Optimizing Cervical Cancer Diagnosis with Correlation-Based Feature Selection: A Comparative Study of Machine Learning Models

Wiwit Supriyanti[1], Sujalwo[2], Dimas Aryo Anggoro[3], Maryam[4], Nova Tri Romadloni[5]
Computer Engineering, Universitas Muhammadiyah Karanganyar, Karanganyar, Indonesia[1]
Informatics, Universitas Muhammadiyah Karanganyar, Karanganyar, Indonesia[2, 5]
Informatics Department, Universitas Muhammadiyah Surakarta, Surakarta, Indonesia[3, 4]

***Abstract*—Cervical cancer remains a significant global health issue, particularly in developing countries where it is a leading cause of mortality among women. The development of machine learning-based approaches has become essential for early detection and diagnosis of cervical cancer. This research explores the optimization of classification algorithms through Correlation-Based Feature Selection (CFS) for early cervical cancer detection. A dataset consisting of 198 samples and 22 attributes from medical records was processed to reduce dimensionality. CFS was used to select the most relevant features, which were then applied to three classification algorithms: Naïve Bayes, Decision Tree, and k-Nearest Neighbor (k-NN). The results showed that CFS significantly improved classification accuracy, with Decision Tree achieving the highest accuracy of 85.89%, followed by Naïve Bayes with 83.34%, and k-NN with 82.32%. These findings demonstrate the importance of feature selection in enhancing classification performance and its potential application in the development of cervical cancer detection tools.**

***Keywords*—Cervical cancer; feature selection; machine learning**

I. Introduction

Cervical cancer is a major global health concern and a leading cause of women's death; the situation is even worse in developing regions due to a lack of health facilities and people's knowledge regarding this deadly disease. People have various assumptions regarding this disease; on the contrary, their knowledge often misleads them on how to prevent the deadly disease or to heal it [1].

For example, a common misconception about cervical cancer is that it only affects women who are sexually active or those with a history of multiple sexual partners. While human papillomavirus (HPV) infection, a key risk factor for cervical cancer, is often transmitted through sexual contact, it is crucial to note that the disease can also develop in women with a limited sexual history with only one partner in conservative marriage or even those who have never been sexually active since HPV can be spread through various ways.

HPV (Human Papillomavirus) is most commonly associated with sexual transmission, but there is a strong possibility that HPV will be transmitted through non-sexual means. One possible way is through direct skin-to-skin contact. HPV can be transmitted when the skin comes into contact with an infected area, even without sexual intercourse. This means that activities such as touching or rubbing areas where the virus is present—like the genital, anal, or oral regions—can lead to transmission. The virus can enter the body through tiny cuts, abrasions, or micro-tears in the skin or mucous membranes, making it possible for someone to contract HPV without having penetrative sex. Since the youth has been experiencing pre-sexual activity, there are strong possibilities that such a viral transmission through this case will eventually occur due to the frequency and possibility that happened after pre-sexual activities.

Another potential, although less common, route of HPV transmission is through indirect contact with contaminated objects. This can happen if personal items such as towels, razors, or undergarments, which an infected person has used, come into contact with broken skin or mucous membranes. While the likelihood of contracting HPV in this manner is lower, it is still a possible mode of transmission. Thus, sharing personal items should be avoided to reduce the risk of spreading the virus.

HPV can also be transmitted from mother to child during childbirth, a process known as vertical transmission. In these cases, the newborn may come into contact with the virus in the birth canal, which can lead to conditions such as recurrent respiratory papillomatosis—a rare disorder where warts grow in the respiratory tract. While such occurrences are uncommon, they highlight the need for awareness that HPV transmission is not solely linked to sexual behavior.

Since HPV is highly infectious, this misconception can lead to a dangerous delay in screening or seeking medical advice, as many women might mistakenly believe they are not at risk since the women do not have enough exposure to information that viral infection can be caused by non-penetrative activities and the virus can remain dormant for years before slowly transforming cellular changes in the cervix.

Furthermore, other factors such as genetics, smoking, and a weakened immune system also play a role in the development of cervical cancer. Raising awareness about the true risk factors and promoting regular screening, such as Pap smears and HPV tests, are crucial in early detection and prevention, regardless of one's sexual history.

Not only the risk of infection that urgently needs to be addressed but simultaneously, we should discuss the fatality of this disease. Naturally, cervical cancer originates in a woman's

cervix, a part of the female reproductive system that is particularly vulnerable to infections and abnormal cellular changes. This vulnerability is due to its location and exposure to the external environment, making it susceptible to open wounds, abrasions, and subsequent infections by viruses or bacteria. HPV infection, in particular, poses a significant risk because it can cause cellular mutations in the cervix that may eventually lead to cancer. Unfortunately, cervical cancer can be deceptively slow in its development, with precancerous changes taking years or even decades to progress into invasive cancer. This gradual progression often results in many women being unaware of the presence of the disease until it reaches an advanced stage, by which time treatment becomes more complex, and outcomes are less favorable.

What makes cervical cancer particularly dangerous is that the symptoms often do not appear until the cancer has progressed significantly. Symptoms like unusual vaginal bleeding, pelvic pain, or discomfort during intercourse may not manifest until the cancer is already advanced, making early detection crucial. Regular screenings, such as Pap smears and HPV tests, are therefore vital for catching any abnormalities at the earliest possible stage. Despite these challenges, advancements in cancer detection technology are continually improving. For instance, the development of the HPV vaccine and the ability to test specifically for high-risk HPV strains have greatly enhanced the ability to prevent and diagnose this type of cancer early on. By identifying the presence of high-risk HPV strains before they lead to cellular changes, healthcare providers can intervene with treatment or increased monitoring to prevent the progression of cervical cancer.

Regular, yearly examinations are strongly recommended for all women, particularly those between the ages of 21 and 65, to detect early HPV infections or precancerous changes. Early diagnosis significantly reduces the risk of cervical cancer progressing to a life-threatening stage and provides the opportunity for effective treatment and management. Education and awareness about the non-sexual transmission of HPV and the importance of regular screenings can help dispel misconceptions and encourage more women to take preventive action against this potentially devastating disease.

What makes cervical cancer particularly dangerous is that the symptoms often do not appear until the cancer has progressed significantly. Symptoms like unusual vaginal bleeding, pelvic pain, or discomfort during intercourse may not manifest until the cancer is already advanced, making early detection crucial. Regular screenings, such as Pap smears and HPV tests, are therefore vital for catching any abnormalities at the earliest possible stage. Despite these challenges, advancements in cancer detection technology are continually improving. For instance, the development of the HPV vaccine and the ability to test specifically for high-risk HPV strains have greatly enhanced the ability to prevent and diagnose this type of cancer early on. By identifying the presence of high-risk HPV strains before they lead to cellular changes, healthcare providers can intervene with treatment or increased monitoring to prevent the progression of cervical cancer.

Regular, yearly examinations are strongly recommended for all women, particularly those between the ages of 21 and 65, to detect early HPV infections or precancerous changes. Early diagnosis significantly reduces the risk of cervical cancer progressing to a life-threatening stage and provides the opportunity for effective treatment and management. Education and awareness about the non-sexual transmission of HPV and the importance of regular screenings can help dispel misconceptions and encourage more women to take preventive action against this potentially devastating disease.

Based on the nature of this virus, several prevention systems should be established. if the sexual educational system is designed systematically and the anti-HPV campaign can reach youth efficiently, the spread of HPV can be contained further. Thus, in reality, the youth does not have enough strong relationships and bonds with stakeholders, and an adequate prevention system for protecting women from HPV has not been established firmly.

Consequently, according to the World Health Organization (WHO), in 2018, there were over 570,000 new cases of cervical cancer, leading to more than 311,000 deaths linked to the disease [2]. This statistic highlights that, despite advances in medical technology, the failure of the cervical cancer prevention system has not successfully yet halted cervical cancer as a remaining significant threat to women's health globally.

From the early situation, the urgency of the significance of early detection of cervical cancer is a top priority for saving women widely in different social and economic classes. By detecting cervical cancer as early as possible, women’s lives would be saved, and they can maintain a proper and standardized quality for women as expected based on the conception of human rights. This statement is strengthened by further research that measures the standardization of human life quality and the development of early-stage detection of HPV and cervical cancer. By detecting cervical cancer earlier, the treatment will work significantly, and the impact on women’s lives is undeniably positive. Undoubtedly, when cervical cancer is identified in its early stages, the chances of successful treatment significantly increase. For example, preventive measures such as Pap smear tests and HPV vaccinations have been successfully efficient in controlling the more profound impacts of cervical cancer in terms of health and social welfare.

Nonetheless, while these strategies have successfully reduced incidence and mortality rates in some developed countries, access to early detection methods in developing nations remains severely limited. The challenges include a lack of medical resources, high diagnostic costs, and insufficient public awareness regarding the importance of early detection [3].

With the advancement of technology and data processing, information technology-based methods—such as machine learning and data mining—are increasingly being utilized in the analysis of medical records to assist in diagnosing and early detection of diseases, including cervical cancer. Classification is one of the most frequently employed techniques in data mining, wherein models are created to predict whether a patient is at high risk of a specific disease, including cervical cancer, based on historical or existing medical data [4]. The application of classification algorithms allows researchers to uncover patterns

within datasets and produce accurate predictions based on the available attributes.

Nevertheless, one of the most significant challenges in medical classification processes, particularly for the early detection of cervical cancer, is the high dimensionality of the data. Medical datasets often contain numerous attributes or features, many of which may be irrelevant for predictions or could introduce noise into the model. This situation is referred to as the high dimensionality problem, which can lead to reduced model performance and an increased risk of overfitting, where the model becomes overly focused on the training data and fails to generalize well to new data [5]. In the context of early cervical cancer detection, attributes within the dataset might include various patient information such as age, pregnancy history, age at first menstruation, and symptoms like fatigue, abnormal bleeding, and abdominal lumps. While some attributes may hold greater significance than others, the sheer number of attributes complicates the classification process and makes it challenging to interpret.

To address the high dimensionality issue, feature selection is employed. Feature selection involves selecting the most relevant and informative subset of features from the dataset while discarding irrelevant or redundant ones. The aim of this process is to simplify the model, enhance interpretability, and, most importantly, improve predictive performance. By utilizing feature selection methods, it is anticipated that classification algorithms can operate more efficiently and accurately [6]. One popular feature selection method used in various classification applications is Correlation-Based Feature Selection (CFS).

High dimensionality in medical datasets presents various challenges, particularly in computation and interpretation. As the number of features increases, the potential combinations in data analysis also rise exponentially. This implies that classification algorithms require more time and computational resources to process datasets with a larger number of features, increasing the likelihood of overfitting. Overfitting occurs when the model is too complex and fits the training data closely, yet fails to generalize to new test data, resulting in poor predictive performance when applied to real-world data [7].

For instance, in a medical dataset aimed at the early detection of cervical cancer, certain attributes may be highly relevant for predicting outcomes, such as the patient's age or pregnancy history. Conversely, there may be attributes that are irrelevant or have a low correlation with the target variable, such as unrelated family health history regarding cervical cancer risk. These extraneous features complicate the model without significantly contributing to predictive accuracy. Thus, implementing feature selection methods to identify the most important attributes for the classification process is essential [8].

Correlation-Based Feature Selection (CFS) is a technique that identifies an optimal subset of features based on the correlation between features in the dataset and the target variable. The fundamental concept behind CFS is that a good subset of features should consist of those that have a strong correlation with the target variable while maintaining a low correlation with one another. In other words, CFS selects features that have a robust relationship with the target variable (i.e., whether a patient is diagnosed with cervical cancer) while avoiding redundant features that provide overlapping information [9].

CFS operates by calculating the Pearson correlation between each feature and the target variable, as well as between the features themselves. The Pearson correlation measures the strength and direction of the linear relationship between two variables. Features with strong positive or negative correlations with the target variable are retained, while those with weak or insignificant correlations are discarded. Additionally, CFS considers the correlation among features to prevent redundancy. Features that are highly correlated with one another are deemed to provide similar information, allowing for one of them to be removed without diminishing model performance [10].

The primary advantage of CFS lies in its ability to reduce data dimensionality without compromising predictive accuracy. By eliminating irrelevant or redundant features, CFS helps classification models become simpler, faster, and more interpretable. This aspect is particularly crucial in medical applications, where the interpretability of the model is a key factor in providing trustworthy diagnoses [11].

In this study, after applying feature selection using CFS, the reduced dataset will be used to train classification models. The three classification algorithms employed are Decision Tree [12], Naïve Bayes [13], and k-nearest Neighbor (k-NN) [14]. These algorithms were selected for their differing approaches to classification problems and their unique strengths in medical data analysis. Each algorithm has its own advantages and limitations, and the findings of this study will assess the performance of each algorithm following feature selection using CFS.

Previous research on early detection of cervical cancer utilizing data mining techniques has yielded positive results. For instance, a study by Ali [4] employed the Random Forest method to analyze medical datasets, achieving a high level of accuracy in predicting cervical cancer. Another study by Rahmi [15] implemented the SMOTE algorithm and Naïve Bayes to address imbalanced data issues in the early detection of cervical cancer in Indonesia. Both studies acknowledged the importance of feature selection as a critical factor in enhancing the performance of classification models.

However, challenges remain unresolved in prior research, particularly concerning high dimensionality and model interpretability. This study aims to address these challenges by applying the Correlation-Based Feature Selection (CFS) method to reduce data dimensionality and by comparing the performance of various classification algorithms to identify the most effective approach for detecting cervical cancer.

II. Literature Review

The optimization of cervical cancer diagnosis has seen significant advancements with the integration of correlation-based feature selection and machine learning models. Recent research [16] emphasized the critical role of selecting key clinical and behavioral criteria in risk assessment. Their methodology combined fuzzy Multi-Criteria Decision Making (MCDM) with machine learning to improve the reliability of diagnostic outcomes, highlighting the importance of reducing data complexity while preserving essential diagnostic features.

Addressing the challenges of high-dimensional datasets, Nithya and Ilango [17] introduced a fused feature selection framework that combines filter and wrapper methods to optimize cervical cancer classification. Their approach effectively mitigates issues such as redundant attributes, irrelevant features, and class imbalance, resulting in improved classification accuracy. This demonstrates the importance of integrated feature selection strategies in handling complex medical datasets.

Deep learning has further enhanced cervical cancer diagnostics, particularly in image-based classification tasks. Tawalbeh [18] conducted a comparative study of six feature fusion techniques applied to deep learning models. By leveraging canonical correlation analysis, they achieved a classification accuracy of 99.7%, underscoring the power of optimized feature fusion in improving performance across multiple cancer classes.

Hasan [8] explored machine learning explainability in cervical cancer classification, employing the Boruta algorithm for feature selection and tools like SHAP for model interpretability. Using Random Forest as a classifier, they achieved an accuracy of 99.85%, demonstrating the value of combining explainable AI techniques with advanced feature selection to enhance diagnostic reliability.

The use of wrapper-based feature selection techniques has also proven effective in cervical cancer diagnostics. Setiawan [19] employed Grey Wolf Optimization (GWO) with classifiers like Naive Bayes and Support Vector Machines. Their findings revealed that NB-GWO outperformed other configurations, achieving an accuracy of 96.30%, highlighting the effectiveness of metaheuristic algorithms in selecting optimal features.

Machine learning has been pivotal in developing self-risk assessment tools for cervical cancer. Ramzan [20] utilized AdaBoost and feature selection algorithms to create personalized risk assessment models. Their technique enabled women to estimate their cervical cancer risk with high accuracy, leveraging demographic and medical history data.

Feature selection techniques, such as mutual information and genetic algorithms, have been extensively studied for their ability to remove irrelevant attributes and enhance model performance. Combining these methods with machine learning has proven particularly effective in addressing issues of overfitting and improving the generalization ability of models.

The application of explainable AI in medical diagnostics has received considerable attention, as demonstrated by Hasan [8]. Their integration of feature importance with interpretable models not only improved diagnostic accuracy but also provided clinicians with insights into the key predictors of cervical cancer, bridging the gap between AI and clinical practice.

Comparative analyses of feature selection methods have highlighted the need for tailored strategies to address the unique challenges of medical datasets. Studies like those of Tawalbeh [18] and Setiawan [19] illustrate the benefits of combining advanced feature engineering with machine learning, fostering the development of robust diagnostic tools.

Together, these studies underscore the transformative potential of correlation-based feature selection and machine learning in cervical cancer diagnosis. By integrating these techniques, researchers have achieved higher accuracy, better interpretability, and improved reliability in diagnostic models, paving the way for early detection and personalized healthcare solutions.

III. Methodology

The methods section of this research includes a series of steps starting with data collection, then moving to data preprocessing, feature selection using the Correlation-Based Feature Selection (CFS) technique, the application of classification algorithms, and finally, the assessment of the classification model's performance [21] (Fig. 1).

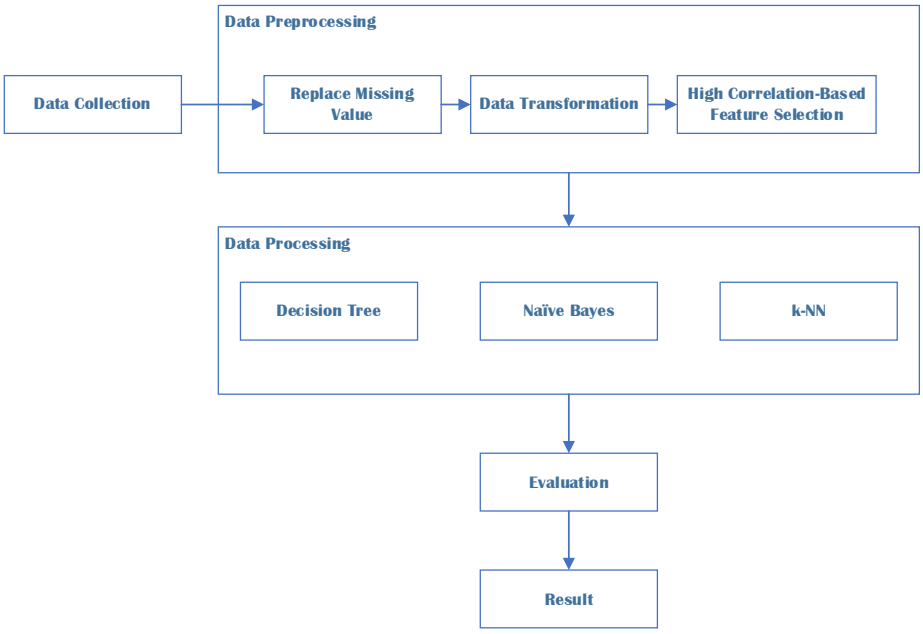

Fig. 1. The flow of research.

A. Data Collection

The data utilized in this research was collected from a hospital located in the area where the researcher resides, comprising a dataset of 198 patient samples who have undergone medical assessments related to cervical cancer. This dataset contains 22 attributes or features that represent various factors that may impact the diagnosis of cervical cancer. These attributes include demographic factors such as age and pregnancy history, as well as clinical symptoms like pelvic pain, abnormal discharge, and bleeding that occurs outside the menstrual cycle. The target variable for this study is the classification of cervical cancer stages, which includes in-situ, early, and advanced stages [22]. Table I shows attributes and variables of cervical cancer.

TABLE I. ATTRIBUTES AND VARIABLES OF CERVICAL CANCER

Variables	Attributes	Description
X1	Age	Age of the patient in years
X2	Marriages	Patient's history of number of marriages
X3	Miscarriages	Patient's history of number of miscarriages
X4	Childbirths	Patient's history of number of deliveries
X5	Age at first marriage	Patient's age at first marriage
X6	Age at first menstruation	Patient's age at first menstruation
X7	Difficulty defecating	0 = no, 1 = yes
X8	Difficulty urinating	0 = no, 1 = yes
X9	Decreased appetite	0 = no, 1 = yes
X10	Pelvic pain	0 = no, 1 = yes
X11	Lower abdominal pain	0 = no, 1 = yes
X12	Weight loss	0 = no, 1 = yes
X13	Nausea	0 = no, 1 = yes
X14	Vomiting	0 = no, 1 = yes
X15	Fatigue	0 = no, 1 = yes
X16	Foul-smelling discharge	0 = no, 1 = yes
X17	Discharge color	milky white, yellowish, greenish
X18	Bleeding outside of cycle	light, heavy
X19	Duration of bleeding	0-7 days, 7-14 days, more than 14 days
X20	Post-coital bleeding	0 = no, 1 = yes
X21	Abdominal lump	0 = no, 1 = yes
X22	Shortness of breath	0 = no, 1 = yes
Y	Cancer stage classification	in-situ, early stage, advanced stage

B. Data Preprocessing

Data pre-processing is an essential phase in preparing datasets to ensure they are suitable for analysis and classification. In this research, the pre-processing stage consists of three primary steps: addressing missing values, transforming the data, and selecting features using the Correlation-Based Feature Selection (CFS) method [23]. A detailed explanation of each step is provided below:

1) Replace missing values: In medical datasets, the occurrence of missing values is common due to incomplete data collection or inconsistencies in patient record documentation. It is essential to address these missing values to ensure the dataset's integrity, allowing the classification model to perform effectively. This study employs two prevalent methods for handling missing values:

a) Row deletion: When a row contains a significant proportion of missing data (for instance, more than 50%), it is removed from the dataset. This approach is taken because a high volume of missing data can compromise subsequent analyses.

b) Imputation: For attributes with only a small number of missing values, imputation techniques are utilized. This method replaces missing values using the information available within the dataset. The imputation process can be done through:

- Mean imputation: for numerical attributes, missing values are substituted with the average value of that attribute.
- Mode imputation: for categorical attributes, missing values are filled with the mode (the most frequently occurring value) of that attribute.

For instance, if there are some missing values in the age attribute, these values are replaced with the average age of the available patients. These techniques help maintain the distribution of the data without introducing significant bias.

2) Data transformation: Once the missing values have been addressed, the subsequent step involves transforming the data to ensure that all attributes are on an appropriate scale and in a format suitable for the classification algorithms being utilized. The data transformation process consists of:

a) Normalization of data: Normalization is essential to ensure that all numerical attributes are consistent in scale, particularly because algorithms such as k-Nearest Neighbor (k-NN) are very sensitive to variations in scale among the attributes. In this research, the Min-Max Scaling method is employed, which adjusts each attribute value to fall within a range of 0 to 1. The formula for normalization is:

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

where x is the original value of the attribute, $x_{\min}$ is the minimum value of that attribute, and $x_{\max}$ is the maximum value. Normalization helps prevent attributes with larger scales from dominating those with smaller scales [24].

b) Data standardization: In addition to normalization, standardization is applied to certain attributes that exhibit distributions significantly different from a normal distribution. Standardization is performed by subtracting the mean of each attribute from its value and then dividing by the standard

deviation of that attribute, as illustrated in the following equation:

$$z = \frac{x-\mu}{\sigma} \quad (2)$$

where z is the standardized value, x is the original value, μ is the mean, and σ is the standard deviation. Standardization ensures that the attributes have a distribution with a mean of 0 and a standard deviation of 1.

3) Correlation-based feature selection (CFS): After the data transformation process, the final step in pre-processing is feature selection using the Correlation-Based Feature Selection (CFS) method. CFS is a feature selection technique that automatically identifies the most relevant attributes in relation to the target variable while minimizing redundancy among the attributes. CFS operates by calculating the correlation between attributes and the target variable, as well as the correlation among the attributes themselves.

The steps in CFS are as follows:

a) Feature correlation with the target variable: CFS computes the Pearson correlation between each feature and the target variable (cervical cancer classes: in situ, early, advanced). Attributes with a high correlation to the target variable are retained, while those with low correlation are discarded.

b) Elimination of redundancy among features: In addition to retaining features that are highly correlated with the target variable, CFS also considers the correlation among the features themselves. Attributes that exhibit high correlation with one another are deemed redundant, and one of them is removed. The goal is to minimize multicollinearity within the dataset.

CFS calculates the "goodness value" (Merit) for each subset of features using the following formula:

$$Merit_S = \frac{k\,.\overline{r_{cf}}}{\sqrt{k+k\,(k-1).\overline{r_{ff}}}} \quad (3)$$

where:

- k is the number of selected features,
- $\overline{r_{cf}}$ is the average correlation between the features and the target variable,
- $\overline{r_{ff}}$ is the average correlation among the features.

The result of the CFS method is a reduction in the number of features from 22 to the 10 most relevant attributes for classification. This feature reduction not only enhances computational efficiency but also helps mitigate the risk of overfitting caused by the inclusion of irrelevant attributes.

C. Classification Algorithms

This study utilizes three popular classification algorithms to develop predictive models: Decision Tree, Naïve Bayes, and k-Nearest Neighbor (k-NN). Below is a detailed explanation of each algorithm:

1) Decision tree: The Decision Tree algorithm constructs a decision tree where each internal node represents an attribute or feature, each branch represents a decision based on the attribute's value, and each leaf node represents a classification outcome. Decisions are made by recursively partitioning the dataset into smaller subsets based on feature values that provide the most information. To select the most informative features, the concept of Information Gain based on Shannon entropy is employed. The formula for calculating entropy is as follows:

$$H(D) = -\sum_{i=1}^{C} p(i) log_2 p(i) \quad (4)$$

where p_i is the probability of class i in the dataset, and C is the total number of classes. The feature with the highest Information Gain is used to split the dataset at each node.

2) Naïve bayes: Naïve Bayes is a probability-based algorithm grounded in Bayes' Theorem. This algorithm assumes that all features are independent of one another, which is often unrealistic in real-world applications. However, in practice, Naïve Bayes continues to yield favorable results, particularly in text classification and large datasets. The basic equation of Naïve Bayes is:

$$P(C_k|X) = \frac{P(X|C_k)\,.P(C_k)}{P(X)} \quad (5)$$

where:

- $P(C_k|X)$ is the probability that sample X belongs to class C_k,
- $P(X|C_k)$ is the probability of observing X given class C_k,
- $P(C_k)$ is the prior probability of class C_k,
- $P(X)$ is the overall probability of the data X.

3) k-Nearest neighbor (k-NN): k-NN is an instance-based algorithm that classifies samples based on their distance to the nearest samples in the dataset. The algorithm works by calculating the Euclidean distance between the unknown sample and all existing samples in the dataset, then selecting the k closest neighbors and determining the majority class among them. The formula for Euclidean distance used in k-NN is:

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad (6)$$

where p and q are two samples in the data space, and n is the number of features.

D. Evaluation

To evaluate the performance of the classification model, this study uses metrics derived from the Confusion Matrix, which includes calculations for accuracy, sensitivity, specificity, and precision [25]. Below are the explanations and formulas used to compute these metrics:

1) Accuracy: Accuracy measures the proportion of correct predictions against the total samples and is calculated using the following formula [26]:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

where:

- *TP* (True Positive) is the number of positive samples correctly classified,
- *TN* (True Negative) is the number of negative samples correctly classified,
- *FP* (False Positive) is the number of negative samples incorrectly classified as positive,
- *FN* (False Negative) is the number of positive samples incorrectly classified as negative.

2) Sensitivity (Recall): Sensitivity measures the model's ability to correctly identify positive samples and is calculated using the following formula:

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

A high sensitivity indicates that the model is effective in detecting positive conditions.

3) Specificity: Specificity measures the model's ability to accurately identify negative samples and is calculated using the following formula:

$$Specificity = \frac{TN}{TN+FP} \quad (9)$$

High specificity means that the model can identify negative samples with a high degree of accuracy.

4) Precision: Precision measures the accuracy of positive predictions, indicating how many of the positive predictions were correct, and is calculated using the following formula:

$$Precision = \frac{TP}{TP+FP} \quad (10)$$

A high precision indicates that the model rarely produces false positive results.

After evaluating all the metrics, a comparative analysis is conducted to determine which algorithm performs best in detecting cervical cancer based on the processed dataset.

IV. Results and Discussions

This study aims to enhance classification performance in the early detection of cervical cancer using the Correlation-Based Feature Selection (CFS) method alongside three classification algorithms: Decision Tree, Naïve Bayes, and k-Nearest Neighbor (k-NN). Following the pre-processing and feature selection processes, the results of this research provide valuable insights into the impact of CFS on the performance of the employed classification algorithms. Below is a discussion of the results from each stage of the model performance evaluation.

A. Results of using Correlation-based Feature Selection (CFS)

The feature selection phase utilizing the CFS method revealed that only 10 out of 22 attributes in the dataset had significant relevance to the target variable (in-situ, early, advanced). The features selected through CFS include those directly related to clinical symptoms and risk factors known to contribute to the development of cervical cancer. Some important selected attributes include:

1) Abdominal lumps
2) Duration of bleeding
3) Bleeding outside the menstrual cycle
4) Color of discharge
5) Fatigue
6) Weight loss
7) Lower abdominal pain
8) Pelvic pain
9) Difficulty urinating
10) Difficulty defecating (specifically for the Decision Tree and k-NN algorithms)

The feature selection using CFS significantly reduced the number of irrelevant or less important attributes, ultimately enhancing the classification model's performance. Before feature selection, the dataset contained 22 attributes, but after selection, this number was reduced to 10. This reduction contributes to decreased computational complexity and minimizes the risk of overfitting commonly associated with high-dimensional datasets.

B. Evaluation of Classification Algorithm Performance

After the dataset was refined through feature selection using CFS, all three classification algorithms were implemented to predict the classification outcomes for cervical cancer. The performance of each algorithm was evaluated using several metrics, including accuracy, sensitivity, specificity, and precision.

1) Decision Tree: The Decision Tree algorithm is highly interpretable, and the results from this study indicate that it provides the highest accuracy compared to the other two algorithms. The maximum accuracy achieved by Decision Tree was 85.89% with all 10 selected attributes. This model also demonstrated a balanced performance in terms of sensitivity (79.3%) and specificity (83.1%), indicating its ability to accurately detect both positive and negative cases.

One of the main strengths of the Decision Tree is its capability to visualize the decision-making process in tree form, facilitating easier interpretation for medical practitioners. By examining the nodes and branches of the tree, doctors can discern which factors most influence the model’s decisions. In this case, attributes such as "abdominal lumps," "bleeding outside the cycle," and "color of discharge" emerged as key features underlying the decision of whether a patient is at risk for cervical cancer. This aligns with previous medical research indicating these symptoms are early warning signs of cervical cancer.

Another advantage of the Decision Tree is its ability to handle diverse data types, whether numerical or categorical, which is common in medical datasets. Furthermore, the process of partitioning the dataset into smaller subsets enables the Decision Tree to effectively manage high-dimensional issues after feature selection is applied.

1) Naïve bayes: Naïve Bayes is an efficient algorithm for handling large datasets and provides rapid predictions. In this dataset, following feature selection, the Naïve Bayes algorithm achieved the highest accuracy of 83.34% with 8 selected attributes. Despite the underlying assumption of independence among features often not being met in real-world applications, this algorithm still managed to deliver good predictive results. However, its sensitivity (72.5%) was somewhat lower compared to other algorithms, indicating that this model occasionally fails to detect some positive cases of cervical cancer.

The low sensitivity suggests that Naïve Bayes tends to have a higher rate of false negatives, which can be dangerous in medical cases like cervical cancer, where patients requiring treatment may go undetected. Nevertheless, the specificity attained (85.9%) is quite high, indicating that this algorithm is proficient in identifying negative cases and avoiding false positives, which is crucial for reducing unnecessary anxiety among patients deemed at low risk.

2) k-Nearest Neighbor (k-NN): The k-Nearest Neighbor (k-NN) algorithm showed significant performance improvement following feature selection using CFS. Before feature selection, k-NN only achieved an accuracy of 46.82%, demonstrating that this algorithm was highly affected by the high dimensionality of the data. However, after feature selection, accuracy increased to 82.32% with 8 selected attributes.

This substantial improvement indicates that k-NN relies heavily on a smaller number of relevant attributes. The k-NN algorithm functions by calculating the distance between new samples and existing samples in the dataset, classifying the new sample based on its nearest neighbors. When too many irrelevant features are included, the distance calculations become less accurate, leading to inaccurate classification results. Therefore, feature selection with CFS allows k-NN to concentrate on the most relevant features, yielding more accurate predictions.

Despite the notable accuracy improvement, k-NN still has a weakness in sensitivity (69.5%), suggesting that it tends to overlook some positive cases of cervical cancer. Additionally, this algorithm is sensitive to data scaling, necessitating proper normalization to ensure all attributes contribute equally to the distance calculations.

C. Comparative Analysis of Classification Algorithms

To provide a clearer picture of the performance of the three tested algorithms in this study, the following table compares the results for accuracy, sensitivity, specificity, and precision:

TABLE II. COMPARISON OF CLASSIFICATION ALGORITHM RESULTS

Algorithm	Accuracy	Sensitivity	Specificity	Precision
Decision Tree	85.89%	79.3%	83.1%	80.5%
Naïve Bayes	83.34%	72.5%	85.9%	74.1%
k-Nearest Neighbor	82.32%	69.5%	84.4%	71.8%

From the Table II, it is evident that the Decision Tree algorithm delivers the best overall performance, with the highest accuracy and relatively good sensitivity, followed by Naïve Bayes and k-NN. Meanwhile, Naïve Bayes excels in specificity, indicating that this algorithm is more accurate in identifying patients who are not at risk of cervical cancer. However, the drawback of Naïve Bayes lies in its sensitivity, which implies a higher risk of missing cases of cervical cancer.

The strength of k-NN lies in its simplicity and effectiveness in classification following feature selection. However, its lower sensitivity suggests that this algorithm tends to miss more positive cases of cervical cancer compared to the others.

D. Implementation in Early Detection of Cervical Cancer

The findings from this study indicate that correlation-based feature selection (CFS) is highly effective in enhancing the performance of classification algorithms in detecting cervical cancer. By reducing the number of irrelevant attributes and focusing solely on the most significant features, classification algorithms can produce more accurate and efficient predictions. This is crucial in medical applications, where diagnostic accuracy is a key factor in determining further actions.

Additionally, the Decision Tree stands out as the most ideal algorithm for use in the early detection of cervical cancer, as it not only provides the highest accuracy but also allows for easy interpretation of the decision-making process. This makes the Decision Tree a valuable tool for medical practitioners to understand the factors contributing to a cervical cancer diagnosis. One of the main benefits of the Decision Tree is its ability to visually represent the classification process, breaking down complex decision-making into a clear and straightforward flowchart. This visual representation helps healthcare professionals trace the sequence of factors leading to a diagnosis, making it easier to explain results to patients and colleagues in an understandable manner.

Another significant advantage of implementing the Decision Tree algorithm is its robustness in handling both categorical and numerical data, which is often the case in medical datasets. This flexibility allows for the integration of a diverse range of patient information, such as demographic factors, test results, and clinical history, into a single cohesive model without requiring extensive data preprocessing. Additionally, Decision Trees are not as sensitive to outliers or missing values as other algorithms, making them highly resilient in real-world medical applications where data quality can vary.

Moreover, Decision Trees facilitate effective feature selection by prioritizing the most important attributes early in

the model, which can lead to more efficient and faster computations. This is particularly beneficial in large-scale healthcare systems or when developing real-time diagnostic applications, as it reduces the processing time and computational resources needed to generate accurate predictions. The interpretability and efficiency of Decision Trees make them ideal for use in telemedicine and remote diagnostic tools, providing a means for timely and accurate assessments even in resource-constrained environments.

Lastly, the transparency of the Decision Tree's decision-making process supports better compliance with healthcare regulations and ethical standards. Because each step of the diagnosis can be easily traced and justified, Decision Tree models can enhance trust between medical professionals and patients and ensure accountability in the diagnostic process. These benefits position Decision Trees as a highly effective and practical solution for implementing early detection systems for cervical cancer, ultimately contributing to improved patient outcomes and more accessible healthcare.

With the results obtained, this research can serve as a foundation for developing information technology-based applications that assist individuals in conducting early screenings for cervical cancer independently. By leveraging proven classification algorithms and effective feature selection methods, such applications can deliver accurate results and help raise public awareness about the importance of early detection of cervical cancer.

V. Conclusion and Future Works

This study demonstrates that the Correlation-Based Feature Selection (CFS) method can significantly enhance classification performance on cervical cancer datasets. The Decision Tree, Naïve Bayes, and k-Nearest Neighbor (k-NN) algorithms all showed substantial improvements in accuracy following feature selection. Among the three algorithms, the Decision Tree achieved the highest accuracy, while k-NN experienced the most significant improvement after feature selection. These results indicate that employing appropriate feature selection techniques can optimize the predictive capabilities of machine learning models, leading to better diagnostic outcomes.

Furthermore, this research highlights the crucial role of feature selection in reducing data dimensionality and eliminating irrelevant or redundant features. By focusing on the most relevant attributes, the CFS method improves the computational efficiency of the models, resulting in faster and more accurate predictions. This optimization is particularly beneficial when dealing with complex datasets, such as those involving medical records or diagnostic tests, where irrelevant data can hinder model performance and lead to misleading results.

Widely, the implementation of these findings has practical applications beyond the academic context, especially in the development of technology-based tools to support the early detection and diagnosis of cervical cancer. By integrating machine learning algorithms enhanced with feature selection methods into digital healthcare platforms, it becomes possible to build robust and reliable diagnostic systems that can be used by healthcare professionals and patients alike. Such systems are particularly valuable in regions with limited access to specialized medical care, where timely diagnosis and treatment can be challenging.

Nonetheless, the use of these advanced diagnostic tools can significantly improve early detection rates, enabling healthcare providers to identify potential cases of cervical cancer at an earlier stage when treatment is more effective and the chances of recovery are higher. Early detection systems can also alleviate the burden on healthcare infrastructure by allowing patients to be triaged more effectively, prioritizing those who need immediate medical attention. Ultimately, this study underscores the potential of machine learning and feature selection techniques to revolutionize cervical cancer screening and contribute to better health outcomes for women worldwide.

Acknowledgment

The authors would like to thank the Ministry of Research and Technology of Indonesia for funding this research under the Research and Development Grant No. 108/E5/PG.02.00.PL/2024.

References

[1] T. L. Ersado, "Cervical Cancer Prevention and Control," in Cervical Cancer - A Global Public Health Treatise, R. Rajkumar, Ed., Rijeka: IntechOpen, 2021. doi: 10.5772/intechopen.99620.

[2] K. Canfell, "Towards the global elimination of cervical cancer," Papillomavirus Res., vol. 8, p. 100170, Dec. 2019, doi: 10.1016/j.pvr.2019.100170.

[3] M. Arbyn et al., "Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis," Lancet Glob. Heal., vol. 8, no. 2, pp. e191–e203, Feb. 2020, doi: 10.1016/S2214-109X(19)30482-6.

[4] M. M. Ali et al., "Machine learning-based statistical analysis for early stage detection of cervical cancer," Comput. Biol. Med., vol. 139, p. 104985, 2021, doi: https://doi.org/10.1016/j.compbiomed.2021.104985.

[5] R. Alsmariy, G. Healy, and H. Abdelhafez, "Predicting Cervical Cancer using Machine Learning Methods," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 7, pp. 173–184, 2020, doi: 10.14569/IJACSA.2020.0110723.

[6] A. AlMohimeed, H. Saleh, S. Mostafa, R. M. A. Saad, and A. S. Talaat, "Cervical Cancer Diagnosis Using Stacked Ensemble Model and Optimized Feature Selection: An Explainable Artificial Intelligence Approach," Computers, vol. 12, no. 10, 2023, doi: 10.3390/computers12100200.

[7] H. Momeni and A. Ebrahimkhanlou, "High-dimensional data analytics in structural health monitoring and non-destructive evaluation: a review paper," Smart Mater. Struct., vol. 31, no. 4, p. 43001, Mar. 2022, doi: 10.1088/1361-665X/ac50f4.

[8] M. Hasan, P. Roy, and A. M. Nitu, "Cervical Cancer Classification using Machine Learning with Feature Importance and Model Explainability," in 2022 4th International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE), 2022, pp. 1–4. doi: 10.1109/ICECTE57896.2022.10114548.

[9] K. Alpan, "Performance Evaluation of Classification Algorithms for Early Detection of Behavior Determinant Based Cervical Cancer," in 2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), IEEE, Oct. 2021, pp. 706–710. doi: 10.1109/ISMSIT52890.2021.9604718.

[10] I. Jain, V. K. Jain, and R. Jain, "Correlation feature selection based improved-Binary Particle Swarm Optimization for gene selection and cancer classification," Appl. Soft Comput., vol. 62, pp. 203–215, Jan. 2018, doi: 10.1016/j.asoc.2017.09.038.

[11] A. H. Gandomi, F. Chen, and L. Abualigah, "Machine Learning Technologies for Big Data Analytics," Electronics, vol. 11, no. 3, p. 421, Jan. 2022, doi: 10.3390/electronics11030421.

[12] A. H. Elmi, A. Abdullahi, and M. A. Bare, "A comparative analysis of cervical cancer diagnosis using machine learning techniques," Indones. J.

Electr. Eng. Comput. Sci., vol. 34, no. 2, pp. 1010–1023, 2024, doi: 10.11591/ijeecs.v34.i2.pp1010-1023.

[13] G. Ou, Y. He, P. Fournier-Viger, and J. Z. Huang, "A Novel Mixed-Attribute Fusion-Based Naive Bayesian Classifier," Appl. Sci., vol. 12, no. 20, p. 10443, Oct. 2022, doi: 10.3390/app122010443.

[14] D. A. Anggoro and N. C. Aziz, "Implementation of K-Nearest Neighbors Algorithm for Predicting Heart Disease Using Python Flask," vol. 62, no. 9, 2021, doi: 10.24996/ijs.2021.62.9.33.

[15] N. S. Rahmi, N. W. S. Wardhani, M. B. Mitakda, R. S. Fauztina, and I. Salsabila, "SMOTE Classification and Random Oversampling Naive Bayes in Imbalanced Data : (Case Study of Early Detection of Cervical Cancer in Indonesia)," in 2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA), 2022, pp. 1–6. doi: 10.1109/ICITDA55840.2022.9971421.

[16] T. Ganguly, P. B. Pati, K. Deepa, T. Singh, and T. Özer, "Machine Learning based Comparative Analysis of Cervical Cancer Risk Classifications Algorithms," in 2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), 2023, pp. 1–7. doi: 10.1109/ACCAI58221.2023.10200617.

[17] B. Nithya and V. Ilango, "Machine Learning Aided Fused Feature Selection based Classification Framework for Diagnosing Cervical Cancer," in 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), 2020, pp. 61–66. doi: 10.1109/ICCMC48092.2020.ICCMC-00011.

[18] S. Tawalbeh, H. Alquran, and M. Alsalatie, "Deep Feature Engineering in Colposcopy Image Recognition: A Comparative Study," Bioengineering, vol. 10, no. 1, 2023, doi: 10.3390/bioengineering10010105.

[19] Q. S. Setiawan, Z. Rustam, and J. Pandelaki, "Comparison of Naive Bayes and Support Vector Machine with Grey Wolf Optimization Feature Selection for Cervical Cancer Data Classification," in 2021 International Conference on Decision Aid Sciences and Application (DASA), 2021, pp. 451–455. doi: 10.1109/DASA53625.2021.9682329.

[20] Z. Ramzan, M. A. Hassan, H. M. S. Asif, and A. Farooq, "A Machine Learning-Based Self-Risk Assessment Technique for Cervical Cancer," Curr. Bioinform., vol. 15, pp. 1–18, 2020, doi: 10.2174/1574893615999200608130538.

[21] N. R. Haddaway et al., "Eight problems with literature reviews and how to fix them," Nat. Ecol. Evol., vol. 4, no. 12, pp. 1582–1589, Oct. 2020, doi: 10.1038/s41559-020-01295-x.

[22] J. J. Tanimu, M. Hamada, M. Hassan, and S. Yusuf Ilu, "A Contemporary Machine Learning Method for Accurate Prediction of Cervical Cancer," SHS Web Conf., vol. 102, p. 04004, May 2021, doi: 10.1051/shsconf/202110204004.

[23] C. H. Bhavani, C. Sarada, A. J. Babu, G. Kumar, and M. Sangeetha, "NGBFA Feature Selection Algorithm-based Hybrid Ensemble Classifier to Predict Cervical Cancer," Int. J. Intell. Syst. Appl. Eng., vol. 12, no. 4, pp. 960–967, 2024, [Online]. Available: https://ijisae.org/index.php/IJISAE/article/view/6318

[24] D. A. Anggoro and W. Supriyanti, "Improving accuracy by applying Z-score normalization in linear regression and polynomial regression model for real estate data," Int. J. Emerg. Trends Eng. Res., vol. 7, no. 11, 2019, doi: 10.30534/ijeter/2019/247112019.

[25] D. A. Anggoro, A. A. T. Marzuki, and W. Supriyanti, "Classification of Solo Batik patterns using deep learning convolutional neural networks algorithm," TELKOMNIKA (Telecommunication Comput. Electron. Control., vol. 22, no. 1, pp. 232–240, 2024, doi: 10.12928/telkomnika.v22i1.24598.

[26] W. Supriyanti and D. A. Anggoro, "Classification of Pandavas Figure in Shadow Puppet Images using Convolutional Neural Networks," Khazanah Inform. J. Ilmu Komput. dan Inform., vol. 7, no. 1, pp. 18–24, 2021, doi: 10.23917/khif.v7i1.12484.