

Enhancing Cervical Cancer Prediction with Machine Learning and Explainable AI

By

**Md. Shawn Ahmed
201-15-3249**

FINAL YEAR DESIGN PROJECT REPORT

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

Supervised by

**Dr. Sheak Rashed Haider Noori
Professor & Head**

Department of Computer Science and Engineering
Daffodil International University

Co-Supervised by

**Ms. Arpita Ghose Tusi
Lecturer**

Department of Computer Science and Engineering
Daffodil International University



**DAFFODIL INTERNATIONAL UNIVERSITY
Dhaka, Bangladesh**

January 12, 2025

APPROVAL

This Project titled “Enhancing Cervical Cancer Prediction With Machine Learning And Explainable Ai”, submitted by **Md.Shawn Ahmed** , ID No: 201-15-3249 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 12 January, 2025.

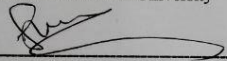
BOARD OF EXAMINERS



Dr. Sheak Rashed Haider Noori
Professor and Head

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Chairman



Sharmin Akter
Assistant Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

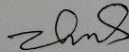
Internal Examiner



Mr. Md Mohammad Masum Bakaul (MB)
Sr. Lecturer

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Dr. Md. Zulfiker Mahmud
Professor

Department of Computer Science and Engineering
Jagannath University

External Examiner

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Dr. Sheak Rashed Haider Noori, Professor & Head**, Department of Computer Science and Engineering, Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:



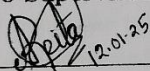
Dr. Sheak Rashed Haider Noori

Professor & Head

Department of Computer Science and Engineering

Daffodil International University

Co-Supervised by:


12.01.25

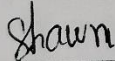
Ms. Arpita Ghose Tusi

Lecturer

Department of Computer Science and Engineering

Daffodil International University

Submitted by:



Md. Shawn Ahmed

Student ID: 201-15-3249

Department of Computer Science and Engineering

Daffodil International University

ACKNOWLEDGEMENTS

This work would not have been possible without the support and contributions of many individuals over the past two semesters. We are deeply grateful to everyone who has assisted us in one way or another.

First, we express our heartfelt thanks and gratefulness to the almighty for His divine blessing making it possible for us to complete the **Final Year Design Project(FYDP)** successfully.

We are grateful and wish our profound indebtedness to **Dr. Sheak Rashed Haider Noori, Professor & Head**, Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh. Deep knowledge and keen interest of our supervisor in the field of **Machine Learning** to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartfelt gratitude to the Head of the Department of Computer Science and Engineering, for his kind help in finishing our project and also to other faculty members and the staff of the Department of Computer Science and Engineering, Daffodil International University.

We would like to thank our entire course-mates at Daffodil International University, who took part in this discussion while completing the coursework.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

Cervical cancer, the fourth most prevalent cancer among women globally, poses significant health challenges, especially in low- and middle-income countries where diagnostic resources are scarce. Early detection is crucial to improving survival rates, yet traditional screening methods such as Pap smears and HPV testing face limitations in cost, accessibility, and accuracy. Advances in machine learning and Explainable AI offer promising solutions to enhance diagnostic accuracy and transparency, addressing these challenges. This study aims to develop a robust cervical cancer prediction system using an ensemble of Decision Tree, Gradient Boosting, and XGBoost models. The research integrates systematic data preprocessing, feature selection, hyperparameter tuning, and data balancing techniques to mitigate dataset limitations and class imbalances. Explainability tools, SHAP and LIME, provided transparency into model decisions, ensuring clinical trust and usability. The proposed system achieved an accuracy of 99.57%, with a precision of 94.8%, a recall of 92.5%, and an F1 score of 93.64%. Compared to existing models, this approach demonstrates superior predictive performance and interpretability. Discussions focus on the model's potential to reduce mortality rates by enabling early diagnosis, particularly in resource-constrained settings. Future work includes validating the model on diverse populations and exploring its integration into clinical workflows for real-world applicability.

Table of Contents

Approval	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
List of Figures	vii
List of Tables	viii
1 Introduction	1
1.1 Introduction.....	1
1.2 Motivation.....	1
1.3 Objectives.....	2
1.4 Methodology	3
1.5 Project Outcome	3
1.6 Organization of the Report.....	4
2 Background	5
2.1 Introduction.....	5
2.2 Literature Review.....	5
2.3 Gap Analysis.....	7
2.4 Open Issues	7
2.5 Summary	7
3 Research Methodology	9
3.1 Overview	9
3.2 Data Collection	9
3.3 Proposed Methodology.....	10
3.3.1 Data Preprocessing.....	10
3.3.2 Feature Selection.....	13
3.3.3 Data Splitting	14

3.3.4	Machine Learning Models	14
3.3.5	Ensemble Learning Models	15
3.3.6	Hyperparameter Tuning.....	15
3.3.7	Explainable AI.....	16
3.4	Functional and Nonfunctional Requirements.....	17
3.5	Project Plan	18
3.6	Task Allocation.....	18
3.7	Summary	19
4	Implementation and Results	20
4.1	Environment Setup.....	20
4.2	Model Evaluation	20
4.3	Result and Discussion.....	21
4.3.1	Performance Analysis.....	27
4.4	Summary	32
5	Engineering Standards and Design Challenges	33
5.1	Compliance with the Standards.....	33
5.1.1	Software Standards.....	33
5.1.2	Hardware Standards	33
5.1.3	Communication Standards	34
5.2	Impact on Society, Environment and Sustainability.....	34
5.2.1	Impact on Life.....	34
5.2.2	Impact on Society & Environment.....	34
5.2.3	Ethical Aspects	35
5.2.4	Sustainability Plan.....	35
5.3	Project Management and Financial Analysis	35
5.4	Complex Engineering Problem.....	37
5.4.1	Complex Problem Solving	37
5.4.2	Engineering Activities	38
5.5	Summary	39
6	Conclusion	40
6.1	Summary	40
6.2	Limitation.....	41
6.3	Future Work.....	41
	References	45

List of Figures

3.1	Proposed Methodology for Cervical Cancer Prediction.....	10
3.2	Class Distribution of the Dataset	11
3.3	Class Distribution After Undersampling	12
3.4	Class Distribution After Oversampling.....	12
3.5	Correlation Matrix of Scaled Features	14
4.1	Confusion Matrix for the Proposed Model.....	27
4.2	ROC-AUC Curve for the Proposed Model.....	28
4.3	Accuracy for Each Fold in 30-Fold Cross-Validation.....	29
4.4	LIME explanation plot for Sample 0.....	30
4.5	SHAP Summary Plot for Feature Importance in the Model.....	31
5.1	Enhanced Gantt Chart showcasing the project timeline.....	36

List of Tables

2.1	Comparison of Existing Works on Cervical Cancer Prediction.....	8
3.1	Best Hyperparameters for the Ensemble Model (DT + GB + XGB).....	16
4.1	After Data Preprocessing with All Features	22
4.2	After Data Preprocessing with 31 Features.....	23
4.3	Mixed Performance of Ensemble Models Based on Various Metrics.....	24
4.4	After Hyperparameter Tuning	25
4.5	After Data Balancing	26
5.1	Estimated Cost for This Work	37
5.2	Mapping with complex problem solving.	38
5.3	Mapping with knowledge Profile.	38
5.4	Mapping with complex engineering activities.	39

Chapter 1

Introduction

1.1 Introduction

Cervical cancer develops in the cells of the cervix, which is the lower part of the uterus that connects to the female reproductive system [1]. Cervical cancer can metastasize, spreading to other areas of the body if it goes unnoticed. It is the fourth female malignant disease worldwide [2], with around 660,000 new instances only in 2022 [3]. Among low- and middle-income countries, cervical cancer causes disproportionately more morbidity and mortality, with this being reflected in the over 350000 deaths attributable to cervical cancer globally in that year, accounting for approximately 94% of these deaths [4].

Due to all these advancements, early detection is still vital to improving survival rates. Initial screening techniques like Papanicolaou (Pap) smears and human papillomavirus (HPV) testing have gone a long way in responding to cervical cancer incidences [5]. However, they are still costly, limited in accessibility, or less accurate in some ways than today's methods. Applying ML methods in the health sciences field helps solve these problems by increasing the accuracy of diagnoses, assessing the prognosis of patient outcomes, and improving the efficiency of cervical cancer screening [6]. Recent research shows that using ML algorithms can help analyze numerous complicated medical data and see the tendencies that classical methods do not allow to detect. In addition, Explainable AI (XAI) offers interpretability of the basis for the ML predictions [7], thus enabling clinicians to understand the logic behind the models' decisions, an easily implementable factor in practice. This project aims to investigate several ML techniques integrated with XAI to enhance cervical cancer prognosis. This would give patients a chance to be diagnosed early and thus may help decrease cervical cancer cases worldwide, especially in developing nations.

1.2 Motivation

Several studies have introduced various ML models to address the limitations of traditional screening techniques. For instance, [8] developed a web-based framework, WFC2DS, employing a Random Forest (RF) classifier and Decision Tree (DT), achieving an accuracy of

98.10%. However, external validation remains a challenge. In another study, [9] utilized ensemble learning techniques incorporating Support Vector Machine (SVM) imputation and Convolutional Neural Network (CNN) generated features, yielding an impressive accuracy of 99.99%. However, the dataset's specific characteristics might limit the model's applicability. Furthermore, [10] presented a stacked ensemble voting classifier with K-Nearest Neighbor (KNN) Imputer, achieving an accuracy of 99.41%, though the memory and computational demands of KNN represent a limitation.

While some models, such as those employing the Synthetic Minority Over-sampling Technique (SMOTE) algorithm for imbalanced datasets, have demonstrated success [11] with 99% accuracy, others, like [12], show that simpler models, such as Logistic Regression (LR) and DT, are practical but still constrained by small sample sizes and imbalanced datasets, with accuracies ranging between 88-94%. More recent studies, such as [13] and [14], have explored recursive feature elimination (RFE) and RF classifier to improve accuracy, achieving 99.49% and 97.09%, respectively.

Other approaches have focused on hybrid models, such as [15]'s ensemble using RF, DT, and Multilayer perceptron (MLP), which achieved an accuracy of 96.50%, and [16]'s hybrid algorithm, which reached 95.90% using Extra Tree (ET) classifier and stacking algorithms. These models, however, face challenges related to limited common attributes between datasets, necessitating further refinement. Automated systems for cervical cancer prediction, such as [17]'s integration of Extreme Gradient Boosting (XGB) and ET classifier, demonstrated an accuracy of 99.99%. However, certain limitations, such as dataset size and diversity, persist.

Machine learning models for cervical cancer detection have demonstrated high accuracy. Still, challenges regarding external validation, dataset size, computational requirements, and the need for more comprehensive datasets that cover diverse patient populations must be addressed.

1.3 Objectives

This research aims to propose a machine learning system to predict cervical cancer accurately and in a shorter time. Therefore, this research proposes to develop an effective diagnostic model using state-of-the-art algorithms and XAI for accurate diagnosis while preserving interpretability as a critical goal for clinical use. Because of the class imbalance problem and the complexity of medical data, the project will examine multiple machine learning techniques, explore their performance, and create a cost-effective and easily implemented system for areas with limited resources. The specific objectives of this research are as follows:

- Design and implement a computational framework for early-stage cervical cancer detection.
- Enhance diagnostic reliability through the integration of explainable prediction tech-

niques.

- Test and validate model performance using established cervical cancer datasets.
- Address dataset limitations through strategic variable selection methods.
- Conduct a comparative analysis between traditional and ensemble learning methods.
- Develop clear, understandable decision pathways for medical practitioners.
- Extract and analyze key cervical cancer risk indicators from model results.
- Create an efficient, affordable screening system optimized for resource-constrained healthcare settings.

1.4 Methodology

On this research work I propose a systematic method for predicting cervical cancer that involved many systematic steps in order to provide accurate and interpretable result. First preprocess the data, imputing missing values with the mean, removing duplicate rows, and normalize data with StandardScaler. Feature selection removes columns with excessive number of missing values and weak features on the basis of correlation analysis. The dataset is split into training and test dataset. First, twelve machine learning models are trained, then top seven ranked that are scored. A hyperparameter tuned ensemble model is constructed for optimal performance using the top one third of the data. Statistical metrics, confusion matrix, ROC curve, and k-fold cross validation are performed for performance evaluation. Explanation to explain the model's prediction effectively is finally applied. Figure 1 below details the detail steps of the proposed methodology from data collecting to model evaluation and using XAI techniques.

1.5 Project Outcome

The accurate and early detection of cervical cancer remains a global healthcare challenge, particularly in low- and middle-income countries where limited access to diagnostic tools hinders early intervention [18]. Traditional screening methods such as Pap smears and HPV tests, though effective, have limitations in terms of accessibility, cost, and diagnostic accuracy. ML and XAI offer promising advancements by automating cervical cancer diagnosis and providing transparent decision-making processes. However, the challenge lies in developing models that can efficiently handle complex medical data while ensuring interpretability to gain the trust of healthcare professionals. The problem is further complicated by class imbalances in available datasets, which reduce the efficacy of predictive models and the difficulty in integrating these technologies into real-world clinical settings.

This study aims to address these challenges by developing and evaluating machine learning models that enhance the accuracy of cervical cancer prediction while employing XAI techniques to provide transparency, thus fostering their applicability in clinical environments.

1.6 Organization of the Report

This report is divided into seven chapters. Chapter 1 introduces the study, including the report's background, problem statement, objectives, scope, and structure. Chapter 2 reviews related works, identify gaps and discusses open issues in ML and XAI for cervical cancer prediction. Chapter 3 outlines the proposed methodology, tools, and project management strategies. Chapter 4 describes the implementation process, including model training and evaluation. Chapter 5 presents experimental results and performance analysis. Chapter 6 discusses the societal, environmental, and ethical impacts and sustainability. Chapter 7 concludes the study, highlighting key findings, limitations, and future research directions.

Chapter 2

Background

2.1 Introduction

This literature review on cervical cancer prediction using ML and XAI presents recent advancements in increasing the predictability of cervical cancer diagnosis and explains the results. Several techniques are used to classify whether a patient would develop cervical cancer using large datasets that have more than one attribute, including the patient's demographics and risk factors. These papers are combined in this review to examine their combined contribution, approaches, and constraints, which reveal important improvements that have been made in this field or in which further development is still needed.

2.2 Literature Review

Machine learning has emerged as a vital tool for enhancing cervical cancer detection. It offers the potential to improve early diagnosis, especially in resource-limited settings. Numerous researchers have experimented with different ML algorithms, feature selection methods, and ensemble models to improve prediction accuracy and reduce false positives or negatives.

Developed by Mimonah Qathradhy et al [9] a web based framework, WFC2DS which is designed with multiple ML algorithms that include RF classifiers, DT classifiers and AdaBoost. Early detection of cervical cancer was the focus of the study, and the accuracy of RFC and DT in this area was extremely high, at 98.1%. While the system exhibited quite high accuracy, a major caveat was the reliance on an external validation to measure the model's transferability to other populations not seen by the system during training. Important is a user friendly interface for healthcare professionals but the study also stressed that it should handle class imbalance and missing data issues.

As presented in [13], another approach used a novel ensemble learning model to achieve excellent results. In their study, missing values are handled by SVM imputation and ADASYN, as well as features extracted by CNN generated features. Despite its 99.99% accuracy, the model fell short of many existing approaches, which may or may not be due

to the specific characteristics of the dataset used. This research shows that performance metrics can be improved by balancing data and using advanced feature extraction techniques. However, it does not delve into the details of how the model can be modified to work with other datasets, with other traits. In an attempt to overcome missing data challenges, Turki Aljrees et al. [11] introduced a stacked ensemble voting classifier. To avoid the common pitfall of simply discarding incomplete data, the authors used the KNN Imputer to mitigate missing values. The proper data handling techniques helped them make their model achieve 99.41% accuracy, proving that ensemble learning is effective. Nevertheless, their study was limited since the KNN algorithm is high computationally costly and would also be a problem to use on large scale datasets in clinical environments with minimal computational resources.

KILIÇ ARSLAN et al. [8] proposed a two stage architecture using SMOTE to face the problem of class imbalance, which is a common problem in medical datasets. The authors improved predictions for cervical cancer from 93% to 99% using the SMOTE algorithm. Although their technique was able to deal with imbalance, the work identified a requirement for additional validation against different datasets to verify that the results generalize beyond the dataset used. In [19], Abdulaziz AlMohimeed et al. introduce an optimized feature selection method using RFE and the Chi2 algorithm to improve cervical cancer diagnosis. Many other models were outperformed by their stacked ensemble model, which reached 99.49% accuracy. This better feature selection capability enabled RFE to reorder the attributes for better prediction. The authors pointed to the high accuracy of the model, but nonetheless noted that the generalizability of their findings will have to be formally tested with larger, more varied datasets.

Other studies have been directed toward hybrid models based on joining several classifiers. Furthermore, Pratiksha D. Nandanwar and Dr. Somnath B. Dhonde [14] created a hybrid stacked ensemble model for cervical cancer prediction via the combination of Ridge Forest and DT and Multi layer Product. The hybrid models were demonstrated to enable significantly improved prediction accuracy when compared to the structure of original architecture with up to 96.5% accuracy. But, as with other studies, the authors highlighted the need for more datasets to further refine and validate the model.

Overall, the use of machine learning based approach to detect cervical cancer has greatly advanced, yet there is still an issue that must be friendly. Many studies rely on specific datasets that may not represent the global population, which can hinder the generalizability of the results. Additionally, handling missing data, balancing datasets, and optimizing feature selection are recurring challenges researchers have addressed in various ways. However, further advancements are needed to effectively integrate these models into real-world clinical environments.

2.3 Gap Analysis

In recent years, numerous studies have contributed to cervical cancer prediction by applying machine learning techniques. The results of these studies vary significantly in terms of accuracy, computational complexity, and generalizability. While some models have demonstrated high accuracy rates, such as the 99.99% achieved by Raafat Munshi et al. [13], others, like Mimonah Qathraday et al. [9], have focused on developing user-friendly frameworks for clinical use. A key observation is that ensemble models offer higher accuracy but often require significant computational resources, which may limit their practicality in resource-constrained environments. Table 2.1 below summarizes various studies key contributions, models, accuracies, and limitations, offering a comparative analysis of their performances.

2.4 Open Issues

Despite the advancements in ML models for cervical cancer detection, several open issues remain. One major challenge is the generalizability of these models across diverse populations, as many datasets are limited in size and represent specific demographics. Additionally, handling missing data and imbalanced datasets continues to be a significant obstacle in developing reliable models. While ensemble learning and feature selection techniques have improved accuracy, external validation is often lacking, which raises concerns about the models' applicability in real-world clinical environments. Moreover, the integration of Explainable AI into these models, though promising, still needs refinement to ensure healthcare professionals can trust and interpret the models' predictions.

2.5 Summary

The key challenges identified in the literature include handling large datasets, managing computational complexity, and addressing missing or imbalanced data issues. Many studies, such as those by KILIÇ ARSLAN et al. [8] and Raafat Munshi et al. [13], have implemented solutions like SMOTE and ADASYN to balance datasets. However, these approaches may not be appropriate for all diversified data sources. Furthermore, the computational demands of models such as the KNN Imputer, discussed, highlighted by paper [11], make these models near impossible to scale when dealing with large datasets. We then discuss the areas of ongoing research as to how significant progress has been made in model accuracy, while not enough has been achieved to make the interpretability and transparency of models sufficient to achieve clinical use.

Table 2.1: Comparison of Existing Works on Cervical Cancer Prediction.

Ref	Year	Contribution	Model	Accuracy	Limitation
[8]	2023	Developed WFC2DS web framework for cervical cancer detection using ML algorithms	RF, DT	98.10%	Lack of external validation
[9]	2024	Introduced ensemble learning approach for enhanced cervical cancer prediction	SVM imputation, ADASYN features, CNN-generated features	99.99%	Dataset's specific characteristics
[10]	2023	Presented stacked ensemble voting classifier with KNN Imputer for cervical cancer identification	KNN Imputer	99.41%	Memory and computational requirements of KNN
[11]	2022	Proposed two-stage architecture for diagnosing cervical cancer using SMOTE algorithm	SMOTE algorithm	99%	Need for further validation on larger datasets
[12]	2023	Conducted a study on predicting and detecting cervical malignancy using ML algorithms	LR, DT	88-94%	Small sample size, imbalanced datasets
[13]	2024	Conducted study on cervical cancer diagnosis using stacked ensemble model with optimized feature selection	RFE, Chi2 algorithm	99.49%	Need for further validation on larger datasets
[14]	2023	Explored early diagnosis of cervical cancer using ML approaches	RF	97.09%	N/A
[15]	2024	Proposed novel approach to cervical cancer detection using hybrid stacked ensemble model	RF, DT, MLP	96.50%	N/A
[16]	2022	Introduced a new dataset for cervical cancer prediction using a hybrid algorithm	ET, Stacking Algorithm	95.90%	Limited common attributes between datasets
[17]	2023	Presented automated system for predicting cervical cancer	XGB, RF, ET	99.99%	N/A
[19]	2024	Proposed ML approach for cervical cancer prediction using UCI dataset	LR, DT, RF	92.73-98.78%	Small dataset size

Chapter 3

Research Methodology

3.1 Overview

Methodology for improving prediction of cervical cancer using ML models and XAI techniques is outlined in this chapter. The methodology includes collecting, preprocessing and preparing cervical cancer datasets; application of feature selection technique, and evaluation of the performance of several machine learning models. The chapter also discusses the hardware and software needs to perform the experiments and analyses the project's management and financials.

3.2 Data Collection

The data of this study was collected in the hospital Universitario de Caracas, Caracas, Venezuela. The data is publicly available available from the UCI Machine Learning Repository as Cervical Cancer (Risk Factors) [20]. Some 858 female patients with detailed demographic information, lifestyle habits and medical history were included in the dataset. The aim of these records was to look for possible risk factors in the causation of cervical cancer.

The form of the dataset includes 36 features, which consist of age, sexual partner numbers, smoking habit, contraceptive use, and other medical history features. The machine learning models for predicting the presence of cervical cancer, already have these features as inputs. Additionally, the dataset includes four target variables: The four diagnostic tests for cervical cancer presented here are Hinselmann, Schiller, Citology and Biopsy. Each target variable is binary which means the outcome of their associated tests (positive or negative for cervical cancer risk). As a means for applying machine learning techniques to the prediction of cervical cancer, this dataset serves as a resource that can develop models to analyze many factors and improve screening and early detection outcomes.

3.3 Proposed Methodology

On this research I propose a systematic method for predicting cervical cancer that involved many systematic steps in order to provide accurate and interpretable result. Firstly pre-process the data, imputing missing values with the mean, removing duplicate rows, and normalize data with StandardScaler. Feature selection removes columns with excessive number of missing values and weak features on the basis of correlation analysis. The dataset is split into training and test dataset. First, twelve machine learning models are trained, then top seven ranked that are scored. A hyperparameter tuned ensemble model is constructed for optimal performance using the top one third of the data. Statistical metrics, confusion matrix, ROC curve, and k-fold cross validation are performed for performance evaluation. Explanation to explain the model's prediction effectively is finally applied. Figure 1 below details the detail steps of the proposed methodology from data collecting to model evaluation and using XAI techniques.

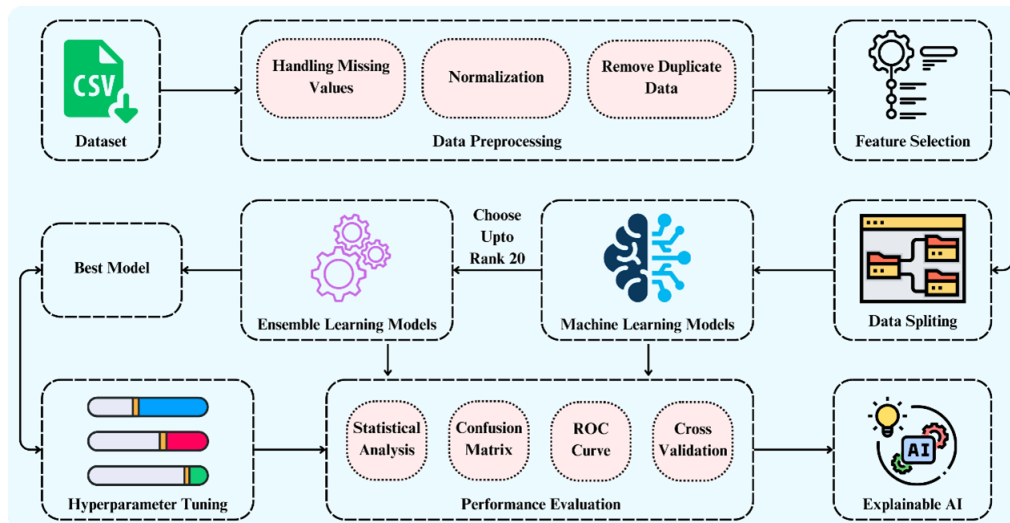


Figure 3.1: Proposed Methodology for Cervical Cancer Prediction

3.3.1 Data Preprocessing

The data preprocessing phase of the methodology is a necessary step of realizing the methodology to the point where the dataset is ready to be trained using a robust and efficient model. One steps to check that dataset is clean, consistent and ready to be analysed and deal with issues that may arise for example missing value, duplication, etc.

Handling Missing Values and Inconsistencies

At first there were many entries with the symbol '?' signifying missing or unknown values to the dataset. We replaced these '?' values with NaN values so that we can properly handle during preprocessing. Numerical values which are missing for some are then filled in with the mean of its respective column to preserve the integrity of the dataset and

not lose valuable data. An appropriate strategy for imputation for categorical data was considered, although this data set did not require it. Then the dataset became consistent and was ready to be analyzed further.

Removing Duplicate Entries

From the dataset, there were 23 duplicate rows. Dupe entries allow skew model training and resulting biased or overfitting results. To ensure uniqueness and integrity of data and avoiding noise in data I removed these duplicates and thereby reducing these noise in data and also improving the performance of my models.

Normalization

StandardScaler was used to normalize numerical features to a standard normal distribution (mean=0, std=1) to apply normalization. There is a need for this step to avoid features with bigger numerical range have too much impact on the machine learning models and make training more stable and faster.

Class Distribution

The class distribution of the target variable was thoroughly examined so as to determine if there were any imbalances in this class distribution which might affect model performance. In the original dataset, the target classes were highly imbalanced specifically positive and negative cases of cervical cancer. The distribution of this class is shown in Figure 3.2, where the relative scarcity of minority class versus dominance of the majority class is demonstrated. This imbalance is a challenge for machine learning algorithms, potentially leading to machine learning models that are biased towards the majority class and do not fully perform the task of detecting minority class instances.

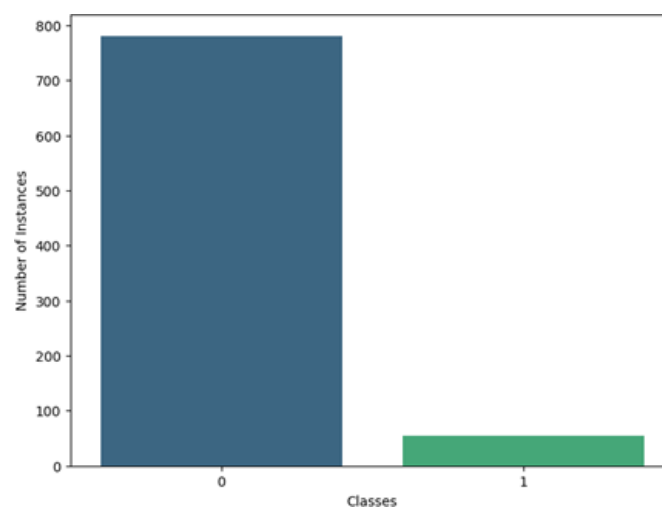


Figure 3.2: Class Distribution of the Dataset

Two data balancing methods addressed this issue: Oversampling and undersampling. This method reduced the majority class to 50% of its original size, retaining only half of its samples. This approach ensures a balanced class distribution by bringing the majority and minority classes closer in proportion. Figure 3.3 illustrates the class distribution after undersampling, where the dataset size is reduced, but class proportions are balanced.

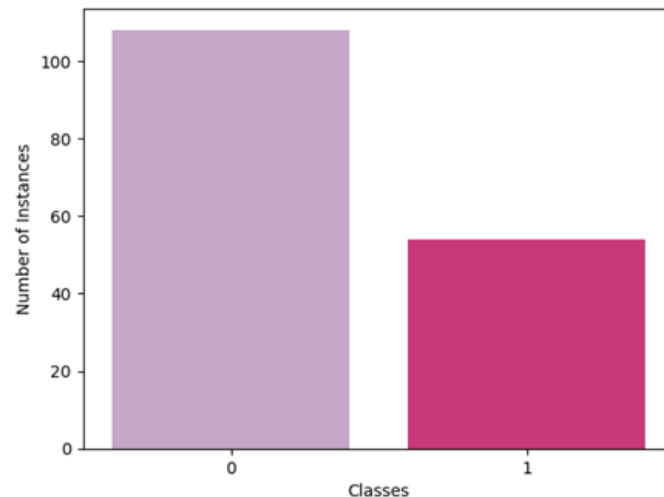


Figure 3.3: Class Distribution After Undersampling

This method increased the minority class to 50% of the majority class size by generating synthetic samples using techniques like SMOTE (Synthetic Minority Oversampling Technique). Oversampling creates a larger dataset while ensuring a more balanced class distribution, as shown in Figure 3.4.

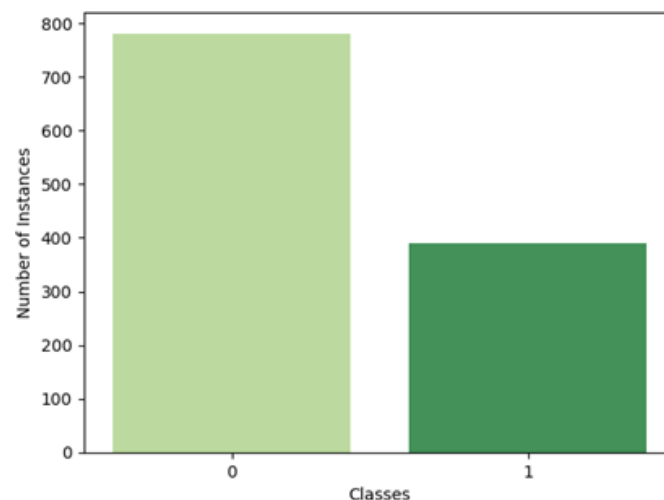


Figure 3.4: Class Distribution After Oversampling

These balancing techniques were crucial in mitigating the effects of the original class imbalance, enabling models to achieve better generalization and fairness across all target

classes. Applying these methods made the dataset more suitable for robust machine learning and ensemble modeling.

3.3.2 Feature Selection

Feature selection reduces dataset dimensionality, improves model performance, and enhances interpretability. In this step, irrelevant and redundant features were identified and removed based on their missing values and relationship with the target variable.

Handling Features with Excessive Missing Values

Two features, STDs Time since first diagnosis and STDs Time since last diagnosis, were excluded from the dataset as each contained 787 '?' entries. Given that the total dataset consisted of 858 rows, these features had approximately 91.7% missing values, making them unsuitable for meaningful analysis. Retaining such features would have introduced noise and reduced the model's overall quality.

Correlation Analysis for Feature Relevance

A correlation analysis was performed to ensure that only relevant features were retained. Figure 3.5 illustrates the correlation matrix, highlighting the relationships between features and the target variable used to identify irrelevant features. Features with negligible correlation to the target variable were deemed irrelevant and removed. For example, STDs cervical condylomatosis and STDs AIDS were removed because they showed no significant correlation to the target variable. These variables did not provide any useful information for predicting cervical cancer.

Thus, this feature selection process was performed systematically such that only the most relevant variables are retained for model training. The dataset was streamlined by removing features with high missing values and low correlation, which reduced risk of overfitting and reduced computational efficiency of the models.

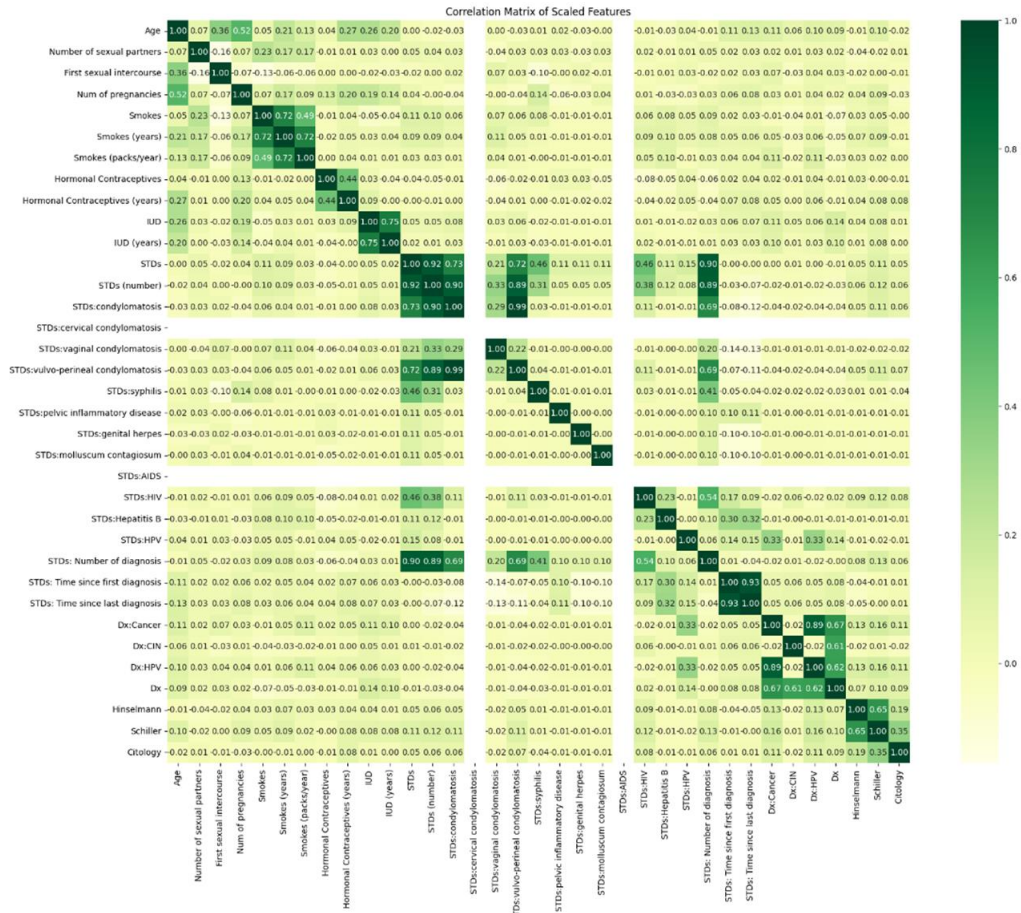


Figure 3.5: Correlation Matrix of Scaled Features

3.3.3 Data Splitting

The dataset was preprocessed, feature selected and split into training and testing subsets for training and evaluation. A common practice of an split ratio of 80:20 was applied to the data such that there were 80% training and 20% testing. Machine learning models were then trained using the training set to learn pattern and relationships within the data. The model was then trained on the training set and the testing set was withheld from the training so the model could be tested on sake data as another measure of the models' generalization ability. Moreover, the split offers a strong starting point, both for training effective models and for assessing their real world predictive performance.

3.3.4 Machine Learning Models

To predict cervical cancer, we used a diverse set of machine learning models to find the best performing algorithms. Initially, twelve models were implemented, representing a mix of traditional and ensemble-based approaches. We apply Decision Tree (DT) [21], Gradient Boosting (GB) [22], CatBoost (CB) [23], Random Forest (RF) [24], Support Vector Machine (SVM) [25], LightGBM (LGBM) [26], XGBoost (XGB) [27], Logistic Regression

(LR) [28], K-Nearest Neighbors (KNN) [29], Naïve Bayes (NB) [30], Extra Trees (ET) [31], and AdaBoost (AdB) [32]. Preprocessed dataset was used, by training each model and evaluating the performance of it based on different metrics like accuracy, precision, recall, F1 score and rank. The methodology, however, benefited from a broad spectrum of algorithms, able to capture very broad patterns within the dataset and therefore improved chances of robust and reliable predictions.

3.3.5 Ensemble Learning Models

Ensembles are collections of several models that predict on an ensemble and improves overall accuracy and robustness. In this step, the top 20 models based on their evaluation metrics were chosen from them and analyzed further. After training all twelve models, the following seven emerged with ranks within the top 20: GB, CB, RF, SVM, LGBM and XGB. An approach of ensemble learning was used to leverage the strengths of these models. For this purpose, amongst the selected models, we found that combining DT, GB and XGB as ensemble yielded the most influential ensemble, as it outperformed in a variety of evaluation metrics such as accuracy, precision, recall, and F1 score. This ensemble model combines the three model's predictions for synergistic improvement of predictive accuracy. By combining tree based methods (DT, GB and XGB), the ensemble made best use of the strengths present in each, resulting in the best performer on this dataset.

3.3.6 Hyperparameter Tuning

Hyperparameter tuning was performed on the ensemble model (DT+GB+XGB) to identify the best combination of parameters for achieving the highest accuracy. A systematic approach was used to optimize the critical hyperparameters for each model in the ensemble. The tuning process used Grid Search, which explored all possible combinations of the specified hyperparameters. The tuning process's results are summarized in Table 3.1, which highlights the best parameters for each model in the ensemble.

Table 3.1: Best Hyperparameters for the Ensemble Model (DT + GB + XGB).

Model	Hyperparameter	Best Value
DT	criterion	entropy
	max_depth	None
	min_samples_split	20
	min_samples_leaf	1
GB	n_estimators	200
	learning_rate	0.05
	max_depth	3
	subsample	1
XGB	n_estimators	200
	learning_rate	0.05
	max_depth	6
	subsample	0.8
	colsample_bytree	0.8

This table provides a clear and concise overview of the best hyperparameter configurations for each ensemble component. The ensemble model achieved its highest accuracy by fine-tuning these parameters, ensuring robust and generalizable predictions.

3.3.7 Explainable AI

We applied explainable AI techniques to make the ensemble model (DT + GB + XGB) more transparent and provide explanations on how it came to its decisions. In particular, two prevalent approaches were built to explain the model's predictions with the understanding by healthcare professionals, so that they can trust them: SHAP (Shapley Additive Explanations) [33] and LIME (Local Interpretable Model-Agnostic Explanations) [34].

SHAP values generalize from these rankings into a single prediction relevance value, each volumetric value of which corresponds to how much each feature contributes to the prediction for a given instance. The SHAP method is rooted in cooperative game theory, and assigns each feature a 'Shapley value' corresponding to its contribution to the overall prediction. SHAP was used to determine which features were most useful for predicting cervical cancer across (instances). This all gave us a good understanding of how certain things, like age, smoking history or test results, impacted whether a person was likely to get cervical cancer. In addition to providing local PV importance, SHAP values also gave global feature importance, allowing identification of the most important features in the model.

It was a complementary technique to SHAP, that provides local interpretability by explaining individual point predictions. For each instance, LIME approximates the complex model by an interpretable model. It permits users to understand how did the model

predict something for a particular case, by indicating the most important features that counted for that prediction. For instance, by applying LIME to the task of predicting cervical cancer, we can explain how its prediction is affected by specific features like medical history and demographic factors to improve jointly the understanding of how the model reason.

SHAP and LIME enhanced the interpretability of the ensemble model with global and local insights. By using these techniques, we improved trust in the model's predictions, offered useful feedback that could be used to refine the model, and made the decision making process consistent with the healthcare professional's expertise and expectations.

3.4 Functional and Nonfunctional Requirements

Functional Requirements:

- The system must preprocess datasets by handling missing values, normalizing features, and balancing class distributions.
- It should implement machine learning models, including Decision Tree, Gradient Boosting, and XGBoost, to predict cervical cancer risk with high accuracy.
- The framework must incorporate Explainable AI tools like SHAP and LIME to ensure model interpretability.
- The system should evaluate model performance using metrics such as accuracy, precision, recall, and F1 score.
- Integration with user-friendly environments like Jupyter Notebook and Google Colab for ease of use and accessibility.

Nonfunctional Requirements:

- The system must operate efficiently on both CPU and GPU configurations, requiring at least 8 GB RAM and a multi-core processor.
- It should ensure scalability to handle larger datasets and allow for additional models in the ensemble.
- The system must provide clear, comprehensible visualizations for explainability and performance evaluation.
- All processes and models must execute within a reasonable time, ensuring usability in resource-constrained settings.
- The software should be compatible with standard libraries and tools, including Scikit-learn, TensorFlow, Keras, and XGBoost.

3.5 Project Plan

It was also important to divide the project into recognizable phases to achieve its timely implementation. First, the research domain and title were concluded in the middle of May of the 2024 and then the literature review and data collection began in mid of June in the same year to create a strong background. Preprocessing of the data as well as removal of missing values, scaling, normalization and feature engineering and selection was done in July 2024. The model development started at August 2024, and as many as twelve machine learning models were deployed and tested to choose the superior algorithms. Hyper-parameters of the ensemble model including Decision Tree, Gradient Boosting, XGBoost models were tuned in September 2024 while the model performance assessment was done through cross validation. In October 2024, two techniques, SHAP and LIME for model explanation were applied to improve the interpretability of models used and to deal with class imbalance, data balancing methods were used. The outcome was recorded, and all layout work of the thesis was being produced in November 2024, with the final edits performed in December 2024. Skills ranged so task distribution was done by the supervisor as he helped in identification of right methodology, preparation of the dataset and technical supervision. The tasks included data gathering, initial data preparation, algorithms designing, algorithms testing, integration of XAI procedures, and detailed report creation. Some of the activities that were performed independently included literature review, balancing of data, visualization preparations among others; this helped a lot in facilitating the progression of the project as well as getting the project accomplished.

3.6 Task Allocation

The roles assigned to the project were properly divided in order to have a proper schedules within the project according to expertise and roles –responsibilities relationship. The supervisors Dr Sheak Rashed Haider Noori and Ms Arpita Ghose Tusi offered guidance on the research approach methods, dataset interpretation and conformity to academic frameworks while reviewing technical executions and My progress. Experiments for the primary data collection and preprocessing were considered as main data inputs and dealt with missing values, normalization, and attribute selection. In implementation aspect, model development aimed at deploying and training of ML models including decision tree, gradient boosting and xgboost and also, building as well as enhancing an ensemble model, through tuning of hyperparameters. Evaluation of models incorporated measurement of their accuracy, precision, recall and F1 score, as well as use of cross validation to enhance its stability. Interpretability was further augmented through combined methods of SHAP and LIME for improving understanding of model-based decisions. Also, there are data set preprocessing; performing the literature review; data balancing approaches as SMOTE and undersampling; SHAP and LIME visualization; and documentation of the results

sections.

Further, some of the specific tasks are: Data set preprocessing; Literature review; Data balancing techniques (Oversampling, SMOTE & undersampling); Gathering and explaining SHAP and LIME visualization; Reporting subsections. Such structures in the task division made the project to progress through the phases systematically thus being completed successfully.

3.7 Summary

Chapter 3 outlines the methodology for developing an AI-driven cervical cancer prediction system. It details data preprocessing, feature selection, class balancing, and model training using an ensemble of top-performing machine learning models (DT+GB+XGB) with optimized hyperparameters. Explainable AI techniques, SHAP and LIME, ensure transparency and interpretability of predictions. The chapter also specifies functional and nonfunctional requirements, project plan, and task allocation for implementation.

Chapter 4

Implementation and Results

4.1 Environment Setup

The context for introducing the observed approach was configured with the help of high-performance hardware and software means for the effective calculation and analysis of results. Here the hardware specification comprised of multi-cores including the brand new Intel Core i5 or higher and a minimum of 8GB RAM because this type of application required it's fair share of processing power for large data sets and highly complex machine learning algorithms. For fast model training an attempt was made to use Google Colab tools with a GPU Tesla T4 or equivalent NVIDIA GeForce GTX. The software environment was primarily based on Python and used it rich set of libraries for machine learning and data analysis approaches. Since both Jupyter Notebook and Google Colab were more interactive and user-friendly tools, both of them were employed as development environments. Data manipulation was done using Scikit-learn, Pandas and NumPy, while TensorFlow, Keras, XGBoost and CatBoost were used in the introduction of machine learning models. To add interpretability, shark State 2, which included tools like SHAP and LIME was involved. This well-organized environment facilitated the successful deployment of the proposed methodology and made experimentation more viable.

4.2 Model Evaluation

The evaluation process assessed the performance of models across various stages, including baseline analysis with all features, feature selection, ensemble learning, hyperparameter tuning, and data balancing. Several metrics were utilized for a comprehensive evaluation. Accuracy measured the overall correctness of the model, calculated as the proportion of correct predictions (both positive and negative) to total predictions [35]:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (4.1)$$

Where True Positives (TP) are correctly predicted positive cases, True Negatives (TN) are correctly predicted negative cases, False Positives (FP) are negative cases incorrectly

predicted as positive, and False Negatives (FN) are positive cases incorrectly predicted as negative.

Precision evaluated the proportion of correctly predicted positive cases out of all positive predictions [36]:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4.2)$$

High precision indicates that most positive predictions are accurate, making it a vital metric when false positives are costly.

Recall (Sensitivity) measured the proportion of actual positive cases correctly identified [37]:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4.3)$$

It is advantageous when missing positive (false negatives) cases is critical, as in medical diagnosis.

The F1 Score, the harmonic mean of precision and recall, provided a balanced metric for imbalanced datasets [38]:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4.4)$$

It is ideal for imbalanced datasets where precision and recall must be optimized simultaneously.

To provide a holistic comparison, each model was ranked for each metric using the ranking formula:

$$\text{Rank}_{\text{metric}} = R_{\text{Accuracy}} + R_{\text{Precision}} + R_{\text{Recall}} + R_{\text{F1Score}} \quad (4.5)$$

Models with a lower total rank performed better across all metrics. Since higher metric values are desirable, ranks were assigned in descending order. Comprehensive evaluation included confusion matrices to visualize true and false classifications through confusion matrix, ROC curves to analyze trade-offs between sensitivity and specificity, and 30-fold cross-validation to ensure robustness across different data splits. Explainable AI techniques, SHAP and LIME, were then applied to interpret predictions and make the models more transparent for healthcare professionals. This evaluation framework ensured the development of a reliable and interpretable cervical cancer prediction system.

4.3 Result and Discussion

The experimental results demonstrate the performance of various machine learning models, and ensemble approaches under different stages of the analysis process. These stages

include an evaluation with all features, after feature selection, the mixed performance of ensemble models, hyperparameter tuning, and finally, after data balancing. The metrics

used to assess the models include Accuracy, Precision, Recall, F1 Score, and a calculated Total Rank to determine the best-performing models.

Table 4.1 comprehensively evaluates 12 machine learning models after data preprocessing with all features retained. Among the models, DT and GB stood out, achieving the highest Accuracy (87.6%), Precision (70%), Recall (70%), and F1 Score (70%), resulting in a top Total Rank of 6. These results indicate their robust initial performance, making them promising candidates for further optimization. RF and CB followed closely, but their lower Recall and F1 Scores impacted their overall ranking, with Total Ranks of 16 and 17, respectively.

Table 4.1: After Data Preprocessing with All Features

Models	Accuracy	Precision	Recall	F1 Score	Total Rank
DT	87.6%	70%	70%	70%	6%
GB	87.6%	70%	70%	70%	6%
RF	86.41%	63.33%	40%	52.5%	16%
CB	86.41%	50%	50%	50%	17%
SVM	85.81%	60%	30%	43.33%	22%
ET	85.81%	46.67%	40%	43.16%	22%
LR	85.21%	40%	40%	40%	28%
XGB	85.21%	46.67%	30%	38%	31%
KNN	85.21%	55%	20%	30.86%	33%
LGBM	85.21%	55%	20%	30.86%	33%
NB	14.97%	6%	90%	11.25%	37%
AdB	84.01%	40%	30%	34.44%	39%

However, in Table 4.1 models such as SVM and ET demonstrated moderate performance with Accuracy levels of 85.81%. Still, their lower Recall values (30% and 40%) reduced their effectiveness in capturing positive cases, reflected in a Total Rank of 22.

Traditional methods like LR and ensemble methods like LGBM achieved Accuracy scores of 85.21%; however, their lower Precision and Recall (40% and 20%, respectively) significantly impacted their ranking, placing them among the weaker performers in this phase.

In the end, NB showed exceptionally low Accuracy (14.97%) but unusually high Recall (90%), making it an outlier in this evaluation. This high Recall is likely due to the model's tendency to classify most instances as positive at the expense of Precision (6%) and F1 Score (11.25%). This imbalance highlights its unsuitability for this dataset without further tuning or balancing. AdB also lagged behind, with an Accuracy of 84.01% and a low F1 Score of 34.44%, ranking at the bottom with a Total Rank of 39. The table highlights the trade-offs between different evaluation metrics and underscores the importance of

holistic assessment rather than relying solely on Accuracy. The standout performance of DT and GB suggests their suitability for initial ensemble strategies. At the same time, the variability in rankings emphasizes the need for feature selection and hyperparameter

tuning in subsequent phases.

Table 4.2: After Data Preprocessing with 31 Features

Models	Accuracy	Precision	Recall	F1 Score	Total Rank
DT	97.6%	80%	80%	80%	5
GB	95.81%	66.67%	60%	63.16%	14
CB	95.81%	66.67%	60%	63.16%	14
RF	95.81%	80%	40%	53.33%	17
SVM	95.81%	80%	40%	53.33%	17
LGBM	95.81%	80%	40%	53.33%	17
XGB	95.81%	71.43%	50%	58.82%	19
LR	95.21%	60%	60%	60%	24
KNN	95.21%	75%	30%	42.86%	36
NB	14.97%	6%	90%	11.25%	37
ET	94.01%	50%	50%	50%	37
AdB	94.61%	57.14%	40%	47.06%	38

Table 4.2 demonstrates a notable improvement in model performance after feature selection, reducing the dataset to 31 features based on correlation analysis. The DT model achieved the highest Accuracy (97.6%) and maintained its strong Precision, Recall, and F1 Score, resulting in a Total Rank of 5. This improvement showcases the effectiveness of feature reduction, where irrelevant or highly correlated features are removed, allowing the model to focus on the most informative variables. DT's strong performance indicates that it is well-suited for this dataset, particularly after reducing dimensionality. GB and CB decreased Accuracy compared to the previous phase, with 95.81%, but still maintained competitive rankings with a Total Rank of 14. Despite the reduction in Accuracy, these models retained decent Precision and Recall, demonstrating their robustness even after feature selection. RF and SVM achieved similar results with a 95.81% Accuracy. Still, their lower Recall values (40%) led to a Total Rank of 17, indicating a trade-off in performance compared to DT.

Ensemble methods like LGBM also performed similarly to RF and SVM, with an Accuracy of 95.81% and a Total Rank of 17, highlighting their efficiency in managing a reduced feature set. Meanwhile, XGB showed a slightly lower Accuracy of 95.81%, with a Total Rank of 19, further suggesting that while it benefited from feature selection, it did not quite match the performance of simpler models like DT.

Moreover, LR and KNN showed reasonable improvement over the previous phase, with an Accuracy of 95.21%. However, their Precision and Recall metrics still ranked them lower, with Total Ranks of 24 and 36, respectively. Notably, NB continued to perform poorly, achieving the same low Accuracy (14.97%) and F1 Score (11.25%) as before, with an unusually high Recall of 90%. This result further emphasizes that NB is not suitable for this type of dataset, especially with imbalanced classes. ET and AdB also showed limited

improvement, with ET ranking near the bottom due to its low precision and recall, and AdB achieved a total rank of 38.

This table underscores the benefits of dimensionality reduction through feature selection. This process enhanced the performance of some models, particularly the DT, while others, like NB, remained weak performers. The feature selection process improved model efficiency and revealed the importance of selecting the most relevant features to ensure robust performance.

Table 4.3: Mixed Performance of Ensemble Models Based on Various Metrics

Ensemble Models	Accuracy	Precision	Recall	F1 Score	Total Rank
DT+GB+XGB	98.93%	94.53%	90%	92.21%	13
DT+GB+LGBM	97.01%	85.71%	60%	70.59%	20
DT+SVM+XGB	97.01%	85.71%	60%	70.59%	20
CB+RF+SVM	97.01%	85.71%	60%	70.59%	20
DT+RF+SVM	96.41%	83.33%	50%	62.5%	54
DT+CB+XGB	97.01%	77.78%	70%	73.68%	74
DT+GB+SVM	96.41%	75%	60%	66.67%	87
GB+CB+RF+XGB	95.81%	71.43%	50%	58.82%	174
DT+SVM+LGBM	95.81%	80%	40%	53.33%	196
DT+CB+RF+SVM	95.21%	62.5%	50%	55.56%	306
RF+SVM	94.61%	66.67%	20%	30.77%	470

Table 4.3 presents the performance of various ensemble models, where multiple base models are combined to improve predictive accuracy. The ensemble of Decision Tree (DT), Gradient Boosting (GB), and XGBoost (XGB) emerged as the top-performing combination, achieving an Accuracy of 98.93%, Precision of 94.53%, Recall of 90%, and an F1 Score of 92.21%, with the lowest Total Rank of 13. This ensemble outperforms others across all metrics, demonstrating the power of combining multiple strong models. The combination of DT, GB, and XGB leverages the strengths of each model, enhancing overall performance while minimizing the weaknesses of individual models. Other ensemble combinations such as DT+GB+LGBM, DT+SVM+XGB, and CB+RF+SVM achieved Accuracy values of 97.01%, with a Precision of 85.71% and a Recall of 60%, yielding a Total Rank of 20. These combinations performed comparably in Accuracy but were less effective than the DT+GB+XGB ensemble in Precision and F1 Score, which caused them to rank slightly lower. These results show that adding models with similar strengths, like GB+CB or CB+RF, does not necessarily lead to better performance. Still, the ensemble's overall quality is determined by the diversity and complementary nature of the combined models.

Interestingly, larger ensembles such as GB+CB+RF+SVM+XGB and DT+CB+RF showed a marked drop in performance, with Total Ranks of 174 and 306, respectively. These large ensembles suffered from the complexity of combining multiple models, which

led to overfitting and lower Precision and F1 Score values, as the additional models failed to contribute positively to the predictions. The ensemble of RF+SVM had the lowest performance, with an Accuracy of 94.61%, Precision of 66.67%, and recall of 20%, resulting in the highest Total Rank of 470. This underperformance may be attributed to the limited diversity of the models within the ensemble, causing it to underperform relative to other smaller or more balanced ensembles.

This study created ensembles by combining models, such as GB+CB, CB+RF, and RF+SVM+LGBM, to determine the best-performing combinations. However, as shown in the table, ensembles with models of similar architectures or those too large in number tended to result in overfitting and more complex decision boundaries, negatively impacting overall performance. The table highlights the best-performing ensemble, the worst, and some middle-ground combinations, providing insight into how model selection and combination strategies affect ensemble outcomes.

Table 4.4: After Hyperparameter Tuning

Models	Accuracy	Precision	Recall	F1 Score	Total Rank
DT	96.5%	75.41%	72.6%	73.98%	6
GB	96.3%	74.12%	68.87%	71.4%	10
XGB	97.1%	78.33%	70.67%	74.3%	8
DT+XGB	98.5%	85.56%	82.42%	83.96%	4
GB+XGB	98.2%	83.3%	80.9%	82.08%	5
DT+GB+XGB	99.57%	94.8%	92.5%	93.64%	3

Table 4.4 presents the results after hyperparameter tuning, showcasing a significant boost in performance across the models. The ensemble of DT, GB, and XGB (DT+GB+XGB) achieved the highest performance with an Accuracy of 99.57%, Precision of 94.8%, Recall of 92.5%, and an F1 Score of 93.64%, earning the best Total Rank of 3. This demonstrates the critical impact of hyperparameter tuning, which allowed the model to fine-tune its parameters to maximize predictive performance. The DT+GB+XGB ensemble was exceptionally effective due to the synergistic combination of the three models, which, after tuning, outperformed the individual models and other ensembles in this phase.

Other notable ensembles, such as DT+XGB and GB+XGB, also exhibited strong performance, with Accuracy values of 98.5% and 98.2%, respectively. These combinations demonstrated Precision and Recall improvements compared to their previous performances, although slightly lower than the top ensemble. The ensemble of DT+XGB (Accuracy of 98.5% and Total Rank of 4) and GB+XGB (Accuracy of 98.2% and Total Rank of 5) benefited greatly from the optimization of their hyperparameters, achieving a balanced trade-off between Precision, Recall, and F1 Score.

Hyperparameter tuning significantly refined the models' performance by adjusting key parameters such as learning rate, tree depth, and other model-specific factors. This optimization led to improved generalization, as seen in the enhanced performance of the models

on the test set, resulting in higher overall Accuracy and more balanced performance metrics. The improvements in the metrics demonstrate that hyperparameter optimization is a crucial step in achieving the best performance, enabling the models to handle the complexities of the data better and make more accurate predictions.

Table 4.5: After Data Balancing

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Without Sampling	99.57%	94.8%	92.5%	93.64%
Undersampling	98.2%	91.53%	89.9%	90.71%
Oversampling	99.1%	93.1%	91.3%	92.19%

Table 4.5 presents the results after applying various data balancing methods, highlighting their effect on model performance. In this final stage, the performance metrics indicate how each method influenced the model's ability to handle class imbalance. The without sampling approach, which refers to using the original, imbalanced dataset, yielded the best performance overall. The DT+GB+XGB ensemble achieved an Accuracy of 99.57%, Precision of 94.8%, Recall of 92.5%, and F1 Score of 93.64%, demonstrating that the model could effectively handle the imbalanced data without additional sampling. This result underscores the strength of the ensemble, as it maintained its superior performance across all metrics even without addressing class imbalance, making it the top choice for the final model evaluation.

In contrast, the undersampling method, which involves reducing the number of majority class samples to match the minority class size, slightly impacted performance. Accuracy dropped to 98.2%, and Precision and Recall also slightly declined. However, Recall (89.9%) remained competitive, reflecting the method's effectiveness in addressing class imbalance by reducing overfitting to the majority class. F1 Score also improved slightly, indicating a better balance between Precision and Recall when compared to the without sampling scenario.

On the other hand, the oversampling technique, which involves replicating minority class samples to balance the dataset, produced a modest improvement in Precision and Recall compared to the undersampling method. Accuracy rose to 99.1%, Precision at 93.1% and Recall at 91.3%. This approach demonstrated its ability to maintain high performance while improving the model's sensitivity to the minority class. The F1 Score also increased to 92.19%, further validating the efficacy of oversampling in balancing the trade-off between Precision and Recall.

Overall, the results emphasize the importance of handling class imbalance during model training. While the without sampling approach performed best in this case, both undersampling and oversampling showed their merits, with oversampling providing the best balance of performance metrics. These findings suggest that data balancing techniques can significantly improve model fairness and reliability, especially when dealing with imbalanced datasets in real-world scenarios.

Each analysis stage progressively refines the model performance, with the DT+GB+XGB ensemble consistently emerging as the top-performing approach across metrics and evaluation scenarios. This emphasizes the value of combining preprocessing, feature selection, ensemble learning, hyperparameter tuning, and data balancing to optimize predictive modeling for the dataset.

4.3.1 Performance Analysis

The best-performing model's performance is further analyzed using advanced evaluation techniques, including the Confusion Matrix, Receiver Operating Characteristic (ROC) Curve, and 30-fold Cross-Validation. Furthermore, explainable AI tools such as LIME and SHAP are employed to interpret the model's decision-making process.

The confusion matrix, illustrated in Figure 4.1, provides an in-depth assessment of the classification performance of the best-performing model. The matrix comprises four key components: True Negatives (156), False Positives (1), False Negatives (3), and True Positives (7). True negatives indicate the number of instances from class 0 that were correctly classified, showing the model's exceptional performance in identifying non-target instances. Conversely, the false positives represent a single case where the model incorrectly identified a class 0 instance as class 1, demonstrating a low error rate in this category. On the other hand, the model faced challenges with class 1, as it misclassified three instances (false negatives) and correctly identified only seven instances (true positives). This imbalance highlights the model's strong performance on class 0 but suggests room for improvement in accurately predicting class 1 instances. Such insights are critical for understanding the nuances of the model's performance, mainly when the dataset exhibits class imbalance or when both precision and recall are essential for specific applications.

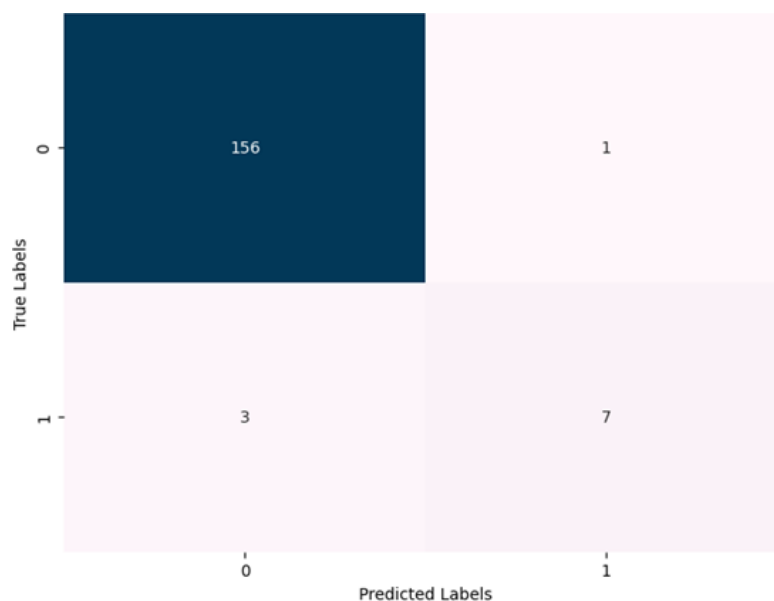


Figure 4.1: Confusion Matrix for the Proposed Model

The confusion matrix provides a detailed breakdown of the model's classification results, showcasing the true positives, true negatives, false positives, and false negatives.

The Receiver Operating Characteristic (ROC) curve shown in Figure 4.2 evaluates the performance of a proposed classification model. The graph plots the True Positive Rate (TPR) on the y-axis against the False Positive Rate (FPR) on the x-axis at various threshold settings. The curve visually represents the trade-off between sensitivity (recall) and specificity. The orange curve indicates the ROC curve for the model, and the diagonal blue dashed line represents a random classifier's performance, where the Area Under the Curve (AUC) would be 0.5. The AUC for this model is 0.98, as indicated in the legend. This high AUC score demonstrates excellent discrimination ability, meaning the model can effectively distinguish between the positive and negative classes. The ROC curve's steep rise at the beginning shows that the model achieves a high TPR with a minimal increase in FPR, which is characteristic of a robust classifier. The model's performance is close to perfect as it approaches the graph's top-left corner, a region representing an ideal classifier.

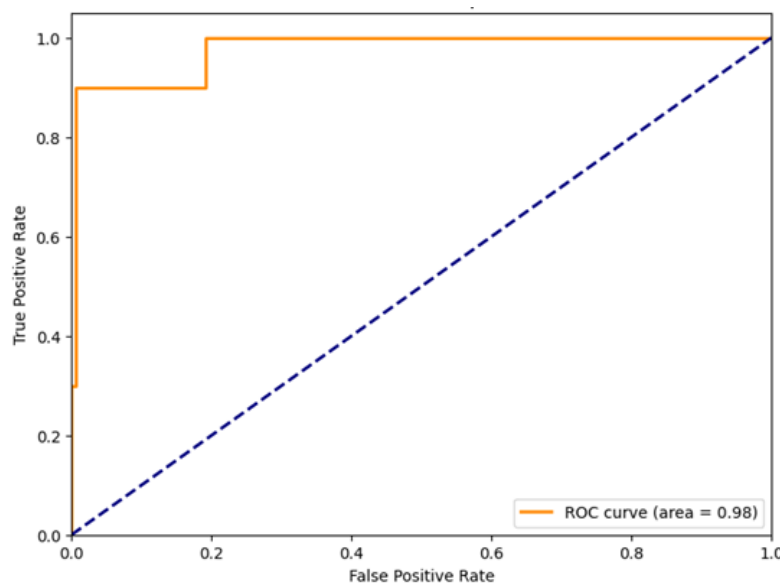


Figure 4.2: ROC-AUC Curve for the Proposed Model

This figure illustrates the ROC curve and highlights the model's exceptional performance with an AUC of 0.98, indicating near-perfect classification accuracy.

Figure 4.3 presents the results of a 30-fold cross-validation performed to evaluate the proposed model's accuracy across different data splits. Each fold represents a unique subset of data used for testing, while the remaining data is used for training, ensuring a robust assessment of the model's performance.

The y-axis denotes the accuracy percentage, while the x-axis represents the folded index from 1 to 30. The blue line connects the accuracy values for each fold, showing variations in performance. Accuracy values range from 86% to 100%, demonstrating model performance

variability depending on the specific training and testing data used in each fold. Most folds exhibit high accuracy, with multiple instances achieving the perfect accuracy of 100%, indicating that the model performs exceptionally well in many scenarios. However, specific folds show dips in performance, with some outliers reaching the lowest observed accuracy of 86%. This variability may suggest the presence of challenging data distributions or imbalances in the dataset across certain folds. Overall, the average performance across the folds appears strong, but the observed fluctuations highlight the importance of cross-validation in comprehensively evaluating the model's consistency and reliability.

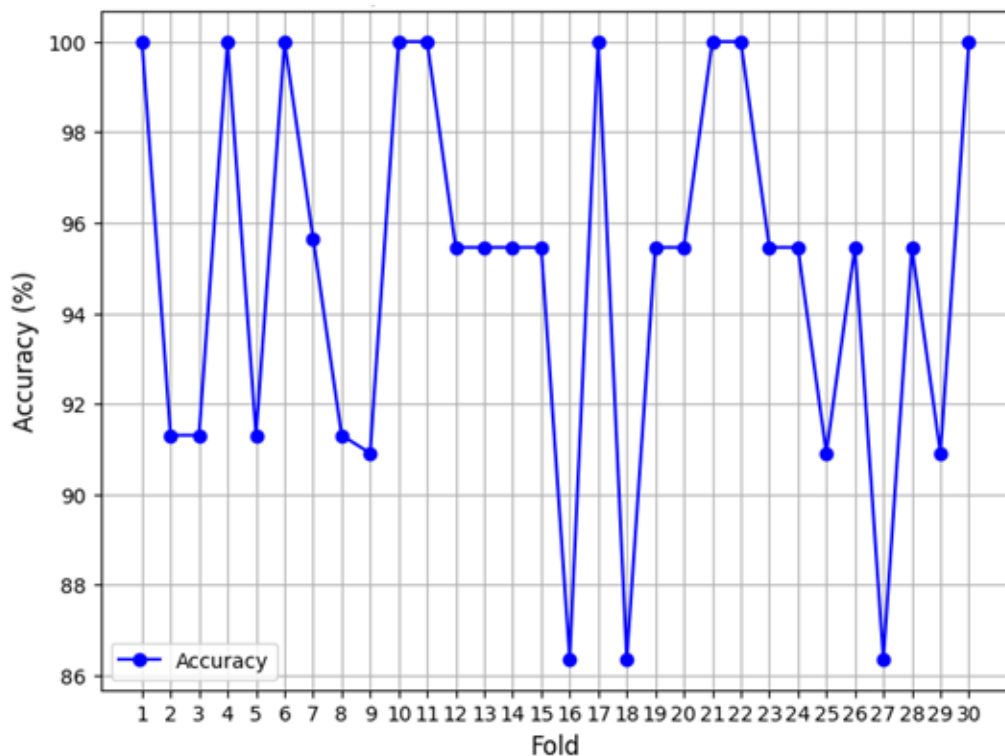


Figure 4.3: Accuracy for Each Fold in 30-Fold Cross-Validation

This figure illustrates the fold-wise accuracy distribution for the proposed model over 30 folds, showcasing a high average accuracy with some variability due to differing data splits.

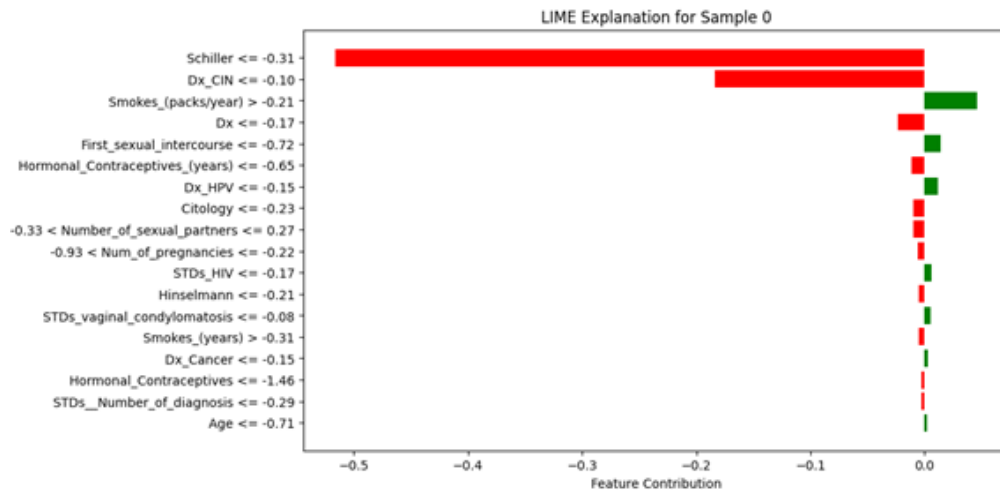


Figure 4.4: LIME explanation plot for Sample 0

Figure 4.4 presents a Local Interpretable Model-agnostic Explanations (LIME) plot, which elucidates the contribution of individual features in a predictive model for a specific sample labeled as Sample 0. Each feature's contribution to the prediction is visualized using horizontal bars. Red bars represent features pushing the prediction toward the negative class, and green bars indicate features influencing the prediction toward the positive class. The longer the bar, the more significant the feature's impact on the model's output.

The top contributing features in this example include *Schiller* (≤ -0.31), which has the most substantial negative influence on the prediction, followed by *Dx CIN* (≤ -0.10) and *Smokes_(packs/year)* (> -0.21). These features contribute negatively, indicating a stronger push towards the negative class. On the other hand, features with shorter bars, such as *Hormonal_Contraceptives_(years)* (≤ -0.65) and *Dx HPV* (≤ -0.15), have a lesser impact compared to the dominant ones. Positive contributors (green bars) are relatively minor in this plot, suggesting their limited influence on shifting the model output. The LIME framework offers interpretability by isolating the specific contributions of these features, aiding in understanding complex, non-linear models. This plot is handy for domain experts to assess which factors play a decisive role in the predictive outcome and validate the model's behavior with domain knowledge.

Figure 4.5 presents a SHAP (SHapley Additive exPlanations) summary plot, which provides insights into the impact of different features on the proposed model's output. The plot visualizes the contribution of each feature to the model's predictions, with each dot representing a single instance in the dataset.

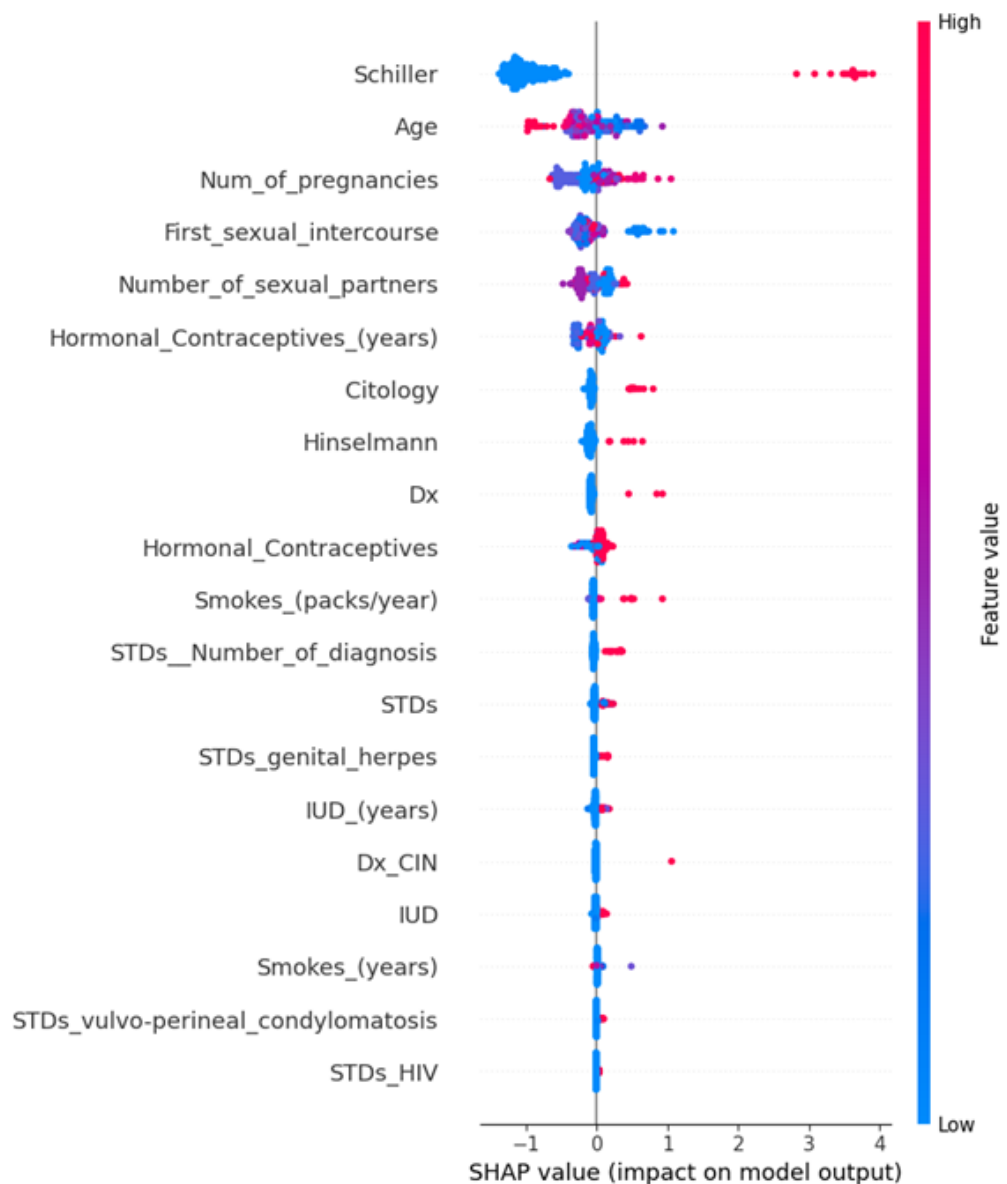


Figure 4.5: SHAP Summary Plot for Feature Importance in the Model

This figure displays the SHAP values for features, illustrating their contribution to the model's predictions. Features are ranked by importance, and feature values are indicated by color coding.

The features are ranked by their average magnitude of impact, with the most influential features at the top. Schiller, Age, and Num of pregnancies are this plot's top three most impactful features. The horizontal axis represents the SHAP value, which indicates the direction and magnitude of each feature's influence on the model's output. Positive SHAP values push the prediction toward the positive class, while negative values move it toward the negative class. Each dot's color corresponds to the feature value, with blue indicating low values and pink/red indicating high values. For example:

- Schiller has a strong influence, with higher SHAP values for specific instances with low feature values (blue).
- Age and Num of pregnancies show varying impacts, with high values (red) generally increasing or decreasing SHAP values depending on their position.

The scatter of dots for each feature reveals how consistently or variably the feature affects predictions. Wider spreads, like for Number of sexual partners and First sexual intercourse, suggest that their influence varies significantly across instances. This SHAP summary plot is crucial for interpreting the model. It highlights which features drive predictions and their relative importance.

4.4 Summary

This chapter presents the detailed results of the conducted experiments and analyses performed throughout the study. Starting with an evaluation of individual models after data preprocessing, the chapter progresses through feature selection, ensemble learning, hyperparameter tuning, and data balancing to achieve optimal performance. Additionally, performance analysis highlights the reliability of the best model using a confusion matrix, ROC curve, and 30-fold cross-validation. Finally, explainable AI techniques like SHAP and LIME are utilized to interpret the predictions, adding transparency to the results.

Chapter 5

Engineering Standards and Design Challenges

5.1 Compliance with the Standards

Construct conformity to engineering specifications is a prerequisite to this project as it depicts dependability, interoperability, and conformity to standards. Reading these standards helps inform the development process and establish the standard by which the effectiveness of the system can be measured in terms of compatibility. It meets different software, hardware, and communication standards to attain sound implementation and for its compatibility with actual applications.

5.1.1 Software Standards

The software pieces of this work were implemented and run with popular frameworks and libraries, following best practices of the modern programming world. Python, the leading language in the entire code, complies with the standard for software engineering and computational tools. Scikit-learn, TensorFlow, Keras, and XGBoost were used in the project as they belong to the open source community leading to frequent update, well documented, and support multiple platforms. Temporary work environments such as Jupyter Notebook and Google Colab meet general evaluation requirements for interactive computation environments in terms of usability. While there was no independent software testing done, following these well developed tools and libraries will ensure the software utilized in the project is credible and efficient.

5.1.2 Hardware Standards

The hardware applied in the project meets the most commonly used benchmarks in the field of performance and compatibility. The CPU such as, Intel Core i5 or comparative and GPU such as, Tesla T4 or NVIDIA GeForce GTX follows the IEEE standard for medical device connectivity that adequately manages medical datasets. The sizes of the memory

and processing capacities meet the basic hardware needs for machine learning algorithms that are advised by manufacturers of the hardware and tools. Also, the system's hardware is compliant with some power efficiency like the Energy Star; hence, the environmentally conscious computations for extended computation tasks.

5.1.3 Communication Standards

This work followed standard procedures to guarantee the efficient and comprehensible dissemination of outcomes and conclusions. Numerical results and graphics of the form of graphs and plots were presented with clear labeling, scaling, and captions in order to optimize their interpretability. Reporting was done following standard format with clear methodologies, results and analyses sections as well as accompanying appendices. The Major advantage of using Jupyter Notebook and Google Colab was the ability to share code, data and the results in one package called .ipynb. These communication practices made it possible for the audience both technical and non-technical to get openness and fitness for use from the project outputs.

5.2 Impact on Society, Environment and Sustainability

5.2.1 Impact on Life

Cervical cancer prediction using ML and XAI significantly improves people's lives. Since it is easier to prevent than treat diseases, finding them at a young age and diagnosing them correctly leads to high survival percentages. This is especially important in developing countries where the tremendous shortage of diagnostic and treatment equipment leads to late diagnosis and treatment. The patients will be able to save costs and reduce stress from treatments that would otherwise be required in the early detection of diseases. It is also important to note that making predictions using an ML model helps self-organize the population's knowledge and improve their health by creating new opportunities for individuals. These indicators demonstrate that artificial intelligence-supported technologies can help save lives and enhance people's standard of living.

5.2.2 Impact on Society & Environment

Societal-level benefits include increased diagnostic accuracy, which assists in a decrease in the burden of cervical cancer. By targeting high-risk populations, resource utilization in public health systems becomes more efficient, and hence, more dollars and healthcare infrastructure can be leveraged. It will reduce the healthcare disparities and unequal treatment of insignificant clients and enhance global health inequalities. Moreover, by using XAI, healthcare specialists and patients gain the trust and acceptance of AI models and contribute to further social integration.

From an environmental point of view, digital and AI-driven solutions decrease the reliance on physical diagnostic tools and lab-based processes, which often involve the dis-

posal of hazardous materials. AI systems reduce the carbon footprint associated with traditional healthcare infrastructure by promoting remote diagnostics and minimizing travel for in-person consultations.

5.2.3 Ethical Aspects

Many considerations are worthy of debate regarding whether healthcare should use AI technologies. They include privacy, confidentiality, informed consent, and issues arising from using algorithms. The privacy of patient information must be maintained by strictly observing the laws governing the protection of personal information and the values of ethics. Educating the patient and making them comprehend how the data will be used makes them concordant with the moves and builds trust. Bias correction in datasets is vital to avoid adverse consequences on some demographics in model outcomes. By implementing XAI, one can understand the reasoning behind AI's decision to prevent the offset of deep learning algorithms, often called black boxes.

5.2.4 Sustainability Plan

An implementation plan of sustainability as an intervention is crucial for its effectiveness and sustainability in a cervical cancer prediction AI system formed by AI. This asymptotic feature comes into focus with the system built to augur well with complex healthcare structures in various territories; it has been developed with an eye on correctly endowed settings and implementations with limited access to resources. The technology is affordable to the target patient population by optimizing computing costs and reducing the general overhead expenses for running the algorithms in the cloud. The education and training will help equip the healthcare workforce with the necessary tools to use and interpret the AI-based diagnostics packages for widespread diffusion and utilization. To sustain its accuracy and reliability, the model must be updated frequently using new data, more frequent medical knowledge, or theories that explain akin phenomena. The same approach can be used to reach out to more people; partnerships with governments, non-governmental organizations, and other international health bodies are paramount in phasing up the solution. These strategies collectively guarantee that the outcomes supplied by the AI system are immediate and that the overall application remains an adequate and relevant instrument for enhancing global health outcomes.

5.3 Project Management and Financial Analysis

Effective project management is critical to the success of any research project, exceptionally one as multifaceted as developing an AI-driven cervical cancer prediction system. The project timeline was meticulously planned and executed using a Gantt chart to track progress, allocate resources efficiently, and ensure the timely completion of each task.

The Gantt chart, presented in Figure 5.1, illustrates the sequential and overlapping tasks spanning from May 2024 to December 2024.

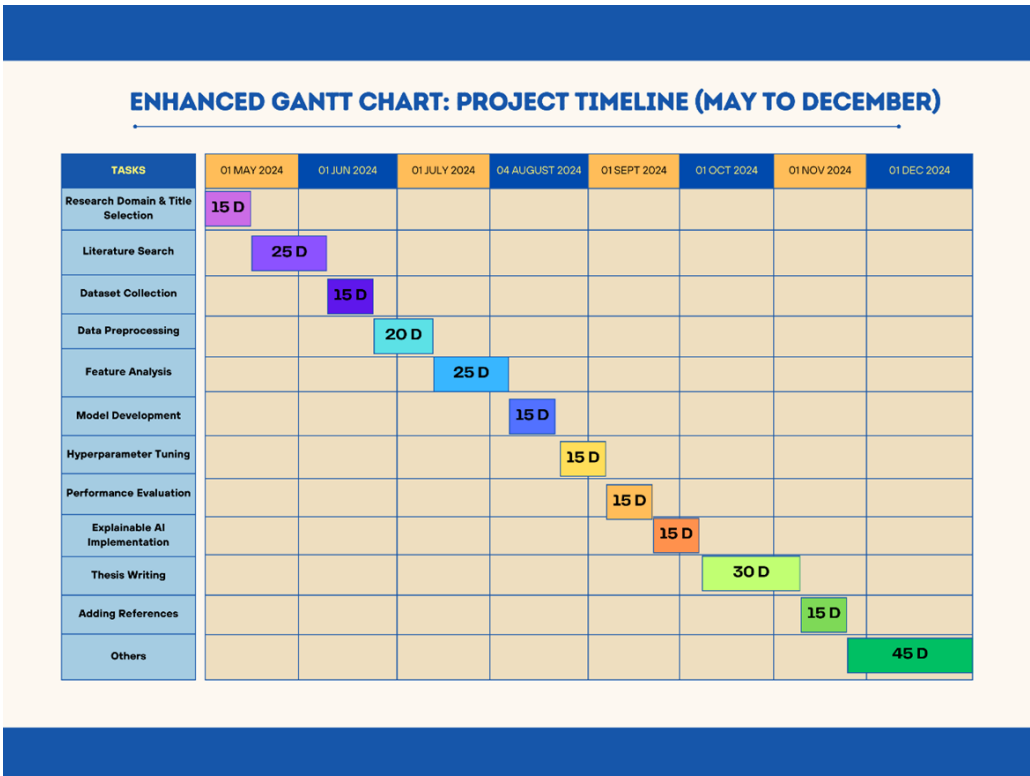


Figure 5.1: Enhanced Gantt Chart showcasing the project timeline

The project commenced in May 2024 with the Research Domain & Title Selection, which was allocated 15 days. This was followed by an extensive Literature Search, lasting 25 days, aimed at understanding existing methodologies and identifying gaps to inform the study’s approach. The Dataset Collection phase, completed in June 2024, was conducted over 15 days to gather relevant, high-quality data for the predictive model.

Subsequently, Data Preprocessing was performed over 20 days in July 2024, ensuring the dataset was cleaned, balanced, and ready for analysis. The Feature Analysis phase spanned 25 days, focusing on identifying critical attributes for practical model training. In August 2024, Model Development commenced, with 15 days dedicated to building robust machine learning models. This was immediately followed by Hyperparameter Tuning, which optimized model performance for 15 days in September 2024. Performance Evaluation, also conducted in September, involved another 15 days of rigorous testing to assess the models’ accuracy, precision, recall, and F1 score. One of the pivotal tasks, Explainable AI Implementation, was carried out in October 2024 over 30 days to ensure transparency and interpretability in the system’s predictions. Concurrently, Thesis Writing began in November 2024, running for 15 days to document the methodology, results, and conclusions. Finally, the project concluded with a 45-day phase for Adding References and miscellaneous tasks to ensure completeness and compliance with academic standards.

An economic analysis was conducted to estimate the cost of resources required for the project. The following table 5.1 outlines the estimated expenses:

Table 5.1: Estimated Cost for This Work

S.No.	Components	Estimated Cost (BDT)
1	Software Licenses and Tools	10,000
2	Hardware and Computing Resources	60,000
3	Research and Literature Access	2,000
4	Technical Support and Consulting	8,000
5	Miscellaneous Expenses	3,000
Total		83,000

5.4 Complex Engineering Problem

Real world engineering tasks pose questions that have solutions based on multiple factors and often address a combination of issues that are technical, social, environmental, and even ethical. These problems are characterized by risk, multi-disciplinary and holistic and all-encompassing view of engineering concepts. This section describes how such a problem has been solved within this research, in particular focusing on explaining how dimensions of problem solving, knowledge profiles, and engineering activities have been identified and described.

5.4.1 Complex Problem Solving

The research maps the solution to the complex problem-solving dimensions, as shown in Table 5.2. The project proves understanding at the knowledge application level (EP1) while adopting enhanced artificial intelligence through Machine Learning and XAI. Thus, it satisfies constrained demands (EP2) for improved accuracy, the ability to interpret results, and reasonable computing time. The depth of analysis (EP3) is done by evaluating models using different criteria hence making use of accuracy, precision, recall, and F1 score. Concerning the experienced real-world problems (EP4), the proposal addresses the task of handling the unbalanced datasets and the problem of ethical processing of the health care data. Last, the project demonstrates Interdependence (EP7) by demonstrating a logical amalgamation of SHAP and LIME as tools whose integration enhances interpretability and trust in Artificial Intelligence decision-making.

Table 5.2: Mapping with complex problem solving.

EP1 Dept of Knowl- edge	EP2 Range of Con- flicting Require- ments	EP3 Depth of Analysis	EP4 Familiarity of Issues	EP5 Extent of Applicable Codes	EP6 Extent of Stake- holder Involve- ment	EP7 Inter- dependence
✓	✓	✓	✓			✓

Mapping with Knowledge Profile for EP1

Table 5.3 shows how EP1 (Depth of Knowledge) is aligned with the knowledge profiles of learning. The work expands on engineering principles (K3) through the process of data pre-processing, algorithm and model construction and assessment. In line with the K4, domain-specific methods, which include ensemble learning and XAI tools, are employed to guarantee clinical relevance and decision transparency. Engineering design (K5) is illustrated through the process of constructing an efficient ensemble model of DT, GB, and XGB. The learning objective (K6) of integrating engineering practices is seen through using state-of-the-art tools such as SHAP and LIME for interpretability and reliability. Moreover, the research also contributes to the advancement of knowledge (K8) through integration of literature review and development of an automated efficiently accurate cervical cancer prediction system.

Table 5.3: Mapping with knowledge Profile.

K3 Engineering Funda- mentals	K4 Specialist Knowl- edge	K5 Engineering Design	K6 Engineering Practice	K8 Research Literature
✓	✓	✓	✓	✓

5.4.2 Engineering Activities

Engineering activities related to this project are aligned with key dimensions which are described in Table 5.4. This project shows how versatile (EA1) resources can be, whether, for example, vigorously used and freely available public databases, or Google Colab with an accessible GPU, or Scikit-learn, TensorFlow, or XGBoost. The functionality and practicability of the developed system is ensured through interactive collaboration between

the involved healthcare specialists and artificial intelligence researchers and developers on an intermediate level (EA2). Moreover, the society and environment impact (EA4) is achieved by targeting early cervical cancer screening in resource poor settings where health outcomes can be significantly enhanced and mortality reduced. Knowledge (EA5) about previous approaches and difficulties related to cervical cancer prediction enabled the project to build efficient ensemble-based one based on the previous literature review.

Table 5.4: Mapping with complex engineering activities.

EA1 Range of re- sources	EA2 Level of Interac- tion	EA3 Innovation	EA4 Consequences for society and environment	EA5 Familiarity
✓	✓		✓	✓

5.5 Summary

In this chapter, the engineering standards of the cervical cancer prediction system, the design considerations, social implications, and the project management on the AI system are also discussed. Implementation of the software, hardware as well as communication standards increase reliability and compatibility of project outcomes, dissemination. Utilizing tested structures such as Scikit-learn, TensorFlow, and XGBoost, coupled with sound hardware requirements, this research passes the engineering criteria for reliability and growth. The overall significance of the project here is towards society since deployment of the algorithm showcase a promise for enhanced detection of cervical cancer in the initial stage, especially in the developing world. Organization will hence experience decreased unnecessary variations in healthcare and efficient use of resources hence reduce their negative impact in the environment through the use of artificial intelligence diagnosing tools. There is data privacy included to ensure the data being received is legal and bias correction and obtaining of informed consent has also been put in place to ensure that results obtained are fair and equitable. Some of the long-term features emphasized in the sustainability plan, include pricing, regular revision in duty, and conformity to preceding healthcare frameworks. Additional information is highlighted to involve stakeholders, governments and other international organizations to expand the scope of the solution. Lastly, the general project management strategy with the help of the provided Gantt chart shows careful planning and implementation. Resource allocation and financial analysis made the project costing efficient while the time line placed most significant stages of the project for efficiency. Due to its compliance with complex problem solving dimensions, knowledge profiles and engineering activities this research provides balanced solution towards real life healthcare issue.

Chapter 6

Conclusion

6.1 Summary

With the incorporation of the AI system, the cervical cancer prediction system enjoys formulated major beneficial breakthroughs in improving diagnosis precision and objectiveness. When incorporating ML and XAI, results were impressive, with an accuracy level of 99.57%, precision of 94.8%, recall of 92.5%, and an F1 score of 93.64%. Despite optimizing hyperparameters, it was found that the Decision Tree, Gradient Boosting, and XGBoost-based ensemble model provided the best overall framework for the system. This approach used an ensemble of individual models to the advantage of the solution to build a highly efficient diagnostic capability that could easily tackle medical data as seen in real-world scenarios. The aspect that makes this system one of the most fascinating is the ability to build interpretability through SHAP and LIME. Supplementary to the paper, the tools offered an additional appreciation of the model, guaranteeing that health-care providers can rely on the results and comprehend the outcomes. This transparency is essential for clinical application, closing the gap between AI technologies and clinical decision-making on the one hand and, on the other, the demand for the practical usefulness of such technologies. As a result, the system has its forecast reliable and explainable, removing one critical obstacle to using AI in diagnosis. The methodological approach used in this study benefited the system. Implementing a feature selection mechanism reduced some of the noises in the dataset, enabling the model to attend to those attributes of paramount importance. Handling of data enhanced minority class instances along with balancing, leading to higher correct prediction across most populations. Adjusting some difficult-to-learn values and hyperparameter tuning improved the model's performance even further, increasing the models' reliability; ensemble learning also decreased the overfitting issue of overfitting. Such methodological developments make it possible to present the system as a universal and easily portable technology suitable for use in various environments of healthcare service delivery, including low-resource areas that may not have access to conventional diagnostics tools. In addition, implementing the system has future implications for health by supplementing the weaknesses of known typical diagnosis pro-

cedures. On the contrary, the AI-based system provides the best solution, which is more affordable and efficient in its capacity to analyze large amounts of transferred data. It is capable of delivering answers promptly and accurately. With early detection from this system, incorporation of health interventions enables improved survival and the alleviation of cervical cancer burden in the whole world.

6.2 Limitation

While this study demonstrated impressive results, certain limitations were evident. The use of openly available data limits the variety of data that become the input to the system, which may impact the system's capability of functioning in a diverse population. The demands for computations, specifically with ensemble methods, present significant obstacles for low-resource settings. While data balancing techniques alleviate some of the difficulties caused by class imbalance, perfect representation of the minority classes is still an important issue. In addition, the absence of clinical trials hinders the practical implementation of the system every day and calls for pilot projects. Addressing data privacy and ethical concerns is also critical, as handling sensitive medical data requires stringent compliance with privacy laws and ethical standards.

6.3 Future Work

As a result of this study, more research should be directed towards enlarging the sample size and extending the types of data used. Replacing random data and expanding the data set to more significant population data from many different areas and demographics will make the model fair for all the various groups of people, helping to give all better treatments. The next logical extension of the work is to evaluate the system in real-world clinical settings to determine the usability and adoption of the system by end-users and the fit of the system into the current clinical workflow. Positive aspects of the models should be sought to allow them to change to new data while enhancing their efficiency and suitability. User-friendly interfaces designed primarily for clinical practitioners and other healthcare professionals in the early stage of product planning will ensure that the system remains as functional as it is easy to use in a clinical context. Furthermore, applying this technology to diagnosis systems of multiple diseases could be the most beneficial approach to diagnosing various medical conditions using the same AI-based system. More actions should also be taken to measure and mitigate adverse environmental impacts by using computational resources most efficiently to make the system less environmentally costly.

References

- [1] Sharmila Pimple and Gauravi Mishra. Cancer cervix: Epidemiology and disease burden. *CytoJournal*, 19:21, 3 2022.
- [2] Amelia Goldstein, Mallory Gersh, Gabriela Skovronsky, and Chailee Moss. The future of cervical cancer screening. *International Journal of Women's Health*, 16:1715–1731, 2024.
- [3] Gulzhanat Aimagambetova, Gauri Bapayeva, Talshyn Ukybassova, Nazira Kamza- yeva, Gulnara Sakhipova, Nasrulla Shanazarov, and Milan Terzic. Risks of cervical cancer recurrence after fertility-sparing surgery and the role of human papillomavirus infection types. *Journal of Clinical Medicine 2024, Vol. 13, Page 6318*, 13:6318, 10 2024.
- [4] Mohan Narasimhamurthy and Santhosh Upadhyaya Kafle. Cervical cancer in nepal: Current screening strategies and challenges. *Frontiers in Public Health*, 10:980899, 11 2022.
- [5] Nina Youneszade, Mohsen Marjani, and Chong Pei Pei. Deep learning in cervical cancer diagnosis: Architecture, opportunities, and open research challenges. *IEEE Access*, 11:6133–6149, 2023.
- [6] Taha Buğra Ç elik, Ö zğ ü r İ can, and Elif Bulut. Extending machine learning prediction capabilities by explainable ai in financial time series prediction. *Applied Soft Computing*, 132:109876, 1 2023.
- [7] Arrabella R. King, Mahmudul Hassan Al Imam, Sarah McIntyre, Catherine Morgan, Gulam Khandaker, Nadia Badawi, and Atul Malhotra. Early diagnosis of cerebral palsy in low- and middle-income countries. *Brain Sciences 2022, Vol. 12, Page 539*, 12:539, 4 2022.
- [8] Serhat KILIÇ ARSLAN, Maruf G Ö GEBAKAN, and Cemil K Ö ZKURT. Cervical cancer prediction using smote algorithm and machine learning approaches. *Iğdır Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 13:747–759, 6 2023.
- [9] Mimonah Al Qathrady, Ahmad Shaf, Tariq Ali, Umar Farooq, Aqib Rehman, Samar M. Alqhtani, Mohammed S. Alshehri, Sultan Almakdi, Muhammad Irfan,

- Saifur Rahman, and Ladon Ahmed Bade Eljak. A novel web framework for cervical cancer detection system: A machine learning breakthrough. *IEEE Access*, 12:41542–41556, 2024.
- [10] Hanen Karamti, Raed Alharthi, Amira Al Anizi, Reemah M. Alhebshi, Ala' Abdulmajid Eshmawi, Shtwai Alsubai, and Muhammad Umer. Improving prediction of cervical cancer using knn imputed smote features and multi-model ensemble learning approach. *Cancers*, 15, 9 2023.
- [11] Turki Aljrees. Improving prediction of cervical cancer using knn imputer and multi-model ensemble learning. *PLoS ONE*, 19, 1 2024.
- [12] Fahima Khanam, M Rubaiyat Hossain Mondal, Md Elias Shawon Bhuiyan, and Yearul Kabir. Prediction of cervical cancer in bangladesh by using machine learning algorithms, 2023.
- [13] Raafat M. Munshi. Novel ensemble learning approach with svm-imputed adasyn features for enhanced cervical cancer prediction. *PLoS ONE*, 19, 1 2024.
- [14] Pratiksha D. Nandanwar and Somnath B. Dhonde. A novel approach to cervical cancer detection using hybrid stacked ensemble models and feature selection. *International Journal of Electrical and Electronics Research*, 11:582–589, 6 2023.
- [15] Seeta Devi, Sachin Ramnath Gaikwad, and R. Harikrishnan. Prediction and detection of cervical malignancy using machine learning models. *Asian Pacific Journal of Cancer Prevention*, 24:1419–1433, 2023.
- [16] Gaurav Kumawat, Santosh Kumar Vishwakarma, and Prasun Chakrabarti. Cervical cancer prediction using machine learning techniques. In *Lecture Notes in Networks and Systems*, volume 817, pages 13–28. Springer Science and Business Media Deutschland GmbH, 2024.
- [17] Shikha Prasher, Leema Nelson, and Vasantha Sandhya. Early diagnosis of cervical cancer using machine learning approaches: Comparative analysis. In *2023 IEEE International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering, RMKMATE 2023*. Institute of Electrical and Electronics Engineers Inc., 2023.
- [18] Jesse Jeremiah Tanimu, Mohamed Hamada, Mohammed Hassan, Habeebah Adamu Kakudi, and John Oladunjoye Abiodun. A machine learning method for classification of cervical cancer. *Electronics 2022, Vol. 11, Page 463*, 11:463, 2 2022.
- [19] Abdulaziz AlMohimeed, Hager Saleh, Sherif Mostafa, Redhwan M.A. Saad, and Amira Samy Talaat. Cervical cancer diagnosis using stacked ensemble model and optimized feature selection: An explainable artificial intelligence approach. *Computers*, 12, 10 2023.

- [20] Rashiduzzaman Shakil, Sadia Islam, and Bonna Akter. A precise machine learning model: Detecting cervical cancer using feature selection and explainable ai. *Journal of Pathology Informatics*, 15:100398, 12 2024.
- [21] Naif Al Mudawi and Abdulwahab Alazeb. A model for predicting cervical cancer using machine learning algorithms. *Sensors 2022, Vol. 22, Page 4132*, 22:4132, 5 2022.
- [22] M. Sangeetha, Vijayan Sugumaran, T. Karthick, and Jagadeesh Puli. High performance cervical cancer detection using gradient boosting algorithm and bayesian optimization. *2023 9th International Conference on Information Technology Trends, ITT 2023*, pages 102–107, 2023.
- [23] Md Shahin Ali, Md Maruf Hossain, Moutushi Akter Kona, Kazi Rubaya Nowrin, and Md Khairul Islam. An ensemble classification approach for cervical cancer prediction using behavioral risk factors. *Healthcare Analytics*, 5:100324, 6 2024.
- [24] Umesh Kumar Lilhore, M. Poongodi, Amandeep Kaur, Sarita Simaiya, Abeer D. Al-garni, Hela Elmannai, V. Vijayakumar, Godwin Brown Tunze, and Mounir Hamdi. Hybrid model for detection of cervical cancer using causal analysis and machine learning techniques. *Computational and Mathematical Methods in Medicine*, 2022:4688327, 1 2022.
- [25] Milad Rahimi, Atieh Akbari, Farkhondeh Asadi, and Hassan Emami. Cervical cancer survival prediction by machine learning algorithms: a systematic review. *BMC Cancer*, 23:1–10, 12 2023.
- [26] Haowen Yan, Gaoting Huang, Zhihe Yang, Yirong Chen, and Zhiming Xiang. Machine learning-based multiparametric magnetic resonance imaging radiomics model for preoperative predicting the deep stromal invasion in patients with early cervical cancer. *Journal of Imaging Informatics in Medicine* 2024 37:1, 37:230–246, 1 2024.
- [27] R Thangamani, M. Vimaladevi, S H Varshini, A Mohammed Aadhil, and K R Sasvath. Cervical cancer prediction using machine learning. *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–7, 6 2024.
- [28] Yanjun Sun, Yuhao Ma, Menghan Cao, Zhiqing Hu, Wei Lin, Mingsheng Chen, and Yuan He. Breast and cervical cancer screening adherence in jiangsu, china: An ecological perspective. *Frontiers in Public Health*, 10:967495, 8 2022.
- [29] N. Meenakshisundaram and G. Ramkumar. Early identification of cervical cancer using k-nearest neighbor (knn). *Proceedings of the 2023 2nd International Conference on Electronics and Renewable Systems, ICEARS 2023*, pages 1005–1010, 2023.

- [30] Michal- Kruczkowski, Anna Drabik-Kruczkowska, Anna Marciniak, Martyna Tar-
czewska, Monika Kosowska, and Mal-gorzata Szczerska. Predictions of cervical cancer
identification by photonic method combined with machine learning. *Scientific Reports*
2022 12:1, 12:1–11, 3 2022.
- [31] Xiaoyuan Chen, Turki Aljrees, Muhammad Umer, Oumaima Saidani, Latifah Al- muqren,
Olfa Mzoughi, Abid Ishaq, and Imran Ashraf. Cervical cancer detection us- ing k nearest
neighbor imputer and stacked ensemble learningmodel. *Digital Health*, 9, 1 2023.
- [32] Emmanuel Ileberi and Yanxia Sun. Machine learning-assisted cervical cancer predic-
tion using particle swarm optimization for improved feature selection and prediction.
IEEE Access, 12:152684–152695, 2024.
- [33] Remah Younisse, Ashraf Ahmad, and Qasem Abu Al-Haija. Explaining intrusion
detection-based convolutional neural networks using shapley additive explanations
(shap). *Big Data and Cognitive Computing 2022, Vol. 6, Page 126*, 6:126, 10 2022.
- [34] Tanya Chowdhury, Razieh Rahimi, and James Allan. Rank-lime: Local model-
agnostic feature attribution for learning to rank. *ICTIR 2023 - Proceedings of the*
2023 ACM SIGIR International Conference on the Theory of Information Retrieval,
pages 33–37, 8 2023.
- [35] Md Alif Sheakh, Mst Sazia Tahosin, Md Maruf Hasan, Taminul Islam, Onamika
Islam, and Md Masum Rana. Child and maternal mortality risk factor analysis using
machine learning approaches. *ISDFS 2023 - 11th International Symposium on Digital*
Forensics and Security, 2023.
- [36] Taminul Islam, Md Alif Sheakh, Mst Sazia Tahosin, Most Hasna Hena, Shopnil Akash,
Yousef A. Bin Jordan, Gezahign FentahunWondmie, Hiba Allah Nafidi, and Mo-
hammed Bourhia. Predictive modeling for breast cancer classification in the context of
bangladeshi patients by use of machine learning approach with explainable ai. *Scientific*
Reports 2024 14:1, 14:1–17, 4 2024.
- [37] Md Alif Sheakh, Mst. Sazia Tahosin, Lima Akter, Israt Jahan, Md Nakibul Islam, Md
Rafiuddin Siddiky, Md Mahadi Hasan, and Sakibul Hasan. Comparative anal- ysis of
machine learning algorithms for ecg-based heart attack prediction: A study using
bangladeshi patient data. [https://wjarr.com/sites/default/files/WJARR-2024-
2928.pdf](https://wjarr.com/sites/default/files/WJARR-2024-2928.pdf), 23:2572–2584, 9 2024.
- [38] Mst Sazia Tahosin, Md Alif Sheakh, Taminul Islam, Rishalatun Jannat Lima, and
Mahbuba Begum. Optimizing brain tumor classification through feature selection and
hyperparameter tuning in machine learning models. *Informatics in Medicine Unlocked*,
43:101414, 1 2023.

Enhancing Cervical Cancer Prediction with Machine Learning and Explainable AI

ORIGINALITY REPORT



PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	1 %
2	Submitted to United International University Student Paper	1 %
3	Submitted to University of Hertfordshire Student Paper	1 %
4	Submitted to Liverpool John Moores University Student Paper	<1 %
5	Nermeen Gamal Rezk, Samah Alshathri, Amged Sayed, Ezz El-Din Hemdan, Heba El- Behery. "XAI-Augmented Voting Ensemble Models for Heart Disease Prediction: A SHAP and LIME-Based Approach", Bioengineering, 2024 Publication	<1 %
6	www.ncbi.nlm.nih.gov Internet Source	<1 %
7	cdn.techscience.cn	