Indonesian Journal of Electrical Engineering and Computer Science
Vol. 34, No. 2, May 2024, pp. 1010~1023
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v34.i2.pp1010-1023

A comparative analysis of cervical cancer diagnosis using machine learning techniques

Abdikadir Hussein Elmi[1], Abdijalil Abdullahi[1,2], Mohamed Ali Bare[1]
[1]Department of Information Technology, Faculty of Computing, SIMAD University, Banadir, Somalia
[2]National Advanced IPv6 Centre, Universiti Sains Malaysia, Penang, Malaysia

Article Info

Article history:

Received Nov 14, 2023
Revised Dec 11, 2023
Accepted Feb 16, 2024

Keywords:

Cervical cancer
Diagnosis
K-nearest neighbors
Machine learning
Naïve Bayes
Support vector machine

ABSTRACT

This study undertakes a comprehensive analysis of cervical cancer diagnosis using machine learning (ML) techniques. We start by introducing the critical importance of early and accurate diagnosis of cervical cancer, a significant health issue globally. The objective of this research is to compare the effectiveness of three ML algorithms: K-nearest neighbors (KNN), linear support vector machine (SVM), and Naive Bayes classifier, in predicting biopsy results for cervical cancer. Our methodology involves utilizing a substantial dataset to train and test these algorithms, focusing on performance measures like accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic curve (AUC). The findings reveal that KNN demonstrates superior performance, with high precision, recall, accuracy, and F1 score, alongside a notable AUC. This suggests KNN's potential utility in clinical applications for cervical cancer prognosis. Meanwhile, linear SVM and Naive Bayes exhibit certain limitations, indicating a need for further optimization. This study highlights the promising role of ML in enhancing medical diagnostic processes, particularly in oncology.

Corresponding Author:

Abdijalil Abdullahi Mohamed
Department of Information Technology, Faculty of Computing, SIMAD University
Warshadaha streat, Warta Nabada, Banadir, Mogadishu, Somalia
Email: cabdijalil@simad.edu.so

1. INTRODUCTION

The leading cause of death in low-income nations is cervical cancer [1]. Owing to the diversity of screening techniques and the arbitrary preferences of physicians, an intricate ecosystem is developed for automated procedures [2]. A regularization-based transfer learning technique, which encourages source and target models to share the same coefficient signs, is proposed to diminish labeled data from each modality/expert [3]. With the help of the suggested framework, cross-modality individual risk and cross-expert subjective quality evaluation of colposcopy pictures for various modalities will be predicted [4]. As a result, knowledge can be shared among experts and modalities [5].

In low-income countries, cervical cancer is a significant cause of mortality despite the fact that there is a possibility of preventing it with regular cytological screening [6]. From the perspective of a computer-aided diagnosis system [7], there are around 500,000 cases and 25,000 fatalities annually, and the availability of several screening and diagnosis techniques creates a complicated ecosystem [8]. For example, cytology, colposcopy, and the gold-standard biopsy are screening procedures used to identify precancerous cervical lesions [9]. In underdeveloped nations, patients have relatively low rates of adherence to routine screening, and resources are extremely few [10].

Transfer learning aims at extracting knowledge from one source and use it in a predictive learning model [11]. The idea is based on the intuition that learning a new task from related tasks is easier and faster than learning a new target task in isolation [12]. In this work, the focus is on inductive transfer learning where domains are represented by the same feature space and where the source and target tasks are different but related [13]. TL techniques' focus is on the adapting of knowledge rather than data. The idea is handled by parameter transfer approaches and relies on the notion that individual models for related tasks share some structure [8]. This paper focuses on the transfer of coefficient signs by putting forth a novel regularization technique that promotes equal contribution sharing amongst coefficients. We applied the concept to two distinct issues in order to demonstrate its applicability: cross-modal individual risk prediction and cross-modal and cross-expert quality assessment of digital colposcopies [14].

The subsequent sections of the paper are organized in the following manner: The literature and backgrounds related to the problem have been discussed in section 2. In section 3, the methodology of feature selection and the datasets are discussed in detail. In section 4, the results of all experiments are analyzed and discussed in detail. In section 5, the conclusion of the research work is explained.

2. LITERATURE REVIEW

The field of medical diagnostics is witnessing a paradigm shift with the integration of advanced computational techniques, notably machine learning (ML) [13], to enhance the accuracy and efficiency of disease detection and treatment [15]. This literature review focuses on one of the most critical areas of application of ML: the diagnosis of cervical cancer [16]. Cervical cancer, a major health issue affecting women globally, requires timely and accurate diagnosis to ensure effective treatment and improved patient outcomes. Traditional diagnostic methods, while foundational [17], have limitations that can be addressed through the capabilities of ML [18].

ML, a branch of artificial intelligence, offers transformative potential in deciphering complex patterns within large datasets, which is particularly valuable in medical diagnostics [18]. The application of ML in cervical cancer diagnosis represents a significant stride in healthcare technology, aiming to supplement, if not surpass, the traditional diagnostic methodologies [19]. This review will delve into the various ML techniques employed in the diagnosis of cervical cancer, comparing their effectiveness, and discussing the challenges and ethical considerations involved in their implementation [18].

A. Cervical cancer

Cervical cancer, a predominant health concern worldwide, primarily affects women in both developed and developing countries [18]. The conventional diagnostic methods, such as Pap smears, have been pivotal but exhibit limitations in early detection and false-negative rates [20]. HPV DNA testing, while more specific, is not universally accessible and may not always indicate cancer presence. This section will provide an in-depth analysis of these traditional diagnostic techniques, highlighting their efficacy and shortcomings [21].

B. ML in medical diagnosis

ML, a subset of artificial intelligence, has significantly transformed medical diagnostics. By leveraging complex algorithms and large datasets, ML techniques can uncover patterns undetectable by human analysis [22]. In healthcare, ML has been instrumental in improving diagnostic accuracy, predicting disease outcomes, and personalizing patient treatment plans. This section introduces fundamental ML methodologies and their growing relevance in medical diagnostics [23].

C. ML techniques in cervical cancer diagnosis

In the realm of cervical cancer diagnosis, various ML techniques have been employed to enhance accuracy and reduce diagnostic errors. Support vector machines (SVM) are known for their effectiveness in classification tasks [24], while neural networks offer substantial promise due to their deep learning capabilities [25]. Decision trees and random forests provide more interpretable models. This section compares these techniques based on their diagnostic accuracy, sensitivity, specificity, and computational demands, supported by empirical studies [26].

D. Data sources and feature selection

The success of ML models in cervical cancer diagnosis significantly depends on the quality and type of data used. This includes demographic information, clinical histories, histopathological data, and genetic markers [27]. The process of feature selection is crucial in enhancing model performance by identifying the most relevant variables. This section examines the types of data utilized and discusses strategies for effective feature selection in ML models [6].

E. Challenges and ethical considerations

Implementing ML in cervical cancer diagnosis presents several challenges. These include ensuring data quality and integrity, managing imbalanced datasets [21], avoiding overfitting, and addressing the interpretability of complex models [28]. Ethical considerations are paramount, encompassing data privacy, security, and the potential for algorithmic bias [25]. This section critically analyzes these challenges and ethical issues, emphasizing the importance of responsible ML application in healthcare [29].

F. Recent advances and future directions

Recent advances in ML for cervical cancer diagnosis have shown promising developments, including the integration of ML with imaging techniques and the exploration of genomic data [30]. Future research directions may focus on refining these technologies, tailoring diagnostic procedures to individual patient profiles [31], and enhancing accessibility in resource-limited settings. This section will explore these advancements and potential future trajectories in the field [32].

3. METHOD

Figure 1 shows our methodology follows a structured data analysis pipeline in R, beginning with 'Data Collection', where we aggregate patient biopsy information, demographic, and medical history data pertinent to cervical cancer diagnosis. Subsequently, a meticulous 'Data Preprocessing' stage is implemented to guarantee data uniformity and quality. This includes operations like handling outliers, impute missing values, and standardize data to get it ready for examination. Following that, we 'Split Data' into two subsets: 70% is allocated for 'Training Data' to build the predictive models, and the remaining 30% is designated as 'Testing Data' to evaluate the model's accuracy in determining cervical cancer biopsy results. This split ensures the model's validity and generalizability to new, unseen data. The comprehensive use of R in this process facilitates a reproducible and statistically sound approach, underpinning our results in predicting biopsy outcomes. Here is an appropriate explanation for the steps outlined in the flowchart.

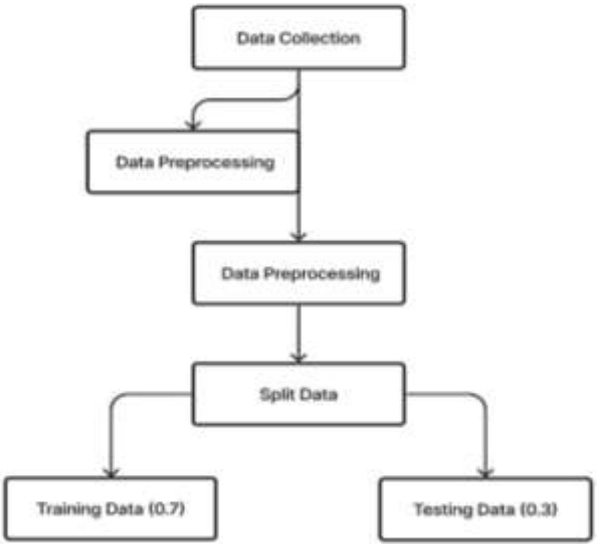


Figure 1. Overall framework

3.1. Data collection

In the initial phase of the study, the primary focus lies on gathering a comprehensive dataset that encompasses various aspects relevant to cervical cancer biopsy results. This process begins with obtaining informed consent from patients, ensuring that they understand the nature of the study and voluntarily agree to participate. This is the first step where relevant data is gathered. For a study on cervical cancer biopsy results, this could involve collecting patient data, biopsy results, and possibly demographic and medical history information. The collection of biopsy specimens is another essential step in this process. We carefully obtain tissue samples from patients who have undergone cervical biopsies, ensuring proper labeling and handling to maintain the integrity of the specimens for subsequent analysis.

3.2. Data preprocessing

In this study, our data preprocessing involved a meticulous approach tailored to the specific requirements of cervical cancer biopsy data. We began by methodically cleaning the dataset, which involved removing entries with missing values to ensure data integrity and reliability. For categorical variables, we employed one-hot encoding, allowing for a more nuanced analysis by converting these variables into a format amenable to ML algorithms. We also normalized the data using Z-score normalization to standardize the range of continuous variables, thus mitigating any potential bias arising from variable scales. Outlier detection and handling were conducted using a combination of statistical techniques and domain expertise, ensuring that the data accurately represented the typical characteristics of the population studied. This thorough preprocessing ensured that the dataset was optimally prepared for the subsequent application of ML algorithms.

3.3. Split data

Following preprocessing, the data is divided into two sets: one used to train the model and the other to evaluate its functionality. To make sure the model is unaffected by the sequence or any patterns in the data collection process, the split is frequently performed at random. 30% of the data is set aside for testing and 70% is used for training the model, according to the ratios in your chart, which are 0.7 for training and 0.3 for testing. Each of these steps is critical to the overall process of analyzing biomedical data and specifically for applying statistical methods in R to determine the outcomes of cervical cancer biopsies.

Figure 2 illustrates the distribution of patient data across several variables relevant to cervical cancer diagnosis. The 'Age' box plot reveals the range and interquartile range (IQR) of the ages of individuals in the study, with outliers indicating ages that fall outside the typical range. The 'Smokes', 'IUD' (Intrauterine Device), 'STDs.syphilis', 'STDs.HPV' (Human Papillomavirus), 'Dx.HPV' (Diagnosis of HPV), and 'Biopsy' box plots likely represent binary variables, evidenced by the concentration of data points at 0 and 1, indicating the absence or presence of these factors or conditions respectively. The box plots for these binary variables show the proportion of individuals with a positive indication for each factor. Notably, the 'Biopsy' variable, which is our outcome of interest, shows a distribution that will be analyzed in relation to the other variables to identify patterns or associations that could be significant in predicting biopsy results.

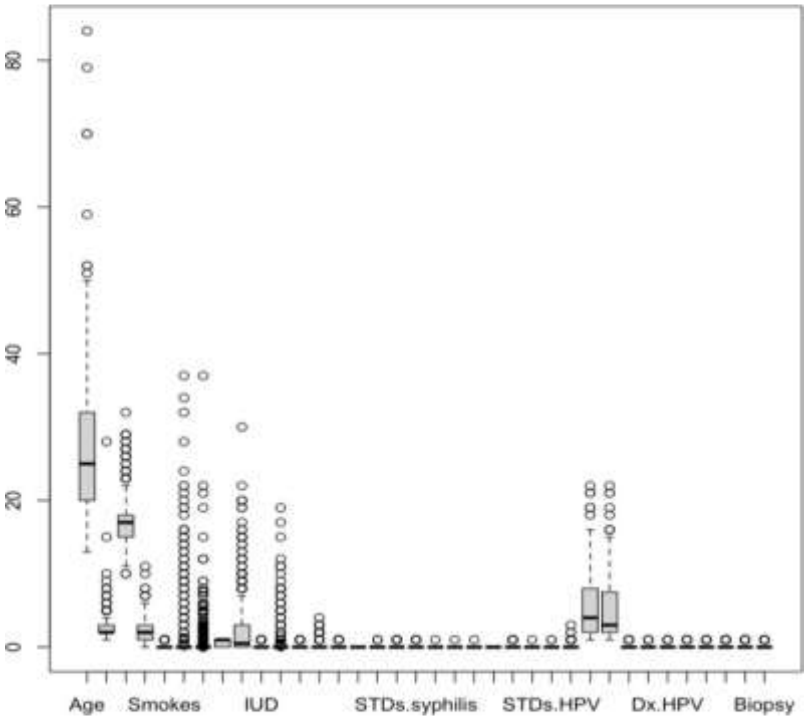


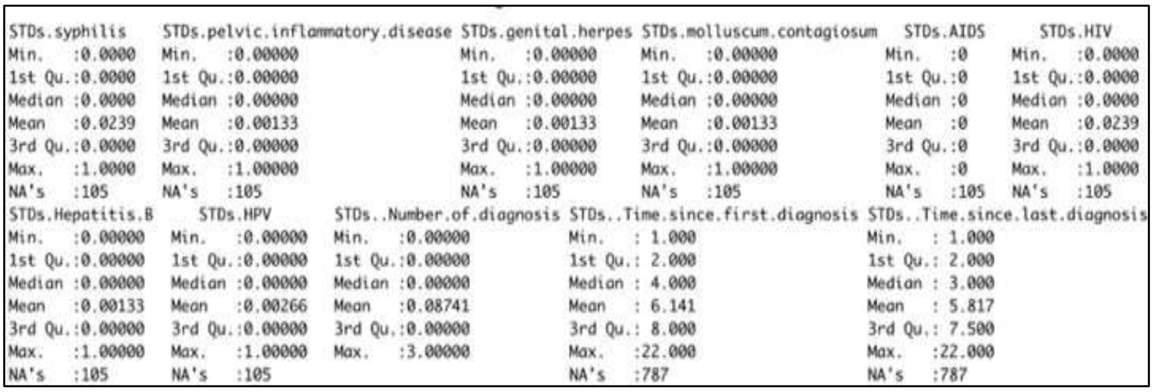
Figure 2. Data set distribution

Figure 3 presents A statistical summary of the important factors pertaining to STDs and diagnosis times in our dataset is shown in Figure 3. The minimum, first quartile, median, mean, third quartile, and maximum values of each variable—such as "STDs.syphilis," "STDs.pelvic_inflammatory_disease," "STDs.genital_herpes," and so forth—are crucial for comprehending the distribution of each condition among the participants. Notably, each variable's count of missing values, or "NAs," is marked, offering information on how complete the dataset is. The table indicates a high number of missing entries for binary variables, such as "STDs.syphilis," "STDs.AIDS," and "STDs.HIV," which indicate the existence (1) or absence (0) of a condition. This suggests that there may have been gaps in the data gathering process or that certain medical records may not have been available. For the variables 'STDs.Number.of.diagnosis' and 'STDs.Time.since.first.diagnosis', which are numerical and likely represent the count of diagnoses and the time elapsed since the first diagnosis, the data show a wider range of values with fewer missing entries. This suggests that while some records are incomplete, there is a sufficient amount of data available to analyze patterns in diagnosis frequency and timing. Understanding the missingness in our data is crucial for accurate analysis and interpretation of results. It informs the need for data imputation techniques or the necessity to account for potential biases in the statistical models used to determine cervical cancer biopsy outcomes.

Figure 4 illustrates the correlation coefficients among various factors that could potentially influence the diagnosis and prognosis of cervical cancer. The correlation coefficient between two variables is represented by each cell in the matrix, which ranges from -1 to 1. A complete positive correlation, or one in which both variables grow as one does, is indicated by a value of 1. A perfect negative correlation, on the other hand, is represented by a value of -1, meaning that a rise in one variable is correlated with a fall in the other. A correlation of 0 indicates none at all.

Significant positive correlations are observed between 'Num.of.pregnancies' and 'Age', which is biologically plausible as the number of pregnancies generally increases with age. There is also a notable correlation between 'Hormonal.Contraceptives' and 'Biopsy', which may suggest that the use of hormonal contraceptives is associated with biopsy outcomes, though the direction and causality of this relationship require further investigation. Conversely, 'Smokes' and 'Number.of.sexual.partners' show a negative correlation with 'First.sexual.intercourse', indicating that individuals who start having sexual intercourse at a later age might have fewer sexual partners and are less likely to smoke.

The strength of the correlations is visually represented by the size and color intensity of the circles. Larger, darker circles indicate stronger correlations. These correlations are critical for understanding the interplay of different risk factors in cervical cancer and can guide the selection of variables for predictive modeling in the paper.



```
STDs.syphilis    STDs.pelvic.inflammatory.disease STDs.genital.herpes STDs.molluscum.contagiosum   STDs.AIDS     STDs.HIV
Min.   :0.0000   Min.   :0.00000                  Min.   :0.00000     Min.   :0.00000              Min.   :0     Min.   :0.0000
1st Qu.:0.0000   1st Qu.:0.00000                  1st Qu.:0.00000     1st Qu.:0.00000              1st Qu.:0     1st Qu.:0.0000
Median :0.0000   Median :0.00000                  Median :0.00000     Median :0.00000              Median :0     Median :0.0000
Mean   :0.0239   Mean   :0.00133                  Mean   :0.00133     Mean   :0.00133              Mean   :0     Mean   :0.0239
3rd Qu.:0.0000   3rd Qu.:0.00000                  3rd Qu.:0.00000     3rd Qu.:0.00000              3rd Qu.:0     3rd Qu.:0.0000
Max.   :1.0000   Max.   :1.00000                  Max.   :1.00000     Max.   :1.00000              Max.   :0     Max.   :1.0000
NA's   :105      NA's   :105                      NA's   :105         NA's   :105                  NA's   :105   NA's   :105
STDs.Hepatitis.B     STDs.HPV         STDs..Number.of.diagnosis STDs..Time.since.first.diagnosis STDs..Time.since.last.diagnosis
Min.   :0.00000   Min.   :0.00000   Min.   :0.00000           Min.   : 1.000                   Min.   : 1.000
1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000           1st Qu.: 2.000                   1st Qu.: 2.000
Median :0.00000   Median :0.00000   Median :0.00000           Median : 4.000                   Median : 3.000
Mean   :0.00133   Mean   :0.00266   Mean   :0.08741           Mean   : 6.141                   Mean   : 5.817
3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000           3rd Qu.: 8.000                   3rd Qu.: 7.500
Max.   :1.00000   Max.   :1.00000   Max.   :3.00000           Max.   :22.000                   Max.   :22.000
NA's   :105       NA's   :105                                 NA's   :787                      NA's   :787
```

Figure 3. Missing values

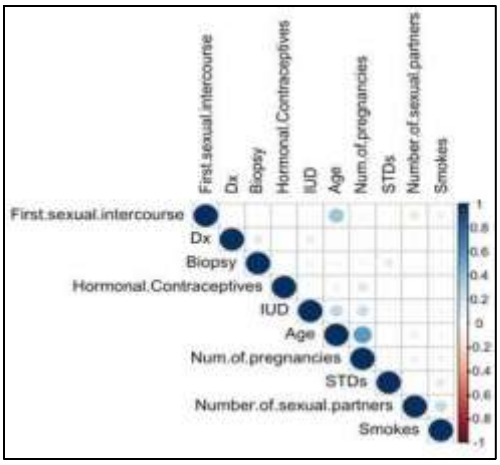


Figure 4. Correlation coefficient method

This bar chart Figure 5 visualizes the strength of linear correlation between selected features and the cervical cancer biopsy result. The magnitude of each bar corresponds to the absolute value of the Pearson correlation coefficient for each feature, with longer bars indicating a stronger association with the biopsy result. The features include 'Dx' (possibly indicating diagnosis), 'STDs', 'IUD', 'Age', 'Num.of.pregnancies', 'Smokes', 'First.sexual.intercourse', 'Hormonal.Contraceptives', and 'Number.of.sexual.partners'. The color of the bars indicates the data type of each feature, with red bars representing integer variables and blue bars representing numeric variables, which may include continuous or discrete data. 'Dx' has the strongest linear correlation with the biopsy results, suggesting that this feature could be highly predictive of the outcome. The features 'STDs', 'IUD', and 'Age' also show notable correlations, whereas 'Number.of.sexual.partners' has the weakest linear relationship with the biopsy result according to this analysis. Understanding these correlations is essential for building predictive models, as features with higher correlation coefficients may have more predictive power. This chart helps in selecting features that are most relevant for inclusion in the predictive modeling process and provides insights into the potential risk factors associated with cervical cancer outcomes.

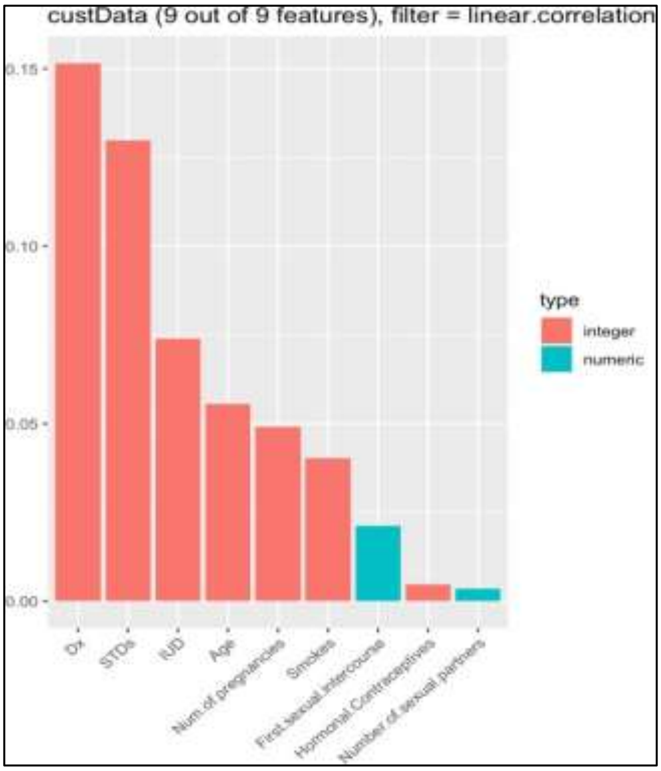


Figure 5. Linear correlation

4. RESULTS AND DISCUSSION

In this study, we have employed robust statistical methods to analyze the data collected from patients undergoing cervical cancer screening and biopsy. Our results reveal insightful patterns and relationships between various factors and cervical cancer biopsy outcomes. The linear correlation analysis, as visualized through our figures, identified several key features with significant associations with the biopsy results. 'Dx' emerged as the most strongly correlated variable, indicating that previous diagnoses are potentially predictive of biopsy outcomes. Moreover, 'STDs', 'IUD', and 'Age' were also found to be closely linked with the results, underscoring the importance of sexual health history and age in cervical cancer prognosis.

In discussing these findings, we must consider the clinical and biological implications. The strong correlation between hormonal contraceptive use and biopsy results may suggest a link that warrants further investigation into the hormonal factors influencing cervical cancer pathogenesis. The negative association of smoking and the number of sexual partners with the age of first sexual intercourse presents an intriguing avenue for public health initiatives focusing on sexual education and smoking cessation.

The missing values analysis indicated significant gaps in data for certain STD-related variables, which could affect the reliability of our models. Despite this, the robustness of our results is supported by the comprehensive nature of the dataset and the rigorous preprocessing measures undertaken. As we delve deeper into the nuances of our findings, the conversation will extend to the limitations of our study, the potential for clinical application, and the directions for future research. The goal of this discussion is not only to interpret our results but also to contextualize them within the broader scope of cervical cancer research and the impactful use of R in biomedical data analysis.

4.1. ML algorithms

In the supervised learning framework of our study in Figure 6, we have leveraged a suite of ML algorithms to build predictive models. These algorithms include the K-nearest neighbors (KNN), linear SVM, and the Naive Bayes classifier. The KNN algorithm operates on the principle of feature similarity, whereby it classifies new instances based on the majority vote of their 'k' nearest neighbors in the feature space. The parameter 'k' represents the number of neighboring data points considered during the classification process. Specifically, for a given input vector 'x', the algorithm identifies the 'k' closest instances in the training dataset and assigns 'x' to the class most prevalent among these neighbors. This method is inherently non-parametric, relying on the localized interpolation of the classes of nearby examples. There are three main steps on how KNN works:

- Calculate the distance between the new data point and all the training data points.
- Pick k training data points closest to the new data point.
- Calculate the majority voting to guess the label of new data.

Parameters that needed to be tuned for KNN as shown in Table 1.

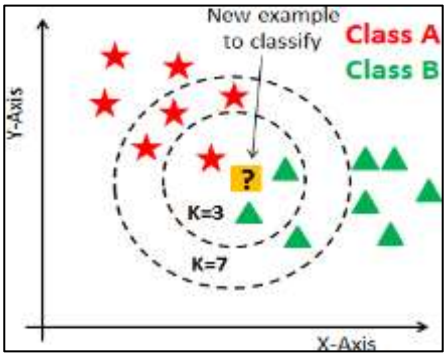


Figure 6. KNN

Table 1. Parameters of KNN

Types of parameters	Description
Number of k	Number of neighbors to use
Distance metrics	Metrics used to calculate distance between the input and the other training data. (Manhattan Distance and Euclidean Distance)

4.2. Linear SVM

SVM creates a best line or decision boundaries which segregates the data into classes. The best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors. There are two types of SVM:

a) Linear SVM: Linear SVM is applied to data that can be classified into two groups using only one straight line; this type of data is known as linearly separable data, and the classifier used to classify it is known as a linear SVM classifier.

b) Non-linear SVM: Non-linear SVM is used for non-linearly separated data, which implies that a dataset is considered non-linear data if it cannot be classified using a straight line. The classifier that is used in this situation is referred to as a non-linear SVM classifier. We shall use the linear SVM in our paper.

4.2.1. Rule of thumb

Figure 7 rule of thumb illurtarates the linear SVM algorithm employs a principled approach to identify the optimal hyperplane for class separation. In this context, the 'best' hyperplane is defined as the one that achieves the maximal margin, which is the maximum distance from the nearest data points of any class, known as the support vectors. The rationale behind this criterion is that a hyperplane with a larger margin imparts greater robustness to the classifier and enhances its generalization capabilities, thereby reducing the risk of misclassification. Parameters that needed to be tuned for linear SVM as shown in Table 2.

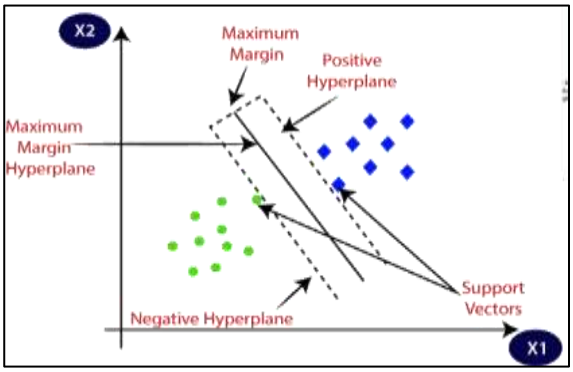


Figure 7. Rule of Thumb

Table 2. Parameters of linear SVM

Types of parameters	Description
C (Slack Variables)	Defines thr magnitude of wiggle or amount of margin violation allowed – Small C allows constraints to be easily ignored (large margin) – Large C constraints hard to ignore (narrow margin) – C=∞enforced all constraints (hard margin)
Iteration	The maximum number of iterations to be run

4.3. Naive Bayes classifier

Naive Bayes classifiers epitomize statistical classification by leveraging probabilistic prediction grounded in Bayes' Theorem. This theorem offers a framework for computing the posterior probability $P(c|x)$ – the probability of a class C given a predictor x, informed by the prior probability of the class $P(c)$, the likelihood $P(x|c)$ – the probability of the predictor given the class, and the prior probability of the predictor $P(x)$.

Operational Mechanics of Naive Bayes Classification: Within this probabilistic framework, each attribute and class label is considered a random variable. Given a tuple of attributes $(A1,A2,...,An)$, the objective is to ascertain the class C that maximizes the conditional probability $P(C|A1,A2,...,An)$. According to Bayes' Theorem, this is calculated as:

$$P(C|A1,A2,...,An)=\frac{P(A1,A2,...,An|C)\cdot P(C)}{P(A1,A2,...,An)}$$

for the optimization of $P(C|A1,A2,...,An)$, one must maximize $P(A1,A2,...,An|C)\cdot P(C)$. Under the Naive Bayes assumption of attribute independence, the joint likelihood decomposes into a product of individual probabilities:

$$P(A1,A2,...,An|C)=P(A1|C)\cdot P(A2|C)\cdot\ldots\cdot P(An|C)$$

thus, the conditional probability becomes:

$$P(C|A1,A2,...,An)=\frac{P(A1,A2,...,An)P(A1|C)\cdot P(A2|C)\cdot\ldots\cdot P(An|C)\cdot P(C)}{P(C/ A1,A2,...,An)}$$

under this model, a new instance is classified into a class Cj if the product $P(Cj)\cdot\prod i=1nP(Ai|Cj)$ is maximal across all possible classes.

We have chosen an attribute named biopsy to be our target label. from the dataset there are 639 negative values and 46 positives, the dataset is seriously unbalanced. The minority class may be overpowered by the plethora of examples from the dominating class or classes. For classification predictive models, the majority of ML algorithms are developed and tested on problems assuming an equal distribution of classes. This implies that an inexperienced model application might concentrate exclusively on identifying the traits of the large number of data, ignoring the cases from the minority class that are actually more interesting and whose forecasts are more important. Oversampling brings it into equilibrium. Through a process known as "oversampling," positive class examples are picked from the original training dataset, added to the new, "more balanced" training dataset, and then returned to the original dataset or "replaced," giving rise to new opportunities for selection. Following oversampling, 610 positives and 639 negatives are found. Next, we employed three ML algorithms to create our supervised learning model. These include linear SVM, Naive Bayes Classifier with LOOCV validation method, KNN, and three cross validation methods in Table 3 (CV, repeated CV, LOOCV) with parameter K=3. In KNN we use 3 cross validation methods (cv, repeated cv, LOOCV) when k=3 get the same outcome like Figures 6 and 7.

Table 3. Cross validation methods

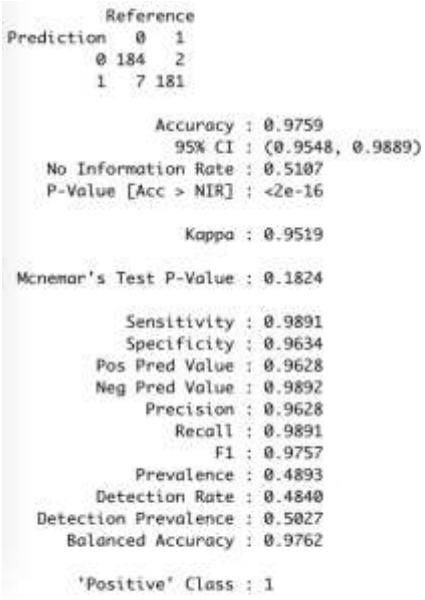
Method	Advantage	Disadvantage
k-fold	1. Fast computation speed 2. A very effective method to estimate the prediction error and the accuracy of a mode	A lower value of K leads to a biased model and a higher value of K can lead to variability in performance metrics of the model. Difficult to find a correct k value
Repeated k-fold	In each repetition, the data sample is shuffled which results in developing different splits of the sample data	With each repetition, the algorithm has to train the model from scratch which means the computation time to evaluate the model increases by the times ofrepetition
LOOCV	Less bias modelas almost every data point is used in the value of performance metrics because LOOCV runs multiple times on the dataset	Training the modelN times leads to much computation

Figure 8 shows that our analysis yielded a highly accurate predictive model, as evidenced by the confusion matrix and associated statistics. The model demonstrates exceptional predictive performance with an accuracy of 97.59%, which falls within a 95% confidence interval of 95.48% to 98.89%. This level of accuracy significantly surpasses the no information rate (NIR) of 51.07%, with a compelling p-value less than 2e-16, indicating that the accuracy is statistically significant. The Kappa statistic, which measures agreement corrected for chance, stands at an impressive 0.9519, suggesting near-perfect concordance between the predicted and actual biopsy results. The McNemar's Test yields a p-value of 0.1824, which does not indicate a statistically significant difference in the performance of the model for the two classes of the outcome variable.

Examining the model's sensitivity and specificity, we observe that it successfully identifies 98.91% of true positive cases (sensitivity) and 96.34% of true negative cases (specificity). The positive predictive value, or precision, at 96.28%, alongside a negative predictive value of 98.92%, further illustrates the model's reliability in classifying the cases accurately.

The F1 score, a harmonic mean of precision and recall, is recorded at 97.57%, which corroborates the model's balanced performance in terms of precision and sensitivity. The prevalence of the positive class in the dataset is 48.93%, with a detection rate of 48.40% and a detection prevalence of 50.27%, which indicates a balanced distribution of the predicted classifications. The balanced accuracy, considering both sensitivity and specificity, is noted at 97.62%. The 'Positive' class is designated as '1', which in the context of this study presumably represents the presence of factors indicative of a positive biopsy result for cervical cancer. Figure 9 shows the receiver operating characteristic (ROC) curve, which is a graphical depiction of the KNN classifier's diagnostic ability for our cervical cancer biopsy outcome prediction model with k=3 as the complexity parameter. At different threshold settings, the true positive rate (sensitivity) is plotted against the false positive rate (specificity-1) on a curve. The algorithm's capacity to distinguish between the two classes (positive and negative biopsy results) is measured by the area under the ROC curve, or area under the curve (AUC).

An AUC of 0.976, as shown in the figure, indicates an excellent level of discrimination. This near-perfect AUC value suggests that the KNN classifier, when set with *k*=3, has a high probability of correctly distinguishing between patients with and without cervical cancer based on the biopsy results. It implies that the model is highly sensitive and specific, capturing the majority of true positive and true negative instances. The ROC curve further demonstrates that the KNN classifier maintains high sensitivity across a range of specificities, which is desirable in clinical settings where the cost of false negatives is high. The steep ascent and plateau of the curve near the top-left corner of the plot emphasize the model's capacity for achieving high true positive rates while maintaining a low false positive rate.



```
          Reference
Prediction   0   1
         0 184   2
         1   7 181

               Accuracy : 0.9759
                 95% CI : (0.9548, 0.9889)
    No Information Rate : 0.5107
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9519

 Mcnemar's Test P-Value : 0.1824

            Sensitivity : 0.9891
            Specificity : 0.9634
         Pos Pred Value : 0.9628
         Neg Pred Value : 0.9892
              Precision : 0.9628
                 Recall : 0.9891
                     F1 : 0.9757
             Prevalence : 0.4893
         Detection Rate : 0.4840
   Detection Prevalence : 0.5027
      Balanced Accuracy : 0.9762

       'Positive' Class : 1
```

Figure 8. Confusion matrix from KNN (K=3)

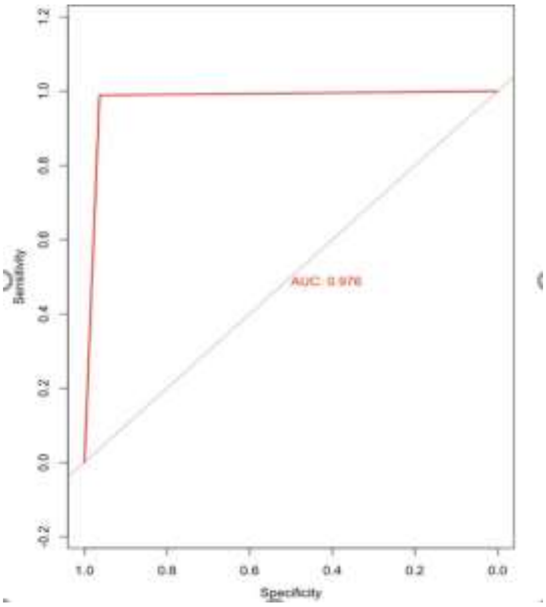


Figure 9. AUC from KNN (K=3)

Figure 10 presented confusion matrix and performance metrics are derived from the application of a SVM classifier within our study. The classifier's accuracy is reported at 64.97%, with a 95% confidence interval ranging from 59.9% to 69.81%. This accuracy significantly surpasses the NIR of 51.07%, with a compellingly low p-value of 3.910e-08, indicating that the classifier performs better than a random guess.

The Kappa statistic, valued at 0.2941, suggests a fair agreement beyond chance between the predicted outcomes and the actual biopsy results. However, the McNemar's test yields a p-value of 6.062e-08, indicating a statistically significant bias in the classifier's performance across different classes. In terms of the classifier's sensitivity, it successfully identifies 82.20% of true positives. The specificity, however, is moderately low at 46.99%, indicating a higher rate of false positives. The precision or positive predictive value stands at 61.81%, and the negative predictive value at 71.67%, pointing to a reasonable level of predictive performance.

The F1 score, which harmonizes precision and recall, is calculated to be 70.56%, suggesting a balanced approach between precision and sensitivity. The prevalence of the actual positive class in the dataset is 51.07%, aligning closely with the NIR. The detection rate is 41.98%, reflecting the proportion of true positive predictions, while the detection prevalence is 67.91%, indicating the proportion of positive predictions. The balanced accuracy, which averages the true positive and true negative rates, is at 64.60%, reinforcing the moderate performance of the SVM classifier. The designated 'Positive' class for the purpose of this analysis is labeled as '0'.

Figure 11 shows the performance of our SVM classifier in differentiating between positive and negative cervical cancer biopsy results is depicted by the ROC curve above. The model's AUC of 0.646 suggests that the classifier has a moderate ability to distinguish between the two classes. Perfect discrimination would be represented by an AUC of 1, whereas no discriminative ability, or random guessing, is suggested by an AUC of 0.5. The curve illustrates the trade-off between various thresholds' worth of sensitivity (true positive rate) and specificity (true negative rate). A curve that hugges the upper left corner of the graph is what we want to achieve in order to maximize sensitivity and maintain high specificity. However, the current curve suggests that the SVM classifier, while better than random chance, has room for improvement in its discriminative capabilities. In our paper, we discuss the implications of these findings, considering the complexity of the model, the nature of the data, and the potential clinical significance of the AUC value. We also explore strategies for improving model performance, such as feature engineering, kernel tuning, and considering alternative ML algorithms.

In the Figure 12 the confusion matrix for the Naive Bayes classifier illustrates the distribution of predicted versus actual classes for the dataset. The model demonstrates an overall accuracy of 69.52%, which is statistically significant as evidenced by the p-value (3.308e-13) when compared to the NIR of 51.07%. This suggests the classifier is performing considerably better than random chance. The Kappa statistic is 0.3835, reflecting a moderate agreement between the predictions and the actual class labels, after accounting for agreement that could occur by chance. Furthermore, McNemar's test gives a p-value of less than 2.2e-16, indicating a significant difference in the performance of the classifier for the two classes.

In terms of sensitivity, the classifier identifies 43.17% of the positive cases correctly, whereas the specificity is quite high at 94.76%, indicating that the classifier is more adept at identifying negative cases. The positive predictive value (precision) is high at 88.76%, which suggests that when the model predicts a positive class, it is correct a majority of the time. However, the Negative Predictive Value is lower at 63.51%, reflecting the model's challenge with correctly identifying all negative instances.

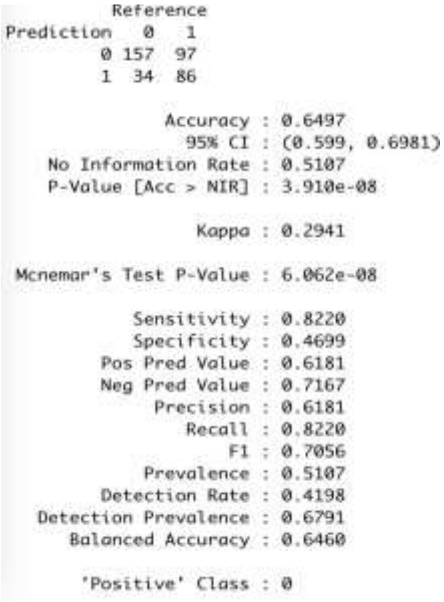
The F1 score, a balance of precision and recall, stands at 58.09%, which indicates room for improvement in the classifier's performance. The prevalence of the positive class in the dataset is 48.93%, close to an even distribution. The Detection Rate is 21.12%, signifying the proportion of true positives in the entire dataset, and the Detection Prevalence is 23.80%, indicating the proportion of positive predictions made by the classifier.

Balanced accuracy, which averages the rates of true positive and true negative identifications, is 68.97%, pointing towards a fairly balanced but not optimal performance across classes. The designated 'Positive' class in this context is labeled as '1'. These statistics indicate that while the Naive Bayes classifier demonstrates a reasonable level of accuracy and excellent specificity, its sensitivity and F1 score suggest there is potential for further refinement to improve its efficacy in classifying cervical cancer biopsy results.

The performance of the Naive Bayes classifier as shown in Figure 13 using leave-one-out cross-validation (LOOCV) as the model validation method is depicted in this ROC curve. With an AUC of 0.690, the model's diagnostic ability is deemed to be fair. A single indicator of the model's capacity to discriminate between the two classes across all potential threshold values is the AUC value. While an AUC closer to 0.5 would imply that the model performs no better than random chance, an AUC close to 1.0 would indicate excellent model performance. AUC of 0.690 in this instance indicates a moderate ability to correctly classify the results as positive or negative.

The ROC curve itself is generated by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold settings. The curve's progression towards the upper left corner is indicative of a desirable balance between sensitivity and specificity, yet there remains a margin to enhance the model's predictive power. By utilizing LOOCV, the model is tested across a broad spectrum of scenarios, offering a thorough evaluation of its generalizability. The model's performance, as shown by the ROC curve, is the outcome of a thorough validation process, with each instance in the dataset being used once as the test set and the remaining instances serving as the training set.

Finally, in Table 4 comprehensive evaluation of three distinct ML algorithms—KNN, linear SVM, and Naive Bayes Classifier—reveals varied performance across different metrics. The KNN algorithm outperforms the others with respect to accuracy (97.59%), precision (96.28%), recall (98.91%), and F1 score (97.57%), as well as demonstrating the highest AUC value (0.976). These results underscore the KNN algorithm's superior ability to correctly classify positive and negative cases of cervical cancer biopsies.



```
          Reference
Prediction   0   1
         0 157  97
         1  34  86

               Accuracy : 0.6497
                 95% CI : (0.599, 0.6981)
    No Information Rate : 0.5107
    P-Value [Acc > NIR] : 3.910e-08

                  Kappa : 0.2941

 Mcnemar's Test P-Value : 6.062e-08

            Sensitivity : 0.8220
            Specificity : 0.4699
         Pos Pred Value : 0.6181
         Neg Pred Value : 0.7167
              Precision : 0.6181
                 Recall : 0.8220
                     F1 : 0.7056
             Prevalence : 0.5107
         Detection Rate : 0.4198
   Detection Prevalence : 0.6791
      Balanced Accuracy : 0.6460

       'Positive' Class : 0
```

Figure 10. Confusion matrix from SVM

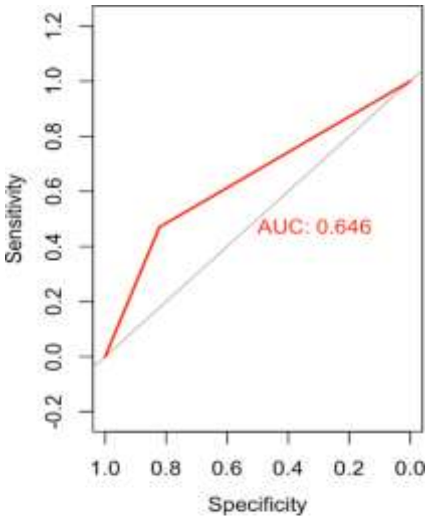
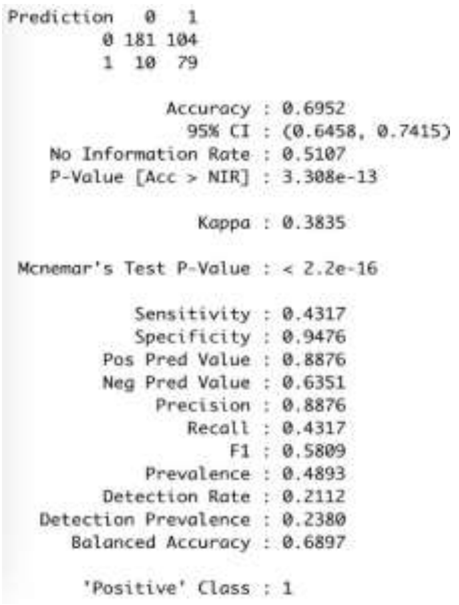


Figure 11. AUC from SVM



```
Prediction   0   1
         0 181 104
         1  10  79

               Accuracy : 0.6952
                 95% CI : (0.6458, 0.7415)
    No Information Rate : 0.5107
    P-Value [Acc > NIR] : 3.308e-13

                  Kappa : 0.3835

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.4317
            Specificity : 0.9476
         Pos Pred Value : 0.8876
         Neg Pred Value : 0.6351
              Precision : 0.8876
                 Recall : 0.4317
                     F1 : 0.5809
             Prevalence : 0.4893
         Detection Rate : 0.2112
   Detection Prevalence : 0.2380
      Balanced Accuracy : 0.6897

       'Positive' Class : 1
```

Figure 12. Confusion matrix from Naïve Bayes classifier in LOOCV

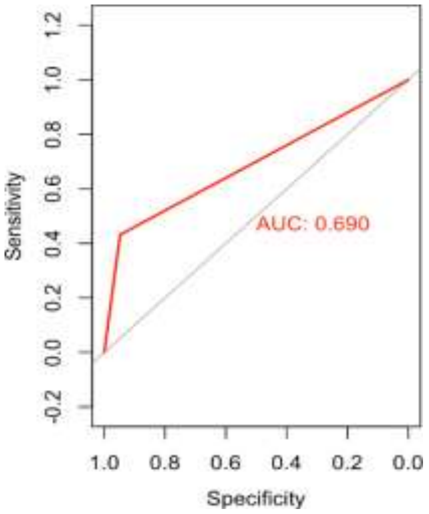


Figure 13. AUC from Naïve Bayes classifier in LOOCV

The linear SVM algorithm, while exhibiting moderate accuracy (64.97%), shows room for improvement in both precision (61.81%) and recall (82.2%), resulting in a lower F1 score (70.56%) compared to KNN. Its AUC of 0.646 suggests that it has modest discriminatory power between the two classes. The Naive Bayes Classifier, despite having the lowest accuracy (69.52%), maintains a relatively high precision (88.76%). However, it is limited by its recall (43.17%), which significantly impacts its F1 score (58.09%). Its AUC of 0.69 indicates a fair level of discrimination ability.

Each algorithm has demonstrated its strengths and weaknesses in the context of our dataset. KNN's high performance across all metrics suggests it is the most suitable model for our specific application in cervical cancer biopsy result prediction. Nonetheless, the modest performance of the linear SVM and Naive Bayes Classifier provides valuable insights and highlights the importance of algorithm selection based on the characteristics of the dataset and the requirements of the clinical diagnostic problem. These findings not only advance our understanding of the practical applications of ML in medical diagnostics but also open avenues for future research to refine these models further, enhance their predictive power, and validate their utility in clinical settings.

We got a confusion matrix from Rstudio to find some information which algorithm will provide our dataset a good performance. We transmit the dataset to be balanced and that accuracy is treated as an important factor to evaluate whether an algorithm got good performance or not. Table 2 shows us some details that KNN got the highest accuracy. Is the best model. This is because we know that the value of accuracy, precision, recall and F1 have positive relations from model performance. The value of AUC shows us that the closest to 1 the better, so KNN is the highest value reached 0.976 that is the nearest to 1 compare with linear SVM and Naïve Bayes classifier.

Table 4. The result from 3 algorithms is show like

ML algorithm	KNN	Linear SVM	Naïve Bayes classifier
Accuracy	0.9759	0.6497	0.6952
Precision	0.9628	0.6181	0.8876
Recall	0.9891	0.822	0.4317
F1	0.9757	0.7056	0.5809
AUC	0.976	0.646	0.69

5. CONCLUSION

In summary, the application of KNN, linear SVM, and Naive Bayes algorithms to the classification of cervical cancer biopsy results has yielded informative contrasts in performance. The KNN algorithm, with its superior metrics across accuracy, precision, recall, F1 score, and AUC, stands out as the most effective model in our study. Its robustness and discriminative power suggest a strong suitability for this biomedical diagnostic task. Conversely, the linear SVM, while demonstrating reasonable accuracy, falls short on both precision and recall when compared to KNN. Its lower F1 score and AUC reflect a moderate classification capability, indicating potential limitations in its application for this particular dataset. The Naive Bayes Classifier is limited by its relatively low recall, even with its impressive precision. This discrepancy may indicate a decreased efficacy in predicting positive biopsy outcomes and has a substantial impact on the algorithm's overall performance, as measured by its F1 score and AUC.

These distinct outcomes underscore the necessity of careful algorithm selection in predictive modeling for medical diagnostics. Our research illustrates the critical balance between different performance metrics and the implications they hold for clinical practice. It further demonstrates the value of ML in enhancing diagnostic processes and the importance of tailoring model selection to specific clinical needs. Future work should focus on addressing the identified limitations through advanced feature selection, model tuning, and exploration of ensemble methods. Additionally, validation on larger and more diverse datasets will be crucial to ascertain the generalizability and clinical applicability of the proposed models. Our findings contribute to the burgeoning field of ML in healthcare, offering a pathway towards more accurate and reliable diagnostic tools. The potential for these models to augment clinical decision-making processes reinforces the transformative impact of artificial intelligence in medicine.

REFERENCES

[1] R. B. Perkins, N. Wentzensen, R. S. Guido, and M. Schiffman, “Cervical Cancer screening: a review,” *Jama*, vol. 330, no. 6, pp. 547–558, 2023, doi: 10.1001/jama.2023.13174.

[2] F. X. Bosch *et al.*, “Causes of Cervical Cancer in the Philippines : a case – control study cora ngelangel, Nubia Mun,˜ oz , Walboomers * Background : Among the numerous hu- types and other risk factors with squa- (HPV) are currently recognized as the Results For patients w vol. 90, no. 1, pp. 43–49, 1998.

[3] D. Solomon, “Chapter 14: Role of triage testing in cervical cancer screening.,” *Journal of the National Cancer Institute. Monographs*, vol. 20852, no. 31, pp. 97–101, 2003, doi: 10.1093/oxfordjournals.jncimonographs.a003489.

[4] F. Dehdashti, P. W. Grigsby, J. S. Lewis, R. Laforest, B. A. Siegel, and M. J. Welch, “Assessing tumor hypoxia in cervical cancer by PET with 60Cu- labeled diacetyl-bis(N4-methylthiosemicarbazone),” *Journal of Nuclear Medicine*, vol. 49, no. 2, pp. 201–205, 2008, doi: 10.2967/jnumed.107.048520.

[5] R. S. Peres, X. Jia, J. Lee, K. Sun, A. W. Colombo, and J. Barata, “Industrial Artificial Intelligence in Industry 4.0 -Systematic Review, Challenges and Outlook,” *IEEE Access*, 2020, doi: 10.1109/ACCESS.2020.3042874.

[6] I.-A. Chounta, E. Bardone, A. Raudsep, and M. Pedaste, “Exploring teachers’ perceptions of Artificial Intelligence as a tool to support their practice in Estonian K-12 education Exploring teachers’ perceptions of Artificial Intelligence as a tool to support their practice in Estonian K-12 education.”

[7] K. Chadaga, S. Prabhu, N. Sampathila, R. Chadaga, S. Swathi, and S. Sengupta, “Predicting cervical cancer biopsy results using demographic and epidemiological parameters: a custom stacked ensemble machine learning approach,” *Cogent Engineering*, vol. 9, no. 1, 2022, doi: 10.1080/23311916.2022.2143040.

[8] J. M. M. Walboomers *et al.*, “Human papillomavirus in false negative archival cervical smears: Implications for screening for cervical cancer,” *Journal of Clinical Pathology*, vol. 48, no. 8, pp. 728–732, 1995, doi: 10.1136/jcp.48.8.728.

[9] T. J. Selman, C. Mann, J. Zamora, T. L. Appleyard, and K. Khan, “Diagnostic accuracy of tests for lymph node status in primary cervical cancer: A systematic review and meta-analysis,” *CMAJ. Canadian Medical Association Journal*, vol. 178, no. 7, pp. 855–862, 2008, doi: 10.1503/cmaj.071124.

[10] K. Papadatou, P. Perros, N. Thomakos, D. Haidopoulos, A. Rodolakis, and V. Pergialiotis, “Biomarkers in Cervical Cancer,” *Hjog*, vol. 21, no. 1, pp. 1–14, 2022, doi: 10.33574/HjoG.0401.

[11] P. D. Blumenthal, L. Gaffikin, Z. M. Chirenje, J. McGrath, S. Womack, and K. Shah, “Adjunctive testing for cervical cancer in low resource settings with visual inspection, HPV, and the Pap smear,” *International Journal of Gynecology and Obstetrics*, vol. 72, no. 1, pp. 47–53, 2001, doi: 10.1016/S0020-7292(00)00329-5.

[12] M. Reuschenbach, N. Wentzensen, M. G. Dijkstra, M. V. K. Doeberitz, and M. Arbyn, “P16INK4a Immunohistochemistry in cervical biopsy specimens a systematic review and meta-analysis of the interobserver agreement,” *American Journal of Clinical Pathology*, vol. 142, no. 6, pp. 767–772, 2014, doi: 10.1309/AJCP3TPHV4TRIZEK.

[13] J. L. Reid *et al.*, “Human papillomavirus oncogenic mRNA testing for cervical cancer screening: Baseline and longitudinal results from the CLEAR study,” *American Journal of Clinical Pathology*, vol. 144, no. 3, pp. 473–483, 2015, doi: 10.1309/AJCPHVD7MIP3FYVV.

[14] M. Badea *et al.*, “Modern interdisciplinary monitoring of cervical cancer risk,” *Romanian Journal of Morphology and Embryology*, vol. 60, no. 2, pp. 496–478, 2019.

[15] A. J. Blatt, R. Kennedy, R. D. Luff, R. M. Austin, and D. S. Rabin, “Comparison of cervical cancer screening results among 256,648 women in multiple clinical practices,” *Cancer Cytopathology*, vol. 123, no. 5, pp. 282–288, 2015, doi: 10.1002/cncy.21544.

[16] M. Boshart, L. Gissmann, H. Ikenberg, A. Kleinheinz, W. Scheurlen, and H. zur Hausen, “A new type of papillomavirus DNA, its presence in genital cancer biopsies and in cell lines derived from cervical cancer,” *The EMBO journal*, vol. 3, no. 5, pp. 1151–1157, 1984, doi: 10.1002/j.1460-2075.1984.tb01944.x.

[17] L. Gortzak-Uzan *et al.*, “Sentinel lymph node biopsy vs. pelvic lymphadenectomy in early stage cervical cancer: Is it time to change the gold standard?,” *Gynecologic Oncology*, vol. 116, no. 1, pp. 28–32, 2010, doi: 10.1016/j.ygyno.2009.10.049.

[18] X. Zhang, B. Bao, S. Wang, M. Yi, L. Jiang, and X. Fang, “Sentinel lymph node biopsy in early stage cervical cancer: A meta-analysis,” *Cancer Medicine*, vol. 10, no. 8, pp. 2590–2600, 2021, doi: 10.1002/cam4.3645.

[19] M. Arbyn *et al.*, “Pooled analysis of the accuracy of five cervical cancer screening tests assessed in eleven studies in Africa and India,” *International Journal of Cancer*, vol. 123, no. 1, pp. 153–160, 2008, doi: 10.1002/ijc.23489.

[20] P. Cafforio *et al.*, “Liquid biopsy in cervical cancer: Hopes and pitfalls,” *Cancers*, vol. 13, no. 16, pp. 1–19, 2021, doi: 10.3390/cancers13163968.

[21] K. U. Petry *et al.*, "Triaging Pap cytology negative, HPV positive cervical cancer screening results with p16/Ki-67 Dual-stained cytology," *Gynecologic Oncology*, vol. 121, no. 3, pp. 505–509, 2011, doi: 10.1016/j.ygyno.2011.02.033.

[22] M. Iwakawa *et al.*, "The radiation-induced cell-death signaling pathway is activated by concurrent use of cisplatin in sequential biopsy specimens from patients with cervical cancer," *Cancer Biology and Therapy*, vol. 6, no. 6, pp. 905–911, 2007, doi: 10.4161/cbt.6.6.4098.

[23] R. G. Pretorius *et al.*, "Colposcopically directed biopsy, random cervical biopsy, and endocervical curettage in the diagnosis of cervical intraepithelial neoplasia II or worse," *American Journal of Obstetrics and Gynecology*, vol. 191, no. 2, pp. 430–434, 2004, doi: 10.1016/j.ajog.2004.02.065.

[24] R. G. Pretorius, Y. P. Bao, J. L. Belinson, R. J. Burchette, J. S. Smith, and Y. L. Qiao, "Inappropriate gold standard bias in cervical cancer screening studies," *International Journal of Cancer*, vol. 121, no. 10, pp. 2218–2224, 2007, doi: 10.1002/ijc.22991.

[25] J. M. Alcocer-González *et al.*, "In vivo expression of immunosuppressive cytokines in human papillomavirus-transfonned cervical cancer cells," *Viral Immunology*, vol. 19, no. 3, pp. 481–491, 2006, doi: 10.1089/vim.2006.19.481.

[26] C. C. Roberts *et al.*, "Detection of HPV in Norwegian cervical biopsy specimens with type-specific PCR and reverse line blot assays," *Journal of Clinical Virology*, vol. 36, no. 4, pp. 277–282, 2006, doi: 10.1016/j.jcv.2006.03.013.

[27] L. Torres-Ibarra *et al.*, "Triage strategies in cervical cancer detection in Mexico: Methods of the FRIDA study," *Salud Publica de Mexico*, vol. 58, no. 2, pp. 197–210, 2016, doi: 10.21149/spm.v58i2.7789.

[28] M. Durst, L. Gissmann, H. Ikenberg, and H. Zur Hausen, "A papillomavirus DNA from a cervical carcinoma and its prevalence in cancer biopsy samples from different geographic regions," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 80, no. 12 I, pp. 3812–3815, 1983, doi: 10.1073/pnas.80.12.3812.

[29] A. Gupta *et al.*, "Human papillomavirus DNA in urine samples of women with or without cervical cancer and their male partners compared with simultaneously collected cervical/penile smear or biopsy specimens," *Journal of Clinical Virology*, vol. 37, no. 3, pp. 190–194, 2006, doi: 10.1016/j.jcv.2006.07.007.

[30] S. Abdul, B. H. Brown, P. Milnes, and J. A. Tidy, "The use of electrical impedance spectroscopy in the detection of cervical intraepithelial neoplasia," *International Journal of Gynecological Cancer*, vol. 16, no. 5, pp. 1823–1832, 2006, doi: 10.1111/j.1525-1438.2006.00651.x.

[31] T. Sasagawa, "Human papillomavirus infection and cervical cancer," *Biomedical Reviews*, vol. 14, pp. 75–93, 2003, doi: 10.14748/bmr.v14.110.

[32] N. A. Obukhova, A. A. Motyko, U. Kang, S. J. Bae, and D. S. Lee, "Automated image analysis in multispectral system for cervical cancer diagnostic," *Conference of Open Innovation Association, FRUCT*, vol. 2017-April, pp. 345–351, 2017, doi: 10.23919/FRUCT.2017.8071332.

BIOGRAPHIES OF AUTHORS

Abdikadir Hussein Elmi is a highly skilled IT and Computer Science professional with a wealth of experience in various data-related roles. Over the past several years, he has successfully transitioned towards Data Science and Machine Learning, honing his expertise in these cutting-edge fields. He holds a Master of Science in Data Science and Analytics from the prestigious University Science Malaysia (USM), where he gained a strong foundation in advanced data analytics and machine learning techniques. Currently, he serves as a Lecturer at SIMAD University, where he passionately shares his knowledge and expertise with aspiring data scientists. Known for his dynamic and engaging teaching style, his research interests revolve around machine learning, artificial intelligence, natural language processing, and neural networks. He can be contacted at email: xayeeysi77@gmail.com.

Abdijalil Abdullahi received a B.Sc. degree in information technology and an M.Sc. degree in networking and data communication from SIMAD University, Mogadishu, Somalia, in 2014 and 2018, respectively, where he is currently pursuing the Ph.D. degree with the National Advanced IPv6 Center (NAv6), Universiti Sains Malaysia (USM). His research interests include software-defined networking, inter-domain routing, and the internet ecosystem. He can be contacted at email: cabdijaliil22@gmail.com.

Mohamed Ali Barre is a highly experienced system administrator with seven-plus years of expertise in managing advanced network systems and teachings of computer science courses. He has strong skills in troubleshooting technical issues related to hardware, network infrastructure, operating systems, CCTV systems, and software installations. He received a Master of Science in networking and data communication and a Bachelor of Science in Information Technology from SIMAD University. His research interests include network security, internet of things, machine learning related to computer networks. He can be contacted at email: eng.barre1@gmail.com.