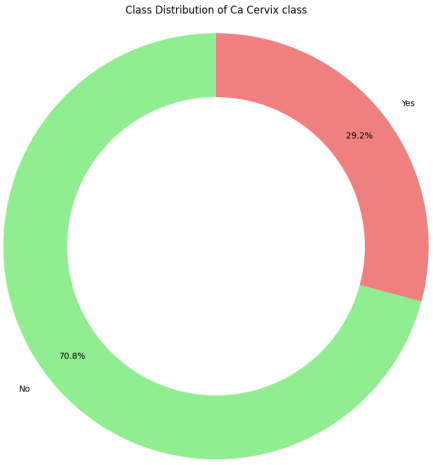


**FIGURE 2**. Dataset class distribution

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

**TABLE 1.** Statistical Significance Testing Results

| Feature | P value | Sign |
|---|---|---|
| Behavior sexual risk | 2,89426E-61 | < 0,05 |
| Behavior eating | 8,60999E-55 | < 0,05 |
| Behavior personal hygine | 9,92502E-41 | < 0,05 |
| Intention aggreagation | 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 significant person | 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 |
| Social support emotionality | 7,30993E-24 | < 0,05 |
| Social support appreciation | 6,79656E-26 | < 0,05 |
| Social support 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 |

TABLE 1 presents the results of statistical significance testing conducted on all 18 variables using p-value analysis at the 0.05 significance threshold. The statistical testing results in TABLE 1 demonstrate that all variables exhibit p-values orders of magnitude below the 0.05 significance threshold, with values ranging from $10^{-19}$ to $10^{-61}$. This extraordinary level of statistical significance indicates that each variable—from behavioral sexual risk to empowerment desires—maintains a robust and reliable association with cervical cancer risk factors. From an inductive reasoning perspective, these findings substantiate the theoretical framework positing that behavioral patterns, intention formation, social normative influences, risk perception, motivational drivers, social support mechanisms, and empowerment dimensions collectively constitute an interconnected system of determinants influencing cervical cancer vulnerability. The uniformly low p-values across all psychosocial domains suggest that no single factor operates in isolation; rather, cervical cancer risk emerges from the synergistic interaction of multiple behavioral and psychosocial dimensions. Consequently, effective risk mitigation strategies must adopt a comprehensive, multi-faceted intervention approach addressing all identified variables simultaneously rather than targeting isolated factors.

TABLE 2 below presents the comprehensive descriptive statistical summary of the dataset following preprocessing. Analysis of TABLE 2 reveals that the cervical cancer dataset comprises 18 numerical features measuring psychosocial and behavioral dimensions, alongside one categorical target variable (ca_cervix). The heterogeneous range of measurement scales—spanning from 1-5 (e.g., norm significant person) to 3-15 (e.g., behavior eating, intention commitment)—indicates that the data originates from diverse psychometric instruments rather than standardized clinical biomarkers. This observation confirms that the study adopts an interdisciplinary approach, integrating social psychology constructs with health sciences to elucidate non-medical determinants influencing cervical cancer risk. The standard deviation values, ranging from 1.186782 (behavior sexual risk) to 4.907577 (norm fulfillment), reveal varying degrees of response dispersion across different psychosocial domains. Notably, variables related to social norms and perceptions exhibit higher variability, suggesting greater heterogeneity in how individuals perceive and internalize these constructs. The necessity for min-max normalization becomes evident from these heterogeneous scales, as failure to standardize feature ranges would result in attributes with larger numeric domains disproportionately influencing the discriminant function calculations, thereby introducing systematic bias into the classification process.

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

**TABLE 2**. Dataset Descriptive Statistical Summary

| Feature | Mean | Std dev | Min | Max |
|---|---|---|---|---|
| Behavior sexual risk | 9.666667 | 1.186782 | 2 | 10 |
| Behavior eating | 12.791667 | 2.361293 | 3 | 15 |
| Behavior personal hygine | 11.083333 | 3.033847 | 3 | 15 |
| Intention aggreagation | 7.902778 | 2.738148 | 2 | 10 |
| Intention commitment | 13.347222 | 2.374511 | 6 | 15 |
| Attitude consistency | 7.180556 | 1.522844 | 2 | 10 |
| Attitude spontaneity | 8.611111 | 1.515698 | 4 | 10 |
| Norm significant person | 3.125000 | 1.845722 | 1 | 5 |
| Norm fulfillment | 8.486111 | 4.907577 | 3 | 15 |
| Perception vulnerability | 8.513889 | 4.275686 | 3 | 15 |
| Perception severity | 5.388889 | 3.400727 | 2 | 10 |
| Motivation strength | 12.652778 | 3.207209 | 3 | 15 |
| Motivation willingness | 9.694444 | 4.130406 | 3 | 15 |
| Social support emotionality | 8.097222 | 4.243171 | 3 | 15 |
| Social support appreciation | 6.166667 | 2.897303 | 2 | 10 |
| Social support instrumental | 10.375000 | 4.316485 | 3 | 15 |
| Empowerment knowledge | 10.541667 | 4.366768 | 3 | 15 |
| Empowerment abilities | 9.319444 | 4.181874 | 3 | 15 |
| Empowerment desires | 10.277778 | 4.482273 | 3 | 15 |

**Comparative Model Performance Evaluation**

FIGURE 3 presents a comparative evaluation of the classification performance between Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) models based on accuracy and Area Under the Curve (AUC) metrics.

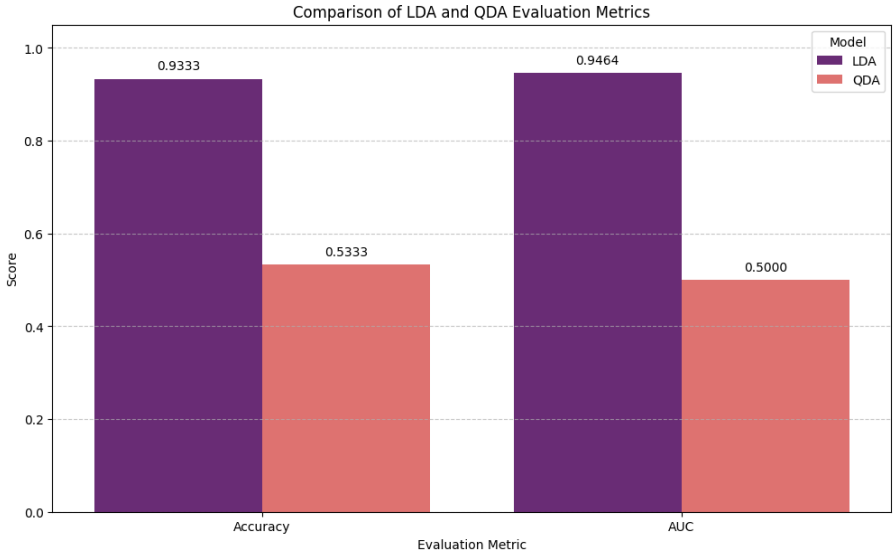


**FIGURE 3**. Comparison of LDA and QDA based on accuracy and AUC

The comparative analysis depicted in FIGURE 3 demonstrates a striking performance disparity between LDA and QDA classifiers. LDA emerges as the superior model, achieving an accuracy of 0.9333 (93.33%) and an AUC of 0.9464, substantially outperforming QDA, which yielded an accuracy of only 0.5333 (53.33%) and an AUC of 0.5000. This extreme performance divergence carries significant theoretical and practical implications. Given that LDA is theoretically optimized for linearly separable data distributions while QDA is designed for nonlinear classification boundaries, the superior performance of LDA strongly suggests that the underlying behavioral and psychosocial risk factors in this dataset exhibit linear separability characteristics. The near-perfect AUC value of 0.9464 for LDA indicates exceptional

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

discriminative capability, demonstrating that the model can correctly rank a randomly selected positive instance higher than a negative instance. Conversely, QDA's AUC of exactly 0.5000 signifies complete absence of discriminative power, performing equivalently to a random classifier.

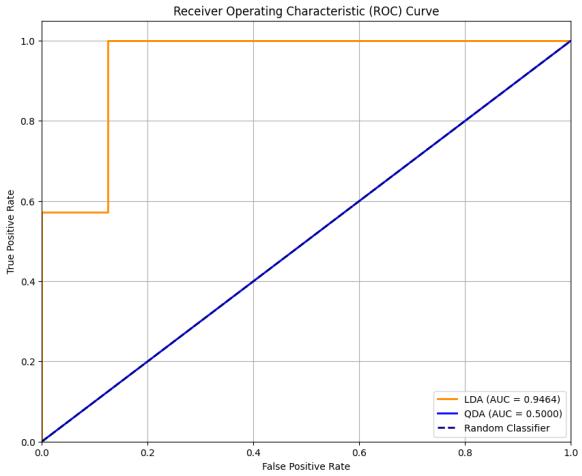


**FIGURE 4**. Comparison of ROC curve of each classifier

The Receiver Operating Characteristic (ROC) curves for the LDA and QDA classifiers are shown in FIGURE 4, which offers a graphic depiction of the trade-off between 1-specificity (False Positive Rate) and sensitivity (True Positive Rate) at different classification levels. The ROC curve analysis presented in FIGURE 4 offers a comprehensive visualization of classifier performance across all possible decision thresholds. The LDA classifier (orange curve) demonstrates superior discriminative capability with an AUC value of 0.9464, indicating that the model possesses a 94.64% probability of assigning a higher risk score to a randomly selected cervical cancer patient compared to a healthy individual. The LDA curve exhibits a pronounced trajectory toward the upper-left corner of the ROC space (coordinates 0,1), which represents the ideal classifier achieving 100% sensitivity with 0% false positive rate. This geometric configuration confirms that LDA maintains high true positive rates while minimizing false positive classifications across most threshold values, reflecting robust class separability and reliable risk stratification capability.

In stark contrast, the QDA classifier (dark blue curve) demonstrates complete predictive failure, with an AUC value of precisely 0.5000 that perfectly overlaps with the diagonal line representing a random classifier. This phenomenon indicates that QDA possesses no discriminative ability whatsoever, performing no better than a coin toss in distinguishing between cervical cancer cases and non-cases. The theoretical implications of this extreme disparity are profound: the data structure under investigation is fundamentally linearly separable, rendering LDA's linear decision boundary assumption highly appropriate. Conversely, QDA's quadratic decision boundary complexity introduces unnecessary model flexibility that, combined with limited sample size, results in severe overfitting and covariance matrix estimation instability. The quadratic boundaries, rather than capturing genuine nonlinear patterns, essentially model random noise in the training data, leading to catastrophic generalization failure on the test set. This finding underscores a critical principle in machine learning: increased model complexity does not inherently translate to improved performance and may, in fact, degrade predictive accuracy when the underlying data structure does not warrant such complexity or when sample sizes are insufficient to support complex parameter estimation.

**Confusion Matrix and Error Analysis**

Having established LDA as the superior model, FIGURE 5 presents the confusion matrix for the LDA classifier, providing granular insights into classification patterns and error distributions.

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

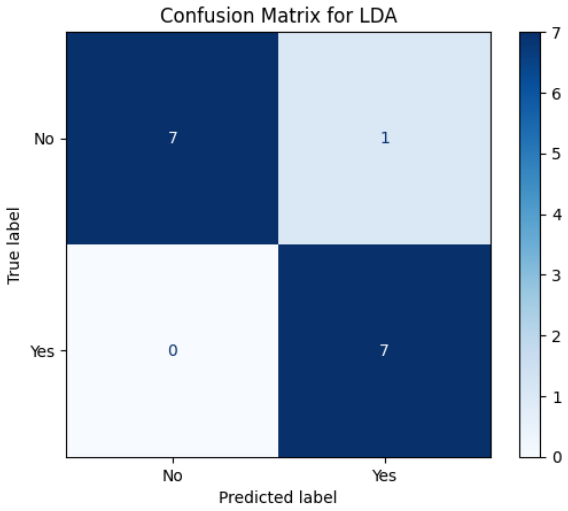


**FIGURE 5**. Confusion matrix of LDA as best model

The confusion matrix depicted in FIGURE 5 reveals a nearly perfect classification performance with minimal misclassification errors. The LDA model successfully identified 7 True Negative (TN) cases and 7 True Positive (TP) cases, demonstrating balanced accuracy across both classes. Critically, the matrix shows zero False Negative (FN = 0) predictions, indicating that the model achieved 100% sensitivity (recall) for the positive class ("yes" - cervical cancer present). This finding carries profound clinical significance: the model failed to miss a single actual cervical cancer case, thereby eliminating Type II errors (false negatives), which represent the most dangerous classification mistake in medical diagnostic contexts. A false negative in cervical cancer screening could result in delayed diagnosis, disease progression, and potentially fatal outcomes.

From a clinical implementation perspective, this error profile renders the LDA model highly suitable for cervical cancer screening applications where the primary objective is to maximize case detection sensitivity. The complete absence of false negatives ensures that no at-risk individual is erroneously reassured, while the minimal false positive rate maintains acceptable specificity. This performance characteristic aligns with the screening principle of "cast a wide net" to capture all potential cases, with subsequent confirmatory diagnostic tests serving to rule out false positives. The model's robustness in minimizing Type II errors while maintaining high overall accuracy (93.33%) validates its potential as a cost-effective, interpretable decision support tool for early cervical cancer risk assessment in resource-constrained healthcare settings.

### Comparative Analysis with Prior Research

Table 3 presents a comprehensive comparison of LDA and QDA performance against previously published machine learning algorithms applied to the same cervical cancer behavioral dataset.

**TABLE 3.** Comparison results between prior research

| Method | Accuracy |
|---|---|
| Naïve Bayes [25] | 91.67% |
| Logistic Regression [25] | 87.5% |
| C4.5 [25] | 78% |
| Decision Stump [25] | 80% |
| Multilayer Perceptron [25] | 82.93% |
| Decision Tree [25] | 77.97% |
| Support Vector Machine [25] | 79.25% |
| KNN (k=3; distance=Manhattan) [14] | 93.06% |
| **Linear Discriminant Analysis** | **93.33%** |
| Quadratic Discriminant Analysis | 53.33% |

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

The comparative evaluation presented in TABLE 3 reveals significant performance heterogeneity across different machine learning algorithms, underscoring the critical influence of algorithm-data compatibility on classification effectiveness. Linear Discriminant Analysis emerges as the empirically superior approach, achieving the highest accuracy of 93.33%, narrowly surpassing K-Nearest Neighbors (KNN) with k=3 and Manhattan distance metric at 93.06% [14]. Naïve Bayes follows closely at 91.67%, while logistic regression achieved 87.5% accuracy. Tree-based methods demonstrated comparatively modest performance, with C4.5 (78%), Decision Stump (80%), and traditional Decision Tree (77.97%) algorithms exhibiting lower predictive accuracy. Support Vector Machine achieved 79.25%, and Multilayer Perceptron neural networks reached 82.93% [25]. Most notably, Quadratic Discriminant Analysis exhibited catastrophic failure with an accuracy of only 53.33%, performing marginally better than random guessing.

The superior performance of LDA can be theoretically attributed to several factors. First, LDA's success indicates that the behavioral and psychosocial risk factor data possesses linearly separable characteristics and likely satisfies the multivariate normality assumption, enabling linear dimensionality reduction to maximize inter-class separation while minimizing intra-class variance. Second, the strong performance of KNN (93.06%) and Naïve Bayes (91.67%) reinforces the argument that the data exhibits strong local structure and that feature independence assumptions are reasonably well-satisfied, respectively. These complementary findings suggest that the dataset contains both global linear patterns (captured by LDA) and local neighborhood structures (exploited by KNN).

Conversely, QDA's dramatic failure at 53.33% accuracy provides compelling evidence that the additional model complexity introduced by class-specific covariance matrix estimation is unwarranted and detrimental. This suggests that, in theory, this dataset satisfies the LDA-fundamental requirement of equal covariance matrices across classes, and QDA's attempt to model different covariance structures for each class results in severe overfitting. The alignment between model assumptions and data characteristics, as demonstrated by LDA's success, proves more critical than algorithmic sophistication in achieving robust generalization.

Collectively, these comparison findings demonstrate that LDA is the best classifier for predicting cervical cancer risk based on behavioral and psychosocial factors, offering an exceptional balance of accuracy (93.33%), interpretability, computational efficiency, and clinical applicability. The model's performance surpasses or matches all previously evaluated algorithms while maintaining theoretical transparency and practical implementability in resource-constrained healthcare environments.

## CONCLUSION

This study demonstrates that Linear Discriminant Analysis (LDA) is a superior and highly practical method for predicting cervical cancer risk based on behavioral and psychosocial factors. The empirical findings reveal that LDA achieved exceptional classification performance with 93.33% accuracy and 0.9464 AUC, surpassing all previously evaluated machine learning algorithms, including K-Nearest Neighbors (93.06%) and Naïve Bayes (91.67%). Critically, LDA attained 100% sensitivity with zero false negatives, effectively eliminating the most dangerous classification error in medical screening contexts. Conversely, Quadratic Discriminant Analysis (QDA) exhibited complete predictive failure (53.33% accuracy), indicating that the underlying behavioral data possesses linearly separable characteristics and that QDA's additional complexity resulted in severe overfitting given the limited sample size. These findings contribute significantly to medical informatics by validating that simpler, assumption-based models can outperform complex algorithms when appropriately aligned with data structure. The study establishes LDA as a cost-effective, interpretable decision support tool suitable for resource-constrained healthcare settings. For healthcare practitioners, integrating this transparent LDA model into routine screening workflows can facilitate early, non-invasive risk stratification without relying on expensive clinical biomarkers. Ultimately, this approach paves the way for developing accessible clinical decision support systems that empower primary care providers to implement targeted, data-driven preventive interventions. Future research should investigate LDA's performance on larger, multi-center datasets and explore ensemble methods combining LDA with complementary algorithms to enhance robustness across diverse population characteristics.

## 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] M. M. Muraru, Z. Simó, and L. B. Iantovics, "Cervical Cancer Prediction Based on Imbalanced Data Using Machine Learning Algorithms with a Variety of Sampling Methods," *Applied Sciences (Switzerland)*, vol. 14, no. 22, Nov. 2024, doi: 10.3390/app142210085.
[3] P. R. Garg *et al.*, "Women's Knowledge on Cervical Cancer Risk Factors and Symptoms: A Cross Sectional Study from Urban India," *Asian Pacific Journal of Cancer Prevention*, vol. 23, no. 3, pp. 1083–1090, 2022, doi: 10.31557/APJCP.2022.23.3.1083.
[4] J. Lu, E. Song, A. Ghoneim, and M. Alrashoud, "Machine learning for assisting cervical cancer diagnosis: An ensemble approach," in *Future Generation Computer Systems*, Elsevier B.V., 2020, pp. 199–205. doi: 10.1016/j.future.2019.12.033.
[5] T. M. Hezron and H. Hasanuddin, "Cervical Cancer with Bulky Tumor: A Case Report," *Journal of Society Medicine*, vol. 4, no. 5, pp. 153–157, May 2025, doi: 10.71197/jsocmed.v4i5.211.
[6] S. Setiawati and Y. Hapsari, "Clinical Manifestations, Diagnosis, Management and Prevention of Cervical Cancer," *Jurnal Biologi Tropis*, vol. 23, no. 4, pp. 382–390, 2023, doi: 10.29303/jbt.v23i4.5594.
[7] N. M. L. D. Paramitha, I. G. A. S. M. Dewi, P. E. Paskarani, and N. W. Winarti, "Clinicopathologic Features of Carcinoma of the Cervix Uteri at Prof. Dr. I.G.N.G. Ngoerah Hospital in 2021-2022," *Indonesian Journal of Cancer*, vol. 19, no. 3, pp. 381–385, Sep. 2025, doi: 10.33371/ijoc.v19i3.1328.
[8] S. Gupta and M. K. Gupta, "Computational Prediction of Cervical Cancer Diagnosis Using Ensemble-Based Classification Algorithm," *Computer Journal*, vol. 65, no. 6, pp. 1527–1539, 2022, doi: 10.1093/comjnl/bxaa198.
[9] S. Zhang, H. Xu, L. Zhang, and Y. Qiao, "Cervical cancer: Epidemiology, risk factors and screening," *Chinese Journal of Cancer Research*, vol. 32, no. 6, pp. 720–728, 2020, doi: 10.21147/j.issn.1000-9604.2020.06.05.
[10] M. Musthofa and M. Anshori, "Comparing Discriminant Analysis Function for Early Prediction of Smartphone Addiction," *Journal of Enhanced Studies in Informatics and Computer Applications*, vol. 2, no. 1, pp. 1–7, 2025, doi: https://doi.org/10.47794/jesica.v2i1.12.
[11] 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.
[12] 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.
[13] S. I. Journal, "Cervical Cancer Cell Prediction using Machine Learning Classification Algorithms Cervical Cancer Cell Prediction using," *Engineering and Scientific International Journal*, vol. 8, no. 1, pp. 25–29, 2021, doi: 10.30726/esij/v8.i1.2021.81006.
[14] M. C. Saletia, M. Anshori, and S. Haris, "Comparing Different KNN Parameters Based on Woman Risk Factors to Predict the Cervical Cancer," *Journal of Applied Informatics and Computing (JAIC)*, vol. 9, no. 5, p. 2179, Oct. 2025, doi: https://doi.org/10.30871/jaic.v9i5.10746.
[15] M. Musthofa and M. Anshori, "Comparing Discriminant Analysis Function for Early Prediction of Smartphone Addiction," *Journal of Enhanced Studies in Informatics and Computer Applications*, vol. 2, no. 1, pp. 1–7, 2025, doi: 10.47794/jesica.v2i1.12.
[16] H. Prasetyo Wibowo, M. Anshori, and M. Syauqi Haris, "the Discriminant Analysis Function Was Implemented To Predict the Presence of Diabetes," *Journal of Enhanced Studies in Informatics and Computer Applications*, vol. 1, no. 2, pp. 47–55, Jul. 2024, doi: 10.47794/jesica.v1i2.10.
[17] 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.
[18] M. Anshori, R. Siwi Pradini, and W. T. Kusuma, "SMOTE Effectiveness and various Machine Learning Algorithms to Predict Self-Esteem Levels of Indonesian Student," vol. 7, no. 2, pp. 175–182, May 2025, doi: 10.21512/emacsjournal.v6.
[19] H. W. Wibowo, M. Hasbi, and M. Anshori, "Use Discriminant Analysis to Identify Eroticism-Related Terms in The Lyrics of Dangdut Songs," *Journal of Enhanced Studies in Informatics and Computer Applications*, vol. 1, no. 1, pp. 17–22, 2024, doi: 10.47794/jesica.v1i1.5.
[20] S. Guo and H. Tracey, "Discriminant Analysis for Radar Signal Classification," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 56, no. 4, pp. 3134–3148, 2020, doi: 10.1109/TAES.2020.2965787.

**Universitas Negeri Jakarta**
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

[21] Andrei Shestserau, “Effectiveness of applying the pareto principle in optimizing business processes and expenses,” *World Journal of Advanced Research and Reviews*, vol. 21, no. 3, pp. 2690–2698, Mar. 2024, doi: 10.30574/wjarr.2024.21.3.0895.

[22] P. Király and G. Tóth, “Being Aware of Data Leakage and Cross-Validation Scaling in Chemometric Model Validation,” *J. Chemom.*, vol. 39, no. 4, Apr. 2025, doi: 10.1002/cem.70026.

[23] M. Anshori and M. S. Haris, “Predicting Heart Disease using Logistic Regression,” *Knowledge Engineering and Data Science*, vol. 5, no. 2, pp. 3016–3019, Dec. 2022, doi: 10.17977/um018v5i22022p188-196.

[24] M. Anshori and M. Musthofa, “Outperforming DNN Using MLP in Water Quality Assessment for Aquaculture,” *Journal of Applied Informatics and Computing (JAIC)*, vol. 10, no. 1, p. 355, Feb. 2026, doi: 10.30871/jaic.v10i1.11835.

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