Cervical Cancer Detection Using Machine Learning

[1]Ms.S. Deepika, [2]M. Sai Suma, [3]A. Sowmya Reddy , [4]Ch. Manideep Reddy

[1]Assistant Professor, Department of Artificial Intelligence,
Anurag Group Of Institutions,
[2,3,4]Student,
[2,3,4] Department of Artificial Intelligence,
Anurag Group Of Institutions, Hyderabad, India,

***Abstract*— Cervical cancer is one of the most deadly diseases in the world among women. It is caused by long term infection in skin cells and mucous membrane cells of the genital area. The likelihood of timely treatment for cervical cancer increases with timely detection of abnormal cervical cells. Automated methods of detecting abnormal cervical cells were established because manual identification requires skilled pathologists and is time consuming and prone to error. The goal of this research paper is to utilize both exploratory data analysis (EDA) and machine learning algorithms to gain a thorough understanding of the risk factors associated with cervical cancer, as well as to develop a dependable approach for detecting the disease. Exploratory data analysis (EDA) is an essential step in understanding the patterns and relationships that exist within a dataset. In the case of cervical cancer, EDA can help us identify the factors that are most likely to be responsible for the development of the disease. By exploring the data, we can gain insight into the various risk factors associated with cervical cancer, including age, sexual activity, family history, and exposure to the human papillomavirus (HPV). The KNeighbors Classifier with n_neighbors = 2 using normalization achieved the highest accuracy of 0.9992, indicating that this model was able to predict cervical cancer with a high degree of accuracy.**

Keywords: Cervical Cancer, Exploratory Data Analysis, Machine Learning, Human Papillomavirus, Normalization.

I. INTRODUCTION

Cervical cancer has been a major cause for death worldwide from the last few decades. It is a third main type of cancer after the lungs and breast cancer among women. It is only cured when detected and treated at its early stage. It is the fourth most frequently diagnosed cancer and the fourth leading cause of cancer death in women, which accounts for 570,000 incident cases and 310,000 deaths each year worldwide. The timely detection of cervical cancer plays a crucial role in enhancing patient outcomes and reducing mortality rates. Machine learning techniques can be employed to examine cervical cancer screening test outcomes and facilitate the early detection of cervical cancer. With the ability to process significant amounts of data, machine learning algorithms can recognize patterns and make precise predictions. These algorithms can be trained with data from screening tests to identify characteristics that may indicate cervical cancer, such as unusual cell morphology or HPV infection. Using machine learning algorithms for cervical cancer detection has a significant advantage: it can automate the screening process, resulting in faster and more accurate diagnosis while lowering costs. Cervical cancer is the most commonly diagnosed cancer in 23 countries in the world and is the leading cause of cancer death in 36 countries, with the vast majority of these countries found in sub-Saharan Africa, Melanesia, South America, and South-Eastern Asia. Certain groups of women like women who smoke who have a weak immune system, who are infected with HIV and/or who have undergone some organ transplant are at higher risk than others to get infected with cervical cancer.

Overall, the use of ML algorithms for cervical cancer detection shows promise as a valuable tool for improving patient outcomes and reducing mortality rates associated with this disease.Our project provides an effective model for improving the performance of using machine learning methods and classification techniques for predicting cervical cancer. It focuses also on the sensitivity and overall accuracy of the model. The cervical cancer dataset contains indicators and risk factors for predicting whether a woman will get cervical cancer. The features include age, number of pregnancies, number of sexual partners, smokes, time since first and last diagnosis and medical history.

The objective is to perform exploratory data analysis that demonstrates the correlations between the parameters that are likely to be responsible for cervical cancer and also to detect 2 cervical cancer using machine learning algorithms. EDA is an important part of data analysis. It is used to investigate the datasets and summarize the main characteristics. EDA is an analysis approach that identifies general patterns in the data. Machine learning algorithms have the potential to assist in the detection of cervical cancer. These algorithms can be trained on a large dataset of patient information, including demographic data, medical history, and test results. By analyzing this data, machine learning algorithms can identify patterns and correlations that are indicative of cervical cancer. One of the primary advantages of using machine learning algorithms for cervical cancer detection is that they can be trained to identify patterns and correlations that may not be immediately apparent to human analysts. Additionally, machine learning algorithms can process vast amounts of data quickly and efficiently, making them ideal for analyzing large datasets.

II. LITERATURE SURVEY

As an initial step, a comprehensive review of the existing literature was conducted.

In a recent study by Gogate U. and Bakel J. in 2018, wireless bio sensors were utilized to measure temperature, heart rate, and oxygen levels in the body using an Arduino Nano board. The data collected from these sensors were processed through an IOT

application called Think-speak to generate predictions and alerts for physicians to manage their patients. The proposed method was tested on cardiac patients and was found to be 95% accurate in predicting disease causes and heart functioning disorders.
Similarly, Rahmani A. M. et. al. in 2015 introduced the concept of a smart e-health gateway for a ubiquitous health monitoring system using IOT. Their idea of developing smart homes and hospitals has gained popularity in recent times due to the internet explosion. The deployment of smart gateways for e-health has the potential to address various healthcare system issues such as reliability, scalability, energy efficiency, interoperability, and accuracy.

The study conducted by Jaswinder Singh and Sandeep Sharma in 2019 aimed to develop and evaluate a machine learning model for cervical cancer prediction using data from the National Cancer Institute's SEER program. To achieve this, the authors utilized several machine learning algorithms, including decision trees, k-nearest neighbors, and support vector machines, to construct their predictive model. The dataset used for the study contained 819 records of cervical cancer patients and 195 records of healthy patients, and feature selection methods were employed to determine the most critical variables for predicting cervical cancer. The performance of different machine learning algorithms was then compared in terms of accuracy, sensitivity, specificity, and AUC-ROC. Furthermore, the paper also highlights some limitations and future directions for research in this field, such as the requirement for validation of the model using data from various populations and the integration of more advanced machine learning techniques.

The paper authored by Surendiran R, Thangamani M, Monisha S, and Rajesh P in 2022 describes the application of machine learning and deep learning techniques for cervical cancer prediction. The authors trained and tested different machine learning models, such as logistic regression, decision trees, random forests, and support vector machines, using a dataset of patient information that included demographic data and medical history. Additionally, they utilized deep learning techniques, such as convolutional neural networks (CNNs) and long short-term memory (LSTM) networks, for the same purpose. The results showed that the LSTM network performed the best, with an accuracy of 92.31% compared to the other models. Overall, the study highlights the potential of using machine learning and deep learning techniques as a tool for early detection of cervical cancer.

III. PROBLEM STATEMENT

Cervical cancer is a significant public health concern globally, with over 500,000 new cases and 300,000 deaths reported each year. Early detection of cervical cancer is essential for effective treatment and improved patient outcomes. Traditional screening methods, such as Pap smear and visual inspection with acetic acid (VIA), have limitations in terms of accuracy and sensitivity. Therefore, there is a need to develop more reliable and accurate methods for cervical cancer detection.

3.1 Existing System

There are several existing solutions for cervical cancer detection including: HPV Testing, Pap Smear, Colposcopy, Biopsy, Molecular Testing. Recent advancements in the field of Artificial Intelligence (AI) have led to the development of computer vision-based screening platforms that use AI algorithms to accurately detect abnormalities in cervical cells from cytology samples or digital colposcopy images. These technologies have the potential to revolutionize cervical cancer screening by providing faster and more accurate diagnoses while reducing the need for human interpretation. AI-powered computer vision-based screening platforms are designed to analyze images taken from Pap smears or digital colposcopy images using machine learning algorithms. These algorithms are trained on large datasets of cervical cell images, enabling them to detect early signs of abnormal cell growth or cancer with a high degree of accuracy. One major advantage of these technologies is that they can process images much faster than human experts, which can greatly reduce the time required for screening and diagnosis.
AI systems can often be biased, and this can lead to false diagnosis. For instance, if the data used to train the algorithm is biased towards detecting cancer in certain women or certain types of cervical cancer, then the system will miss out on some cases or not accurately identify some cases. The use of AI systems for cervical cancer screening and diagnosis raises several ethical concerns, such as confidentiality issues, autonomy, and informed consent. Although AI algorithms have been employed to detect cervical cancer, the accuracy of the diagnosis is ultimately dependent on the quality of data used to train the algorithm. If the data is not representative of the large and diverse population of women in the general population, the diagnosis accuracy could be compromised.

3.2 Proposed System

Machine learning (ML) has shown promise in healthcare applications, including cancer detection. The aim of this project is to develop an ML-based system for the detection of cervical cancer using a textual dataset of patients' demographic and clinical information. The dataset will include features such as smoking history, age, STDs status, biopsy results, and other relevant information.

The proposed ML-based system will use various ML algorithms, such as logistic regression, decision trees, random forests, and support vector machines (SVM), to analyze the dataset and classify patients as having cervical cancer or not. The system will also use feature selection and engineering techniques to identify the most significant features that can predict cervical cancer accurately. The developed ML-based system could be used as a non-invasive and cost-effective screening tool for cervical cancer, especially in low-resource settings where traditional screening methods are not readily available. The project's success will be determined by the accuracy and sensitivity of the developed ML-based system in detecting cervical cancer compared to existing screening methods.

Overall, the proposed solution aims to develop a reliable and accurate ML-based system for the detection of cervical cancer that can improve patient outcomes and reduce the burden of the disease. The use of demographic and clinical information instead of images can make the screening process less invasive and more accessible to a wider population.

IV. SPECIFICATIONS

The development of software for the detection of cervical cancer using machine learning is a complex task that requires the use of various programming languages and libraries. Python is a popular programming language for data science and machine learning applications, and its powerful libraries make it an ideal choice for this project. Scikit-learn is a machine learning library in Python that provides various algorithms and tools for pattern recognition, classification, regression, and clustering. Pandas is a library that provides data structures for efficient data analysis and manipulation. NumPy is a library for numerical computations and matrix operations in Python. Matplotlib is a library for creating 2D plots and visualizations. The model will be evaluated using various performance metrics, such as sensitivity, specificity, precision, and F1 score. Overall, the use of Python and its popular libraries will enable the development of a robust and effective software for the detection of cervical cancer, which can potentially save countless lives by providing early and accurate diagnoses.

Software Specifications

- Programming language – Python
- IDE – Google Colaboratory
- Libraries: pandas, Numpy, Matplotlib, Seaborn, Scikit-learn, Warnings, LabelEncoder, Smote, catboost, ipywidgets.

Hardware specifications

- Processor - Intel i3 and above
- RAM - 4GB and Higher
- Hard Disk - 500 GB Min

V. EXPLORATORY DATA ANALYSIS

Exploratory data analysis is an important first step in data analysis. It is an analysis approach that identifies general patterns in the data. These patterns include outliers and features of the data that might be unexpected. Cervical cancer has been a major cause for death worldwide from the last few decades. To perform exploratory data analysis that demonstrates the correlations between the parameters that are likely to be responsible for Cervical cancer. The required dataset has been collected from Kaggle.

Dataset: https://www.kaggle.com/cervical-cancer-riskclassification

As the dataset plays a major role in analysis, we need to predict patterns using available data. The data is necessary as inputs to the analysis, which is specified based upon the requirements of those directing the analysis. Exploratory Data Analysis that demonstrates the effect of certain features like Hormonal contraceptives, intrauterine devices (IUD), sexually transmitted diseases (STD), etc. There are around 36 features include number of pregnancies, number of sexual partners, smokes ,time since first diagnosis, time since last diagnosis, STD, IUD, Biopsy, Schiller, Hinselmann, Cytology,etc. and there were 858 cases and related medical history in the dataset.

5.1 Effect of features on Target variables

Hormonal Contraceptives-. The use of hormonal contraceptives may increase the risk of cervical cancer. Performing correlation between each feature helps in finding what feature plays an important role in detecting cervical cancer. There is a positive correlation between the use of hormonal contraceptives and cervical cancer.

Intrauterine devices- IUDs are small devices inserted into the uterus that can effectively prevent pregnancy. IUDs may be associated with a reduced risk of cervical cancer, they do not provide protection against other sexually transmitted infections. Women who utilized IUDs had a lower incidence of cervical cancer.

Sexually transmitted diseases- Certain STD's can increase a persons' risk of developing several cancer types. STDs such as HPV, Chlamydia trachomatis and herpes simplex virus (HSV) could potentially be linked to cervical cancer. Correlation matrix of each feature represents that there is no correlation, it means no association between smokes and number of diagnoses, smokes and cytology, STD and Schiller, etc. There exists negative correlation between certain features like STD and AIDS, STD and HPV, biopsy and genital herpes.

5.2 Distribution of Target Variables

Biopsy- Based on a sample of 858 cases, it was found that the biopsy test for cervical cancer was negative in approximately 93.6% of women, indicating the presence of abnormal or cancerous cells in the cervix. On the other hand, around 6.4% of women had positive biopsy results, suggesting the absence of abnormal or cancerous cells in the cervix.

Hinselmann- This test can be precancerous and can be treated before they develop into cervical cancer. A negative Hinselmann test result, on the other hand, does not guarantee that a woman is cancer-free, and regular screening is still recommended for all women at risk of cervical cancer.

Schiller- Based on a sample of 858 cases, it was found that the Schiller test for cervical cancer was negative in approximately 91.4% of women. On the other hand, around 8.6% of women had positive Schiller test results. The conclusion is that a positive Schiller test result does not necessarily mean that a woman has cervical cancer.

Cytology- Cytology, Schiller, Hinselmann, Biopsy are the four main target features of cervical cancer. Among 858 cases, around 94.9% of women tested negative in cytology and 5.1% of women tested positive.

5.3 Feature Importance using RandomForest Classifier and ExtraTrees Classifier

Random Forest is a popular machine learning algorithm that can be used for classification and regression tasks.The feature importance score is calculated by measuring the decrease in impurity or entropy of the tree when a particular feature is used for splitting the data. The more a feature reduces impurity, the more important it is.

Extra Trees Classifier is a type of ensemble learning method used for classification tasks.. In Extra Trees, the splitting of nodes is done randomly and with a larger number of trees than in Random Forest. To reduce the variance in the scores, Extra Trees Classifier constructs more decision trees than Random Forest Classifier and then averages the feature importance scores over all the trees.

5.4 Feature Scaling

Feature scaling is the process of scaling or normalizing the range of values of independent variables or features in a dataset to make them comparable and consistent. In cervical cancer detection, raw feature scaling involves using the original values of the features, such as the patient's age or the HPV status, without any transformation. Normalized feature scaling, on the other hand, involves transforming the values of the features to a new range. There are several methods for normalizing features, such as min-max 22 scaling, z-score standardization, and robust scaling, each of which has its own advantages and disadvantages. Raw and normalized feature scaling are used to ensure that the input features are all on a similar scale and to improve the performance of the machine learning algorithm.

VI. RESULTS

This section presents the results of the experimentation task conducted using various classifiers. Different ML classifiers such as Support Vector Classifier (SVC), Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, Extreme Gradient Boosting Classifier (XGB), Multi-Layer Perceptron Classifier were deployed for the detection of cervical cancer.

Model Name	Feature Scaling	Accuracy	Precision	Recall	F1 Score
SVC	Raw	0.8964	0.90	0.89	0.89
SVC	Normalized	0.9720	0.97	0.97	0.97
Logistic Regression	Raw	0.9626	0.96	0.96	0.96
Logistic Regression	Normalized	0.9603	0.96	0.96	0.96
KNN(n=2)	Raw	0.9953	0.99	0.99	0.99
KNN(n=2)	Normalized	0.9992	1.00	1.00	1.00
Decision Tree	Raw	0.9836	0.98	0.98	0.98
Decision Tree	Normalized	0.9836	0.98	0.98	0.98
Random Forest	Raw	0.9735	0.97	0.97	0.97
Random Forest	Normalized	0.9735	0.97	0.97	0.97
Gradient Boosting	Raw	0.9856	0.96	0.97	0.97
Gradient Boosting	Normalized	0.9756	0.97	0.98	0.98
XGB Classifier	Raw	0.9843	0.97	0.98	0.98
XGB Classifier	Normalized	0.9843	0.97	0.98	0.98
MLP Classifier	Raw	0.9852	0.99	0.99	0.99
MLP Classifier	Normalized	0.9782	0.97	0.97	0.98

Table 6.1 Comparison of performance of algorithms

Table 6.1 compares the performance of different classification models on a given dataset. The metrics compared are accuracy, precision, recall, and F1-score, for both raw and normalized data. The table shows that K Neighbors Classifier with n_neighbors=2 performs the best among all models with an accuracy of 0.9953 for raw data and an accuracy of 0.9992 for normalized data. The DecisionTree Classifier also performs well with an accuracy of 31 0.9836 for both raw and normalized data. The other models also have high accuracy, but their performance varies slightly based on whether the data is raw or normalized.

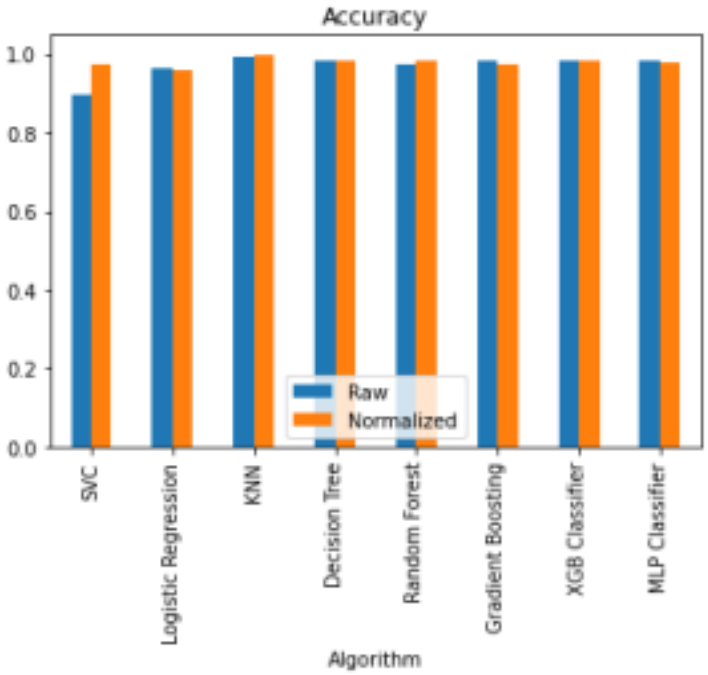

Fig 6.1 Accuracy graph comparing the algorithms

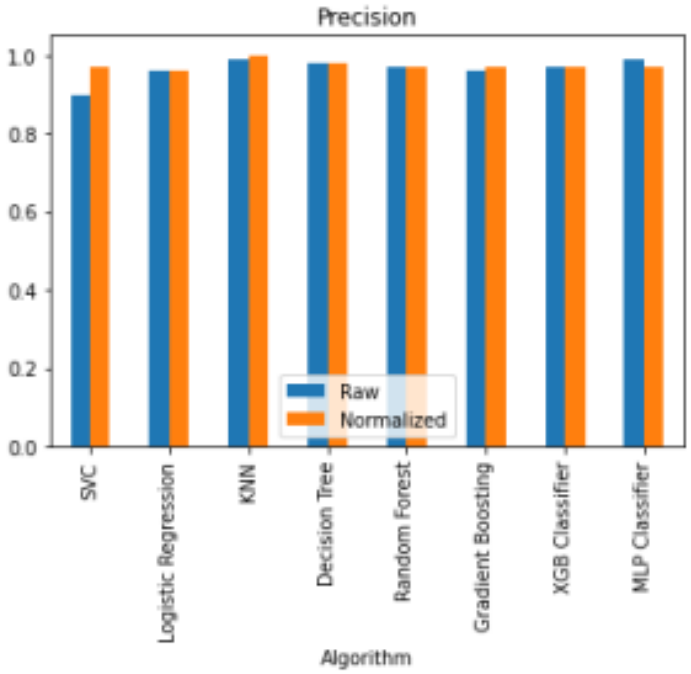

Fig 6.2 Precision graph comparing the algorithms

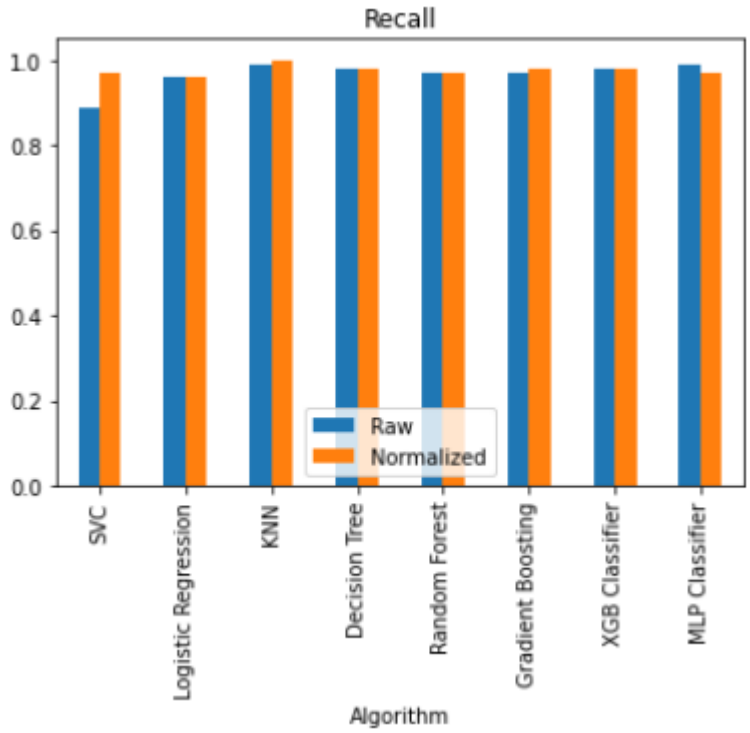

Fig 6.3 Recall graph comparing the algorithms

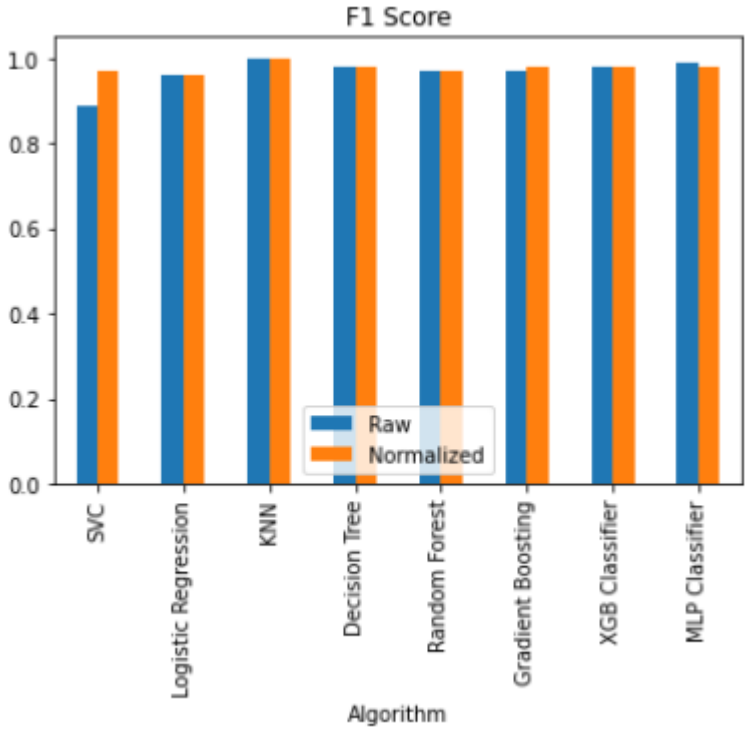

Fig 6.4 F1 score graph comparing the algorithms

VII. CONCLUSION

Based on the results of the various machine learning models applied to detect cervical cancer, it can be concluded that the models performed quite well overall with high levels of accuracy, precision, recall, and F1-score. The models used are: Support Vector Machine (SVM), Logistic Regression, K Neighbors Classifier, Decision Tree Classifier, and Random Forest Classifier, Gradient Boosting Classifier, Extreme Gradient Boosting (XGB) Classifier, Multi-Layer Perceptron Classifier. Each model was tested twice: once on the raw data and once on normalized data.

The results show that all models perform very well on this dataset, with high accuracy scores ranging from 0.8964 to 0.9992. The K Neighbors Classifier with n_neighbors = 2 using normalization achieved the highest accuracy of 0.9992, indicating

that this model was able to predict cervical cancer with a high degree of accuracy. It is important to note that while these models show promise for detecting cervical cancer, further research and validation is necessary before implementing them in real-world medical settings. Additionally, machine learning models should not replace standard screening methods such as Pap smears and HPV tests, but rather be used as an additional tool to aid in early detection and prevention.

REFERENCES

[1] Aditya Arora, Anurag Tripathi, Anupama Bhan, "Classification of Cervical Cancer Detection using Machine Learning Algorithms", 2021 6th International Conference on Inventive Computation Technologies, Coimbatore, India, February, 2021.
[2] Ling Zhang, Le Lu, Isabella Nogues, Ronald M. Summers, Shaoxiong Liu, Jianhua Yao, "Deep Convolutional Networks for Cervical Cell Classification", IEEE Journal of Biomedical Health Information, Nov 2017.
[3] Nasser H. Sweilam, A.A. Tharwat and N.K. Abdel Moniem, "Support Vector Machine for Diagnosis Cancer Disease: A comparative study", Egyptian Informatics Journal, vol. 11, no. 2, pp. 81-92, December 2010.
[4] Y.E. Kurniawati, A.E. Permansari and S. Fauziati, "Comparative study on data mining classification methods for cervical cancer prediction using pap smear results", 2016 1st International Conference on Biomedical Engineering (IBIOMED), pp. 1-5, 2016
[5] R. Vidya and G.M. Nasira, "Prediction of Cervical Cancer using Hybrid Induction Technique: A Solution for Human Hereditary Disease Patterns", Indian Journal of Science and Technology, August 2016.
[6] Jaswinder Singh, Sandeep Sharma, "Prediction of Cervical Cancer Using Machine Learning Techniques", International Journal of Applied Engineering Research (ISSN) , vol 14, no. 11, pp. 2570-2577, 2019.
[7] Min Chen, Yixue Hao, Kai Hwang, Lu Wang, Lin Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities", IEEE Access, vol. 5, pp. 8869-8879, April 2017.
[8] Kangkana Bora, Manish Chowdary, Malay Kumar Kundu, Anup Kumar Das, "Automated Classification of Pap Smear Images to Detect Cervical dysplasia", Computer Methods and Programs in Biomedicine, vol. 138, pp. 31-47, January 2017.
[9] R.S. and S. Rajesh Kumar, "Detection and Classification of Cancer from Microscopic Biopsy Images Using Clinically Significant and Biologically Interpretable Features", Journal of Medical Engineering, vol. 2015, august 2015.
[10] Leila Allahquoli, Antonio Simone Lagana, Afrooz Mazidimoradi, Hamid Salehiniya, Veronika Gunther, Vito Chiantera, Shhirin Karimi Goghari, Mohmmad Matin Ghiasvand, Azam Rahami, Zohre Momenimovahed and Ibrahim Alkatout, "Diagnosis of Cervical Cancer and Pre-Cancerous Lesions by AI: A Systematic Review", MPDI, Diagnostics, November 2022.