Journal of Theoretical and Applied Information Technology

15th October 2023. Vol.101. No 19
© 2023 Little Lion Scientific

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A MACHINE LEARNING-BASED CLASSIFIER FOR ANTICIPATING RISK FACTORS ASSOCIATED WITH CERVICAL CANCER

ANUSHA R [1], DR SRINIVAS PRASAD[2]

[1] Research Scholar, GITAM Deemed to be University, Department of Computer Science and Engineering, Andhra Pradesh, India

[2]Professor, GITAM Deemed to be University, Department of Computer Science and Engineering, Andhra Pradesh, India

ABSTRACT

Cervical cancer ranks as the second most prevalent form of cancer among women, but it is also highly preventable. Numerous studies have underscored the widespread lack of knowledge surrounding cervical cancer and its prevention. In developing countries, medical students represent the future healthcare professionals who can play a pivotal role in increasing public awareness and assessing knowledge about symptoms and risk factors. The present research focused on investigating the causes of cervical cancer and preventive methods for women who have already been diagnosed with this life-threatening disease. Each year, a significant number of women receive a diagnosis of cervical cancer, resulting in a substantial loss of lives worldwide. The Human papillomavirus (HPV) has been recognized as a significant risk factor for cervical cancer, and fortunately, effective prevention measures are available to a large extent. However, most cervical cancer cases are detected in economically disadvantaged countries where organized HPV screening or vaccination programs are lacking. High-income countries that have implemented comprehensive screening programs have successfully reduced the prevalence and mortality rates of cervical cancer by 50% over the past three decades. Fertility-preserving surgical approaches are now considered the standard treatment for women diagnosed with initial-stage, mild-risk cervical cancer. As cervical cancer remains a significant health concern for women, implementing a comprehensive strategy for prevention and control is paramount in our efforts to eliminate this disease. This research aims to introduce an ensemble-based classifier that enhances the early detection and prediction of cervical cancer. Through comparative analysis with other base classifiers, our proposed classifier demonstrates superior accuracy and performance.

Keywords: *Cervical Cancer, Screening, Risk Factors, HPV, Ensembled Enabled Classifier*

1. INTRODUCTION

Cervical cancer represents a substantial and urgent menace to women's health, being one of the most prevalent types of cancer affecting women worldwide [1, 2]. The discovery of a unique cause has fueled efforts in the prevention and control of cervical cancer. In May 2018, the World Health Organization (WHO) made a global appeal to combat cervical cancer, prompting an enthusiastic response from more than 70 countries and international academic institutions[3, 5]. Building upon this momentum, the WHO launched a comprehensive plan on November 17, 2020, aimed at expediting the eradication of cervical cancer and providing a roadmap for future prevention and control efforts [1]. This initiative highlights the commitment of 194 nations to collaborate in the global fight against cervical cancer.

Cervical cancer poses a significant threat to women's health and can have fatal consequences. In recent years, there has been a disturbing rise in the number of young women affected by this type of cancer, with the incidence rate increasing from 25% to 45%. According to the World Health Organization (WHO) and the International Agency for Research on Cancer (IARC), a staggering 930,000 new cases were reported worldwide in 2020, making it the second most prevalent cancer among women. It's important to note that this issue is not limited to low-income countries alone, as even well-established nations are grappling with high cervical cancer rates despite having widespread awareness campaigns in place.

Globally, in the year 2020, there were approximately 600,000 reported cases of cervical cancer, tragically resulting in a high mortality rate, with half of the affected women succumbing to the disease. This cancer ranks as the fourth leading cause of death among women worldwide. Alarmingly, around 88% of these deaths occurred predominantly in both developed and undeveloped countries, highlighting the global impact of this disease. Furthermore, the death rate from cervical cancer is 20% higher in Economically disadvantaged countries compared to developed nations. Another concerning trend is the increasing incidence of breast cancer, particularly in undeveloped countries. This type of cancer has been documented in approximately 42 countries, with a significant concentration observed in several African nations that are grappling with the burden of breast cancer.

Africa demonstrates the highest rates of native occurrence and fatality from cervical cancer, with rates around 7 to 10 times higher compared to New Zealand, Australia, North America, and Western Asia [14]. Based on the data from the national cancer centre in 2015, nearly 900 new cases were diagnosed and 500 deaths were reported [6]. In the past twenty years, China has observed a progressive rise in the incidence and mortality rates of cervical cancer [15].

Low socioeconomic status, smoking, early marriage (before the age of 18), human papillomavirus (HPV) infection, and having multiple sexual partners are all well-established risk factors for developing cervical cancer. These factors contribute to an increased likelihood of developing cervical cancer, along with a history of multiple sexual partners and frequent childbirths [4]. HPV has been identified as the primary cause of cervical cancer, particularly among individuals aged 16 to 45. Numerous research studies have indicated that various factors such as consistent sexual partners, early sexual activity, and sexual activity before each encounter can increase the risk of cervical cancer [4].

The human body consists of hundreds of thousands of cells that continually undergo a process of random movement and dissociation to generate new cells. This process occurs rapidly in early life, enabling growth, while in later stages of life, cells disengage more effectively to replace dead cells. Cancer emerges when cells in the body lose control over their normal functions. The cervix, which serves as the connection between the vagina and the uterine body, can be divided into two regions: the endocervix (near the uterus) and the exocervix (near the vagina). The majority of cervical cancers originate from the cells lining the cervix, particularly the uterine cervix, which is a smaller section of the uterus. The cervix comprises two types of cells: squamous cells and glandular cells, which meet at an area called the transformation zone. The location of this transformation zone changes over time and after childbirth. Cervical cancer often develops within this transformation zone. It is worth noting that normal cervical cells commonly exhibit precancerous characteristics before progressing into cancerous cells.

Cervical cancer, ranked as the second most common illness among women can be effectively prevented and treated when detected at an early stage. However, women, particularly those from low socioeconomic backgrounds, face barriers to accessing cervical cancer screenings such as Pap tests. Lack of awareness and shyness prevent some individuals from recognizing the signs and symptoms and accessing necessary healthcare services. This leads to inadequate screening and treatment. It is crucial to raise public awareness about the risk factors associated with cervical cancer and promote prevention measures. Regrettably, there is a scarcity of available data on cervical cancer risk factors specifically applicable to Indian settings, leading to a decrease in public consciousness concerning early identification and preventive measures. The development of cervical cancer is seen in Figure 1. Biology, anatomy, medicine, physiology, and other healthcare-related sciences provide graphic depictions of the uterus and cervical cancer's various stages.

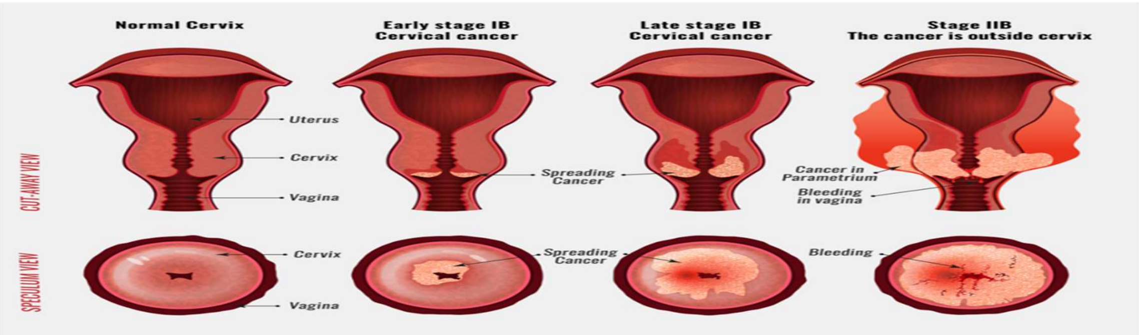


Figure 1: Development Of Cervical Cancer

The main aim of this article is to predict the risk factors of cervical cancer using machine learning and the following explains why we chosen ML.

Improving Early Detection: The key to excellent healthcare is early intervention. Cervical cancer has a better prognosis when detected in its early stages. With its ability to detect small similarities in disparate datasets, machine learning holds the possibility of uncovering nuanced markers that presage the start of this disease. By integrating ML-driven prediction models, we want to revolutionise early detection, increasing survival rates and improving the quality of life for those afflicted.

Preventing Avoidable Losses and Prompt testing and treatments addresses healthcare inequities, but it also functions as an effective strategy in decreasing avoidable losses among disadvantaged groups.

With the use of machine learning approaches, the field of cervical cancer prediction has undergone a significant transformation. To uncover complex patterns from varied datasets, researchers have laboriously investigated a variety of ML methods, including Support Vector Machines, Random Forests, Neural Networks, Deep Learning techniques and more. These efforts have produced some extremely remarkable results, such as the discovery of delicate biomarkers, precise risk assessment, and customised prediction models. Researchers have improved the accuracy and reliability of cervical cancer prediction, enabling healthcare practitioners to implement preventative measures. They did this by combining clinical data, genetic information, and lifestyle variables.

Additionally, research contributions go beyond algorithmic advancements. The creation of user-friendly user interfaces and the incorporation of decision support systems have made it possible to incorporate machine learning predictions into clinical processes with ease. A multidisciplinary approach has been cultivated via collaborative efforts involving data scientists, healthcare practitioners, and policymakers, creating a comprehensive knowledge of the difficulties and potential in cervical cancer prediction. In addition to improving the prediction models, this confluence of knowledge has opened the door for ethical issues, patient privacy, and fair healthcare delivery.

2. RELATED WORK

According to studies, cervical cancer primarily affects young individuals between the ages of 17 and 30, indicating a lack of awareness about the disease and its associated risks. Extensive research has been conducted to eradicate cervical cancer, but achieving a 100% success rate has proven challenging. Within the field of homeopathy, researchers have directed their efforts toward discovering a cure, yet certain homeopathic practitioners [3] may perceive cervical cancer as an untreatable condition based solely on its name, disregarding the contrary evidence provided by biopsy results. Homeopathy, grounded in the principle of treating the patient rather than the disease, can be applied to cervical cancer as well. Although cervical cancer poses significant challenges in terms of its progression, metastasis, and associated complications, it should be approached similarly to other chronic ailments. In a living human being, all physiological processes occur continuously and harmoniously. The majority of cancer patients exhibit a mixed miasmatic state, thus necessitating the use of combined miasmatic medications for both constitutional treatment and symptomatic relief. Various medications may be employed for the treatment of cervical cancer [4, 5, 6, 7].

The dataset obtained from the Kaggle repository includes the following attributes related to cervical cancer risk:

Age: The patient's age when the data was collected.

Number of sexual partners: Denotes the total count of distinct partners the patient has engaged in sexual activity with

First sexual intercourse: Represents the age at which the patient had their first sexual experience.

Number of pregnancies: Indicates the total number of pregnancies the patient has had.

Smokes: A binary variable (1 or 0) indicating whether the person is a smoker or not.

Smokes (years): Represents the duration in years for which the person has been smoking.

Smokes (packs/year): Indicates the average number of packs of cigarettes smoked per year by the individual.

Hormonal Contraceptives: A binary variable (1 or 0) indicating whether the person is using hormonal contraception or not.

Hormonal Contraceptives (years): Represents the duration in years for which the person has been using hormonal contraception.

These are some of the variables present in the dataset, and there may be additional attributes that were not mentioned.

The target variables for diagnosing cervical cancer in this paper are as follows:

Hinselmann: This variable represents the outcome of the Hinselmann test. A value of 0 indicates no

cancer, while a value of 1 indicates the presence of cancer.
Schiller: This variable represents the outcome of the Schiller test. Similarly, a value of 0 indicates no cancer and a value of 1 indicates the presence of cancer.
Citology: This variable represents the outcome of the Citology test. A value of 0 indicates no cancer and a value of 1 indicates the presence of cancer.
Biopsy: The Biopsy variable represents the outcome of the biopsy test. A value of 0 indicates no cancer, while a value of 1 indicates the presence of cancer.

The focus of this paper centers around the target attribute of Biopsy. The presented algorithm demonstrated superior performance compared to the other two algorithms examined in the study.

3. PROPOSED METHOD

Cancer ranks as the second most common cause of mortality worldwide, resulting in the loss of approximately 9.6 million lives in 2019. The development of cancer occurs through a complex, multistage process, ultimately leading to the formation of malignant tumors. Early detection plays a crucial role in increasing treatment efficacy, improving survival rates, reducing morbidity, and minimizing healthcare costs. Analyzing the intricate ecosystem of screening and diagnosis processes from the perspective of computer-aided diagnostic (CAD) systems presents significant challenges. These challenges are further exacerbated in underdeveloped nations due to limited access to computer resources.

One of the key diagnostic challenges faced by individuals who do not undergo regular screening is determining the most effective screening technique and assessing their risk. Historically, screening approaches have relied heavily on the expertise and experience of healthcare professionals. Surveys can aid in identifying high-risk groups and avoiding unnecessary screenings. Implementing a risk-based strategy contributes to the resolution of cancer-related issues.

Based on a neoteric report from the World Health Organization (WHO), Cervical cancer is positioned as the fourth most pervasive type of cancer. [12]. This malignancy holds significant dangers, particularly in comparison to other types of cancer. The initial cause of cervical cancer has been linked to the contraction of the human papillomavirus (HPV) through sexual contact. Various types of HPV exist, with types 18 and 16 being strongly associated with cancer. HPV types 6 and 11 are considered low-risk as they primarily induce cysts on the surface. However, they are the most prevalent types and can lead to the development of cancerous tissues in the surrounding area.

Neural networks have proven to be effective and efficient methods for cancer detection. Researchers have developed ensemble approaches using multiple linear models and demonstrated their applicability in diverse settings, yielding positive results. Other approaches, such as enhanced genetic algorithms, artificial neural networks (ANN), and hierarchical clustering, have also been explored for cancer detection. In the paper mentioned, the cancer dataset is categorized, and the findings indicate accuracy levels ranging from 80% to 95%.

The proposed research strategy encompasses four key elements: the training methodology, the selection of a predictive model (PMS), the utilization of a training data collection, and the inclusion of a testing dataset. The theoretical context for the intended research is perceivably presented in Figure 2, outlining the overall structure and progression of the proposed approach. To effectively accomplish specific objectives, the architectural design is divided into four distinct stages, with each stage performing critical functions. This framework serves as a blueprint for the implementation and assessment of the research, providing a clear and organized roadmap for the study.

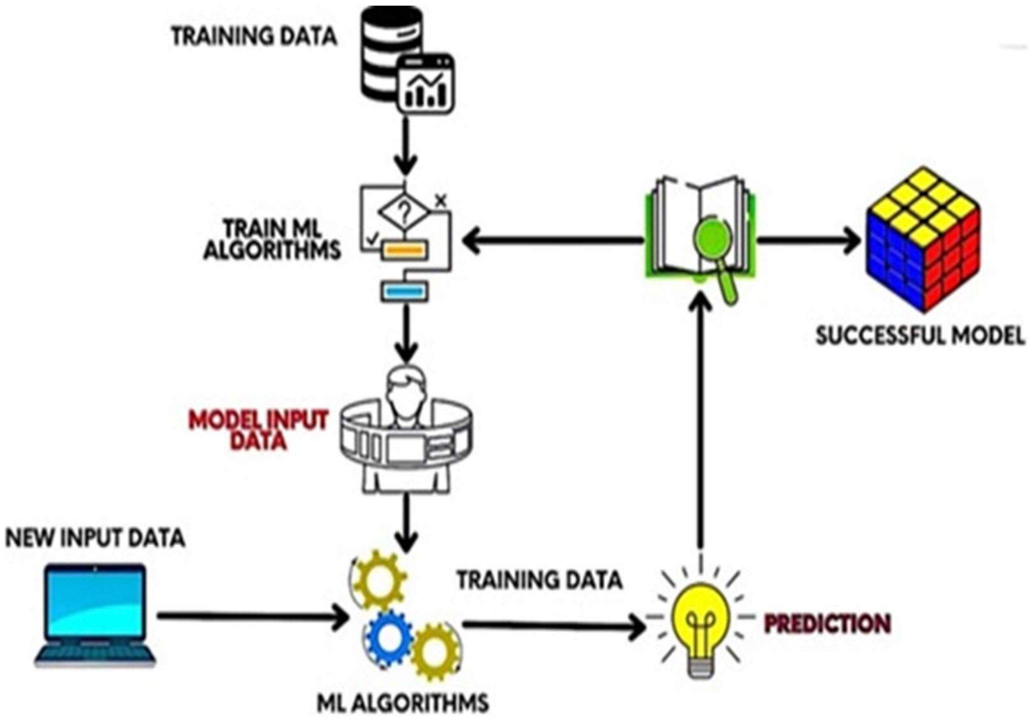


Figure 2: The Architecture Of The Proposed Model

Additional details about data acquisition for the study can be found in the research datasets section, which provides comprehensive information on how the datasets were obtained. The Data Pre-processing section addresses the steps involved in preparing the dataset for machine learning applications. [11]. This section delves into the steps taken to clean, transform, and organize the data to ensure its suitability for machine learning algorithms.

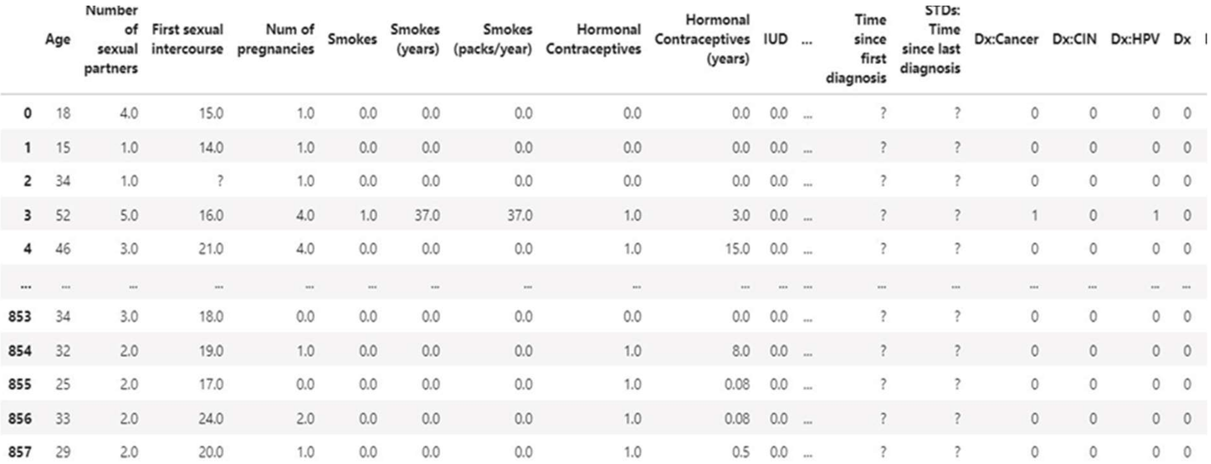
This study focuses on the use of Machine Learning methods to predict cervical cancer using publicly available data sources. It emphasises the use of machine learning techniques to improve the accuracy of early detection and risk assessment in various populations. The scope includes the following major elements:

Application of ML techniques: The paper explores into the use of several ML techniques, such as Support Vector Machines, decision trees, to detect cervical cancer. It investigates their effectiveness in analysing various datasets, such as clinical data, genetic markers, and lifestyle variables.

Data Preprocessing and Feature Selection: The work covers dataset preprocessing and feature selection to improve the performance of ML models in predicting cervical cancer risk.

3.1 Data Set

The dataset employed in this study, named “Cervical Cancer Risk Factors for Biopsy”, has been sourced from the Kaggle repository. It encompasses a diverse range of information, including medical histories, racial and ethnic backgrounds, and lifestyle details, about 5000 individuals. Due to privacy concerns, some patients have chosen not to answer certain questions, resulting in missing values within the dataset. The collection consists of 5000 instances, each characterized by 32 properties. Specifically, the dataset comprises 32 variables and encompasses the medical histories of 858 female patients [13]. Additionally, the collection features the medical histories of 10,000 female patients, encompassing a total of 32 factors such as age, IUD use, smoking habits, STDs, and more. For a visual representation of the dataset, please refer to Figure 3.



	Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	Hormonal Contraceptives	Hormonal Contraceptives (years)	IUD	...	Time since first diagnosis	STDs: Time since last diagnosis	Dx:Cancer	Dx:CIN	Dx:HPV	Dx
0	18	4.0	15.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	?	?	0	0	0	0
1	15	1.0	14.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	?	?	0	0	0	0
2	34	1.0	?	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	?	?	0	0	0	0
3	52	5.0	16.0	4.0	1.0	37.0	37.0	1.0	3.0	0.0	...	?	?	1	0	1	0
4	46	3.0	21.0	4.0	0.0	0.0	0.0	1.0	15.0	0.0	...	?	?	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
853	34	3.0	18.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	?	?	0	0	0	0
854	32	2.0	19.0	1.0	0.0	0.0	0.0	1.0	8.0	0.0	...	?	?	0	0	0	0
855	25	2.0	17.0	0.0	0.0	0.0	0.0	1.0	0.08	0.0	...	?	?	0	0	0	0
856	33	2.0	24.0	2.0	0.0	0.0	0.0	1.0	0.08	0.0	...	?	?	0	0	0	0
857	29	2.0	20.0	1.0	0.0	0.0	0.0	1.0	0.5	0.0	...	?	?	0	0	0	0

Figure 3: Dataset On Risk Factors For Cervical Cancer

3.2 Pre-processing of Dataset

During the pre-processing phase, data undergoes various operations such as cleaning, transformation, and reduction. The quality of the processed data plays a crucial role in the overall success of a project. Data impurities, such as noise, outliers, redundant information, and missing values, can adversely affect the analysis [14]. In this study, the dataset was cleaned by eliminating outlier observations and filling in missing values. Additionally, the data transformation step was performed to prepare the data for mining. The research incorporates a combination of steps including creating a concept hierarchy, selecting relevant attributes, normalizing the data, and discretizing it. Analyzing larger datasets can be challenging, and therefore, a data reduction strategy is employed. The objective of this strategy is to minimize data processing and storage expenses while improving storage efficiency. To address overfitting in machine learning models, dimension reduction techniques were implemented. Specifically, the EBA (Ensembled Boosting Algorithm) technique was applied, as it offers value in reducing dimensionality and improving model performance.

3.3 SVM

For the model to effectively operate, it needs to identify a position in space that goes beyond the limitations of three dimensions. Multiple hyperplanes may exist that define this position, but our objective is to find the one with the maximum margin, which refers to the largest distance between data points from different classes. By achieving this, the categorization and support for additional measured values become more straightforward. The Support Vector Machine (SVM) method is employed to construct a decision boundary in a high-dimensional or potentially infinite space. This technique is valuable for tasks such as data classification, regression, feature extraction, and filtering. SVMs can be categorized into two distinct types.

- *Linear SVM:* The Linear Support Vector Machine (SVM) is a classification method used to classify data that can be separated into two categories using a straight line. This particular SVM classifier is applied to data that is "linearly separable," meaning that a clear boundary can be drawn to separate the two categories. Figure 4 provides a visual representation of the data with 1-D labeling, showcasing the linear separation achieved by the Linear SVM in Figure 4.

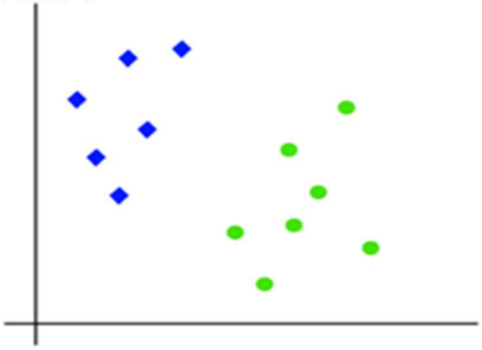


Figure 4: One-Dimensional Label Classification of Data

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

- *Non-Linear SVM*: Non-Linear SVM is a suitable approach for data that cannot be linearly separated. When a dataset cannot be effectively separated using a straight line, it is categorized as non-linear. In such cases, a Non-Linear SVM classifier is employed to perform classification tasks. Figure 5, illustrates a non-linear arrangement of 1-dimensional data, highlighting the need for a non-linear classifier to accurately classify such data.

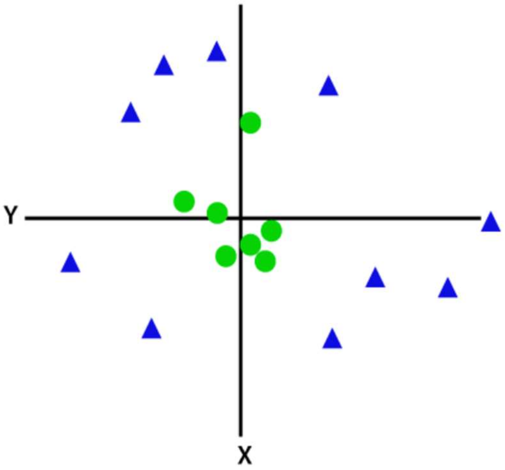


Figure 5: Non-Linear SVM Classification of Data into 1-D Labels.

3.4 Decision Tree

The Decision Tree (DT) algorithm is versatile in handling issues encompassing both classification and regression tasks. The name "tree" is given to the DT due to its resemblance to the branches of a tree. Similar to a tree structure, the decision tree starts at the "root node" and branches out. The branches in the decision tree represent the various possible outcomes at decision nodes, which are further referred to as leaf nodes once a final decision is made. This tree-like structure reflects the hierarchical nature of the decision-making process in the algorithm.

Let's analyze the decision tree constructed using the data of cervical cancer prediction as an illustration. The training data is initially divided into groups based on the first split during the decision tree construction process is referred to as the root node, which evaluates all attributes and features. As there are three features, three possible splits are taken into account. A cost function is subsequently utilized to compute the accuracy cost for each split. The split with the lowest cost is selected, in the case here, the number of sexual partners. This process is recursive, meaning that the same method is applied to the formed groups. This algorithm is referred to as a "greedy algorithm" due to the iterative and incremental nature of its decision-making process, driven by a strong desire to maximize accuracy.

Classification: C = sum (pk * (1 — pk))

In the context of classification, the formula for calculating the classification error or impurity measure is defined as $C = \Sigma(pk * (1 - pk))$, where pk represents the probability of each class in the dataset. This formula is commonly used to assess the impurity or misclassification rate in decision tree algorithms and other classification models. By minimizing this measure, we aim to achieve a more accurate and reliable classification outcome.

The Gini score is a metric that assesses the effectiveness of a split in segregating the target classes within the groups generated by the split. It is calculated by determining the proportion of inputs in a group that belongs to the same class, denoted as pk. A Gini score of 0.5 signifies an equal distribution (50/50) of classes, indicating the purest node. In binary classification, if all inputs in a group belong to the same class, pk is either 1 or 0, resulting in a Gini score of 0. Figure 6 illustrates the construction of a decision tree for a cervical cancer prediction dataset, demonstrating the process of selecting splits based on Gini scores to maximize the segregation of response classes.

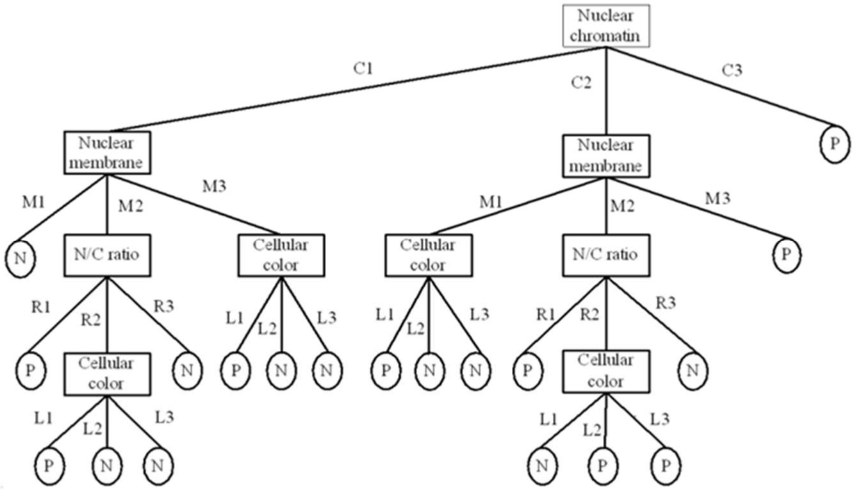


Figure 6: Construction of DT for Predicting CC

3.5 Ensembled Cervical Decision Tree Algorithm (ECDTA)

The proffered algorithm consists of two phases. In the first phase, a decision tree is constructed using the dataset obtained from the Kaggle repository. This decision tree is put up by recursively selecting the input variable that provides the highest information gain, and each edge that connects the nodes in the decision tree corresponds to a distinct input variable. The objective of the decision tree is to forecast the value of the target variable.

In the second phase, an ensembled boosting algorithm is enforced on the decision tree obtained in the first phase. This algorithm prognosticates the model's accuracy by using a confusion matrix. The suggested algorithm has attained the topmost accuracy compared to existing classifiers.

The model's process flow is illustrated in Figure 7, depicting the sequential steps involved in constructing the decision tree and using it for accurate prediction. The decision tree construction is based on the information gain criterion, and the boosting algorithm enhances the predictive accuracy of the model.

.

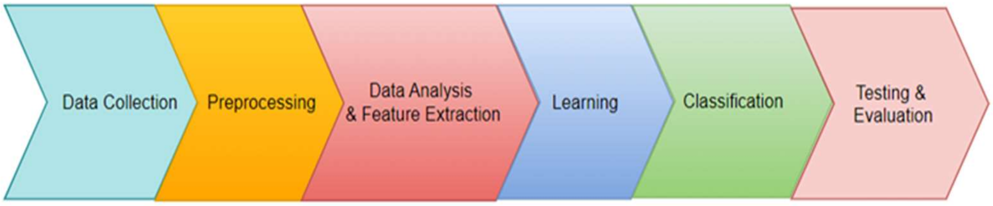


Figure 7: Workflow Model of Proposed Classifier

The proposed model begins with the data collection phase, where a dataset is obtained from the Kaggle repository. The dataset contains more than 2000 records and includes more than 35 attributes. The dataset also contains a target variable, which will be used to predict the outcome of the proposed algorithm.

Before proceeding with the algorithm, it is important to address any noisy data present in the dataset. Noisy data refers to data that is erroneous, inconsistent or contains outliers. To ensure accurate predictions, the noisy data needs to be identified and removed from the dataset.

By removing the noisy data, the proposed algorithm aims to enhance the accuracy of the model. Clean and reliable data is crucial for the success of any predictive algorithm, as it allows for more accurate and meaningful insights to be derived from the data.

Indeed, to address the issue of noisy data, preprocessing techniques need to be applied to the dataset. Preprocessing comprises various steps such as cleansing the data, transformation of the data, and extraction of features to ensure that the data is in a suitable format for analysis.

In the proposed model, the data obtained in the second phase has undergone preprocessing to remove any noisy data. Various Python methods and techniques have been utilized to extract features specifically related to cervical cancer. With the target variable being biopsy, the focus is on predicting the accuracy of the model based on the 36 attributes present in the dataset.

Comparative analysis has been performed with two other models, and the proposed model has demonstrated the highest accuracy among them. This indicates that the proposed model is effective in predicting the outcomes related to cervical cancer based on the selected attributes. As data analytics continues to play a crucial role in various domains, including health-related product buying and pharma, the proposed model contributes to the advancements in this field by providing accurate predictions for better decision-making.

The data is divided into a training set and a testing set, with the training set being used to train the classifier, the testing set evaluates the model's performance. Only the data from the training set should be used for predicting the accuracy of the model. Figure 8 provides a visual representation of the distinction between the training set and the testing data, highlighting the separation between the two datasets and their respective roles in the model evaluation process.

Classification is a process of categorizing a large number of items into distinct groups based on certain criteria. This can be done using various techniques such as function rules, class breaks, or formulas. In supervised classification, known class maps and attributes are used to establish the classification criteria, while unsupervised classification is performed when there are no known instances of the class available. Clustering is a popular method used in unsupervised classification, where data points are grouped based on their similarities. This technique is commonly employed in various domains, including retail product affinity analysis and fraud detection. In the proposed model, a combination of supervised learning and clustering is utilized to predict the accuracy of the model. The supervised learning component leverages known class information to guide the classification process, while clustering helps identify patterns and relationships within the data.

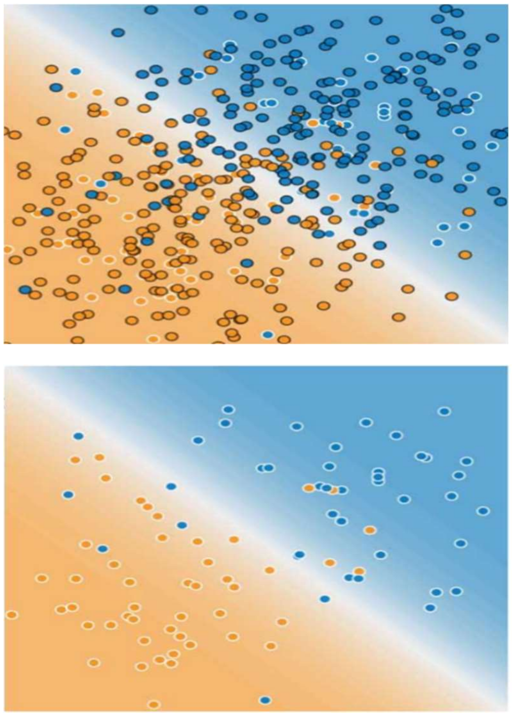


Figure 8: Training Set and Testing Set

To assess the information gained from the dataset, the concept of entropy is employed. The entropy of the dataset, denoted as D, is calculated to measure the impurity or uncertainty within the data. This measure plays a crucial role in determining the attributes that contribute the most to the classification process.

By incorporating both supervised learning and clustering techniques, the proposed model aims to enhance the accuracy of predictions and provide valuable insights for decision-making

$$Entropy(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

We determine the information gain from entropy for the cervical data which is:

$$Gain(D,A) = Entropy(D) - \sum_{j=1}^{v} \frac{|D_v|}{|D|} Entropy(D_v)$$

Best split is based on the attribute with best normalized information gain

$$Split(D, A) = \sum_{j=1}^{v} \frac{|D_v|}{|D|} \times \log_2 \left(\frac{|D_v|}{|D|} \right)$$

The gain ratio is obtained from the split and gain is later used to construct the Decision tree:

$$GainRatio(D, A) = \frac{Gain(D, A)}{Split(D, A)}$$

Figure 9 visually illustrates the progression of cervical cancer, starting from the normal stage and advancing toward the later stages

.

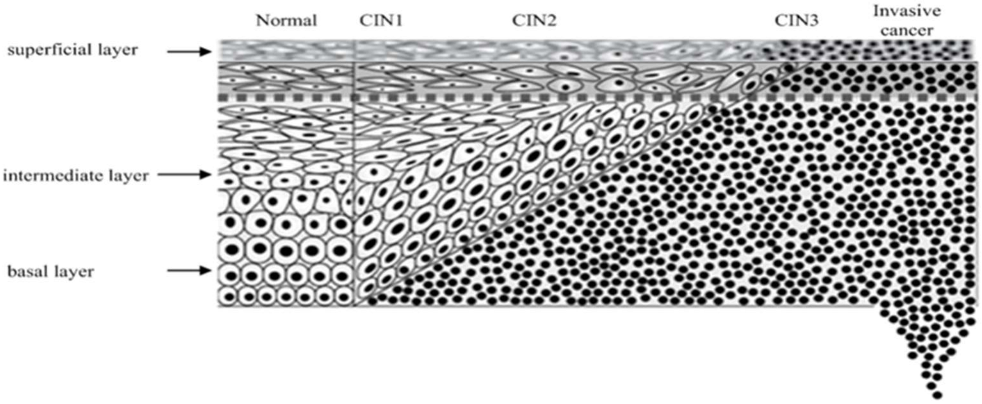


Figure 9: Cervical Cancer from Normal to Massive Level

The architecture of the proposed model is depicted in Figure 10.

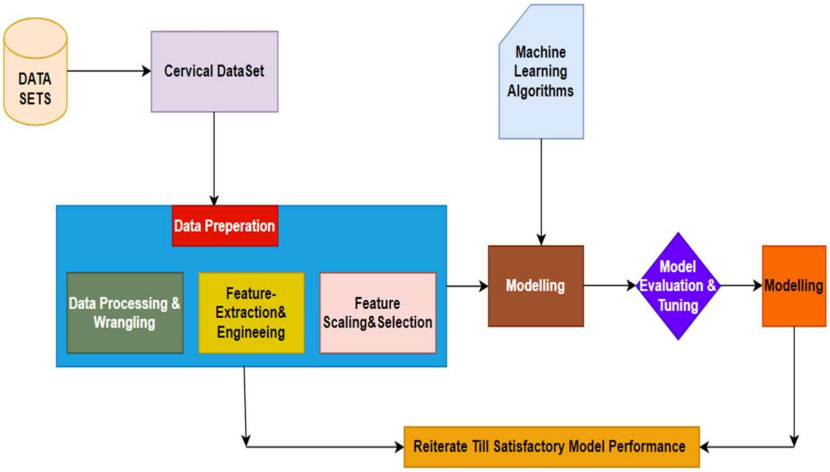

DATA SETS
Cervical DataSet
Machine Learning Algorithms
Data Preperation
Data Processing & Wrangling
Feature-Extraction& Engineeing
Feature Scaling&Selection
Modelling
Model Evaluation & Tuning
Modelling
Reiterate Till Satisfactory Model Performance

Figure 10: The model architecture

In the first phase of the proposed model, data is collected from various repositories. The collected data undergoes pre-processing, which involves cleaning, transforming, and organizing the data. Each sample in the data is then labeled as either positive (+ve) or negative (-ve) based on the classification task.

The labeled data is then used for training and testing purposes.

Predicting the highest accuracy for a classifier model can be achieved by following a set of rules that can be applied to any research subject

The workflow involves several steps. In the first step, we develop a proposed classifier. In the second step, we upload the cervical cancer prediction data. The third step entails training the proposed classifier using our best model. Moving on to the fourth step, we conduct testing to evaluate the performance of the classifier. Finally, we visualize the predictions made by the classifier.

3.6 Proposed Algorithm:

Phase -I

```
Construction of Decision Tree: (Input
Training Dataset Cervical_Cancer)
 Splitting(Cervical Cancer)
Construct D-tree ()
For records in X in the same class:
return
else: for every attribute X:
build the best split for splitting int0 C1, C2,
C3.
recursively call Split(C1)
recursively call Split(C2)
recursively call Split(C1)
then call Gain(D,A),Split(D,A),GainRatio(
        D,A)
Calculate Nuclear Chromatin
if Nuclear Chromatin is of y[c]¡=0.00 to
1.025:
 return C1
elif Nuclear Chromatin is of y[c]¡=1.026 to
1.125:
return C2
else:
return C3
return Decision_Tree.
stop

return Decision_Tree.
stop
```

Phase 2 Algorithm Now call the function ECDTA ()

```
Z= z. ECDTA_Classifier()
Z.fit(X train, Y train)
then call ConstructTree()
P= Generate confusion matrix (Y test, Y
pred)
Find Accuracy n = No of True Predictions /
Total Predictions
return P
Stop
```

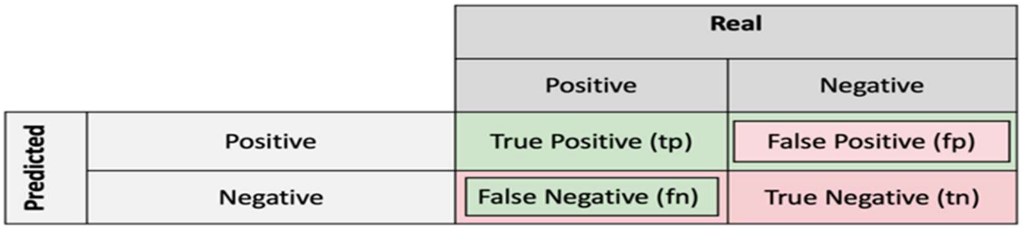
4. RESULTS AND DISCUSSION

Specific guidelines should be followed in research to assess the accuracy of predictive models, including measures such as sensitivity, specificity, and F-score. The sensitivity indicates the ability of a test to correctly identify specific conditions, such as pre-cancer. However, high sensitivity may result in a lower overall accuracy as it may detect fewer cases of pre-cancer. On the other hand, specificity measures the ability to correctly identify cases without the condition, such as the normal cervix. A high specificity may result in fewer cases of the normal cervix being detected, potentially affecting the accuracy of the model. Balancing both sensitivity and specificity is crucial to achieve the best predictive accuracy. These measures can be calculated based on the parameters obtained from a confusion matrix, as depicted in Figure 11.

Sensitivity (%) = TP/(TP+FN) where TP: True Positive and FN: False Negative

Specificity (%) = TN/(FP+TN) where TN: True Negative and FP: False Positive

Accuracy (%) = (TP + TN)/ (TP+FN+TN+FP)



		Real	
		Positive	Negative
Predicted	Positive	True Positive (tp)	False Positive (fp)
	Negative	False Negative (fn)	True Negative (tn)

Figure 11: The Generalized Confusion Matrix to Find Specificity, Sensitivity.

Figure 12 illustrates the distribution of individuals who are susceptible to developing cervical cancer based on their age. The count of individuals within different age groups provides insights into the prevalence of cervical cancer risk across different age ranges.

Figure 13 presents the count of sexual partners among individuals who are at a higher risk of developing cervical cancer. This information helps in understanding the association between the number of sexual partners and the risk of developing cervical cancer.

Figure 14 displays the count of individuals who are at risk of cervical cancer based on the age at which they had their first sexual intercourse. This data offers valuable insights into the potential correlation between the age at first sexual intercourse and the risk of cervical cancer.

These figures contribute to a better understanding of the demographic and behavioral factors linked to the risk of developing cervical cancer, helping researchers and healthcare professionals in designing targeted prevention and intervention strategies.

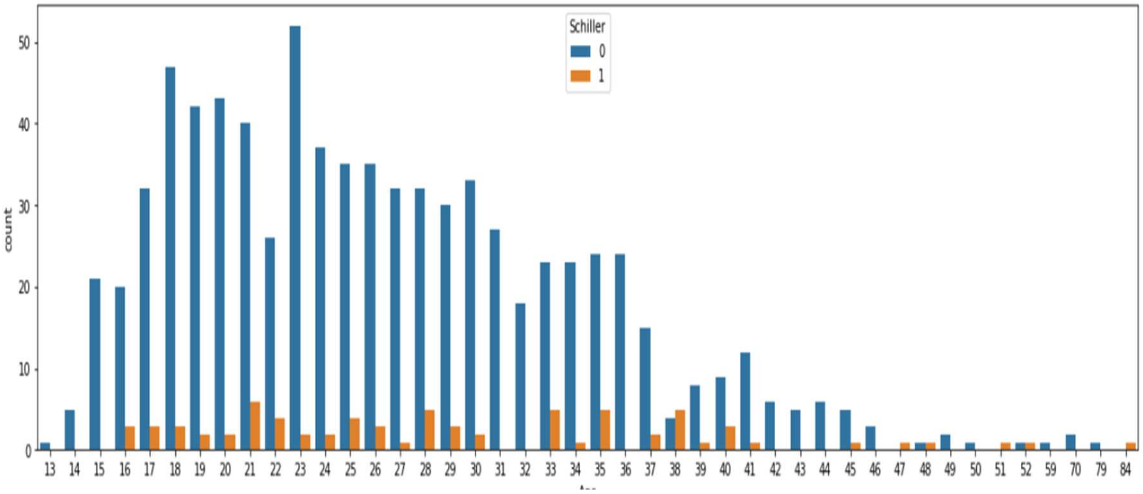


Figure 12: The Count Vs Age

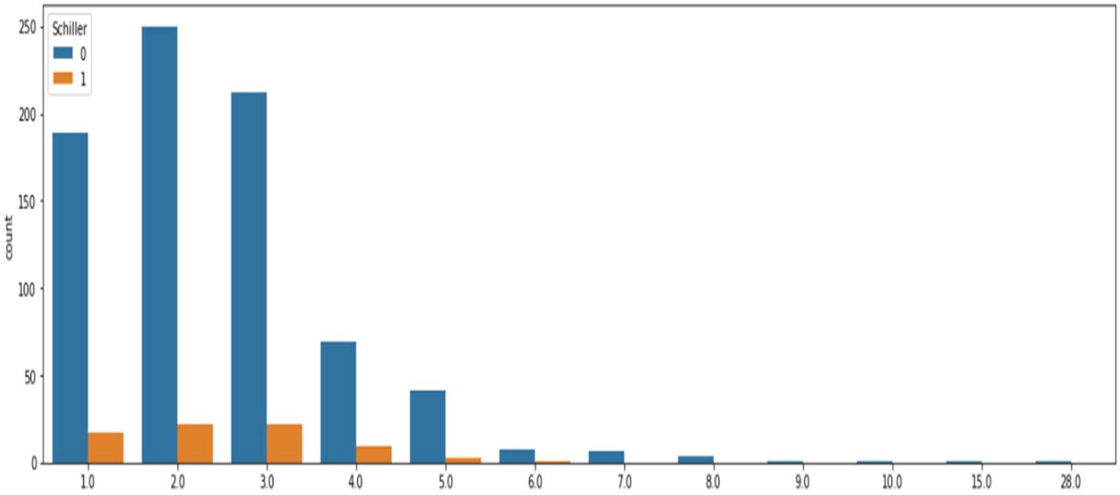


Figure 13: The Count Vs No of Sexual Partners

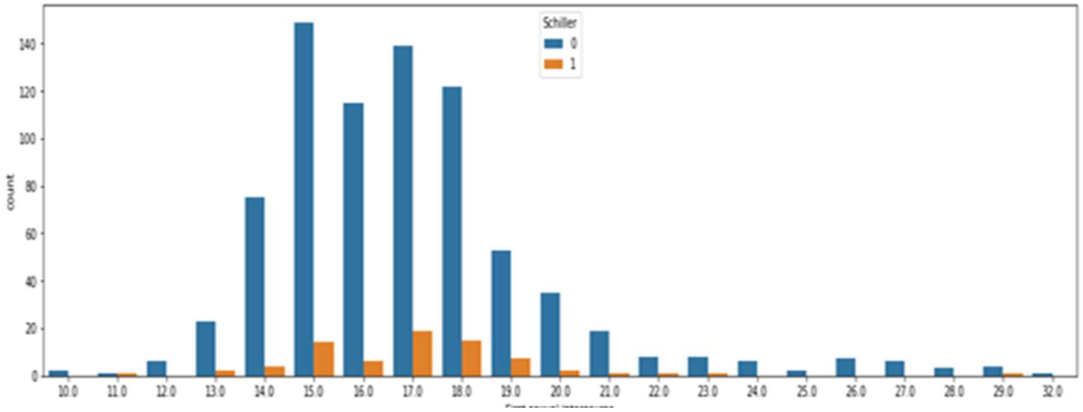
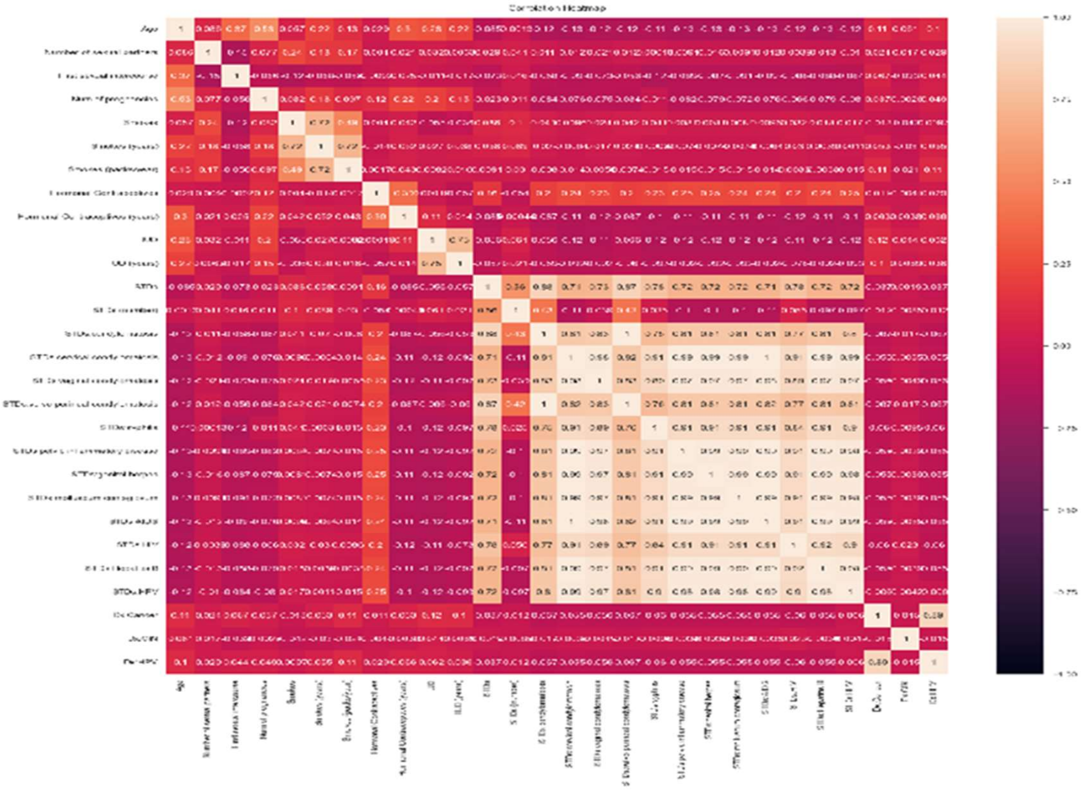


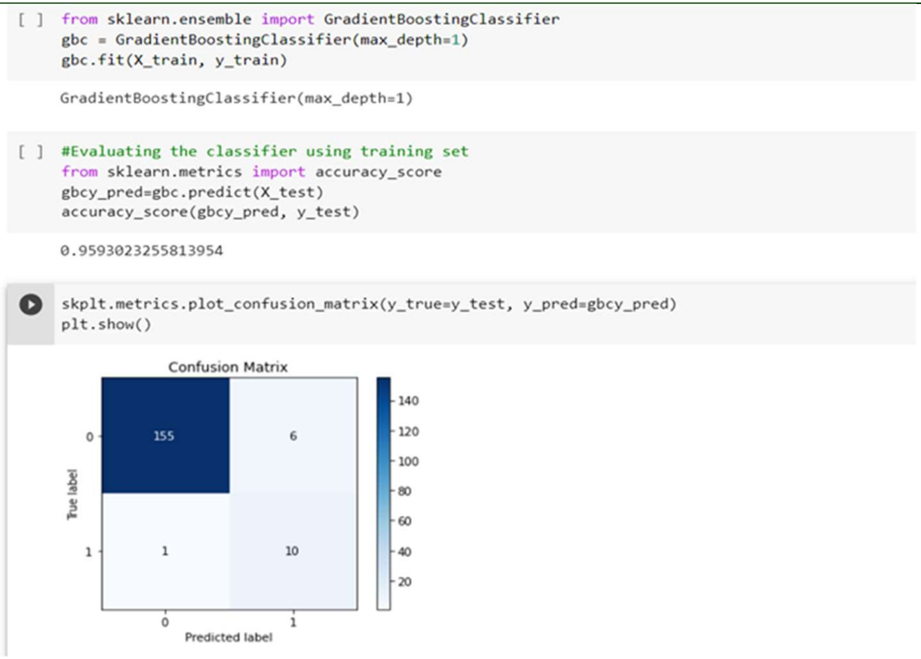
Figure 14: The Count Vs First Sexual Intercourse

Based on the Decision tree a heat map is generated which is used for predicting the accuracy using the following testing parameters like sensitivity[16], specificity, and accuracy of the proposed model which is shown in Figure 15 and Figure 16 shows the accuracy of the proposed model which achieved approx..95%.



.

Figure 15: Generated Heat Map



```
[ ] from sklearn.ensemble import GradientBoostingClassifier
    gbc = GradientBoostingClassifier(max_depth=1)
    gbc.fit(X_train, y_train)

    GradientBoostingClassifier(max_depth=1)

[ ] #Evaluating the classifier using training set
    from sklearn.metrics import accuracy_score
    gbcy_pred=gbc.predict(X_test)
    accuracy_score(gbcy_pred, y_test)

    0.9593023255813954

    skplt.metrics.plot_confusion_matrix(y_true=y_test, y_pred=gbcy_pred)
    plt.show()
```

Figure 16: Shows the sample code for Predicting the Accuracy.

Table 1 presents the confusion matrix obtained from the tested data using the Decision Tree algorithm.

Table 1: The Dt Classifier's Accuracy Rate

Label	Precision	Recall	F1-Score	Support
0	0.86	0.80	0.83	199
1	0.82	0.87	0.84	203
Accuracy			0.84	402
Macro Avg	0.84	0.84	0.84	402
Weighted Avg	0.84	0.84	0.84	402

Table 2 displays the confusion matrix derived from the tested data using SVM:

Table 2: The SVM Classifier's Accuracy Rate

Label	Precision	Recall	F1-Score	Support
0	0.86	0.76	0.81	199
1	0.79	0.88	0.83	203
Accuracy			0.82	402
Macro Avg	0.83	0.82	0.82	402
Weighted Avg	0.83	0.82	0.82	402

Table 3 showcases the confusion matrix acquired from the tested data using the Proposed Classifier.

Table 3: The Proposed Classifier's Accuracy Rate

Label	Precision	Recall	F1-Score	Support
0	0.95	0.88	0.92	199
1	0.89	0.96	0.92	203
Accuracy			0.92	402
Macro Avg	0.92	0.92	0.92	402
Weighted Avg	0.92	0.92	0.92	402

Based on the information presented in the aforementioned tables, it can be inferred that the recommended classifier exhibits the topmost accuracy compared to the existing classifiers, as depicted in Figure 17.

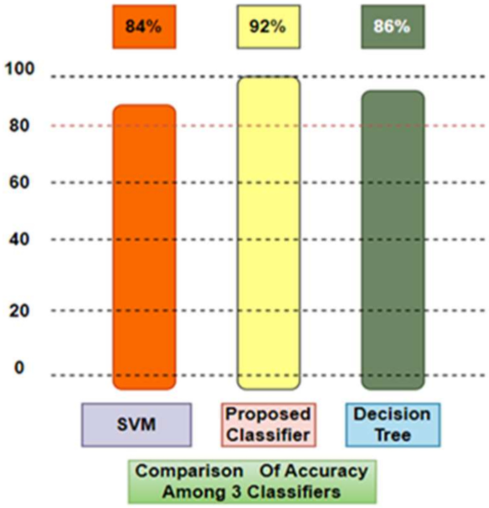


Figure 17: The Comparison of 3 Classifiers.

5. CONCLUSION & FUTURE WORK

Finally, the cumulative scientific advances in the field of machine learning-based cervical cancer prediction are definitely transformational. The combination of state-of-the-art algorithms, big datasets, and cross-disciplinary partnerships has enhanced the precision, usability, and significance of predictive models. We recognise the potential to change the paradigm from reactive treatment techniques to proactive preventative tactics by highlighting these scientific breakthroughs. The fight against cervical cancer is still ongoing, but the advances gained by ML research give people optimism that one day, early detection will be a crucial component of women's healthcare. In the future, ongoing research efforts will play a crucial role in improving current models, investigating fresh data sources, and tackling the changing issues in this important subject.

The research developed a mathematical machine learning model to analyze different scenarios of cervical cancer. A set of eight body parameters was utilized for prediction exploration. The study examines the progression of cervical cancer and its associated risk factors. Before delving into statistical tools, machine learning, and detection methods, the authors identify research gaps. Classification models were crafted using the proposed classifier, decision trees, and SVM. The performance of each technique was evaluated, considering optimal conditions. The article scrutinizes the reliability and effectiveness of these approaches based on the gathered data. The proposed algorithms consistently demonstrated strong performance in accuracy, precision, and other metrics for determining cervical cancer risk and type. The experiment's results highlight the superiority of the proposed algorithm compared to alternative machine learning-based methods. This comprehensive analysis could also evaluate the diagnostic utility of fully automated feature extraction in future studies. To advance cervical cancer prediction research, incorporating new technologies like blockchain and deep learning, along with socio-demographic factors such as region and education level, is crucial. Educational institutions can play a role in promoting healthcare awareness among the families they serve.

REFERENCES:

[1] W. Luo, “Predicting cervical cancer outcomes: statistics, images, and machine learning,” *Frontiers in Artificial Intelligence*, ,vol. 4, article 627369, 2021.

[2] A. Jajodia, A. Gupta, H. Prosch et al., “Combination of radiomics and machine learning with diffusion-weighted MR imaging for clinical outcome prognostication in cervical cancer,” *Tomography*, vol. 7, no. 3, 2021, pp. 344–357.

[3] D. Ding, T. Lang, D. Zou et al., “Machine learning-based prediction of survival prognosis in cervical cancer,” *BMC Bioinformatics*, vol. 22, no. 1, 2021, pp. 331–331.

[4] L. Akter, M. M. Ferdib -Al-Islam, M. S. Islam, M. R. Al-Rakhami, and Haque, “Prediction of cervical cancer from behavior risk using machine learning techniques,” *SN Computer Science*, vol. 2, no. 3, 2021.

[5] A. Arora, Tripathi, and Bhan, “Classification of cervical cancer detection using machine learning algorithms,” in *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, 2021, pp. 1075–1098.

[6] R. Weegar and K. Sundström, “Using machine learning for predicting cervical cancer from Swedish electronic health records by mining hierarchical representations,” *PLoS One*, vol. 15, no. 8, 2020, pp. 237911–237911.

[7] T. R. Ramesh, U. K. Lilhore, M. Poongodi, S. Simaiya, A. Kaur, and M. Hamdi, “Preditive analysis of heart diseases with machine learning approaches,” *Malaysian Journal of Computer Science*, 2022, pp. 132–148.

[8] S. Kim, S. Lee, C. H. Choi et al., “Machine learning models to predict survival outcomes according to the surgical approach of primary

radical hysterectomy in patients with early cervical cancer," *Cancers*, vol. 13, no. 15, 2021, p. 3709.

[9] D. T. S. Patel, "A cross sectional study to estimate delay in diagnosis and treatment of tuberculosis (TB) among patients attending urban health centre in an urban slum area," *Public Health Review: International Journal of Public Health Research*, vol. 5, no. 1, 2018, pp. 1–7.

[10] L. Gupta, A. Edelen, N. Neveu, A. Mishra, C. Mayes, and Y. K. Kim, "Improving surrogate model accuracy for the LCLS-II injector frontend using convolutional neural networks and transfer learning," *Machine Learning: Science and Technology*, vol. 2, no. 4, 2021, pp. 1245–1265.

[11] Kondratenko Y., Atamanyuk I., Sidenko I., Kondratenko G., Sichevskyi S. Machine Learning Techniques for Increasing Efficiency of the Robot's Sensor and Control Information Processing. Sensors. 2022;22:1062.,doi: 10.3390/s22031062.

[12] Hall M.A. Correlation-Based Feature Selection for Machine Learning. The University of Waikato; Hamilton, New Zealand: 1999.

[13] Dokduang K., Chiewchanwattana S., Sunat K., Tangvoraphonkchai V. A comparative machine learning algorithm to predict the bone metastasis cervical cancer with imbalance data problem. Recent Adv. Inf. Commun. Technol. 2014; 10:93–102.

[14] Ghoneim A., Muhammad G., Hossain M.S. Cervical cancer classification using convolutional neural networks and extreme learning machines. Future Gener. Comput. Syst. 2020;102:643–649. doi: 10.1016/j.future.2019.09.015.

[15] Ashok B., Aruna P. Comparison of Feature selection methods for diagnosis of cervical cancer using SVM classifier. Int. J. Eng. Res. 2016;6:94–99.

[16] Pathuri, S.K.; Anbazhagan, N. Feature-Based Sentimental Analysis on Product Review System Using CUDA-BB Algorithm. Int. J. Emerg. Trends Eng. Res. 2018, 8, 6380–6386.