V.Mohanraj[1], J.Senthilkumar[2], Y.Suresh[3], B.Valaramathi[4], S.Sivanantham[5]

Professor,Information Technology,Sona College of Technology,Salem,Tamilnadu,India (1)
Professor,Information Technology,Sona College of Technology,Salem,Tamilnadu,India (2)
Professor,Information Technology,Sona College of Technology,Salem,Tamilnadu,India (3)
Professor,Sofware and Systems Engineering,School of CSE and IS,VIT,Vellore,Tamilnadu,India (4)
Professor, CSE, Saveetha School of Engineering, SIMATS, Chennai, Tamilnadu, India (5)
ORCID:1. 0000-0002-5254-9279 2.0000-0002-9320-316X 3. 0000-0002-8764-8966 4. 0000-0002-2540-121X 5. 0000-0001-5215-8211

doi:10.15199/48.2024.07.19

Prediction of Preliminary Tests for Cervical Cancer using Artificial Intelligence Models

Abstract. *Nowadays, Aritificial Intellgience (AI) based models are extensively used in the medical science for early detection of choronic diseases. AI model plays a vital role in detecting cervical cancer in women at early stage. Cervical cancer is abnormal growth of cells in the cervix. Vagina is connected to uterus through the cervix. Mostly, various strains of Human papillomavirus (HPV) cause the infection over the cervix. A prolonged virus infection over cervix causes some cervical cells become cancer cells. It is difficult to dectect early sign of the cervical cancer. The proposed method explores cervical cancer detection and provides information on the necessary tests to be taken.The initial level of testing is achieved by getting information from users directly and processing it using a Decision Tree based classifier model. The classifier provide information on the mandatory tests that have to be taken. Then the secondary level of testing is carried out using Deep Convolution Neural Network model over a Colposcopy image of the cervix to identify the tumor region in the cervix. The model predicts the causes of cervical cancer based on the collected user information. The performance of the algorithm is evaluated based on Test accuracy, Recall, and precision. The highest cervical cancer prediction accuracy is achieved through AI model comprising Decision Tree and Deep Convolution Neural network model.*

Streszczenie. *Obecnie modele oparte na sztucznej inteligencji (AI) są szeroko stosowane w naukach medycznych do wczesnego wykrywania chorób kosmówkowych. Model AI odgrywa kluczową rolę w wykrywaniu raka szyjki macicy u kobiet we wczesnym stadium. Rak szyjki macicy to nieprawidłowy rozrost komórek szyjki macicy. Pochwa jest połączona z macicą poprzez szyjkę macicy. Zakażenie szyjki macicy powodują głównie różne szczepy wirusa brodawczaka ludzkiego (HPV). Długotrwała infekcja wirusowa szyjki macicy powoduje, że niektóre komórki szyjki macicy stają się komórkami nowotworowymi. Trudno jest wykryć wczesne objawy raka szyjki macicy. Proponowana metoda bada wykrywanie raka szyjki macicy i dostarcza informacji na temat niezbędnych badań, które należy wykonać. Początkowy poziom badań osiąga się poprzez bezpośrednie uzyskanie informacji od użytkowników i przetworzenie ich przy użyciu modelu klasyfikatora opartego na drzewie decyzyjnym. Klasyfikator dostarcza informacji na temat obowiązkowych badań, które należy wykonać. Następnie przeprowadza się drugi poziom badań, wykorzystując model sieci neuronowej o głębokim splocie na podstawie obrazu szyjki macicy z kolposkopii w celu zidentyfikowania obszaru nowotworu w szyjce macicy. Model przewiduje przyczyny raka szyjki macicy na podstawie zebranych informacji od użytkownika. Wydajność algorytmu ocenia się na podstawie dokładności testu, przypomnienia i precyzji. Najwyższą dokładność przewidywania raka szyjki macicy osiąga się dzięki modelowi AI obejmującemu drzewo decyzyjne i model sieci neuronowej Deep Convolution.* (***Przewidywanie wstępnych testów na raka szyjki macicy przy użyciu modeli sztucznej int***eligencji)

Keywords: Cervical Cancer, Deep Convolution Neural Network, Decision Tree Classifier, Accuracy, Precision
Słowa kluczowe: Rak szyjki macicy, sieć neuronowa o głębokim splocie, klasyfikator drzewa decyzyjnego, dokładność, precyzja

Introduction

Nowadays, there is increased number of women affected by cervical cancer due to improper habits such as smoking, sexual intercourse with multiple partner, and many pregnancies.The tissue that links the uterus and vagina is called the cervix, and it has the shape of a drum or cylinder. The cervix, which is found at the base of the uterus, is primarily made of fibromuscular tissue.These new cells will mature, become aged, and then it will die. At the moment, new cells take their place and replace them. This is called a normal cell cycle. A lump of tissues will be formed due to a large number of these new cells are produced abnormally or when the old cells don't die and are called a tumor. When these cells in the cervix are shown to genital human papillomavirus (HPV), our immune scheme normally prevents the virus from serious injure but on occasion, the virus continues to live for years. In the long run, this virus can lead to the changeover of normal cells into cancerous cells on the surface of the cervix. Primarily the cervix region is classified as endocervix, ectocervix, and external os.

Earlier stage detection and recognition of cervix cancer are very important to save a life. The prediction of cervical cancer at an early stage is very rare. Cervical cancer mainly occurs in women and periodic checkup is required to eliminate the serious impact of cancer. The usage of AI models in the medical field has been increased in recent days[1].Nowadays, many new technologies and prediction system has been developed using the AI models to enhance the life of human. Mostly, cervix cancer can be concerned with unwanted habits and improper health maintenance. HPV testing is lacking in many countries mainly because of a lack of awareness about cancer [2]. Thus, the detection of tumors has to be ensured promptly with great precision. This detection can be attained possibly via taking up a series of tests periodically such as pap smear test, colposcopy, skiller, cytology, and biopsy. It spends a significant period of time when a series of tests done manually. In this fast-moving world, people hesitate to spend so much time doing such tests. So, the usage of AI model assist the people to test their body condition without having to discuss with a third person. The information processed from users is determined based on the major causes of cervical cancer. The second level of testing is to provide additional information stating the region affected and tests to be taken.

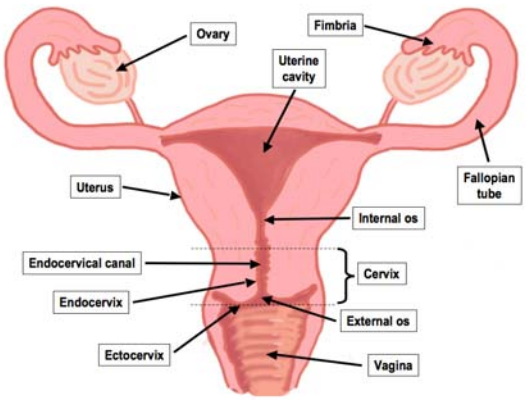

Fig. 1.The Structure of Cervix

The prediction is made using about 4000 colposcopy images classified into three types based on the region affected. Still, numerous research findings are being made to locate cervical cancer quickly with greater accuracy and reduced manual effort for computation.

In the human female reproductive system, the lower part of the uterus is called as Cervix. The structure of the Cervix is shown in Figure1.

The cervix's job is to allow menstrual blood to flow from the uterus into the vagina during sexual activity.,the cervix directs the sperms into the uterus [3]. This section reviews only the most important parts of the cervix such as the Endocervix, Ectocervix, External os. As shown in the Figure 2, The region of the cervix visible from the inside of the vagina is called the ectocervix. Typically, consistent, non-keratinizing squamous epithelium lines the whole length of the vagina and the majority of the ectocervix. Because the adult squamous epithelium contains glycogen, it can absorb the iodine solution and fails the Schiller test. The epithelium becomes Schiller test-positive when it does not absorb the iodine solution. When not pregnant, the smooth, somewhat pink cervical squamous epithelium is visible. It gets more vascular and develops a bluish tint throughout the pregnancy condition.

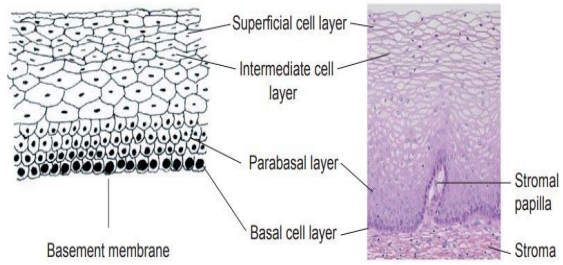

Fig. 2. Squamous Epithelium of Ectocervix

An external os is the center of the ectocervix that opens to allow a route between the uterus and vagina. The endocervical canal act as a tunnel through the cervix from the uterus into the external os. The endocervical canal is lined with columnar epithelium. The Columnar Epithelium of endocervix is shown in Figure 3. The squamous epithelium of the cervix is taller than the canal, which consists of a single layer of cells. It has a reddish appearance, and stromal vascularity can penetrate its thin single layer. The squamous epithelium of the cervix is taller than the canal, which consists of a single layer of cells. It has a reddish appearance, and stromal vascularity can penetrate its thin single layer It meets the endometrial epithelium at its upper limit and the squamous epithelium at the Squamo-Columnar Junction (SCJ) at its lower limit.

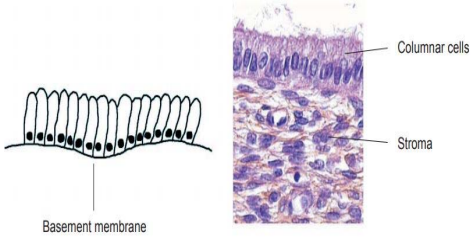

Fig. 3. Columnar Epithelium of Endocervix

The intersecting edge between the endocervix and ectocervix is called the transformation zone. Depending on age, prior infections, and exposure to female hormones, the transformation zones may be either wide or narrow. The Squamo-columnar junction (SCJ) and transformation zones are fundamental landmarks of the transformation process. Human Papilloma virus (HPV) affecting transformation zones may lead to cervical cancer[3].

Literature survey

The primary risk factor for cervical cancer is smoking: women who smoke have a twice-higher risk of developing the disease than non-smokers. Women who smoke have been found to have tobacco byproducts in their saliva or cervical mucus. Medical researchers ensure that these ingredients damage the complete or partial Smoking damages the DNA of cervix cells and may increase the risk of cervical cancer. It also weakens the immune system's ability to combat viral infections. [4]. The other major reasons are starting to have sexual intercourse before the age of 16 soon after the start of their periods, multiple sexual partners, Intake of birth control pills especially for longer than years. Suffering a diluted immune system and women who are having a sexually transmitted infection can also be a major reason for being affected by cervical cancer. Women whose intake doesn't contain adequate vegetables and fruits may be at a higher level of risk for cervical cancer. Cervical cancer might also occur through inheritance that is if any of their parents had cervical cancer then there is a chance of having cancer.

The symptoms of cervical cancer

The primary signs and symptoms of cervical cancer include vaginal bleeding with string odor, vaginal discharge tinged with blood, douching or a pelvic examination, blood spots or light bleeding during or after periods, menstrual bleeding that is heavier and lasts longer than usual, pain or discomfort during and after sexual activity, bleeding after menopause, and persistent pelvic and back pain. Any one of these symptoms requires medical attention right away and should be evaluated by a physician.

Early detection of cervical cancer

Cervical Cancer is a critical cancer disease that happening from the cervix. This cancer arises because of the abnormal growth of cells on the cervix and spreads to all other parts of the body. It is critical most of the time (90%). All women, especially those aged 30-50 years, should take a gynecological test to determine a precancerous lesion exists in the cervix part. Currently, three early detection tests are available. They are HPV test, Visual Inspection with Acetic acid (VIA), and PAP test. The HPV test detects infections caused by HP Virus and identifies precancerous cuts in the cervix, which can enlarge into cervical cancer if the infections left untreated. This test concludes only the presence of the HPV virus in the cervix. This test has been conducted on a collected sample of cells from the cervix. These collected samples of cells are sent to a medical laboratory for analysis. In some countries, the HPV test is done or administered by the woman herself and the sample is sent to the laboratory for testing. However, these infections clear on their own within two-three years. Cervical changes that lead to cancer take years often 9-10 years or more.

The more widely accessible early detection test for cervical cancer is the Pap test [4]. A medical professional will perform a gynecological examination and obtain a cervix sample for testing. For analysis, these samples are sent to a medical laboratory. This test can detect precancerous or cancerous lesions as well as changes in the cervical cells. Detecting cervical cancer at an earlier stage with a Pap smear provides a greater possibility for a complete cure. This Pap test can also find deviations in cervical cells that hint cervical cancer may happen in the future. The Pap test results can be normal or abnormal.

Normal means there is no abnormal growth of cells is found in the cervix. If the result is stated abnormal then there is the presence of abnormal cells and it might be any one of the levels such as Atypia, moderate, mild, or severe dysplasia.VIA is an active and low-cost screening test. VIA is a test to detect precancerous cervical lesions. The test requires gynecological testing by a medical worker, who applies acetic acid to the cervix part that observes whether there are any infections or changes happen in the cells. These test results will be provided instantly. People below age 21 need not require a PAP test. People between the ages of 21-29 need to do a PAP test once in 3 years. People between ages 30-65 need to take both PAP and HPV tests once every 5-6 years. 65 and above do not require a PAP smear test [5].

Many researchers are involved in usage of AI models to classify cervix cells into either positive or negative cervical cancers. In this section, the detailed study of existing AI model used in predicting the cervical cancer cell is discussed. Researchers have been using AI model built from different Machine Learning and Deep Learning Algorithms for predicting the cervical cancer among women community.

Fernandes et a [6] have introduced a computational system for cervical cancer prediction. Authors have come up with the combined approach of dimensionality reduction and deep learning based classifiers such as SVM, KNN, and Decision tree.

Vidyal et al.[7] proposed hybrid AI model comprising unsupervised algorithm (K-Means), supervised classification algorithm such as Random Forest Tree and CART. Authors have inferred that prediction accuracy of the model is higher when Random Forest Tree is used in combination with Random Forest Tree.

Singh and Sharam [8] have used the UCI repository of cervical cancer dataset as input to AI model. Authors have used 6 different classifier as part of the AI model for the prediction. Authors have pre-processed the data of repository and extracted features which are considered to be more relevant input to the classifier. Authors have identified that the prediction accuracy of Decision Tree algorithm is better than other classifiers.

Manika et al. [9] have applied Extreme Gradient Boosting (XGBoost) algorithm to predict cervical cancer cell in cervix. Authors have handled the missing data using pre-processing technique. Further, authors have used regularization technique to avoid overfitting of the model.

Riham Alsmariy et al.[10], have used AI model comprising of 3 different classifiers Decision Tree, Random Forest, and Logistic regression. Authors have used the reference dataset of UCI and applied well known sampling technique called SMOTE to avoid imbalance in the dataset. Futher, authors have applied PCA for dimensionality reduction in the dataset. Authors have infered from results that their model is achieving accuracy of 91% in predicting the cervical cancer.

Devi et al. [11] have proposed ANN based architecture for classifying the cervical cells to either normal or abnormal cells. Authors have claimed that ANN based prediction is better than manual screening methods such as LCB and Pap semar.

Jia et al [12] have applied YOLO algorithm to accurately detect infected cervical cells. Authors have used s3pool algorithm for feature extraction and acheiving better generalization of model. A default k-means algorithm in the YOLO3 is replaced with k-means++ to achieve better generalization of the model. Authors have carried our post processing of the model with NMS algorithm to achieve better detection accuracy of the cells in uncertain sitaution.

Geetha et al. [13] have applied CNN model with connected layer of 24 X 1 over the Pap Smear images to detect cancerous cells in cervex. The model achieves the classfication accuracy of 90.2%.

Merlind and Sathiaseelan [14] have carried out research study on analyzing the classification accuracy of algorithms KNN,SVM,MLP, Decision Tree and Navie Bayes, which is applied over UCI cervical cancer data repository towards predicting the cancer cells in cervix. Machine learning model are implemented using sciklit-learn python package.

Athinarayanan et al., [15] used the SVM algorithm to find the cancer cells from the input images of the patient. Further, authors have compared the classification accuracy of SVM based model with KNN and ANN.

Ghoneim et al., [16] have used hybridization of Convolutional Neural Networks (CNN) and Extreme learning machine (ELM) algorithm for the detection of cervical cancer cells from the screening method results. Authors have used the CNN for feature extraction from the input image and ELM model is built for classification of cervical cell to either cancerous or non-cancerous cells. The model has achieved the classification accuracy of 91.2% on this problem.

Taha et al.[17] have used hybridized AI model comprising of CNN and SVM. Authors have pre-trained CNN model to extract essential features from data set. All the extracted features from CNN model are fed to the SVM for the classification. SVM is trained to classify the cervical cell to either cancerous or non-cancerous based on the feature match.

From the survey, it is inferred that many researchers have opted for hybridized AI model to perform feature extraction and classification task over the image dataset. Most of the existing methods taken a small subset for the prediction of cells and therefore, final conclusion cannot be reached to decide on highest system's accuracy achieved by the hybridized deep learning model.

Proposed methodology

Most cases of cervical cancer result in death and are only diagnosed in the advanced stages of the illness. So, the development of a system that can predict this type of cancer is required. Early detection of cervical cancer can save women from death.

Phase I Prediction

To perform the preliminary tests prediction, the data set and algorithm selection are the key factors. The features of data used for prediction are mentioned below in below Table 1.

Table 1. List of Features and its data types

Features	Datatype
Age	Float
Number of Sexual Partners	Float
Number of pregnancies	Float
First sexual intercourse	Float
Smokes	Object
Smokes(Years)	Float
Smokes(packs/year)	Float
Hormonal Contraceptives	Object
IUD (Years)	Float
IUD	Object
Hormonal Contraceptives(Years)	Float
STDs	Object
STDs(number)	Float
STDs: Vulvo-perineal Condylomatosis	Object
STDs: Syphilis	Object
STDs: Vaginal	Object

Condylomatosis	
STDs: Condylomatosis	Object
STDs: Cervical Condylomatosis	Object
STDs: Pelvic Inflammatory Disease	Object
STDs: Genital Herpes	Object
STDs: Molluscum contagiosum	Object
STDs: Time since first diagnosis	Object
STDs: HIV	Object
STDs: Hepatitis-B	Object
STDs: AIDS	Object
STDs: Number of diagnosis	Float
STDs: Time since first diagnosis	Float
STDs: HPV	Float
Dx: CIN	Object
Dx: HPV	Object
Dx: Cancer	Object
Biopsy	Object
Schiller	Object
Cytology	Object
Colposcopy	Object

Preprocessing is the first stage of any machine learning project. Preprocessing is the removal of unwanted data or processing of raw data to a desirable format. The percentage of missing values in the UCI repository dataset is calculated.The column with a high percentage of missing value is eliminated using this graph. The categorical and numerical data are split in order to add the missing values. Normally the missing values are replaced with the mean, median, and mode of that particular column. Since this is medical data, the proposed work used a separate decision tree algorithm to fill the missing values in the dataet.The Exploratory data analysis is carried out to identify the significant columns. The Decision Tree Classification algorithm is implemented over the data set.

Decision Tree Classifier:

In a decision tree, each internal node represents a "test" on an attribute, including chance event outcomes; each branch indicates the test's result; and each leaf node indicates a class label. The structure resembles a flowchart. Classification rules are represented by the paths from the root to the leaf node. The data set is converted to a desirable format that is from CSV format to data frames using the read_csv function in pandas. The input and output data are split using the slicing function drop in pandas. 23 columns are chosen as input features after the elimination of less significant features. The 4-preliminary test for cervical cancer prediction is considered as output columns they are Colposcopy, Biopsy, Schiller, and cytology.Colposcopy is a practice to closely observe the vulva, vagina and cervix for the symptoms of disease. Initial test for identifying the uterine cervix cancer is

- Schiller's test in which the cervix part is dyed with an aqueous solution of iodine and potassium iodide. If the tissues show as brown means healthy tissue and tissue appears as white or yellow means cancerous issues.
- Biopsy is a sample of tissue taken from the particular portion of a body to examine more closely.
- Cytology is the microscopic examination of cell samples. Cancer disease can be diagnosed by looking at single cells and small clusters of cells.

Each of these tests is considered an output feature and a separate decision tree algorithm is executed over the data set. This set is split in the ratio of 70:30 corresponding to train: test. The train data set is used to construct the decision tree model which is a tree-like structure in which each leaf is a decision based on which results are predicted. The test data are those which is used to test the accuracy of the model. For successful construction of the model, the parameters such as ratios criterion, max_depth, max_features, random_state, and splitter are properly tuned. Each of these parameters is evaluated using the GridSearchCV function under sklearn package to identify the best-suited values for the parameters. Each of the models is executed separately and the results have collaborated.

Phase II Prediction

The results of Phase I prediction tell us the mandatory tests that have to be taken immediately. Among the four tests except for Colposcopy, all are laboratory level tests whereas Colposcopy is a test where microscopic images of the cervix are used for prediction of transformation zone in the cervix. So, in the second phase of prediction, the Colposcopy image of the cervix is used and the transformation zones are identified.

The Colposcopy image is classified into three types based on the region where the cervix is affected. The three types are Endocervix, Ectocervix, External os. Three types are shown in Figure 5, 6 and 7.About 4000 Colposcopy images are used for prediction.

Convolution Neural Network

The secondary level prediction is done using Convolution Neural Network (CNN). The neurons that comprise CNNs have bias functions and learnable weights. Each neuron performs bitwise dot product operation after receiving some inputs from the raw image pixels (data set). This network extracts a single score function from the input of raw image pixels. These networks take advantage of the fact that the architecture is more functional way because it accepts all types of images.

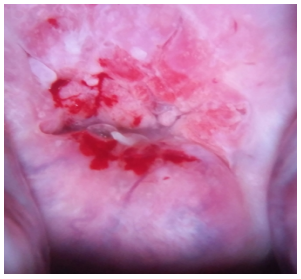

Fig. 5: Ectocervical

Fig. 6. Endocervical

Fig/ 7 Endocervical Region: invisible

In particular, unlike a regular Neural Network, the CNN neuron layers are organized in three dimensions. Three main layers are used to construct the convolution neural network architecture. They are the Convolution layer, Pooling, and Fully Connected layer. These three layers mine the important characteristics in the image which supports the algorithm in predicting the output. The preprocessing of the images is a mandatory step for the efficient execution of the algorithm.

The Image preprocessing is done using the Image data generator which includes functionalities like resale, shear_range, horizontal_flip, and zoom_scale. Then a sequential model is initialized in order to add the hidden layers and filter the images. This model uses two convolution and pooling layers in order to extract features. It then performs a flatten operation, in which the extracted models are converted to vectors, concatenated into a single long vector, and sent to a fully connected layer for feature interpretation. There are about 80 hidden layers are added and three output layers with an activation function for three class estimation. Model is compiled based on three parameters optimizer, loss, and metrics. The Adam optimizer is chosen because it adjusts the learning rate throughout the training process. Adam optimizer equations are given below

(1) $V_t = \beta_1 * V_{t-1} - (1- \beta_1) * g_t$

(2) $S_t = \beta_2 * S_{t-1} - (1- \beta_2) * g_t^2$

(3) $Wt = - \acute{n} \frac{V_t}{Sqrt(S_t + €)} * g_t$

Where: ń : Initial learning rate; g_t : Gradient at time t along w^j; V_t : Exponential Average of gradients along W_j; S_t : Exponential Average of squares of gradients along W_j; β_1, β_2 - Hyper parameters

The most widely used kind of loss function is categorical cross-entropy. The accuracy score on the validation set that the model is trained on is determined by the metric accuracy. The model is trained using the fit() function. To achieve the highest accuracy, the model is run for roughly 50 epochs. The number of times the model will cycle over the data is known as the number of epochs

.**Algorithm**

Step 1 : Import the following packages such as NumPy,Pandas,sklearn,Matplotlib
Step 2: Import the dataset by using read_csv function
Step 3: Preprocessed the data by removing the features with maximum missing values.
Step 4: Make use of decision tree algorithm to fill the missing values in features.
Step 5: Transform the dataset into data frames.
Step 6: Split the dataset into train and test in the ratio 70:30
Step 7: Create an object model for the decision tree classifier and fit the dataset to the model.
Step 8: After predicting the model, algorithm got the accuracy of 94.0%. To increase the accuracy of the model, Grid Search CV is used to get the best parameters for this model.
"Criterion" : gini
"max_depth" : 4
"max_features" : auto
"Random state" : 123
"Splitter" : best
Step 9: Predicted the model that has best parameters and got the accuracy of 97%.
Step 10: if the primary prediction results stating "colposcopy is required", then the user can provide their colposcopy image of the cervix to the secondary level prediction.
Step 11: Based on the colposcopy image provided, the exact transformation region of the cervix will be predicted.
Step 12: if the tests other than colposcopy is "required", then consultation of doctor is required

Experimental setup and results

The reason of this the investigations are to select the best classification method for cervical cancer prediction. For predicting the required preliminary tests for cervical cancer, the proposed method has used four classification algorithms. The first algorithm that was tested with data is Random Forest.

Table 2. Evaluation Metrics of the Classification Algorithms

Model	**Accuracy**	**F1 Score**	**Recall**	**Precision**
Logistic Regression	0.734	0.427	0.734	0.734
Random Forest	0.936	0.539	0.939	0.936
KNN	0.710	0.428	0.710	0.710
Decision Tree	0.97	0.487	0.97	0.952

Random Forest

When training the data with Random Forest, the accuracy and recall score of the algorithm is 93.6%. A random forest is a supervised classification algorithm that applies average measures to improve the predictive accuracy after fitting several decision tree classifiers on different subsamples of the dataset.. The parameters used in this classifier are

n_estimators : The number of trees in the forest.
Criterion : The function to measure the quality of a split.
Random state : It controls the randomness of the bootstrapping of the samples used when building

Logistic Regression

When training the data with Logistic Regression, the accuracy of the algorithm is 73.4%. Logistic regression is the most popular Machine learning algorithm to predict the probability value of a target variable. The class of the target variable is dichotomous, which denotes there would be only binary classes, 1(yes) or 0 (No). The following equation used in logistic regression is

(4) $y = e^{(b0 + b1*x)} / (1 + e^{(b0 + b1*x)})$

y : Predicted Output.
b0 : Bias or Intercept term.
b1 : Coefficient for single input value (x).

K-Nearest Neighbor

When training the data with KNN, the accuracy of the algorithm is 71.0%. K nearest neighbors are the most popular ML algorithm which classifies new cases based on a similarity measure and also stores all available cases. The parameters used in this classifier are

N_neigbours : No of neigbours.
P : Power parameter for the Minkowski metric. Where Manhattan distance for P1, and Euclidean distance for p=2.

Decision Tree Classifier

When training the given data set using a Decision tree classifier, the accuracy of the algorithm is 97%. By using GridSearchCV, the best parameters have been found for the algorithm to train the data. Among algorithms, the Decision Tree Classifier gave the best accuracy score, recall score, and precision. The Decision Tree Classifier algorithm is found to be providing higher accuracy because of its structure developed during the model development. The selection of parameters made during the model construction mainly contributed to the improvement in the accuracy of the algorithm. The number of decision leaves was decided based on the input features. When the KNN algorithm worked based on the grouping of similar types, the decision tree works on decision leaves.

The generalized concept of KNN made it less suitable for the dataset. Though the random forest algorithm is the cluster of decision trees, the higher intercorrelation between input features made it unsuitable for the dataset. The Logistic regression is a generalized algorithm and the simple structure and functioning of the algorithm did not support the dataset. Because of these factors, a decision tree is found to be more appropriate for the dataset. AI models are become popular in the prediction process of medical sciences [18][19]..

Performance metrics

The reason for investigations are to select the best classification method for cervical cancer prediction. To measure the quality of the machine learning algorithm, Evaluation metrics are used. Results shown in Table 1

Accuracy

The key metrics used to evaluate any classification algorithm are accuracy, recall, and precision. Accuracy is defined as the ratio of accurately predicted observations to the total observations for the test data. The accuracy metric primarily depends on the correlation between input and output features.

TP -True Positive is correctly classified the cell as cancer cell.
FN- False Negative is incorrectly classified the cell as no cancer.
FP- False Positive is incorrectly classified the cell as cancer.
TN-True Negative is correctly classified the cell as no cancer.

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

The Decision Tree Classifier algorithm was found to be providing higher accuracy because of its structure developed during the model development. The selection of parameters made during the model construction mainly contributed to the improvement in the accuracy of the algorithm. Accuracy comparison is mentioned in the Figure 8.

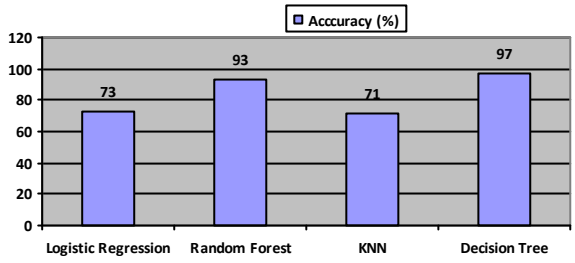

Fig. 8 Accuracy

Recall

A recall is defined as the number of accurately predicted positive observations divided by all predicted observations in actual class. It is an important metric to determine the best algorithm for the dataset.

(6) Recall = (TP) / (TP + FN)

Recall comparison is mentioned in the Figure 9.

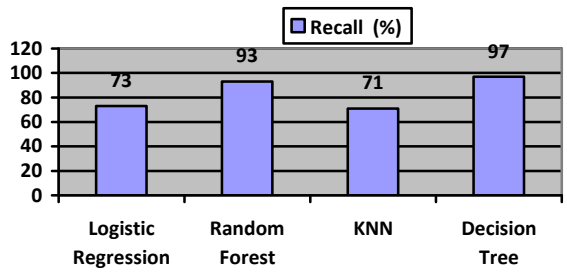

Fig. 9 Recall

Precision: It is the important metric in the prediction of cervical cancer and it is defined as the ratio of correctly predicted positive observations or performances to the total predicted positive observations. Precision comparison is mentioned in the Figure 10.

(7) Precision = TP / TP + FP

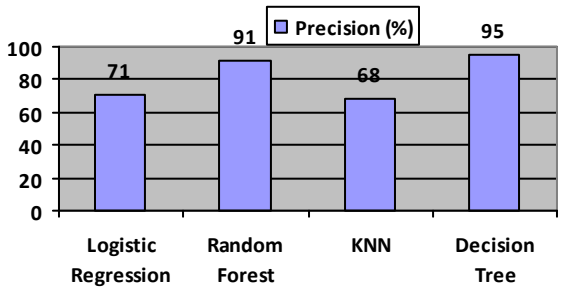

Fig. 10. Precision

F1 Score

It is a weighted average of Precision and Recall. F1 score comparison is mentioned in the Figure 11.

F1Score=

(8) (2 X Precision X Recall) / (Precision + Recall)

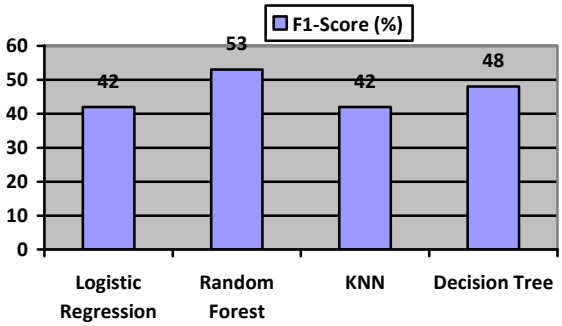

Fig. 11. F1 Score

Convolution Neural Network

The Accuracy in the prediction of transformation zone using CNN is 97.07%. The accuracy of the algorithm is found to be gradually increasing when increasing the number of epochs. The accuracy of the CNN algorithm is not only based on epochs but also relies on a number of hidden layers managed in the CNN. The hidden layer in the neural network mainly concentrates on the important features of the image. This helps in improving the accuracy of the algorithm. There are about 80 hidden layers being used to extract the important features of the image. So, each time an epoch is executed, all the 4000 images are made to pass through the 80 hidden layers in the algorithm for feature extraction. Hence, when a test image is being provided, the predictions are made based on the trained model. Figure 12 show the accuracy of CNN model predicting cancerous cervical cell over the change of epochs value.

Conclusion

One of the main reasons for death to women is cervical cancer. Thus, the objective of the proposed method is to develop a system that would help women to have a periodic check-up and get updates about their health-related issues.

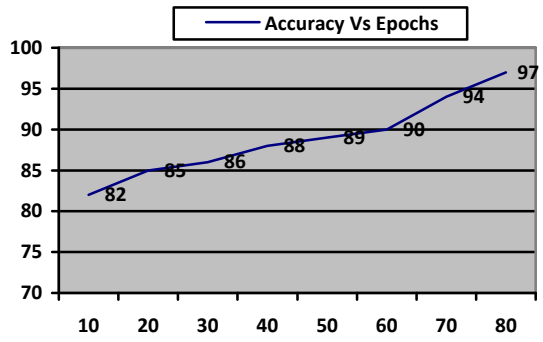

Fig 12 Epochs Vs Accuracy

This application will be of great help to people and also will act as a guide for what are the steps that have to be taken regarding their body condition. Since the application is developed by taking common people into consideration the details fetched from users are at a basic level and it will be easy to access. The algorithms are processed in such a way that it is above 80% accuracy and trustworthy for people to use it. Given the increased accuracy of the cervical cancer cell classification, the proposed hybridization of the CNN system and decision tree classifier aids the doctor in making decisions regarding the patient's continued care more quickly.

REFERENCES

1. J. M. Yamal, M. Guillaud, E. N. Atkinson, M. Follen, C. MacAulay, S. B. Cantor,et al., "Prediction using hierarchical data: Applications for automated detection of cervical cancer," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 8, pp. 65-74, 2015.
2. S. Subramanian, R. Sankaranarayanan, P. O. Esmy, J. V. Thulaseedharan, R. Swaminathan, and S. Thomas, "Clinical trial to implementation: Cost and effectiveness considerations for scaling up cervical cancer screening in low- and middle-income countries," *Journal of Cancer Policy*, vol. 7, pp. 4-11, 2016.
3. K. J. Sales, "Human papillomavirus and cervical cancer," in *Cancer and Inflammation Mechanisms: Chemical, Biological, and Clinical Aspects, ed: John Wiley & Sons,* 2014, pp. 165-180.
4. H. Ramaraju, Y. Nagaveni, and A. Khazi, "Use of Schiller's test versus Pap smear to increase detection rate of cervical dysplasias," *International Journal of Reproduction, Contraception, Obstetrics and Gynecology*, vol. 5, pp. 1446-1450, 2017.
5. Schiffman M, Wentzensen N.,"A suggested approach to simplify and improve cervical screening in the United States", *Journal of lower genital tract disease* ;20(1):PP:1-7,2016.
6. K. Fernandes, D. Chicco, J. S. Cardoso, and J. Fernandes, "Supervised deep learning embeddings for the prediction of cervical cancer diagnosis," *Peer J Computer Science*, vol. 4, p. e154, 2018.
7. R. Vidyal and G. M. Nasira2, "Prediction of Cervical Cancer using Hybrid Induction Technique: A Solution for Human Hereditary Disease Patterns", Indian Journal of Science and Technology, August 2016.
8. Singh J., Sharma S. Prediction of Cervical Cancer Using Machine Learning Techniques. *Int. J. Appl. Eng. Res.* 2019;**14**:2570–2577.
9. Manika J., Richa G and Rajiv S,"Cervical Cancer Risk Prediction using XGBoost Classifer", In the proceedings of 7th International Conference on Signal Processing and Communication,10.1109/ICSC53193.2021.9673474, 2021
10. Riham Alsmariy, Graham Healy, Hoda Abdelhafez," Predicting Cervical Cancer using Machine Learning Methods", International Journal of Advanced Computer Science and Applications, Vol. 11, No. 7, pp.173-184, 2020.
11. Devi, M. A., Ravi, S., Vaishnavi, J., & Punitha, S,"Classification of Cervical Cancer Using Artificial Neural Networks.", *Procedia Computer Science*, 89, 465–472,2016.
12. Jia, D., He, Z., Zhang, C. *et al.* Detection of cervical cancer cells in complex situation based on improved YOLOv3 network. *Multimed Tools Appl* **81**, 8939–8961 (2022). https://doi.org/10.1007/s11042-022-11954-9.
13. Geetha,K, Aarthi,S, Sasikaladevi,N and Mala,C" An Automated Cervical Cancer Detection Mechanism Using Pap Smear Images",In the proceedings of4th EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing,2023
14. Merlin, D. & Sathiaseelan, Dr. (2021). Improved Classification Accuracy for Identification of Cervical Cancer. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. 245-258. 10.32628/CSEIT217633
15. Athinarayanan S., & Midthe., S, " Classification of Cervical Cancer Cells in Pap Smear Screening Test", *ICTACT Journal on Image and Video Processing*,Vol06,No.04,PP:1234–1238,2016.
16. Ghonemi, A., Muhamand, G., & Hossain, M. S,"Cervical cancer classification using convolutional neural networks and extreme learning machines", *Future Generation Computer Systems,*Volume 102,PP: 643-649,2020.
17. Taha . B. Dias. J. and Werghi N,"Classification of Cervical Cancer Using Pap-Smear Images: *A Convolutional Neural Network Approach. Department of Electrical and Computer Engineering*, 1(d), 698–706,2017.
18. Amanuel Kahsay , Paweł Regulski , Piotr Derugo "AI-based control techniques for maximum power point tracking of photovoltaic systems using a boost converter", Przegląd elektrotechniczny,pp.1-6,Vol.11,2023
19. Saravana Ram, Akilandeswari J, Vinoth Kumar M "HybDeepNet:A Hybrid Deep Learning Model for detecting Cardiac Arrhythmia from ECG Signals", *Information Technology and Control,* Vol.5,No.2,pp. 433-444, 2023