

A MACHINE LEARNING BASED APPROACH TO PREDICT CERVICAL CANCER

BY

Jubail Hossain Omar

ID: 201-15-3321

This Report Presented in Partial Fulfillment of the Requirements for the
Degree of Bachelor of Science in Computer Science and Engineering.

Supervised By

Shah Md. Tanvir Siddiquee

Assistant Professor

Department of CSE

Daffodil International University

Co-Supervised By

Tania Khatun

Assistant Professor

Department of CSE

Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

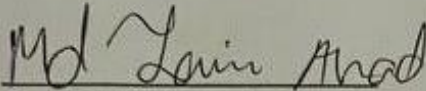
DHAKA, BANGLADESH

JANUARY 2024

APPROVAL

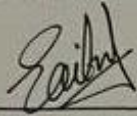
This Project titled A Machine Learning Based Approach to Predict Cervical Cancer, submitted by **Jubail Hossain Omar**, ID: **201-15-3321** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 24 January, 2024.

BOARD OF EXAMINERS



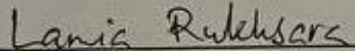
Dr. Md. Taimur Ahad
Associate Professor & Associate Head
Department of CSE
Daffodil International University

Chairman



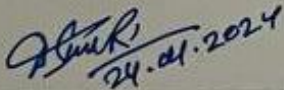
Mr. Saiful Islam
Assistant Professor
Department of CSE
Daffodil International University

Internal Examiner 1



Lamia Rukhsara
Senior Lecturer
Department of CSE
Daffodil International University

Internal Examiner 2



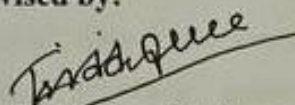
Dr. Abu Sayed Md. Mostafizur Rahaman
Professor
Department of CSE
Jahangirnagar University

External Examiner 1


DECLARATION

I hereby declare that this project has been done by us under the supervision of Shah Md. Tanvir Siddiquee, Assistant Professor, Department of CSE Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree.

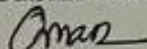
Supervised by:


Shah Md. Tanvir Siddiquee
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:


Tania Khatun
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:


Jubail Hossain Omar
ID: -201-15-3321
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

I begin by extending my heartfelt gratitude to the Almighty for blessing me and enabling me to successfully complete my final year project/internship.

My sincere appreciation goes to **Shah Md. Tanvir Siddiquee, Assistant Professor** of the Department of Computer Science and Engineering at Daffodil International University, Dhaka. Their extensive knowledge and keen interest in the field of "Machine Learning" were instrumental in guiding us throughout this project. Their unwavering patience, scholarly guidance, continuous encouragement, dedicated supervision, constructive criticism, valuable advice, and thorough review of multiple drafts at every stage played a crucial role in the completion of this project.

I would also like to express my deepest thanks to Dr. Sheak Rashed Haider Noori, the Head of the Department of CSE, for his invaluable assistance in bringing my project to fruition. my gratitude extends to all the faculty members and staff of the CSE department at Daffodil International University.

I am grateful to my fellow classmates at Daffodil International University, who engaged in discussions and provided support during the course of my project.

Lastly, I would like to acknowledge and show my utmost respect for the unwavering support and patience of my parents.

ABSTRACT

Recently, there has been a significant increase in the prevalence of physical illnesses, including cervical cancer disease, drawing considerable attention due to its impact on a large population. The severity of the illness can be better understood by analyzing differences between normal and affected diagnostic reports. With numerous studies focused on understanding cervical cancer disease, there are promising opportunities for advancing diagnostic techniques. In this study, I propose the utilization of algorithmic models for early identification and raising awareness of potential threats. My straightforward approach is suitable for predicting simple cases of cervical cancer disease illness in real-world scenarios. We have collected the dataset from Kaggle dataset. We employed various classifiers, including Random Forest (RF), Logistic Regression (LR), Gradient Boosting (GB), K-Nearest Classifier (KNN), Adaboost Classifier (ABC), Decision Tree (DT), Support Vector Machine (SVM) and Gaussian Naïve Bayes (GNB). Notable results were achieved, with the K-Nearest Classifier (KNN), Adaboost Classifier (ABC) standing out as the most accurate, achieving an impressive accuracy rate of 97.33%. Through an experimental investigation and a review of recent findings, I confirmed that the K-Nearest Classifier (KNN), Adaboost Classifier (ABC) performed exceptionally well, accurately predicting cervical cancer disease with an accuracy rate of 97.33%.

Keywords: Cervical Cancer Disease, Algorithm, Model, Accuracy.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	ii
Declaration	iii
Acknowledgements	iv
Abstract	v
List of figures	ix
List of tables	xi
CHAPTER	
CHAPTER 1: INTRODUCTION	1-3
1.1 Introduction	1
1.2 Motivation	1
1.3 Rationale of the Study	2
1.4 Research Questions	2
1.5 Expected Output	2
1.6 Project Management and Finance	2
1.7 Report Layout	3
CHAPTER 2: BACKGROUND	4-6
2.1 Preliminaries/Terminologies	4
2.2 Related Works	4

2.3 Comparative Analysis and Summary	5
2.4 Scope of the Problem	6
2.5 Challenges	6
CHAPTER 3: RESEARCH METHODOLOGY	7-12
3.1 Research Subject and Instrumentation	7
3.2 Data Collection Procedure	7
3.3 Statistical Analysis	8
3.4 Proposed Methodology	9
3.5 Implementation Requirements	11
CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION	13-37
4.1 Experimental Setup	13
4.2 Experimental Results & Analysis	22
4.3 Discussion	35
CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY	38-40
5.1 Impact on Society	38
5.2 Impact on Environment	38
5.3 Ethical Aspects	39
5.4 Sustainability Plan	39

CHAPTER 6: SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH	41-42
6.1 Summary of the Study	41
6.2 Conclusions	41
6.3 Implication for Further Study	41
REFERENCES	43-44

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.1: Methodology of Cervical Cancer Disease	9
Figure 3.2: Correlated Features of Cervical Cancer Dataset	11
Figure 4.1: Random Forest Classifier	14
Figure 4.2: Decision Tree	15
Figure 4.3: Gaussian Naïve Bayes	16
Figure 4.4: Logistic Regression Classifier	17
Figure 4.5: SVC Classifier	18
Figure 4.6: Gradient Boosting Classifier	20
Figure 4.7: K Nearest Classifier	21
Figure 4.8: Adaboost Classifier	22
Figure 4.9: Experimental Results of Training Dataset	24
Figure 4.10: Confusion Matrices of Training SVM	25
Figure 4.11: Confusion Matrices of Training RF	25
Figure 4.12: Confusion Matrices of Training KNN	26
Figure 4.13: Confusion Matrices of Training DT	26
Figure 4.14: Confusion Matrices of Training LR	27
Figure 4.15: Confusion Matrices of Training GNB	27
Figure 4.16: Confusion Matrices of Training GB	28
Figure 4.17: Confusion Matrices of Training ABC	28
Figure 4.18: Experimental Results of Training and Testing Dataset	30
Figure 4.19: Confusion Matrices of Testing SVM	30
Figure 4.20: Confusion Matrices of Testing RF	31
Figure 4.21: Confusion Matrices of Testing KNN	31
Figure 4.22: Confusion Matrices of Testing DT	31
Figure 4.23: Confusion Matrices of Testing LR	32
Figure 4.24: Confusion Matrices of Testing GNB	32

Figure 4.25: Confusion Matrices of Testing GB	32
Figure 4.26: Confusion Matrices of Testing ABC	33
Figure 4.27: Experimental Results of Training Testing and Combine Dataset	35

LIST OF TABLES

TABLES	PAGE NO
Table 2.1: Comparative Analysis	5
Table 4.1: Performance Evaluation of Training Dataset	23
Table 4.2: Performance Evaluation of Training and Testing Dataset	29
Table 4.3: Performance Evaluation of Training, Testing and Combine Dataset	33

CHAPTER 1

INTRODUCTION

1.1 Introduction

Living with cervical cancer disease, a condition marked by hormone deficiencies and performance deterioration, poses daily challenges. Early detection of this prevalent issue remains a critical concern, with timely diagnosis being paramount. Machine learning emerges as a promising tool for predicting cervical cancer by analyzing a wealth of authorized health data and patient diagnostics. My study delved into patient medical records to uncover crucial indicators of the condition, leveraging these findings to identify cervical cancer. While collaborative efforts among academics have aimed to develop machine learning algorithms for this purpose, their methods have often proven unreliable. We propose an alternative approach to enhance illness prediction, differentiating between supervised learning, which relies on labeled data to generate outputs from inputs, and unsupervised learning, which uncovers hidden patterns and information using unlabeled data.

1.2 Motivation

Numerous academic institutions have embarked on creating machine learning algorithms aimed at disease identification within the human body, including conditions like cervical cancer. However, it became evident that their methods lacked accuracy and smoothness in predicting cervical cancer. In response, we propose our innovative approach to enhance the body's ability to forecast illnesses. These machine learning methods fall into two distinct categories, with supervised learning relying on labeled data to generate outputs from inputs through input-output pairings, while unsupervised learning leverages unlabeled data to uncover hidden patterns and information. The developed technique focuses on anticipating the onset of cervical cancer disease in individuals with suspected or ongoing conditions, aiming to provide a more reliable solution.

1.3 Rationale of the Study

My research has yielded a predictive model for identifying cervical cancer in humans, a condition increasingly affecting our society. Recognizing the scarcity of diagnostic resources and information, particularly in our economically challenged nation, where assessing symptoms and diagnosing cervical cancer prove to be expensive, we have turned to machine learning as a potential solution.

1.4 Research Question

- How effective are the algorithms within this model?
- What is the probability of successful cervical cancer detection for individuals, whether afflicted by the condition or not?
- What methods can be employed to anticipate the early onset of cervical cancer?
- What advantages does our proposed model bring to the table?
- In what real-life scenarios can this research find application?
- What is the expected project timeline and progression?

1.5 Expected output

As the prevalence of cervical cancer continues to rise, uncertainty surrounds its presence in individuals. By scrutinizing diagnostic reports, we offer a proactive approach to predict and identify this condition. Our method not only aids in detecting cervical cancer but also enhances decision-making and ensures accurate evaluation of outcomes. Furthermore, it has the potential to measure life satisfaction and address associated issues while simultaneously increasing public awareness of cervical cancer. The efficiency of our model allows for rapid assessment of the condition, providing a valuable tool in addressing this health concern.

1.6 Project Management and Finance

I proposed not only practical but also cost-effective for everyday use, holding significant potential for addressing cervical cancer within our nation. While the implementation of

common tools is essential for the practical application of the prediction process in real-life scenarios, employing high-configuration tools can yield the most optimal results and ensure the smooth functioning of our model. However, even with basic tools, the feasibility of our approach remains viable, making it accessible to a broader range of users.

1.7 Report Layout

The structure of the report encompasses the following key sections:

Background Study: Providing an in-depth exploration of the context and relevant research in the field of cervical cancer.

Research Methodology: Detailing the approach, tools, and techniques used to conduct the study and develop the proposed model.

Experimental Results and Discussion: Presenting the findings, outcomes, and a comprehensive discussion of the research results.

Summary, Conclusion, and Future Analysis: Summarizing the key takeaways, drawing conclusions, and outlining potential directions for future research.

References: Citing the sources and literature used to support the study and its findings.

CHAPTER 2

BACKGROUND STUDY

2.1 Preliminaries

Machine learning methods play a pivotal role in identifying the distinct patterns of cervical cancer disease architecture. Our focus lies in the evaluative examination of patients' diagnostic reports within this domain. To accomplish this, we employ a range of techniques, including SVM, GNB, RF, LR, GB, KN, ABC and DT. This section delves into the exploration of these machine learning models, drawing upon the collective research efforts of various experts in the field, as elaborated in the following segment.

2.2 Related Works

The utilization of machine learning classifiers in the context of cardiac disease diagnosis has proven to be a promising approach [1]. Machine learning algorithms, often employing tree structures for decision models, are applied for their effectiveness in this domain [2]. Juneja et al. [3] conducted a comprehensive survey on the Indian demographic, revealing factors that elevate the risk of cervical cancer, including multiple sexual partners, unhygienic menstrual practices, early marriages, and unhealthy lifestyle choices. Their study emphasized the significant correlation between Human Papillomavirus (HPV) and cervical cancer. Ratul et al. [4] employed machine learning models, achieving an impressive 93.33% accuracy in predicting cervical cancer risk through hyperparameter tuning. Ijaz et al. [5] utilized chi-square testing for feature extraction, employing techniques like DBSCAN, iForest, SMOTE, and SMOTETomek to handle class imbalance and outliers. Alquran et al. [6] used PCA and CCA to classify Pap smear images effectively. Yang et al. [7] developed a machine learning-based cervical cancer risk prediction model, identifying age, number of pregnancies, smoking, and contraceptive use as significant risk factors. Rothberg et al. [8] focused on personalized screening models, demonstrating higher sensitivity than current guidelines. Curia F. [9] proposed an ensemble model with explainable black box methods, enhancing accuracy and interpretability in cervical cancer risk

prediction. Li et al. [10] used machine learning for prognosis prediction in lung adenocarcinoma, emphasizing the potential of personalized treatment planning. Kourou et al. [11] explored machine learning algorithms in cancer prognosis, highlighting the importance of feature selection and integration with genomic and imaging data. Huang et al. [12] developed a deep learning algorithm for predicting lung cancer risk from low-dose CT scans, showcasing high accuracy and reducing unnecessary screenings. Collectively, these studies underscore the potential of machine learning in advancing cervical cancer risk assessment, personalized treatment planning, and prognosis prediction.

2.3 Comparative Analysis and Summary

The comparative analysis presented in Table 2.1 highlights the discrepancies between our research and existing studies, shedding light on the efficacy of our approach.

Table 2.1: Comparative Analysis

Author	Algorithm used	Results
Juneja et al. [3]	SVM, GNB	SVM 89%
Ratul et al. [4]	MLP, DTC, RFC, SVM	MLP 93.33%
Ijaz et al. [5]	DBSCAN, SMOTE, CCPM	SMOTE 80.3%
Alquran et al. [6]	SVM, DT, RF	SVM 95%
Our Proposed Model	SVM, GNB, RF, LR, GB, KNN, ABC and DT	ABC 97.33%

2.4 Scope of the Problem

The task at hand revolved around streamlining and simplifying the diagnosis process for cervical cancer. Given the extensive body of machine learning-related research associated with our proposed model, our primary objective was to maximize accuracy. Despite the limited room for refinement within the existing procedure, the concept was to implement user-friendly technology in order to reduce the frequency of cervical cancer diagnoses, making the process more accessible and efficient.

2.5 Challenges

The material proved exceptionally user-friendly and immensely practical in our use. Upon completing the data collection phase, a meticulous manual examination of the dataset for any missing information becomes necessary. Our commitment to precision in handling this dataset is unparalleled, ensuring that no detail is overlooked or omitted.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research Subject and Instrumentation

To maximize accuracy with our dataset, we harnessed a diverse array of algorithms and hybrid models. Essential to our efforts were cutting-edge configuration tools complemented by top-tier GPUs, ensuring optimal performance. Our toolkit incorporated the Python programming language, alongside associated tools like Jupyter Notebook, Google Colaboratory, and Anaconda. This suite of resources empowered us to seamlessly develop and execute Python code directly within the browser, enhancing the efficiency and versatility of our data analysis and model implementation.

3.2 Data Collection Procedure

The dataset, sourced from kaggle, comprising 858 rows and 36 columns for training dataset and 749 rows and 34 columns for testing dataset. Among these columns, the diagnostic attribute played a pivotal role in categorizing the prevalence of cervical cancer disease, while each individual trait proved crucial for identifying this condition. Patients were classified into two groups denoted by 0 and 1, representing the occurrence and absence of cervical cancer. In the training set, 80% of the applicants were selected, while the remaining 20% constituted the test set, facilitating comprehensive model development and assessment for cervical cancer prediction for each dataset.

Null value handling

In data preprocessing, replacing null values with zero is a common technique to handle missing data effectively. This approach is particularly applicable in situations where zero is a meaningful and plausible value for the variable in question. By substituting null values with zero, the dataset remains consistent, and it ensures that numerical calculations and analyses involving those variables proceed smoothly. However, it is essential to consider the context of the data and the specific characteristics of the variable to determine whether

zero is a suitable replacement. While this method provides a straightforward solution, it may not be appropriate for all scenarios, and careful consideration of the dataset's nature is necessary to make informed decisions about handling missing values.

Feature Selection

Recursive Feature Elimination (RFE) is a feature selection technique commonly used in machine learning to enhance model performance and interpretability. The process involves iteratively fitting a model, ranking features based on their importance or contribution to the model, and eliminating the least significant features. This recursive procedure continues until the optimal subset of features is determined, resulting in a model with improved efficiency and reduced complexity. RFE is particularly useful in scenarios where the dataset contains a large number of features, helping to identify the most relevant variables for predictive modeling. This technique contributes to better model generalization, reduced overfitting, and increased interpretability by focusing on the most informative features during the selection process. Univariate feature selection is a technique in machine learning and statistics that involves evaluating and selecting individual features based on their individual performance in isolation, without considering the interactions between features. It assesses each feature independently, ranking them according to certain statistical measures such as p-values, information gain, or correlation coefficients. By selecting features that demonstrate the strongest correlation with the target variable or have the highest relevance, univariate feature selection aims to enhance model performance, reduce dimensionality, and mitigate the risk of overfitting. This method is particularly useful when dealing with datasets with a large number of features, as it helps identify the most informative variables for building accurate and efficient models.

3.3 Statistical Analysis

The analysis section is an essential component of research projects, and in our case, it played a pivotal role in the development and evaluation of the algorithms employed. We opted to utilize a CSV file as the foundation for our training and testing dataset, necessitating several preparatory steps before it could be effectively used. These

preparations encompassed various actions, including pre-processing and data collection. We have collected the dataset from Kaggle dataset. We employed various classifiers, including Random Forest (RF), Logistic Regression (LR), Gradient Boosting (GB), K-Nearest Classifier (KNN), Adaboost Classifier (ABC), Decision Tree (DT), Support Vector Machine (SVM) and Gaussian Naïve Bayes (GNB). Notable results were achieved, with the K-Nearest Classifier (KNN), Adaboost Classifier (ABC) standing out as the most accurate, achieving an impressive accuracy rate of 97.33%. Through an experimental investigation and a review of recent findings, we confirmed that the K-Nearest Classifier (KNN), Adaboost Classifier (ABC) performed exceptionally well, accurately predicting cervical cancer disease with an accuracy rate of 97.33%.

3.4 Proposed Methodology

Flow chart:

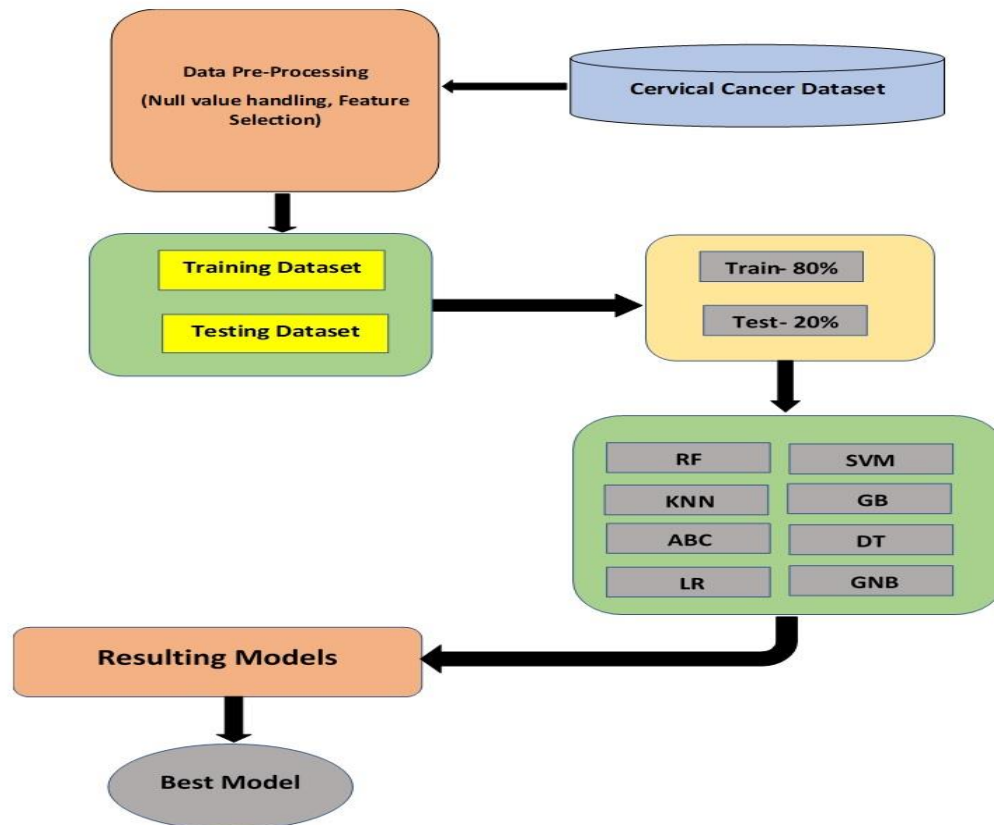


Figure 3.1: Methodology of cervical cancer Disease

In this segment, we harnessed the power of a process diagram to predict cervical cancer disease effectively. Our initial steps revolved around the presentation of the training and testing datasets for our system, followed by the implementation of critical data pre-processing methods such as Null value handling and Feature Selection. The allocation of 80% -20% for training and testing ensured a robust evaluation process. Subsequently, we executed various machine learning algorithms and meticulously assessed their results. The models employed in this phase were subjected to outcome analysis to determine their effectiveness in predicting cervical cancer disease. The model, illustrated in Figure 3.1, encapsulated our research journey, offering insights into the most effective techniques employed in our study.

The intricate connections between two variables were explored in a correlation subplot, which illuminated how one variable's behavior shifted in response to changes in another. The degree of interdependence between variables played a crucial role in determining the likelihood of one factor being accurately predicted from another. This deepened understanding of the dataset has improved our ability to identify the key factors that influence cervical cancer. Figure 3.2 presented a comprehensive view of all the traits associated with the predicted property "Cervical Cancer Disease," shedding light on the interrelationships within the dataset and paving the way for more accurate predictions.

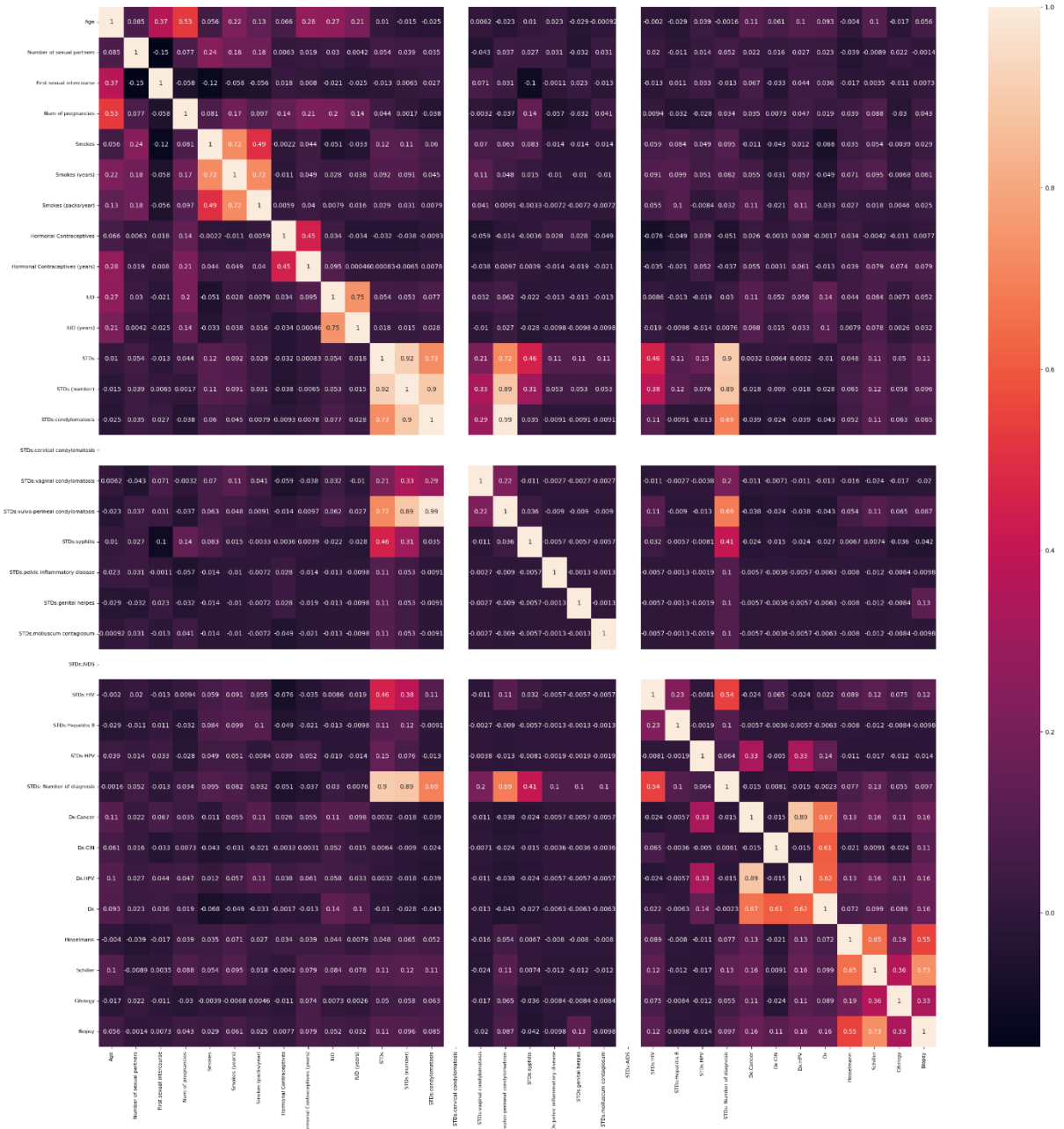


Figure 3.2: Correlated Features of cervical cancer Disease Dataset

3.5 Implementation Requirements

To assess and train our proposed model effectively, a reliable source of data is imperative. The initial step involves the meticulous cleaning of the dataset to ensure smooth operations. The dataset undergoes a comprehensive processing process employing various filtering

methods, culminating in a pristine dataset ready for analysis. Subsequently, vital data pre-processing methods are implemented, including the application of the Feature selection. The dataset is then partitioned into two subsets, with 80% allocated for training and the remaining 20% for rigorous testing. This rigorous testing involves the practical implementation of diverse machine learning algorithms, which are meticulously evaluated to determine their predictive efficacy. Subsequently, the models employed in the process are subjected to a thorough outcome analysis to determine their effectiveness in predicting the target variable. Moving forward, the data analysis stage is essential to lay the foundation for the learning process. Model learning and the fitting of predictive techniques are integral components, paving the way for the next crucial step: model evaluation through voting to identify the model with the highest accuracy. This meticulous selection process ensures the most effective model is chosen for deployment, ultimately maximizing the model's performance and its utility in practical applications.

CHAPTER 4

Experimental results and discussion

4.1 Experimental Setup

In this paper, a supervised learning method based on training and testing was utilized. The classification model was constructed using the training dataset, where the algorithm learned patterns and relationships within the data. Subsequently, the trained model was applied to the testing dataset to predict outcomes or classify new instances. The specific deep learning and machine-learning algorithm employed in this study will be elaborated upon in the subsequent sections.

4.1.1 Classifier Algorithms

SVM, GNB, RF, LR, GB, KNN, ABC and DT methods are some of the classifiers we've created [13].

Random Forest

The Random Forest classifier is a powerful and versatile machine learning algorithm that has gained immense popularity for both classification and regression tasks. It operates by creating an ensemble of decision trees, where each tree is constructed using a random subset of the training data and a subset of the available features. This technique introduces variability and decorrelates the individual trees, mitigating overfitting and improving the model's generalization performance. In classification, the Random Forest combines the results from these decision trees through a majority vote, while in regression, it computes the average of the individual tree predictions. One of the key advantages of Random Forest lies in its ability to handle high-dimensional data, maintain robustness against outliers, and provide feature importance for model interpretability. The algorithm is less prone to overfitting compared to single decision trees, thanks to its inherent bagging (Bootstrap Aggregating) and feature bagging components. Random Forest is particularly useful when dealing with complex and noisy datasets, and it's less sensitive to hyperparameter tuning

than other algorithms. Additionally, the Random Forest can identify influential features and provide insights into their contribution to the model's predictive power. Its robust performance, scalability, and flexibility have made it a popular choice across various domains, including finance, healthcare, and image analysis [20]. However, the trade-off for its power and versatility is increased computational cost and complexity, which can be a consideration for real-time or resource-constrained applications. Nonetheless, the Random Forest remains a reliable workhorse in machine learning, delivering accurate predictions and valuable insights for diverse problem-solving scenarios [18]. The notion is depicted in Fig. 4.1 below.

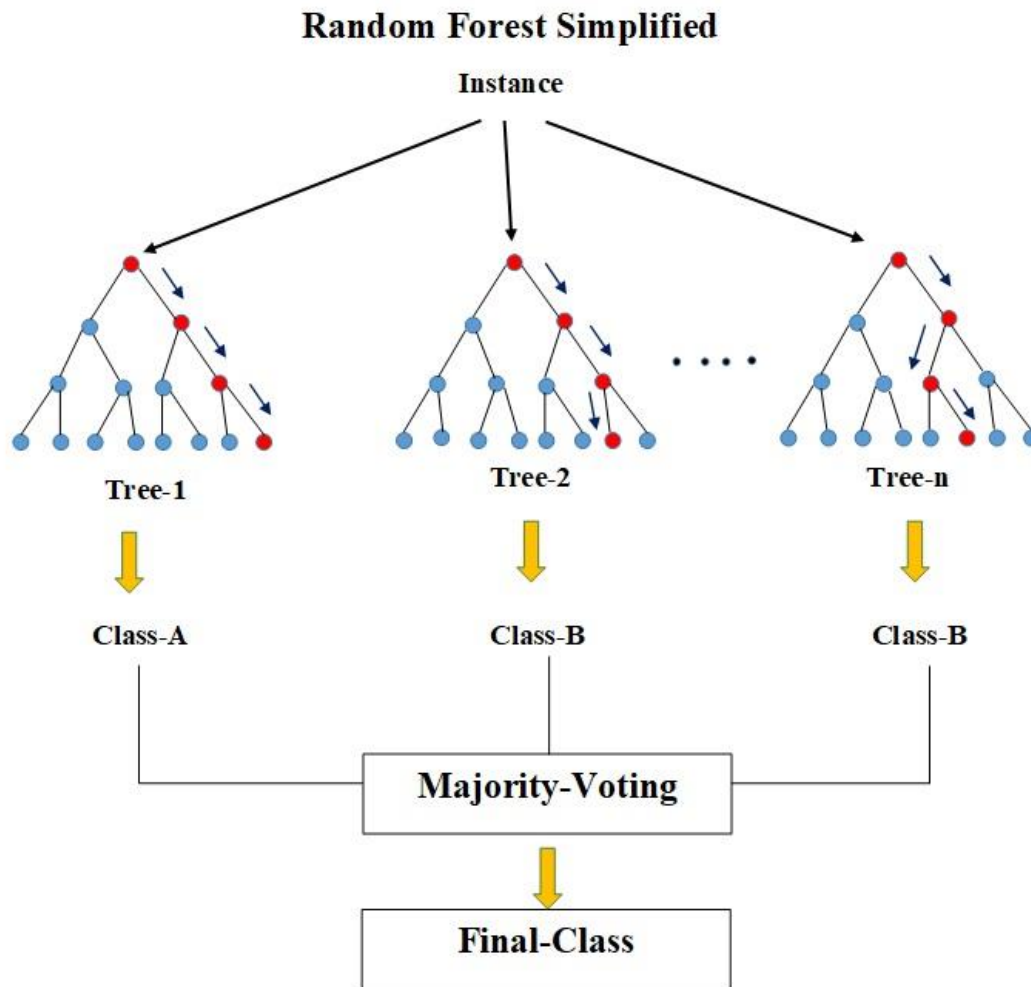


Figure 4.1: Random Forest Classifier

Decision Tree

To assign a classification to an instance, we start by examining the feature represented by the base of the tree node. Then, we follow a branch of the structure that corresponds to the value of that feature. The Decision Tree technique, which just needs two number Classes, is one of the most effective and well-known prediction techniques. Each inner node of a decision tree, a structure of data with an ordered structure where every node in the leaf hierarchy denotes a distinct class, represents an attribute test. On the basis of decision trees, a tree structure known as DT is frequently utilized. The approach may be used to solve classification and regression issues. As the tree grows from the root node, the "splitting" procedure is utilized to select the "Best Features" or "Best Attributes" from the prospective characteristics pool. It is typical to compute two extra metrics, "Entropy," as indicated in (2), and "Data Gain," as mentioned in (3), in order to find the "Best Attribute". Entropy analyzes the consistency of a dataset, whereas collecting data measures the pace at changes that occur in the volatility of attributes [14]. The notion is depicted in Fig. 4.2 below.

$$E(D) = -P(\text{positive})\log_2 P(\text{positive}) - P(\text{negative})\log_2 P(\text{negative}) \quad (4.1)$$

$$\text{Gain}(\text{Attribute } X) = \text{Entropy}(\text{decision Attribute } Y) - \text{Entropy}(X, Y) \quad (4.2)$$

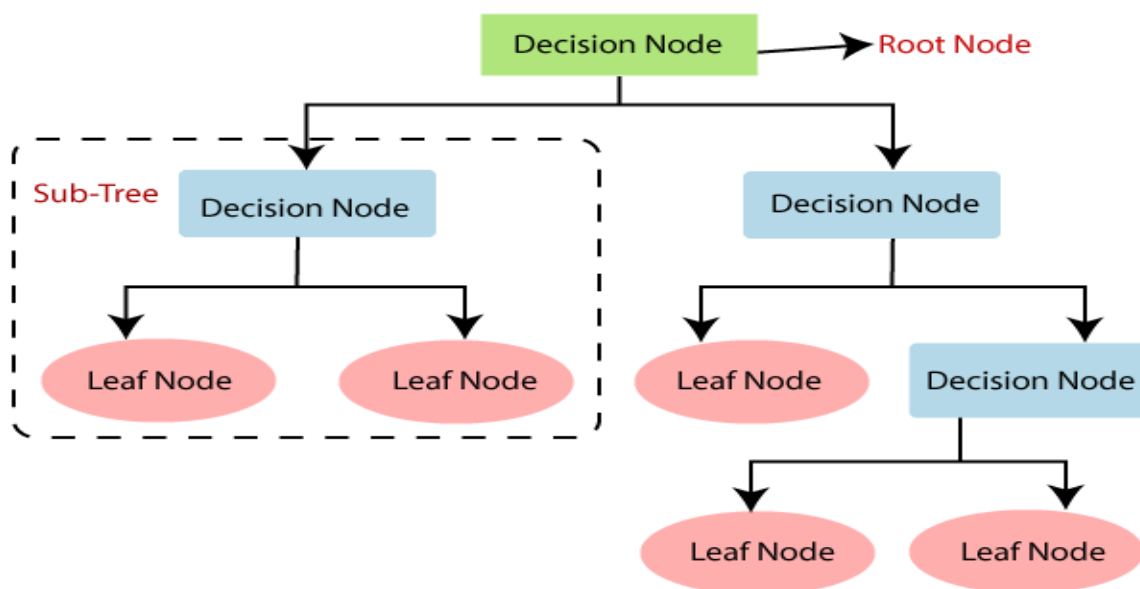


Figure 4.2: Decision Tree

Gaussian Naïve Bayes

The term "GNB" refers to a group of Bayes' Theorem-based algorithms for classification that calculate the probability of an event happening given the probability that another event could also happen. Each algorithm in this group is predicated on the fundamental tenet that any two attributes being identified are unrelated to each other (equation 4.3) [19]. The concept is shown in Fig 4.3 below.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (4.3)$$

The constant value is taken to represent a Gaussian distribution for every characteristic in Gaussian NB. The term "Normal distribution" is often used interchangeably with $w_{(i^{th})}$.

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu_y)^2}{2\pi\sigma^2}\right) \quad (4.4)$$

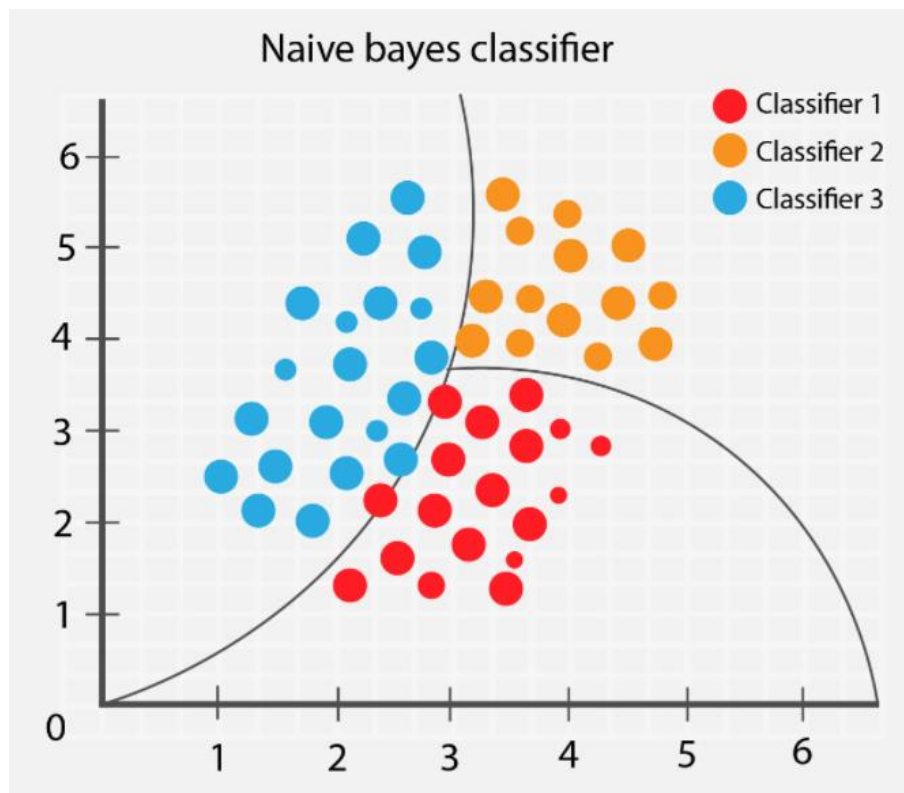


Figure 4.3: Gaussian NB Classifier

Logistic Regression

Logistic Regression is a widely utilized and interpretable machine learning classifier that excels in binary and multiclass classification tasks. Unlike linear regression, which predicts continuous values, logistic regression models the probability of an instance belonging to a particular class using the logistic function (sigmoid). It estimates the odds of an event occurring and maps them to a range between 0 and 1, allowing it to provide clear class separation. The model is trained by minimizing the logistic loss or cross-entropy loss through iterative optimization techniques like gradient descent [21]. Logistic Regression is advantageous for its simplicity, quick training, and ease of interpretation. It can handle both linear and non-linear relationships between features and the target variable through polynomial or interaction terms. While primarily a binary classifier, it can be extended to multiclass problems through techniques like one-vs-rest or softmax regression. One limitation is its susceptibility to overfitting when dealing with high-dimensional data or complex relationships, which can be mitigated through regularization techniques like L1 (Lasso) or L2 (Ridge) regularization. Despite its simplicity, logistic regression is a valuable tool in various domains, including healthcare (predicting disease outcomes), finance (credit risk assessment), and natural language processing (text classification), and it serves as a foundational model in many machine learning pipelines due to its transparency and effectiveness [16] [17]. The concept is shown in Fig 4.4 below.

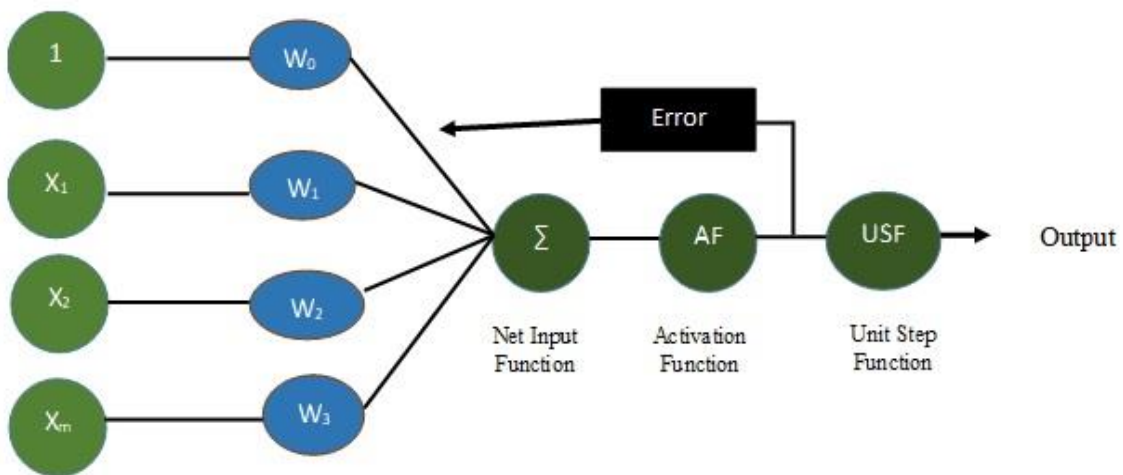


Figure 4.4: Logistic Regression Classifier

Support Vector Machine

Regression and classification problems may both be resolved using the Support Vector Classifier (SVC). However, categorization issues are where artificial intelligence is most frequently applied. The SVM approach looks for a straight line, or judgment limit, that divides the region into categories in all n variables in order to properly categorize fresh data points [22]. A hyperplane is this highest utility bound. Using SVM, which chooses the most extreme locations and vectors, a hyperplane may be created. As a result, the word "support vector," which is used to describe these severe situations, is where the technique's name, "support vector machine," comes from [15]. The Support Vector Classifier (SVC)'s working procedure is depicted in Fig 4.5.

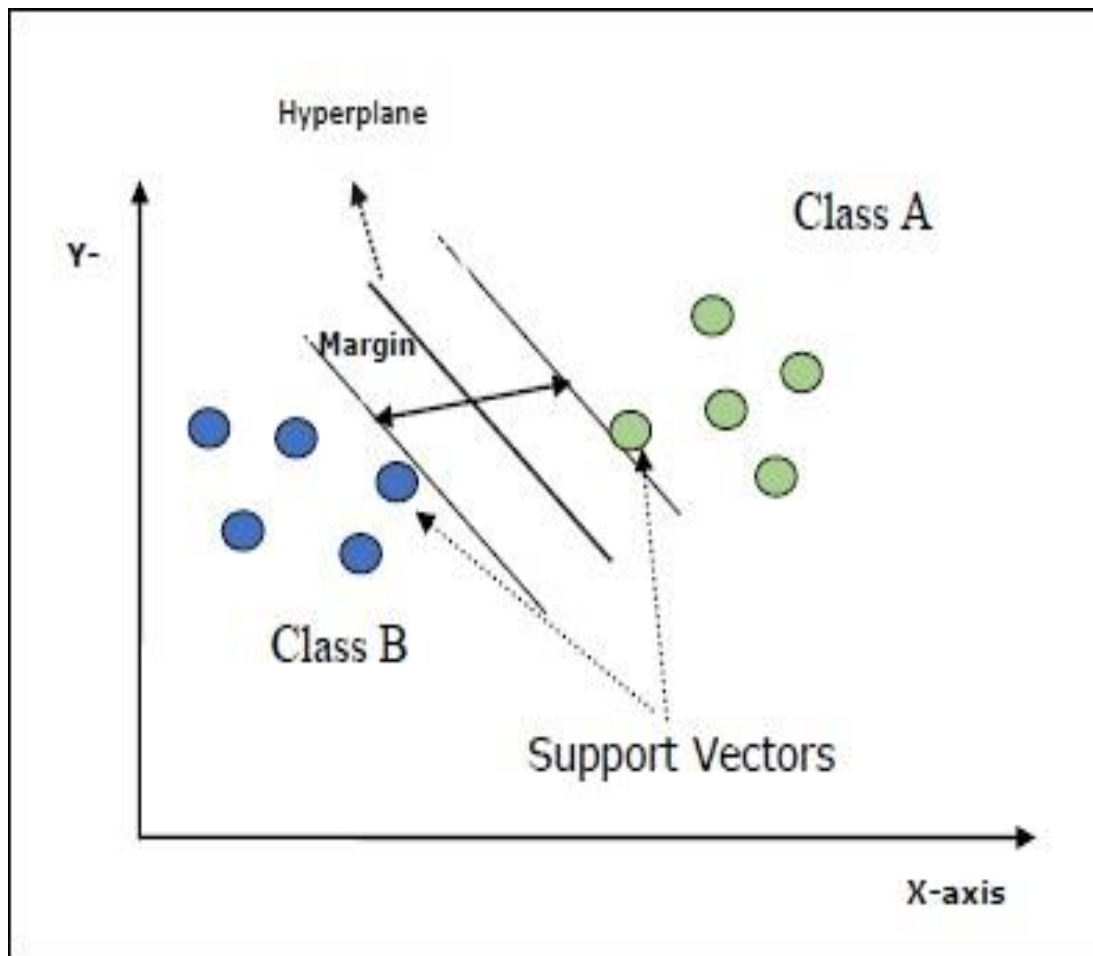


Figure 4.5: SVC classifier

Gradient Boosting

The Gradient Boosting Classifier is a powerful and versatile machine learning algorithm that excels in predictive modeling, particularly in classification tasks. It operates by iteratively building a strong predictive model through the combination of multiple weak models, typically decision trees, in a sequential manner. At each iteration, the algorithm focuses on the misclassified data points from the previous stage, assigning them greater importance. This iterative process allows the algorithm to continuously refine its predictions, ultimately creating a robust ensemble model [25]. One of the key advantages of the Gradient Boosting Classifier is its ability to handle complex, high-dimensional data and capture intricate relationships between variables. By combining the outputs of multiple weak learners, it can achieve superior predictive performance. However, this power comes at a computational cost, and training a Gradient Boosting model can be more time-consuming compared to some other algorithms. To mitigate the risk of overfitting, careful hyperparameter tuning and cross-validation are essential when implementing Gradient Boosting. The choice of the learning rate, the number of boosting iterations (trees), and the maximum depth of trees are critical factors that influence the model's performance. In practice, Gradient Boosting is widely used in various fields, including data mining, finance, and biology, due to its effectiveness in addressing complex classification challenges and producing accurate results [23] [24]. Its versatility and robustness make it a valuable tool for both beginners and experienced data scientists aiming to tackle a wide range of classification tasks. The Gradient Boosting (GB)'s working procedure is depicted in Fig 4.6.

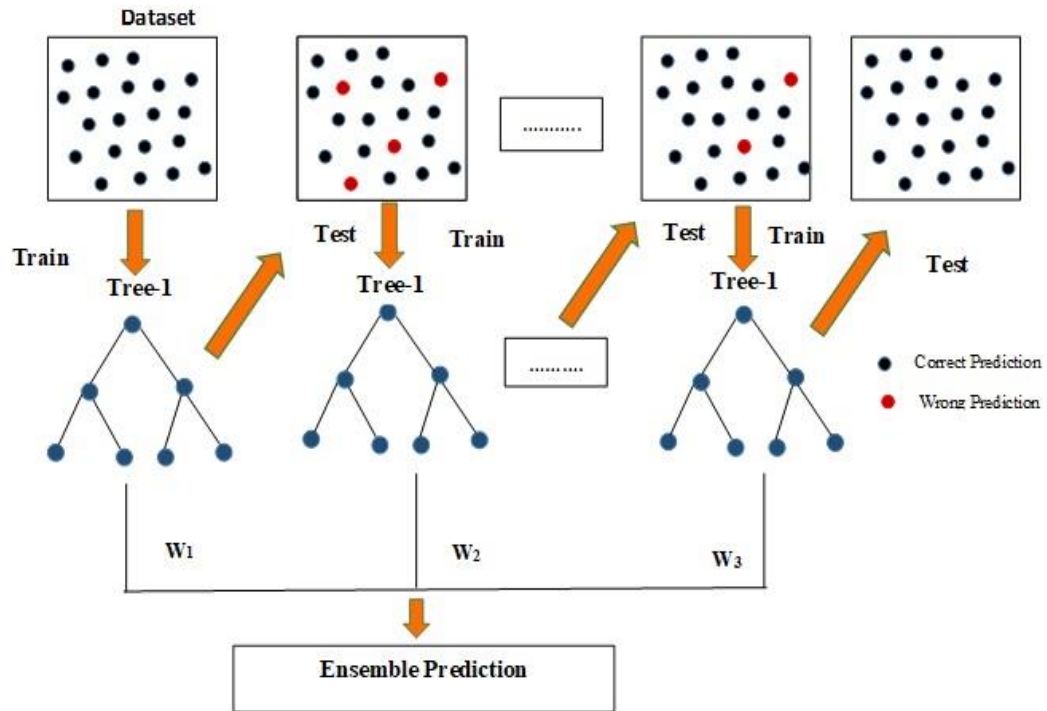


Figure 4.6: Gradient Boosting Classifier

K-Nearest

The K-Nearest Neighbors (KNN) classifier is a widely used and intuitive machine learning algorithm for classification tasks. It operates on the principle that similar data points tend to belong to the same class. In the KNN algorithm, an input data point is classified based on the majority class among its K nearest neighbors in the feature space. The choice of K , the number of neighbors to consider, is a critical hyperparameter that impacts the algorithm's performance. KNN is non-parametric and does not make strong assumptions about the underlying data distribution, making it applicable in various scenarios. Its simplicity and ease of implementation make it a popular choice for introductory machine learning tasks [25] [26]. However, KNN's computational efficiency can be a limitation for

large datasets, as it requires calculating distances between the data point in question and all other data points in the dataset. Moreover, KNN's performance is sensitive to the choice of distance metric, and the curse of dimensionality can affect its accuracy as the number of features or dimensions increases. To address these challenges, techniques such as feature selection, dimensionality reduction, and careful hyperparameter tuning are often employed in conjunction with KNN. Despite its limitations, KNN remains a valuable tool for many classification problems, particularly when the dataset is manageable in size and the algorithm's assumptions align well with the underlying data distribution. Fig 4.7, which is below, illustrates the idea.

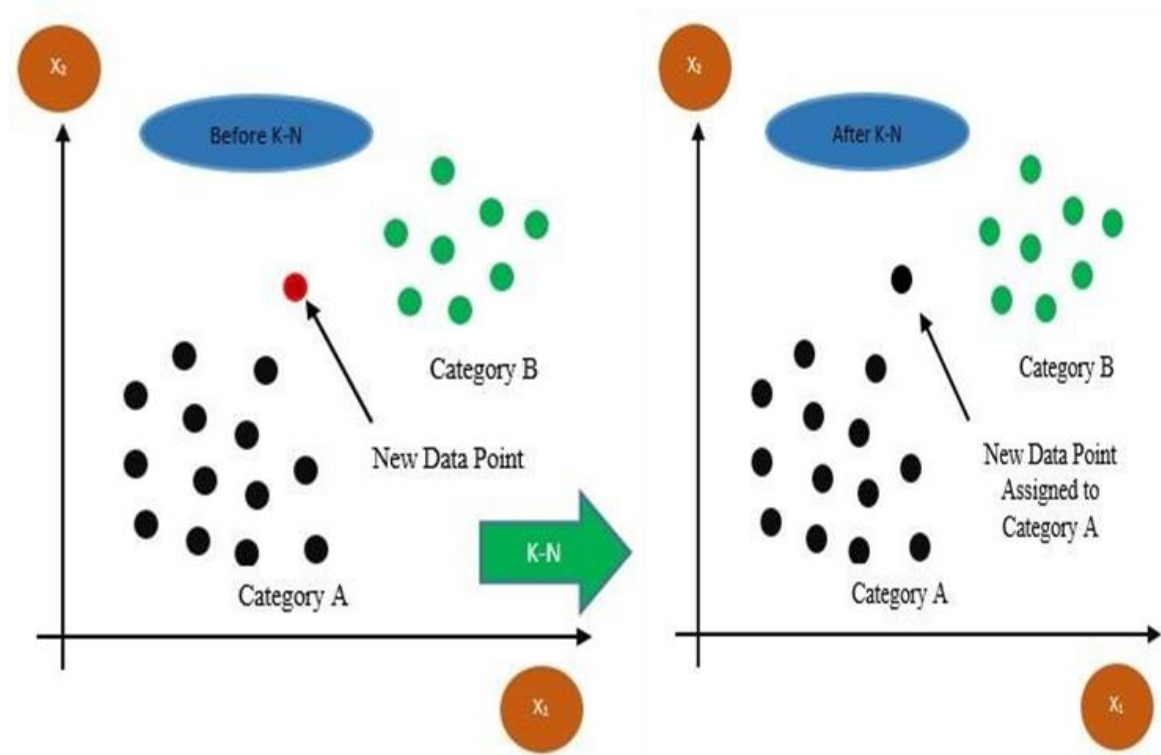


Figure 4.7: K-Nearest Classifier

Adaboost

The AdaBoost (Adaptive Boosting) classifier is a powerful ensemble learning method designed to enhance the performance of weak classifiers by combining them into a robust and accurate model. AdaBoost operates iteratively, sequentially adjusting the weight of each training instance based on the accuracy of the previous weak classifiers. This means

that instances that are misclassified receive higher weights, allowing subsequent weak classifiers to focus on them and improve their classification accuracy. The final prediction is then made by combining the weighted outputs of these weak classifiers. One of AdaBoost's strengths lies in its adaptability to different classification problems, as it can work with a wide range of base classifiers, typically decision stumps or shallow decision trees. It's particularly effective in addressing complex datasets and overcoming issues such as overfitting, as it gives more emphasis to challenging data points during training. Moreover, AdaBoost is known for its ability to handle high-dimensional feature spaces effectively [27] [28]. While AdaBoost is a powerful algorithm, it's not immune to outliers or noisy data, which can adversely affect its performance. However, its capacity to mitigate these issues is strengthened by its sequential learning process. By leveraging AdaBoost's combination of weak learners, it often results in a strong and accurate classifier that is widely used in various fields, including face detection, text classification, and bioinformatics, where high performance and adaptability are essential. The notion is depicted in Fig 4.8 below.

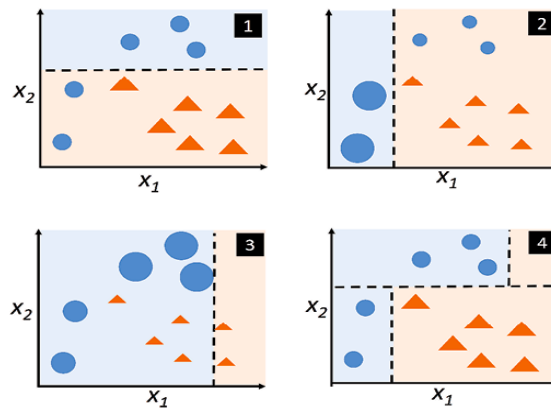


Figure 4.8: Adaboost Classifier

4.2 Experimental Results & Analysis

In this phase of the study, the evaluation of existing models played a pivotal role in assessing the efficiency of the proposed model targeting cervical cancer prediction using the designated dataset. The process commenced with the initial implementation of the chosen dataset, followed by a rigorous examination to identify and rectify missing or

erroneous data points, ensuring the dataset's integrity. A diverse range of machine learning algorithms was subsequently deployed, and their performances meticulously analyzed. For the proposed algorithms, a comprehensive assessment was conducted through confusion matrices, which included key metrics such as Accuracy, Precision, Recall, and F-1 Score providing a holistic view of their predictive capabilities. Additionally, traditional algorithms underwent the same scrutiny, further enabling a comparative analysis. A total of eight distinct traditional classifiers were harnessed, and the resulting outcomes, thoroughly assessed, facilitated the identification of the most effective approaches for predicting cervical cancer disease. This comprehensive evaluation process served as a critical step in gauging the performance of the proposed model and fine-tuning its predictive accuracy for practical application.

In the evaluation of various machine learning algorithms for a specific task, the SVM and Random Forest (RF) models achieved the highest accuracy at 94.76%. While SVM exhibited lower precision, recall, and F-1 score, RF demonstrated better precision and recall, balancing trade-offs between false positives and false negatives. K-Nearest Neighbors (KNN) and Decision Tree (DT) models also performed well, with KNN achieving the highest accuracy at 95.34% and DT exhibiting a high recall of 81.81%. Logistic Regression (LR) outperformed others with the highest accuracy of 96.51% and balanced precision, recall, and F-1 score. Gaussian Naive Bayes (GNB) showed high recall but lower precision. Gradient Boosting (GB) and AdaBoost (ABC) demonstrated competitive performance with balanced precision, recall, and F-1 score. These results suggest that the choice of algorithm depends on the specific trade-offs desired in precision and recall for the given task. The visualization is shown in Table 4.1 and Figure 4.9.

Table 4.1. Performance Evaluation of Training Dataset

Algorithm	Accuracy	Precision	Recall	F-1 Score
SVM	94.76	75.00	27.27	40.00
RF	94.76	58.33	63.63	60.86
KNN	95.34	80.00	36.36	50.00
DT	95.34	60.00	81.81	69.23
LR	96.51	72.72	72.72	72.72
GNB	91.86	43.47	90.90	58.82

GB	95.93	64.28	81.81	72.00
ABC	95.34	61.53	72.72	66.66

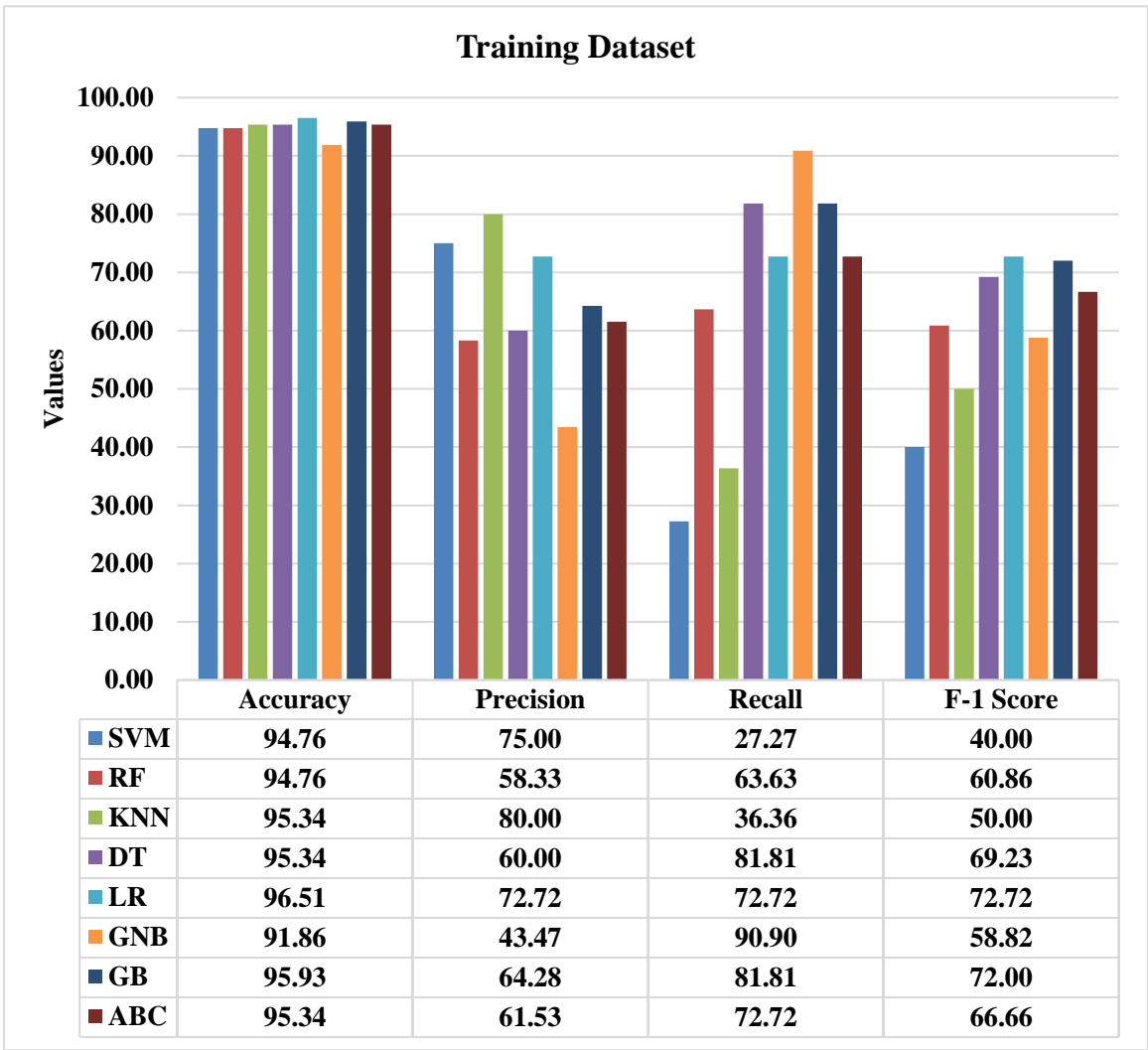


Figure 4.9: Experimental Results of Training Dataset

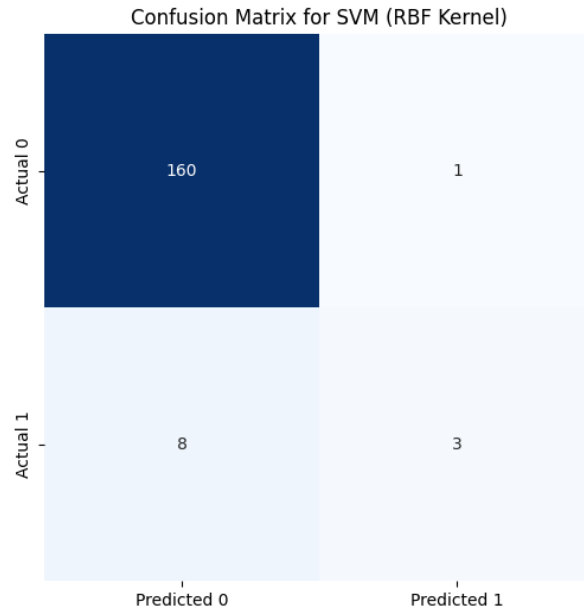


Figure 4.10: Confusion Matrices of Training SVM

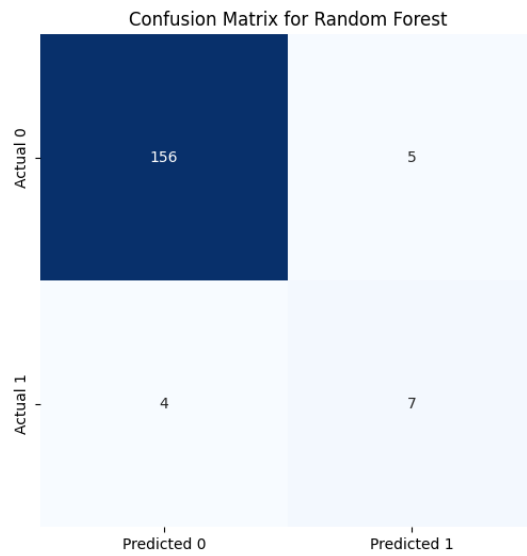


Figure 4.11: Confusion Matrices of Training RF

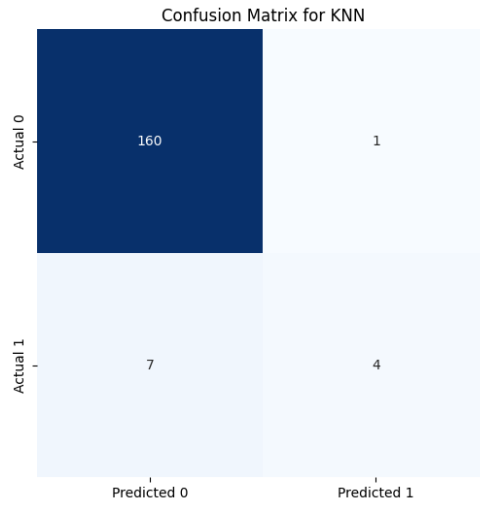


Figure 4.12: Confusion Matrices of Training KNN

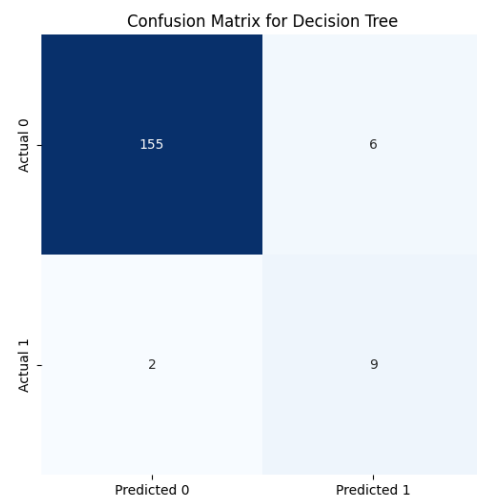


Figure 4.13: Confusion Matrices of Training DT

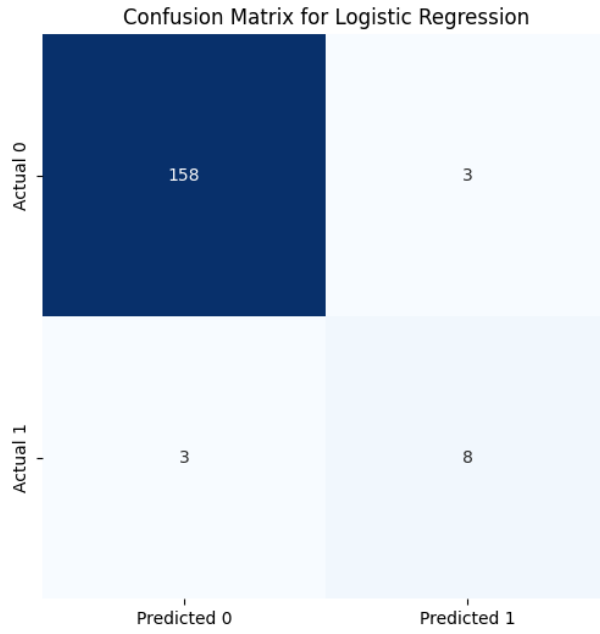


Figure 4.14: Confusion Matrices of Training LR

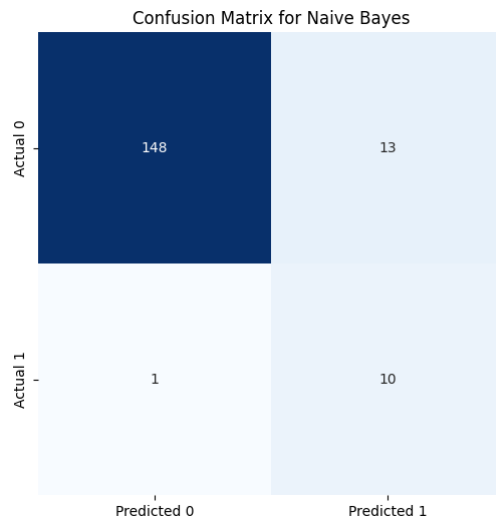


Figure 4.15: Confusion Matrices of Training GNB

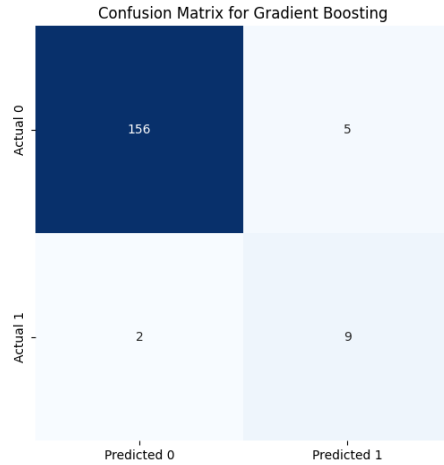


Figure 4.16: Confusion Matrices of Training GB

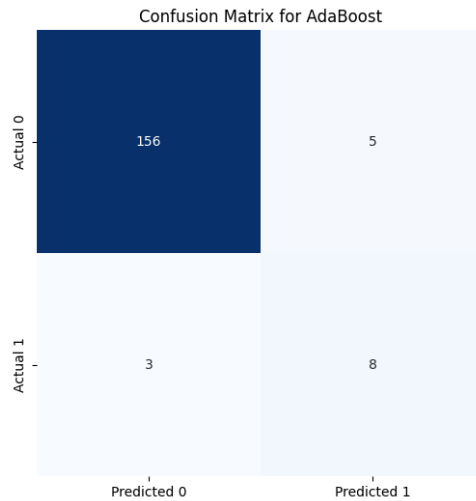


Figure 4.17: Confusion Matrices of Training ABC

In the evaluation of machine learning algorithms on both training and testing datasets, several observations emerge. Support Vector Machine (SVM) exhibits improved performance on the testing set, demonstrating a higher accuracy of 96.66% and enhanced recall at 85.71%, compared to its training performance. Random Forest (RF) also shows a similar trend, with a testing accuracy of 96.00% and notable recall, despite a decrease in precision. K-Nearest Neighbors (KNN), Decision Tree (DT), Logistic Regression (LR), Gradient Boosting (GB), and AdaBoost (ABC) maintain consistent or slightly improved performance on the testing set, reflecting their generalization ability. However, Gaussian

Naive Bayes (GNB) experiences a noticeable decline in accuracy and precision on the testing set, indicating potential limitations in its application to new data. Overall, these results underscore the importance of evaluating models on separate testing sets to assess their real-world performance. The visualization is shown in Table 4.2 and Figure 4.18.

Table 4.2. Performance Evaluation of Testing Dataset

Algorithm	Accuracy	Precision	Recall	F-1 Score
SVM (Training)	94.76	75.00	27.27	40.00
SVM (Testing)	96.66	60.00	85.71	70.58
RF (Training)	94.76	58.33	63.63	60.86
RF (Testing)	96.000	54.540	85.710	66.660
KNN (Training)	95.34	80.00	36.36	50.00
KNN (Testing)	97.33	66.66	85.71	75
DT (Training)	95.34	60.00	81.81	69.23
DT (Testing)	97.33	66.66	85.71	75.00
LR (Training)	96.51	72.72	72.72	72.72
LR (Testing)	96.00	54.54	85.71	66.66
GNB (Training)	91.86	43.47	90.90	58.82
GNB (Testing)	85.33	24.13	99.99	38.88
GB (Training)	95.93	64.28	81.81	72.00
GB (Testing)	96.00	54.54	85.71	66.66
ABC (Training)	95.34	61.53	72.72	66.66
ABC (Testing)	97.33	66.66	85.71	75.00

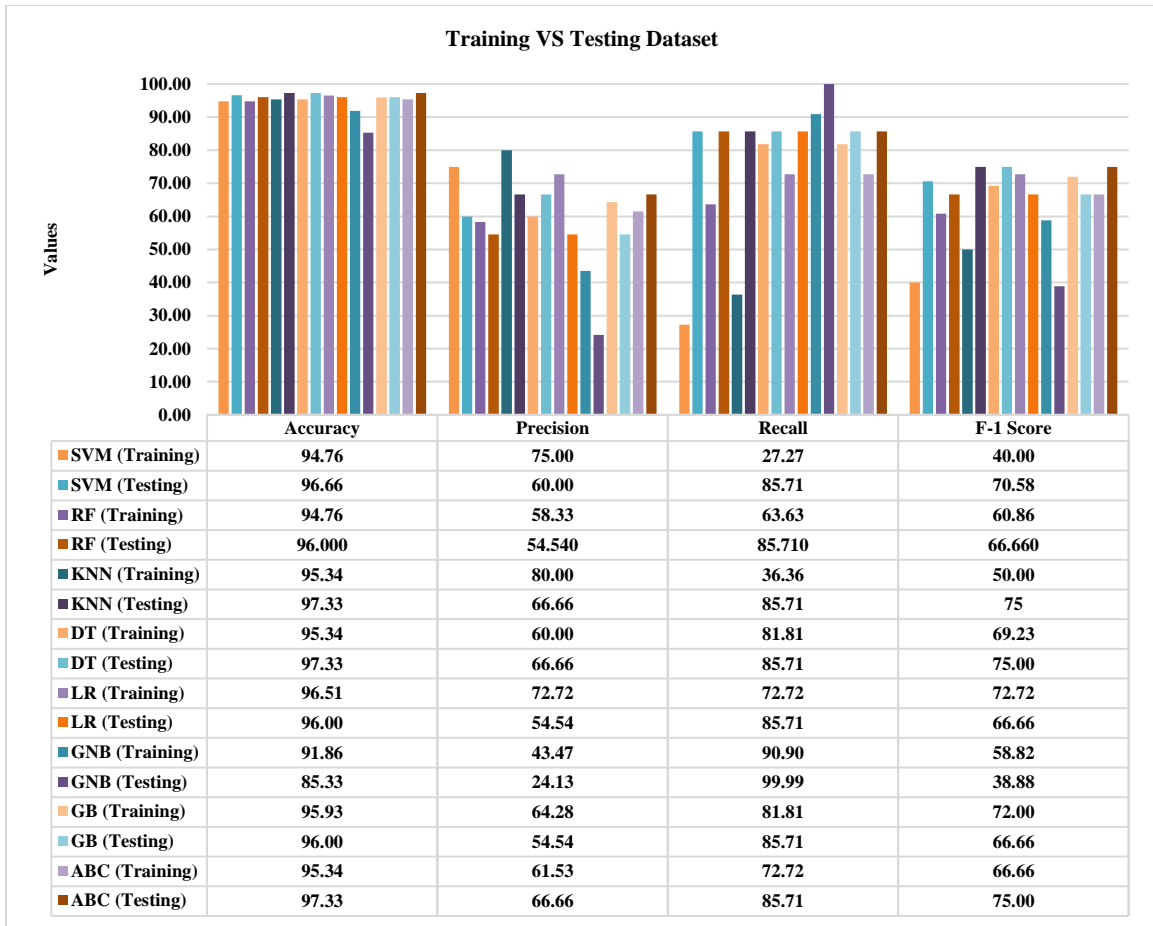


Figure 4.18: Experimental Results of Training and Testing Dataset

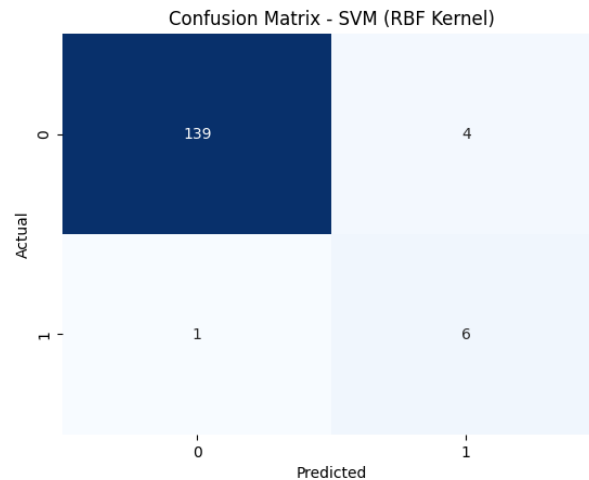


Figure 4.19: Confusion Matrices of Testing SVM

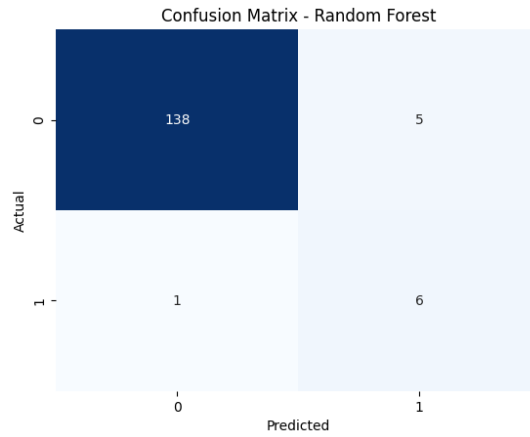


Figure 4.20: Confusion Matrices of Testing RF

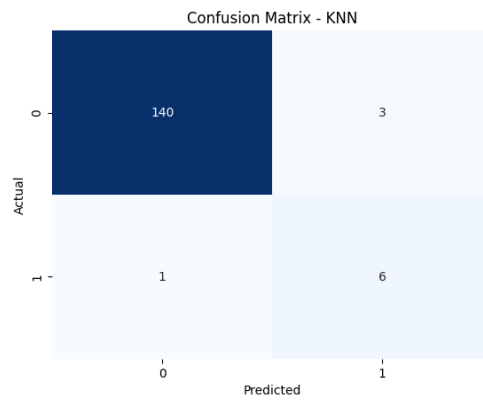


Figure 4.21: Confusion Matrices of Testing KNN

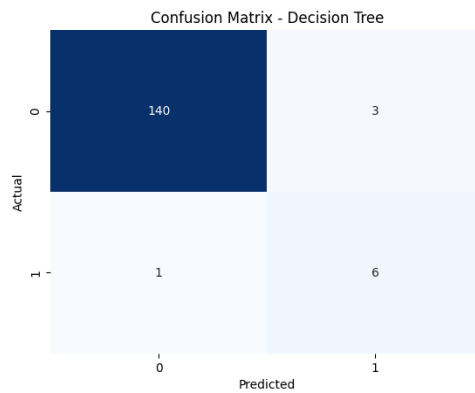


Figure 4.22: Confusion Matrices of Testing DT

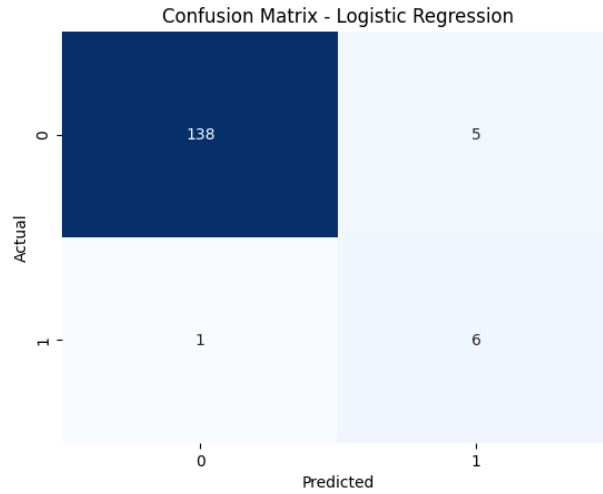


Figure 4.23: Confusion Matrices of Testing LR

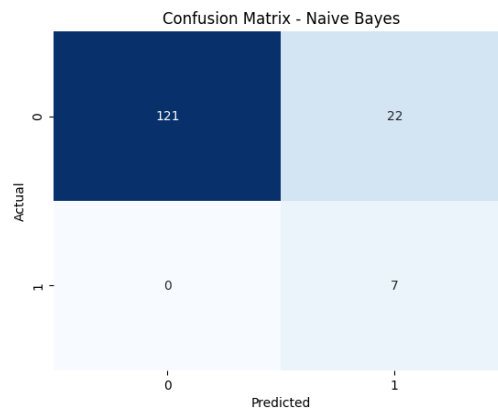


Figure 4.24: Confusion Matrices of Testing GNB

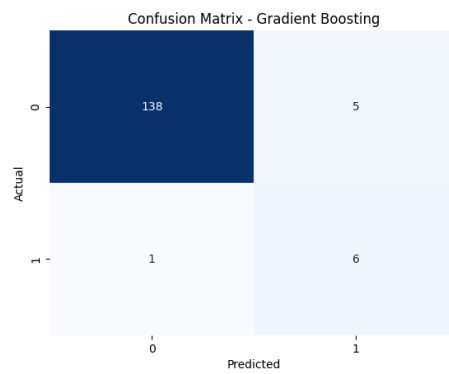


Figure 4.25: Confusion Matrices of Testing GB

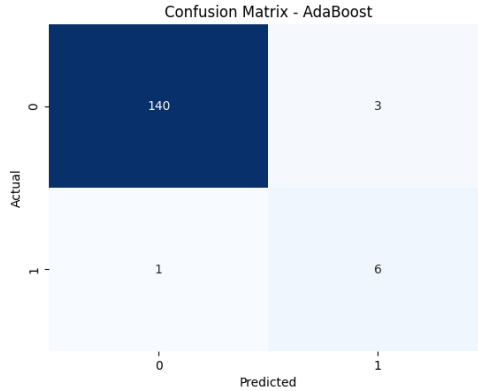


Figure 4.26: Confusion Matrices of Testing ABC

In the assessment of machine learning algorithms across training, testing, and combined datasets, consistent trends and variations are observed. Support Vector Machine (SVM) displays a notable improvement in recall on the testing set compared to its training performance, maintaining this trend in the combined dataset. Random Forest (RF) and K-Nearest Neighbors (KNN) exhibit comparable accuracy and recall across all sets, emphasizing their stable performance. Decision Tree (DT) and Logistic Regression (LR) models demonstrate consistent precision and recall but experience a slight decrease in testing accuracy. Gaussian Naive Bayes (GNB) faces challenges, particularly in precision, showing notable declines across all sets. Gradient Boosting (GB) and AdaBoost (ABC) models maintain stability in performance metrics across training, testing, and combined datasets. These findings underscore the importance of comprehensive evaluation, considering both training and testing outcomes, to ensure robust model performance. The visualization is shown in Table 4.3 and Figure 4.28.

Table 4.3. Performance Evaluation of Training, Testing and Combine Dataset

Algorithm	Accuracy	Precision	Recall	F-1 Score
SVM (Training)	94.76	75.00	27.27	40.00
SVM (Testing)	96.66	60.00	85.71	70.58
SVM (Combine)	96.66	60.00	85.71	70.58
RF (Training)	94.76	58.33	63.63	60.86
RF (Testing)	96.000	54.540	85.710	66.660
RF (Combine)	96.66	60.00	85.71	70.58
KNN (Training)	95.34	80.00	36.36	50.00
KNN (Testing)	97.33	66.66	85.71	75

KNN (Combine)	97.330	66.660	85.710	75.000
DT (Training)	95.34	60.00	81.81	69.23
DT (Testing)	97.33	66.66	85.71	75.00
DT (Combine)	97.33	66.66	85.71	75.00
LR (Training)	96.51	72.72	72.72	72.72
LR (Testing)	96.00	54.54	85.71	66.66
LR (Combine)	96.00	54.54	85.71	66.66
GNB (Training)	91.86	43.47	90.90	58.82
GNB (Testing)	85.33	24.13	99.99	38.88
GNB (Combine)	88.66	29.16	99.99	45.16
GB (Training)	95.93	64.28	81.81	72.00
GB (Testing)	96.00	54.54	85.71	66.66
GB (Combine)	96.00	54.54	85.71	66.66
ABC (Training)	95.34	61.53	72.72	66.66
ABC (Testing)	97.33	66.66	85.71	75.00
ABC (Combine)	97.33	66.66	85.71	75.00

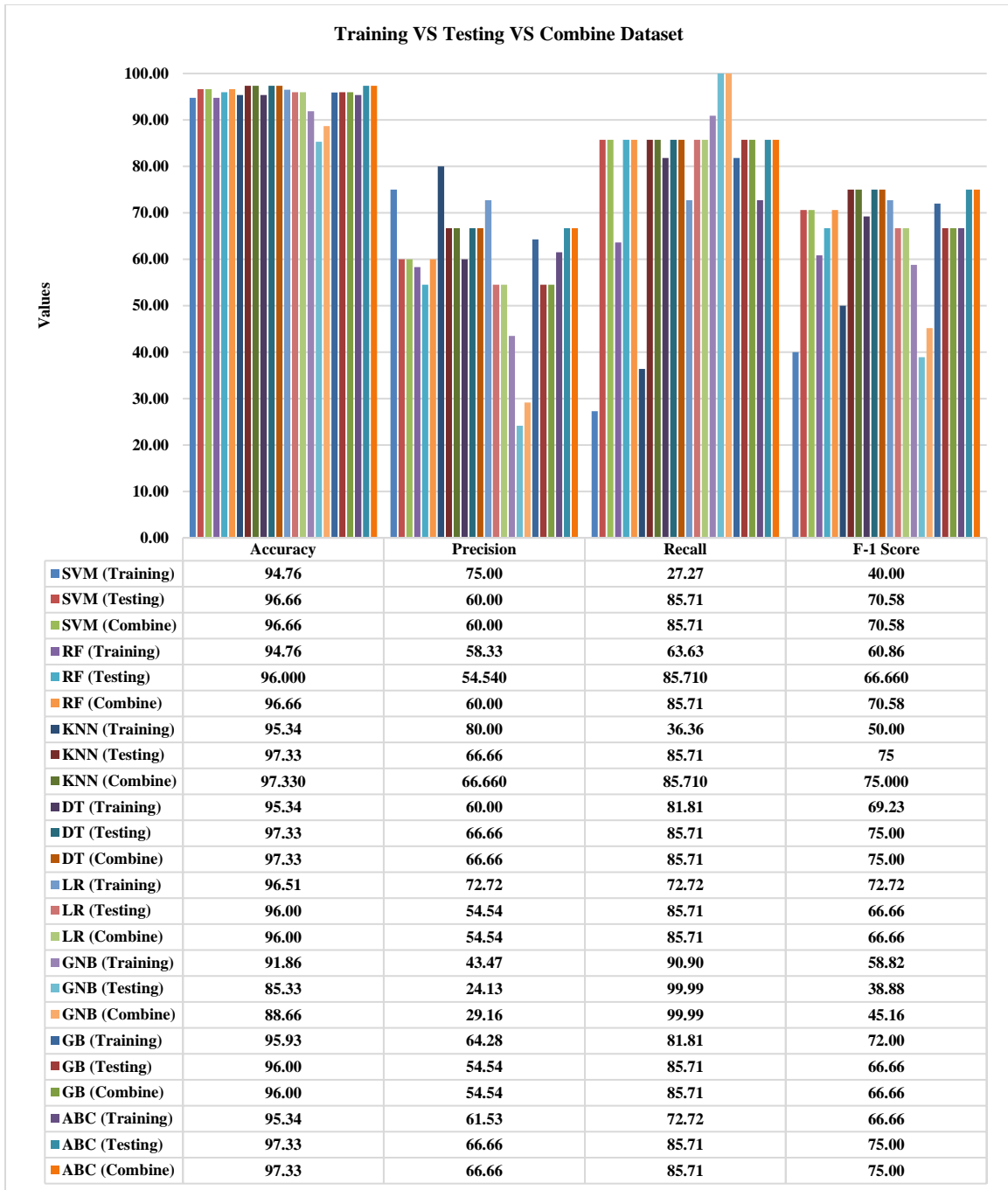


Figure 4.27: Experimental Results of Training, Testing and Combine Dataset

4.3 Discussion

In this phase, we will clarify the evaluation framework for our proposed model, considering key performance metrics such as accuracy, precision, recall, and F-1 score.

4.3.1 Accuracy

This section explores the concept of accuracy, which refers to the percentage of predictions made using testing data that were correct or exact. Accuracy is a measure of the model's correctness, comparing its predictions to the actual real-world measurements. It focuses on a single variable and primarily addresses intentional errors, making it one of the most straightforward and widely used evaluation techniques for any model. Ensuring the accuracy of our models is a crucial aspect of model validation and performance assessment.

$$\text{Accuracy} = \frac{\text{TruePositive} + \text{TrueNegative}}{\text{TruePositive} + \text{FalsePositive} + \text{TrueNegative} + \text{FalseNegative}}$$

4.3.2 Precision

This section addresses precision, which measures the proportion of positively predicted observations that actually occurred. Precision reflects the true positive rate, highlighting the actual percentage of instances when the model correctly predicted true positive outcomes. It's important to note that while a strong recall is desirable for many models, it can sometimes be misleading if not considered in the context of precision and other performance metrics.

$$\text{Precision} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}}$$

4.3.3 Recall

This section discusses recall, which is the proportion of actual positive data points correctly predicted by the model. Recall is crucial in determining the model's ability to capture true positive instances, and it establishes the ratio of all positive labels to the predicted positives. While high accuracy is generally desirable, it's essential to recognize that it can sometimes be misleading if not assessed alongside other important metrics like recall.

$$\text{Recall} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}}$$

4.3.4 F-1 Score

This section discusses the evaluation metrics of accuracy and recall, emphasizing their relevance in assessing a model's performance. Key metrics to consider are the recall and accuracy ratios, which provide insights into the model's ability to correctly identify relevant instances and overall accuracy. It's important to note that if the mean of the harmonic mean of these metrics is relatively low, it may indicate that the model's performance is not optimal, warranting further improvements.

$$F - 1 \text{ Score} = 2 * \frac{\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

Our recommended strategy presents numerous significant benefits, both from an economic and social perspective. Rooted in a real-life dataset, our model was meticulously crafted to investigate and discern the critical components and characteristics of individuals afflicted with cervical cancer. This research bestows a multitude of societal advantages, foremost among them being the capacity to educate and raise awareness about the prevalence of cervical cancer and available preventive measures. Our model's precision in diagnosis and regular monitoring facilitates the early initiation of treatment, enhancing individuals' ability to make informed healthcare decisions and anticipate potential affliction. Notably, the streamlined and efficient nature of our method significantly reduces time and computational demands, simplifying disease prediction with remarkable accuracy. Our comprehensive data analysis employs advanced diagnostic techniques to uncover the underlying factors contributing to cervical cancer. On a societal level, we aspire to witness the widespread acceptance and implementation of our recommended approach. By disseminating knowledge and promoting proactive healthcare practices, we aim to create a more informed and health-conscious society. The ultimate goal is to empower individuals to take charge of their well-being, thereby mitigating the impact of cervical cancer and other health-related challenges. In summary, our model offers a promising avenue for not only precise disease prediction but also for the betterment of public health and healthcare awareness on a broader scale.

5.2 Impact on Environment

Our proposed paradigm holds exceptional promise for remote and underserved areas, offering simplified diagnostic methods that can effectively reduce complexity and save time. Its straightforward and non-invasive nature ensures that the environment will reap the benefits without any adverse effects. With our model, individuals in remote regions

need not travel to urban centers to determine whether they have cervical cancer or not, making healthcare more accessible and convenient. This predictive model, which also forecasts likely outcomes, can seamlessly complement a patient's diagnostic report, alleviating concerns about the cost of local treatment or affordable cervical cancer identification. Its user-friendly design ensures that people at any skill level can utilize it with ease.

Through the implementation of our recommended model, the potential to ascertain the presence of cervical cancer in a patient becomes a reality, significantly enhancing the political and social healthcare landscape. We firmly believe that the adoption of our proposed model will usher in a substantial advancement in the realm of medical scientific technology, ultimately improving the quality of healthcare services and medical technology across the board.

5.3 Ethical Aspects

Before implementing our system, it is imperative to take ethical precautions to safeguard against the inadvertent disclosure of private information, diagnostic outcomes, or unintended humor. The diagnosis and treatment of cervical cancer, both in the real world and in forthcoming research endeavors, stand to benefit from our recommended approach, as this issue transcends geographical boundaries and affects a global population. Our method empowers individuals, whether they are patients or well-informed individuals, to anticipate the onset and progression of their cervical cancer condition, offering a valuable tool for proactive healthcare management on a global scale.

5.4 Sustainability Plan

We have the assurance that our proposed model can seamlessly integrate with the global technology landscape for cervical cancer illness diagnosis and research. We are confident that women who are at risk of developing cervical cancer will greatly benefit from our recommended approach. With the necessary resources and support, we are enthusiastic about extending our assistance to underserved rural areas. Our proposed paradigm is

designed to be practical and enduring, making a lasting impact on healthcare accessibility and cervical cancer management.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary of the Study

In this thought-provoking essay, we harness the power of algorithms to gauge the impact potential of individuals, offering a reliable means to foresee future developments through our model. The diagnostic technique employed holds significant promise, enabling the prediction of an individual's potential influence. This foresight is not only beneficial for individuals who might mistakenly believe they need to be cervical cancer-aware, but also for understanding the various stages of cervical cancer. Our recommended methodology extends its advantages to the field of medical diagnosis, bolstered by the utilization of well-established, swift-to-implement, low-training, and highly accurate algorithms. This multifaceted approach empowers individuals and healthcare professionals alike, fostering a deeper understanding of health-related concerns and their potential impact.

6.2 Conclusions

In our present-day world, characterized by its blend of simplicity and advanced technology, access to cutting-edge innovations is virtually universal. Our proposal leverages this technological landscape to make the process of predicting cervical cancer disorder in individuals remarkably swift and straightforward. With the potential to benefit individuals worldwide, we are committed to ensuring the practicality and continued enhancement of our model, with plans to introduce additional features and address broader healthcare concerns in the future. This vision, founded on the current state of technology, sets the stage for an ever-evolving and progressive approach to healthcare and disease prediction.

6.3 Implication for Further Study

As humans, mortality is an inherent part of our existence, and we grapple with numerous illnesses daily. While cervical cancer disease affects many of us, some possess the tools to

combat it. Residing in a developing nation, we have access to advanced and precise diagnostic and therapeutic technologies. These advancements have streamlined the process of diagnosing cervical cancer illness, making it faster and more efficient. In our pursuit to provide innovation, we aspire to see our approach embraced by others. We have continually refined existing algorithms for enhanced performance and are committed to expanding our offerings in the future, fostering progress in healthcare and disease management.

Reference:

- [1] "Cervical Cancer Risk Classification", Last Accessed: December 2017, Available: <https://www.kaggle.com/datasets/loveall/cervical-cancer-risk-classification>.
- [2] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: theory and practice," *Neurocomputing*, vol. 415, pp. 295–316, 2020.
- [3] Juneja A. et al. "A survey on risk factors associated with cervical cancer." *Indian Journal of Cancer*, 40(1):15–22, 2003.
- [4] Ishrak Jahan Ratul et al. "Early risk prediction of cervical cancer: A machine learning approach." In 2022 19th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON). IEEE, 2022.
- [5] Ijaz Muhammad Fazal, Attique Muhammad, and Son Youngdoo. "Data-driven cervical cancer prediction model with outlier detection and over-sampling methods." *Sensors*, 20(10):2809, 2020.
- [6] Lilhore U. K. et al. "Hybrid model for detection of cervical cancer using causal analysis and machine learning techniques." *Computational and Mathematical Methods in Medicine*, 2022.
- [7] Yang W., Gou X., Xu T., Yi X., and Jiang M. "Cervical cancer risk prediction model and analysis of risk factors based on machine learning." In *Proceedings of the 2019 11th International Conference on Bioinformatics and Biomedical Technology*, pages 50–54, 2019.
- [8] Rothberg M. B. et al. "A risk prediction model to allow personalized screening for cervical cancer." *Cancer Causes Control*, 29:297–304, 2018.
- [9] Curia F. "Cervical cancer risk prediction with robust ensemble and explainable black boxes method." *Health and Technology*, 11(4):875–885, 2021.
- [10] Li Y. et al. "A large cohort study identifying a novel prognosis prediction model for lung adenocarcinoma through machine learning strategies." *BMC cancer*, 19:1–14, 2019.
- [11] Kourou K. et al. "Machine learning applications in cancer prognosis and prediction." *Computational and structural biotechnology journal*, 13:8–17, 2015.
- [12] Huang P. et al. "Prediction of lung cancer risk at follow-up screening with low-dose ct: a training and validation study of a deep learning method." *The Lancet Digital Health*, 1(7):e353–e362, 2019.
- [13] V. Lahoura, H. Singh, A. Aggarwal et al., "Cloud computing-based framework for breast cancer diagnosis using extreme learning machine," *Diagnostics*, vol. 11, no. 2, p. 241, 2021.
- [14] Nahar, Nazmun, and Ferdous Ara. "Liver disease prediction by using different decision tree techniques." *International Journal of DataMining & Knowledge Management Process* 8, no. 2 (2018): 01-09
- [15] Aljahdali, Sultan, and Syed Naimatullah Hussain. "Comparative prediction performance with support vector machine and random forest classification techniques." *International journal of computer applications* 69, no. 11 (2013).

- [16] L. Mary Gladence, M. Karthi, V. Maria Anu. "A statistical Comparison of Logistic Regression and Different Bayes Classification Methods for Machine Learning" ARPN Journal of Engineering and Applied Sciences, ISSN 1819-6608, Vol -10, No-14, August 2015.
- [17] "Logistic Regression for Machine Learning", Accessed: August 6, 2021, Available: <https://www.capitalone.com/tech/machine-learning/what-is-logistic-regression/>
- [18] Ghosh, Pronab, Asif Karim, Syeda Tanjila Atik, Saima Afrin, and Mohd Saifuzzaman. "Expert cancer model using supervised algorithms with a LASSO selection approach." International Journal of Electrical and Computer Engineering (IJECE) 11, no. 3 (2021): 2631
- [19] Shorove Tajmen, Asif Karim, Aunik Hasan Mridul, Sami Azam, Pronab Ghosh, Alamin Dhaly, Md Nour Hossain. "A Machine Learning based Proposition for Automated and Methodical Prediction of Liver Disease". In April 2022 The 10th International Conference on Computer and Communications Management in Japan.
- [20] Aunik Hasan Mridul, Md. Jahidul Islam, Mushfiqur Rahman, Mohammad Jahangir Alam, Asifuzzaman Asif. "A Machine Learning-Based Traditional and Ensemble Technique for Predicting Breast Cancer", In December, 2022. Conference: 22th International Conference on Hybrid Intelligent Systems (HIS 2022) online, 2022At: Auburn, Washington, USA.
- [21] Bentéjac, Candice, Anna Csörgő, and Gonzalo Martínez-Muñoz. "A comparative analysis of gradient boosting algorithms." ArtificialIntelligence Review 54, no. 3 (2021): 1937-1967.
- [22] hou, Zhi-Hua. Ensemble methods: foundations and algorithms. CRC Press, 2012.
- [23] emmens, Aurélie, and Christophe Croux. "Bagging and boosting classification trees to predict churn." Journal of Marketing Research 43, no. 2 (2006): 276-286.
- [24] Wang, Yizhen, Somesh Jha, and Kamalika Chaudhuri. "Analyzing the robustness of nearest neighbors to adversarial examples." In International Conference on Machine Learning, pp. 51335142. PMLR, 2018.
- [25] Drucker, Harris, Corinna Cortes, Lawrence D. Jackel, Yann LeCun, and Vladimir Vapnik. "Boosting and other ensemble methods." Neural Computation 6, no. 6 (1994): 1289-1301.
- [26] Sharma, Ajay, and Anil Suryawanshi. "A novel method for detecting spam email using KNN classification with spearman correlation as distance measure." International Journal of Computer Applications 136, no. 6 (2016): 28-35.
- [27] Pasha, Maruf, and Meherwar Fatima. "Comparative Analysis of Meta Learning Algorithms for Liver Disease Detection." J. Softw. 12, no.12 (2017): 923-933.
- [28] Islam, Rakibul, Abhijit Reddy Beeravolu, Md Al Habib Islam, Asif Karim, Sami Azam, and Sanzida Akter Mukti. "A Performance Based Study on Deep Learning Algorithms in the Efficient Prediction of Heart Disease." In 2021 2nd International Informatics and Software Engineering Conference (IISEC), pp. 1-6. IEEE, 2021.

Plagiarism Report:

A MACHINE LEARNING BASED APPROACH TO PREDICT CERVICAL CANCER

ORIGINALITY REPORT

18%	12%	9%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	5%
2	www.mdpi.com Internet Source	1%
3	Rithvik Hariprasad, T M Navamani, Tejas Ravindra Rote, Ishita Chauhan. "Design and Development of an Efficient Risk Prediction Model for Cervical Cancer", IEEE Access, 2023 Publication	1%
4	"Advances in Data and Information Sciences", Springer Science and Business Media LLC, 2024 Publication	<1%
5	Submitted to University of Greenwich Student Paper	<1%
6	Anagha Patil, Tanmay Arsanias, Rishabh Nahar, Anish Mohite, Aditya Trivedi. "Startup Funding App using Flutter and Machine Learning", 2023 International Conference on Advanced	<1%