TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 01/01 pp1–22
DOI: 10.26599/TST.2025.9010150
Volume x, Number x, xxxx xxxx

A New Enhanced Predictive Model for Cervical Cancer Classification based on CatBoost and CNN Algorithms

Ra'ed M. Al-Khatib*, Ahmad T. Al-Taani, Qasem Al-Radaideh, Sally Haddad, and Heba A. Maabreh

Abstract: Cervical cancer is a disease affecting women worldwide, including in impoverished countries, and it poses a significant threat to their reproductive health. The main drawback of this cancer is that it shows no early signs. Additionally, social factors and individual errors during manual testing hinder screening for this disease. Therefore, developing a model using machine learning and deep learning is a significant and trustworthy option that offers greater support than traditional detection methods. In this research, we used two powerful algorithms (CatBoost and CNN) to tackle the problem of cervical cancer classification. Two datasets are used: the "*Risk Factors Cervical Cancer*" and "*Malhari*", where the first dataset is textual data. After applying the CatBoost algorithm, we reached an accuracy of 100% for "*Kag Risk Factors Cervical Cancer*" data. The second dataset has two types of images ("*Pap Smear*" and "*Colposcopy*"), where we reached an accuracy of 87.2% for the "*Pap Smear*" type, and 91.1% for the "*Colposcopy*" type by applying the newly developed CNN algorithm.

Key words: Artificial Intellegince (AI); Convolutional Neural Network (CNN); Cervical Cancer; Pap Smear; Colposcopy, CatBoost algorithm; Image Classification; Synthetic Minority Over-Sampling Technique (SMOTE)

1 Introduction

Cervical cancer remains and continues to pose a major global public health issue, significantly impacting women's health and well-being. Despite the progress made in screening and preventative methods, it is linked to the causes that lead to death among women around the world. Gaining a comprehensive understanding of the complex factors that contribute to the risk of cervical cancer is crucial for effective prevention, detection, and intervention strategies.

According to the World Health Organization (WHO), malignancies in women, such as breast cancer, ovarian cancer, and cervical cancer, are leading causes of premature death among women worldwide[1, 2]. Cervical cancer, or cancer of the cervix, is a malignant tumor that arises from the uncontrolled growth of cells in the lowest part of the uterus. It is known as the cervix, without adhering to the normal process of cell division[3]. The WHO has proved through statistics that over 270,000 women were victims of cervical cancer, where the death rate in poor countries was the highest, with a percentage equal to 85%, as the numbers worsened until 444,500 new cases of cancer were recorded every year[4, 5].

Numerous publications have clarified that cells in our

• Ra'ed M. Al-Khatib, Ahmad T. Al-Taani, Sally Haddad, and Heba A. Maabreh are with the Department of Computer Sciences, Faculty of Information Technology and Computer Sciences, Yarmouk University, Irbid, 21163, Jordan. E-mail: raed.m.alkhatib@yu.edu.jo; ahmadta@yu.edu.jo; sallyhaddad56@yahoo.com; heba.maabreh198@gmail.com.
• Qasem Al-Radaideh is with the Department of Information Systems, Faculty of Information Technology and Computer Sciences, Yarmouk University, Irbid, 21163, Jordan. E-mail: qasemr@yu.edu.jo.
* To whom correspondence should be addressed.
Manuscript received: 2024-07-10; revised: 2025-03-04; accepted: 2025-09-22

body continuously grow and divide throughout our lives, supported by various genes. Sometimes, this process is hindered by some variables such as drug intake, smoking, and certain behaviors. When this happens, the aging cells in the human body fail to undergo apoptosis at the appropriate time. The timing of the new cells is inconvenient. These cells can aggregate and develop into a neoplastic mass[6]. Subsequently, cancer cells disseminate to other tissues. Cervical cancer is a prevalent type of cancer that specifically impacts women[7].

Cancer has been recognized as the primary cause of death worldwide, resulting in 8.8 million fatalities in 2015. Cancer develops through the conversion of healthy cells into tumor cells, typically occurring in multiple stages that usually begin with a precancerous lesion and ultimately result in the formation of a malignant tumor. Early cancer diagnosis significantly increases the chances of responding effectively to treatment. It enhances the likelihood of survival, lowers the cost of treatment, minimizes potential complications associated with early intervention, and substantially improves patients' quality of life. Early detection and prevention of cancer are vital for improving health outcomes. Postponing medical treatment[8, 9], numerous screening and diagnostic procedures can be challenging, similar to other diseases.

A computer-aided diagnostic (CAD) system perceives a complicated ecology. In developing countries, there is a significant shortage of resources. Furthermore, there are instances when patients exhibit indifference towards regular examinations. Therefore, the main challenge in the diagnostic process is identifying the most suitable test, which involves developing a plan and accurately assessing the specific risks associated with each patient. The majority of the screening procedures produced exceptionally high results, which are related to the doctor's expertise and subjective judgment, where the user's text is issued[10, 11].

Statistics reveal that cervical cancer poses a significant risk to many women in today's society. In 2014, the prevalence of this disease among women in the United States reached a total of 12,578 cases, out of which 4,115 resulted in fatalities[12]. This corresponds to a fatality rate of approximately 32%. Cancer data provides information on the occurrence and characteristics of cancer cases. Cervical cancer datasets play a significant role in advanced data mining techniques[13].

Statistical data show that cervical cancer directly affects women's reproductive systems. The understanding of cervical cancer was assessed through cross-sectional research. Cervical cancer occurs when normal cells in the cervix become malignant cells. The human papillomavirus (HPV) is a virus that impacts the reproductive system of sexually active individuals, causing inflammation of the cervix in both males and females[14]. The immune system of an individual who is exposed to HPV is generally able to fend off the virus effectively[15]. Although HPV may infect women, it only persists for a prolonged period in a small percentage of cases. In certain instances, it may cause the outer layer of the cervix to develop cancerous cells[16]. So, we recommend that every woman aged 30 to 49 undergo a screening test at least once during her lifetime. The use of screening tests and early detection methods significantly contributes to lowering the death rate among individuals diagnosed with cancer. Cervical cancer is the second deadliest disease, surpassed only by breast cancer. In 2018, the number who died from cervical cancer was around 311,000 women[14, 17].

In the rest of this paper, the succeeding sections are organized as follows. Section 2 presents the main state-of-the-art methods published recently in the literature related to the works that solved the cervical cancer problem. Section 3 provides details of the proposed method for cervical cancer classification. Section 4 discusses and explains the main results using evaluation metrics. Section 5 has the conclusion of this research work with some future directions.

2 Related works

Cervical cancer is one of the four most common cancers, and the WHO considers it a leading cause of death for women worldwide, especially in low-income countries. Because machine learning and artificial intelligence are trustworthy and very effective solutions for detecting diseases and cancers, they have been adapted to be relied upon in classifying cervical cancer.

2.1 Studies of Cervical Cancer Risk Factor Dataset

The authors in[5] developed various algorithms to develop a predictive model for individuals with the disease. The primary algorithm employed is a Decision Tree (DT) model, which evaluates the likelihood of

cancer development. In addition, the researchers combined the Recursive feature elimination (RFE) with Least Absolute Shrinkage and Selection Operator (LASSO) algorithm to identify specific features that play a role in disease prediction[18]. Based on the Risk Factors dataset, the results showed that using the decision tree approach in conjunction with RFE and SMOTETomek achieved an impressive accuracy rate of 98.72%, and a sensitivity of 100%. Furthermore, this research proved that the use of the decision tree is effective in achieving more accurate classification[5, 19].

In[16], the researchers used a succession of machine learning (ML) algorithms to extract and categorize cervical cancer-related statistics. The ML-based algorithms that were utilized for classification include: Naïve Bayes (NB), Random Forest (RF), Decision Tree (DT), multi-layer perceptron (MLP), *k*-Nearest Neighbor (KNN), Sequential Minimal Optimization (SMO), Multilayer Perceptron (MLP) Neural Network (NN), and Simple Logistic Regression (SLR). These machine learning algorithms were developed to determine if the data indicated a diagnosis or a state of health. The dataset used in this study included risk factors for cervical cancer, with a primary focus on evaluating the effectiveness of using ML-based algorithms in the analysis process. The researchers developed a confusion matrix with an accuracy of 96.40%, where the RF approach obtained better results in the classification test than other ML algorithms.

In[7], the authors offered four data mining algorithms for identifying cervical cancer: Artificial Neural Networks (ANN), Random Forests (RF), logistic regression or Logistic Model Trees (LMT), and Decision Trees (DT). These techniques were applied to develop association rules that identify factors linked to cancer. The cervical cancer dataset was separated into four classes: Hinselmann, Schiller, Cytology, and Biopsy. Based on the Biopsy dataset, the DT, RF, and LMT models achieved over 99.6% in accuracy. The RF algorithm attained the highest Recall with an F1_score of 99.8%, while the DT classifier performed better than other models utilized on the Cytology dataset. In the Hinselmann dataset, both the RF and LMT classifiers surpassed the ANN when comparing the results in terms of all accuracy matrices.

In[13], the authors employed the Decision Tree and cost-sensitive Decision Tree, which are machine learning algorithms to categorize cervical cancer using the Risk Factors dataset. Their categorization was based on the accurate identification of positive cases. The suggested model yields more precise outcomes in both binary and multi-class categorizations. The True Positive (TP) rate for this model is 0.429, while a typical decision tree in a binary class job obtained a TP rate of 0.160.

In[20], the authors employed a two-stage process using machine learning algorithms, where they first extracted features and then classified them. They used a multi-layer perceptron (MLP) neural network in order to extract deep features[21]. They used VGG-19, ResNet-50, and ResNet-34 deep networks to feed the MLP model. The classification layers in the two networks have been removed from the CNN model, and the outputs go through a flattening layer before being fed into the MLP. Adam optimizer was also used to improve the performance of this model[22]. The Herlev benchmark database was utilized to implement the algorithm, resulting in accuracy rates of 99.23% for the two-class case, and 97.65% for the seven-class case.

In[23], the authors utilized a texture descriptor called Modified Uniform Local Ternary Patterns (MULTP) combined with an optimized multi-layer feed-forward neural network in a two-stage approach for the classification of "*Pap Smear*" images[24]. Following picture segmentation, the texture features were collected, and the neural network's architecture was optimized via the genetic algorithm[25]. Tested and evaluated on 917-image from the Herlev dataset, this technique surpassed state-of-the-art methods in accuracy, which demonstrated resilience to rotation and suitability for online applications due to its short runtime. Consequently, the optimization framework and MULTP descriptor can be applied to additional deep learning and computer vision challenges.

Table 1 summarizes the related works for the main research papers. These works used different types of Cervical Cancer Risk Factors (CCRF) datasets.

2.2 Malhari Cervical Cancer Dataset Studies

The researchers in[26] introduced the Malhari dataset, which was presented to support progress in the diagnosis and screening of cervical cancer. A large amount of varied input data is necessary to drive the development of automated diagnostic systems. Basically, the Malhari dataset aims to significantly advance automated classification algorithms for cervical cancer diagnosis by providing a

Table 1 Summary of related work for Cervical Cancer using the Risk Factors (CCRF) dataset.

Ref.	Year	Method	Dataset	Accuracy
[5]	2022	DT algorithm	Risk Factors dataset	98.72%
[16]	2020	NB, DT, KNN, SMO, RF, MLP with Neural Network, SLR	Cervical Cancer Risk Factor dataset	RF algorithm was the best algorithm with accuracy = 96.40%
[7]	2019	ANN, RF, LMT, and DT	Cervical Cancer dataset	DT, RF, and LMT achieve accuracy over 99%
[13]	2017	DT and Cost-Sensitive DT	Risk Factors dataset	Cost-Sensitive Decision Tree in TP rate (0.429), when compared with (0.160) for a typical decision tree
[20]	2023	pre-trained deep networks and a Multi-Layer Perceptron (MLP) neural network	Herlev benchmark database	99.23% accuracy for the two-class case, and 97.65% accuracy for the seven-class case
[23]	2022	Texture Extraction and Modified Uniform Local Ternary Patterns (MULTP) descriptor	Herlev database has 917 digital images of "*Pap Smear*" type	Optimized MLP got 98.63% on 5-folds, and 96.72 on 10-folds accuracy rate

comprehensive collection of cervical cancer images. In these Malhari datasets, a total of 866 photos were found, which are divided into three classes: 2 classes (453 images), 3 classes (132 images), and 4 classes (301 images)[26]. A Deep Convolutional Generative Adversarial Network (DCGAN) is a type of GAN that uses convolutional layers in both the discriminator and generator networks to generate images during training[27]. However, they did not find good results for this Malhari cervical cancer dataset.

From discussing literature works, there are significant gaps in the current research on cervical cancer categorization and detection, which can be successfully filled by our suggested approach, as follows:

(1) Combining Various Data Types: Prior research frequently focused on either unstructured image data (such as "*Pap Smear*" or "*Colposcopy*" pictures) or structured data (such as patient risk factors). This study will bridge the gap between these two data modalities by integrating the CNN for image-based data classification and the CatBoost algorithm for structured data classification. Our research in this integration will improve the model's capacity to offer a thorough diagnostic framework, which was absent from earlier related works studies.

(2) Better Management of Class Imbalance: Because positive cases are underestimated, asymmetry in cervical cancer datasets is a serious problem. Existing of our proposed research will use techniques such as SMOTE technique, but nothing is known about how well such functions perform with more sophisticated classifiers. The final results of our study achieved better classification performance than conventional methods by using CatBoost's built-in capacity to handle class imbalance and SMOTE technique to balance the data.

(3) Improved Model Efficiency: Our suggested CatBoost model is deployed efficiently, and the experimental results proved that it outperforms earlier models such as Random Forest and LMT models. Our proposed model, based on CatBoost, obtained 100% accuracy on the structured patient risk factors dataset. Similarly, the CNN model shows its resilience in managing image-based classification tasks by exhibiting excellent accuracy on image datasets (91.1% for "*Colposcopy*" and 87.2% for "*Pap Smear*").

(4) Relevance to Real Life: Our work will tackle practical cervical cancer screening issues, including low diagnostic accuracy, reliance on manual interpretation, and resource limits in developing locations, by merging machine learning with deep learning methodology. The suggested approach will be included in automated diagnostic instruments to improve early identification and reduce death rates.

Consequently, this work will be the first attempt to use CNN and CatBoost together for cervical cancer diagnosis, utilizing their distinct advantages for structured and unstructured data, respectively. The smooth fusion of these methodologies produces a unique hybrid solution that is accurate and flexible enough to accommodate a variety of data circumstances.

3 Proposed Method

The sequence of our research is designed to address many factors in cervical cancer and is divided into four stages: identification of the problem, data collection,

processing, and analysis of the results, as presented in Figure 1. This flowchart provides a structured research approach, ensuring that each step is planned and executed methodically.

3.1 Datasets

3.1.1 Risk Factors Cervical Cancer Dataset

In the proposed model, we used a dataset that contains cervical cancer data collected in Venezuela from the University Hospital of Caracas. It is available as the Risk Factors dataset at the University of California, Irvine (UCI) in the machine learning repository[28]. It contains 858 patient medical records with 36 characteristics, as represented in Table 2 that are showing patient habits and demographic data. In addition to their previous medical records, the number of cancer patients was 55 people, since there were 4 target variables in the data set, which are Hinselmann, Schiller, Cytology, and Biopsy. The rows of the columns "$Dx:Cancer$" and "$Dx:CIN$" are added together in a new column called "*new_target*", which is presented in Eq. (1). This is because CIN is classified as a precancer, indicating that the patient has a high chance of developing invasive cancer. We took this action because there were only 18 positive results in "$Dx:Cancer$" on its own, which is a rather small amount, and 9 positive results in "$Dx:CIN$". When these two columns were combined, 27 cases of cancer diagnoses were produced. Consequently, a more accurate assessment of the risk factors would result. As illustrated in Eq. (1), the "*new_target*" is ultimately designated as the target variable[28].

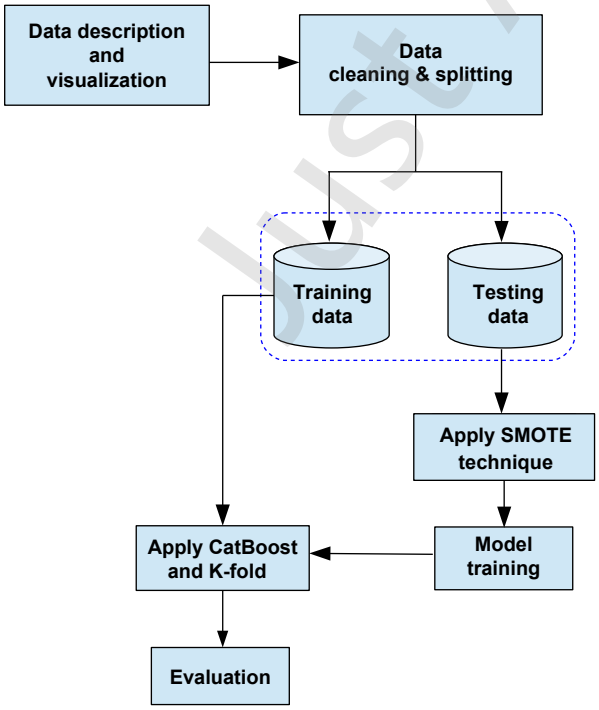

Fig. 1 General overall research design.

Table 2 The main characteristics and features of the cervical cancer dataset with the number of null (NaN) values.

No.	Characteristics (Features)	Data Types	Null (NaN)
(1)	Age	Int.	0
(2)	Number of sexual partners	Int.	26
(3)	First sexual intercourse (age)	Int.	7
(4)	Number of pregnancies	Int.	56
(5)	Smokes	Bool.	13
(6)	Smokes (years)	Int.	13
(7)	Smokes (packs/year)	Int	13
(8)	Hormonal contraceptives	Bool.	108
(9)	Hormonal contraceptives (years)	Int.	108
(10)	IUD	Bool.	117
(11)	IUD (years)	Int.	117
(12)	STDs	Bool.	105
(13)	STDs (number)	Int.	105
(14)	STDS: Condylomatosis	Bool.	105
(15)	STDS: Cervical condylomatosis	Bool.	105
(16)	STDS: Cervical condylomatosis	Bool.	105
(17)	STDS: Vulvo-perineal condylomatosis	Bool.	105
(18)	STDS: Syphilis	Bool.	105
(19)	STDS: Pelvic inflammatory disease	Bool.	105
(20)	STDS: Genital herpes	Bool.	105
(21)	STDS: Molluscum contagiosum	Bool.	105
(22)	STDS: AIDS	Bool.	105
(23)	STDS: HIV	Bool.	105
(24)	STDS: Hepatitis B	Bool.	105
(25)	STDS: HPV	Bool.	105
(26)	STDS: Number of diagnoses	Int.	0
(27)	STDS: Time since first diagnosis	Int.	0
(28)	STDS: Time since last diagnosis	Int.	0
(29)	Dx: Cancer	Bool.	0
(30)	Dx: CIN	Bool.	0
(31)	Dx: HPV	Bool.	0
(32)	Dx:	Bool.	0
(33)	Hinselmann	Bool.	0
(34)	Schiller	Bool.	0
(35)	Cytology	Bool.	0
(36)	Biopsy	Bool.	0

$$new\_target = Dx:Cancer \cup Dx:CIN \quad (1)$$

This conclusion allows the model to be updated to better reflect the associations in the data that indicate the presence of cancer. Furthermore, using this variable as an objective enables effective evaluation of the model's performance by comparing the accuracy of predictions to actual data values, which is critical in medical diagnostic applications. Table 2 represents the features of the used dataset with the number of null (NaN) values.

Two different sorted data types make up the features: Integer and Boolean. While the "Boolean" data type represents one of two potential values (often indicated by "yes" or "no"), the 'Integer' data type represents full Integers without any fractional components.

(A) Data Preprocessing The process of selecting data involves finding and obtaining relevant datasets that contain risk factors related to cervical cancer. These datasets may include clinical data, lifestyle factors, health history, and demographic information. The key is to choose data that is thorough and relevant to the study's goals, providing a solid foundation for analysis.

Data preparation is an essential stage that involves preparing raw data for analysis by cleaning, standardizing, and converting it into a suitable format[29]. To facilitate additional analysis, this can involve resolving missing values, eliminating outliers, normalizing scales, and encoding categorical variables.

The dataset consists of 36 attributes, including 4 attributes that are used as targets for classification; the remaining 32 attributes are, as we discovered in attributes No. (27 and 28), which represent: (STDs: time since first diagnosis and STDs: time since last diagnosis), respectively. The percentage of missing values reached 91% of the column. So, we neglected these two attributes, as they will not have any good effect on the classification process. As we have 30 attributes remaining for the classification that contain missing data, we used the "RandomForestRegressor" algorithm to fill in the missing values, and then we removed the duplicated values, resulting in 835 records. This method entails predicting missing values based on the existing attributes by utilizing the algorithm's capacity to identify patterns and correlations within the data[30]. Therefore, the pre-processing stage takes the following steps in the forecasting process:

(1) Identify columns or features with missing values: typically represented as null values "NaNs", or placeholders in the dataset.

(2) Data Splitting: Divide the dataset into complete data (with all features and target) and data with missing values (features but missing values).

(3) Training "RandomForestRegressor" algorithm: Train the model using complete data, treating missing values as the target variable and other features as predictors.

(4) Predicting Missing Values: Use the trained model to predict missing values in the dataset subset with missing values.

(5) Filling Missing Values: Replace missing values with predicted values from the "RandomForestRegressor" algorithm.

(6) Evaluation: Optionally compare predicted values with real values for evaluation or assess the impact on subsequent tasks, such as modeling.

We used two parts of the data set: the first part was for training, which contained 80% of the total number of records, where the number of training data used was 668 records, and the other part was used for testing, as it contained 20%, where the number of testing data used was 167 records. Figure 2 shows the breakdown of the dataset.

There are two classes of classification regarding the target characteristic. "Biopsy" is classified to (0 or 1), where the number 0 represents the negative result, which is not having cervical cancer, and the number 1 represents the positive result, which is having cancer. To solve class imbalance, data augmentation methods were employed for a cervical cancer dataset, utilizing the power of the SMOTE technique, which stands for Synthetic Minority Over-Sampling Technique. This is especially helpful for medical datasets, where a given

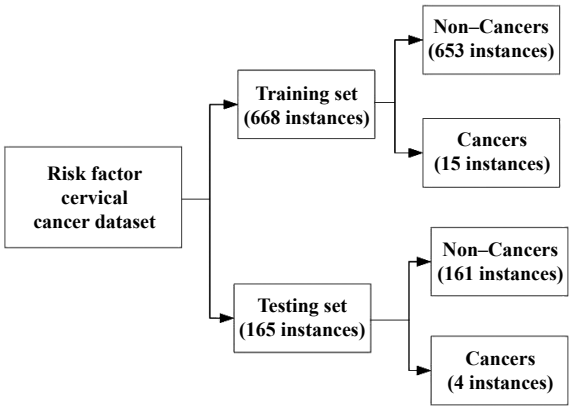

Fig. 2 Train and test split of risk factor cervical cancer Dataset.

illness or outcome may be far less common than another. Predictive models may perform better when a more balanced dataset is produced with the use of augmentation approaches. We noticed that the distribution of groups with and without cervical cancer was extremely unbalanced, as the non-infection rate was 98%, while the infection rate was only 2%, as shown in column two of Table 3. We noticed that much of the previous research did not pay attention to this problem.

Therefore, we adapted the SMOTE technique to balance the dataset due to a significant class imbalance. Algorithm 1 provides a description of how the SMOTE function works. Additionally, Table 3 illustrates the comparison of class distribution in the dataset both before and after applying the SMOTE technique.

(B) Algorithm 1. SMOTE technique

1) Pre-processing Phase: making the dataset ready for SMOTE and training for machine learning steps.

Identify Imbalance Step: Examine the target variable's class distribution to identify the minority and majority classes. The class with fewer samples is designated as the minority class. Handle missing values by imputing or removing missing data using different techniques to ensure dataset completeness.

Encode categorical variables by converting them into numerical representations using one-hot or label encoding. Feature scaling is done by normalizing or standardizing numerical features to ensure they are on a comparable scale, which is particularly crucial for distance-based class distribution techniques like the SMOTE technique.

2) SMOTE during training phase: See Figure 3 for the framework of the SMOTE technique. To balance the class distribution, create synthetic samples for the minority class by following the four steps:

(a) Find k-Nearest Neighbours: by determining each minority class sample employing a distance metric (such as the Euclidean distance) to determine the k-nearest neighbours within the same class, typically k=5.

(b) Random Selection: Choose one or more neighbours randomly from the listed k-nearest neighbours.

(c) Synthetic Sample Generation: Interpolate between the chosen neighbour and the original sample to produce synthetic data points. For feature f_i as in Eq. (2):

$$f = f_{\text{original}} + \text{rand}(0,1) \times (f_{\text{neighbor}} - f_{\text{original}}) \quad (2)$$

(d) Repeat: Continue until the intended distribution of classes is reached.

3) Model training phase: Use the augmented dataset, which contains synthetic examples, to train a machine learning model. If you have not already divide the data into training and testing sets. Train a model using the resampled training set (with data that has been SMOTE-applied). Use any appropriate approach to train the model (e.g., decision trees, support vector machines, neural networks). To prevent overfitting, validate the model's performance during training using methods such as cross-validation.

4) Testing phase: Assess the generalizability and performance of the trained model. Assess the model

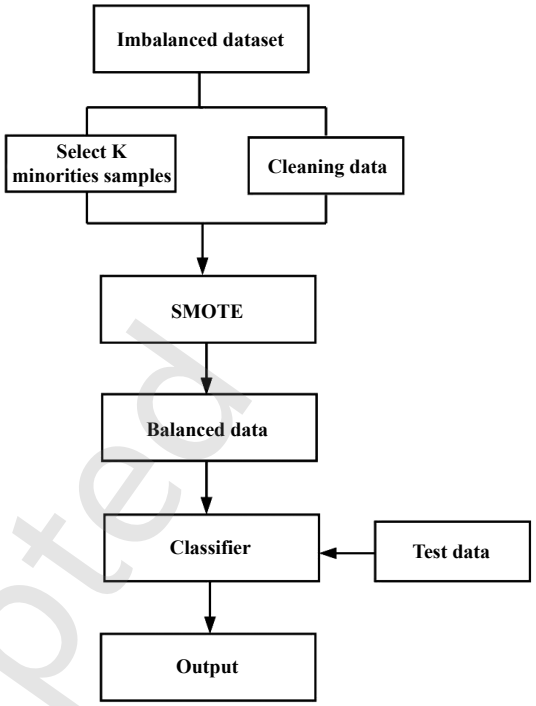

Fig. 3 Main steps of the SMOTE technique.

Table 3 Results before and after using the SMOTE technique to address the imbalanced data.

	Before using SMOTE technique (imbalance-data)	After using SMOTE technique (balance-data)
Negative Result (Class 0)	645	645
Positive Result (Class 1)	14	645

using the unaltered test set. The original class distribution ought to be preserved in the test set. Calculate assessment measures like ROC-AUC, F1_score, Recall, Accuracy, and Precision. Examine the model's performance on unknown data, paying special attention to the minority class.

3.1.2 Malhari Dataset

This data is compiled and named the "Malhari" dataset, taken from a hospital in India[1]. The images of the "Malhari" dataset were taken from 32 patients for research and development purposes. Pictures of an LBC stands for Liquid-Based Cytology, were taken from Mendeley, and the "Colposcopy" dataset consists of four images taken for each patient. Additionally, there are images from a "*Pap Smear*" test, which includes 10 photo corrections[1].

Therefore, the "Malhari" data consists of two different categories ("*Colposcopy*" and "*Pap Smear*"). Figure 4 illustrates the distribution of "Malhari" datasets, which contain seven distinct cell categories distributed as follows: i) The "Colposcopy" is known as Cervical Intraepithelial Neoplasia (CIN)[31], which has three cases: CIN1, CIN2, and CIN3, which represent three cases of dysplasia defect (mild, moderate, and severe cancer), respectively, taken from "*Colposcopy*" dataset. ii) There are four categories of "*Pap Smear*" results[32], which are HSIL, LSIL, NILM, and SSC. The first part has the HSIL and LSIL that indicate high-grade and low-grade squamous intraepithelial lesions, respectively. Then, the NILM indicates a negative result. The last type is SSC, which is a definite cancer. The model is trained, and then an evaluation of the proposed model is performed based on using two different types of "Malhari" datasets.

Figure 5 represents the "Colposcopy" images from the "Malhari" dataset. The first image represents a negative case, while the last two images depict positive cases.

Figure 6 shows the "*Pap Smear*" images, with the first two images illustrating negative cases and the last two images illustrating positive cases.

Data Preprocessing for "Malhari" dataset

(1) Data Selection.

Realistic cervical image data collected from hospitals is vital for examining cells and developing automated systems for diagnosing cervical cancer, supported by various publicly available datasets. Here, we use the "Malhari" dataset, which consists of seven different cell categories, each containing 453 images, as in Figure 4. Each image is classified according to the previously mentioned seven categories obtained from "*Colposcopy*" and "*Pap Smear*". Figure 4 shows the distribution of the "Malhari" dataset. Two portions of the image collection are used: the first portion is for training, and the other portion is for the testing process. The percentage of data used in the testing is 30% for both types of total data size, while the other data is for training (70%). To effectively classify images, we need to normalize and scale them due to their unequal sizes.

(2) Image Segmentation.

Data augmentation and standardization RGB data with a range $[0-1]$ for floats, and $[0-255]$ for Integers. The photos are then scaled to 300×300 pixels using the Matplot library, as shown in Figure 7.

We divided the used data into three sets: training, validation, and testing. Table 4 shows the division and the number of images in each part.

(3) Categorical data.

The three classes for the first type ("*Colposcopy*") are: CIN1, CIN2, and CIN3. They are given the

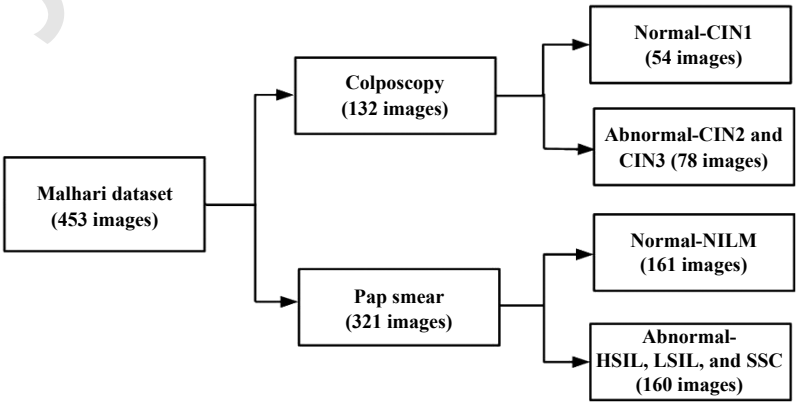

Fig. 4 The Distribution of "*Malhari*" dataset: ("*Colposcopy*" and "*Pap Smear*").

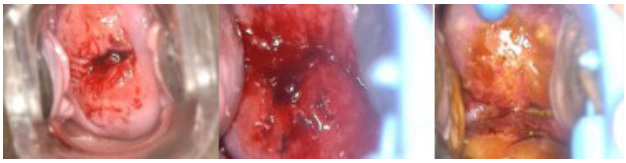

Fig. 5 Images taken from "*Colposcopy*" dataset.

following labels: CIN1: 0, CIN2:1, CIN3:2, and the four classes for the second type ("*Pap Smear*"): NILM, HSIL, LSIL, and SSC. They are given the following labels: NILM:0, HSIL:1, LSIL:2, and SSC:3.

(4) Data Augmentation.

The used images are enhanced to increase the size of the dataset. As indicated in Table 5, eight distinct processes are carried out on the training data cell images and their values. The fill_mode, shear_range, zoom_range, height_shift_range, width_shift_range

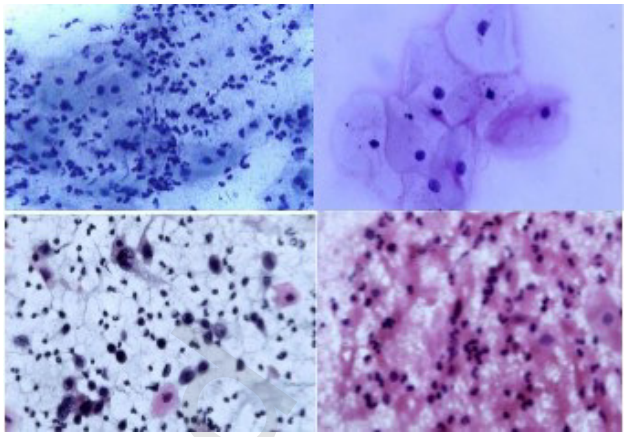

Fig. 6 Images taken from "*Pap Smear*" dataset.

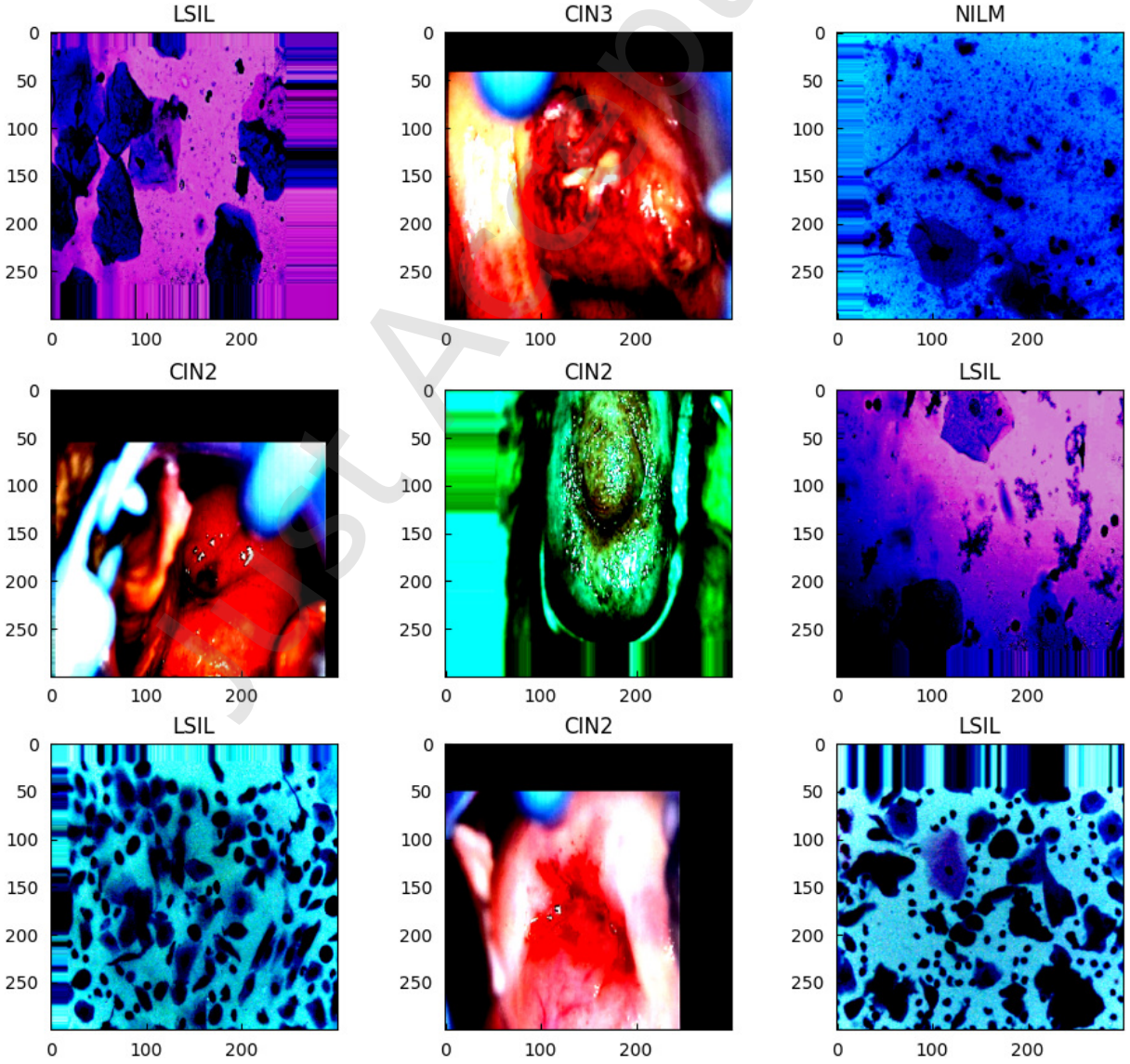

Fig. 7 Collection of cervical cancer images after data augmentation process.

Table 4 Distribution of "Malhari" images with respect to the method and types ("*Colposcopy*" and "*Pap Smear*").

	Colposcopy			Pap Smear			
|I|	CIN 1	CIN2	CIN3	HSIL	LSIL	SSC	NILM
Training	37	35	19	28	56	28	112
Validation	8	7	4	6	12	6	24
Testing	9	8	5	6	12	6	25
Total images	54	50	28	40	80	40	161

Table 5 Strategies for the augmentation process applied to the dataset.

Method of data augmentation	Values
Shear_range	0.2
Zoom_range	0.2
Rescale	1/255
Height_shift_range	0.2
Width_shift_range	0.2
Samplewise_std_normalization	True
Fill_mode	Nearest
Samplewise_center	True

operations, samplewise_center, and samplewise_std_ normalization are used to enhance the training photos.

3.2 Catboost Algorithm for Risk Factors Cervical Cancer

CatBoost is an abbreviation for (Category Boosting), as it belongs to a family called algorithmic boosting algorithms. It is a special case of the decision tree, which builds an effective model for classification by using a group of weak trees. Then, in each step, it merges the trees repeatedly, which gradually leads to improvement in the model's performance. Also, step by step, this algorithm is used in binary classification, as in our case here, by using what is called log loss as an alternative to the loss function. It is also possible to build a custom loss function and thus determine the final class through predictive results. Coming from individual trees that belong to the ensemble. The following are the primary steps in training the developed CatBoost classifier:

- Step 1: Initialize the model with the proportion logarithm of frequencies for different classes found in the training data sample.
- Step 2: Compute the gradients of the loss function for the models of class predictions.
- Step 3: Organize categorical features by their statistical impact on the target variable, and sort them accordingly.
- Step 4: Develop decision trees using a variant of the gradient boosting algorithm, where each iteration involves constructing a new tree by recursively dividing the training data.
- Step 5: Employ *L*2 regularization and Newton's method type regularization on the leaf values of the trees to prevent over-fitting.
- Step 6: Utilize a weighted voting technique to make class predictions, in order to combine the outputs of the ensemble's trees.
- Step 7: Assess the loss function to evaluate and measure the model's performance.
- Step 8: Halt iterations if the model's performance fails to improve on the validation dataset after a specified number of attempts.
- Step 9: Return to step 2 to continue refining the model[33].

Given that a used dataset is manipulated as follows in Eq. (3):

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}, \tag{3}$$

where $x_i \in \mathbb{R}^d$ is the feature vector of risk factors for the i^{th} patient. $y_i \in \mathbb{R}$ is the target variable (e.g., cervical cancer classification or risk score). n is the number of samples in the dataset.

The CatBoost model predicts the target y as follows in Eq. (4):

$$\hat{y}_i = F(x_i;\, \Theta) = \sum_{m=1}^{M} \eta \cdot T_m(x_i;\, \Theta_m) \tag{4}$$

where $T_m(x_i;\, \Theta_m)$ is the output of the $(m)^{th}$ decision tree in the CatBoost model, parameterized by Θ_m. M is the total number of trees. $\eta \in (0,\, 1]$ is the learning rate that determines how much each tree contributes to the overall model. $\Theta = \{\Theta_1, \Theta_2, \ldots, \Theta_M\}$ is the parameters of all decision trees.

3.2.1 Objective Function

CatBoost minimizes a specific loss function $\mathcal{L}(\Theta)$, which could be:

(1) For classification, the cross-entropy loss is used as in Eq. (5).

$$\mathcal{L}(\Theta) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)] \quad (5)$$

(2) For regression, the mean-squared error is used as in Eq. (6).

$$\mathcal{L}(\Theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (6)$$

3.2.2 Handling Categorical Features

CatBoost uniquely handles categorical features, where $x_c \in \mathcal{X}_c$ by encoding them using *ordered statistics* and *target-based encoding* as in Eq. (7).

$$\text{Encoded Value} = \mathbb{E}[y \mid x_c, \text{ prior data}] \quad (7)$$

This ensures no data leakage and accounts for dependencies between categorical features and the target.

3.2.3 Final Risk Factor Model

The model's output is a probability score $P(y=1 \mid x)$ for classification, representing the likelihood of cervical cancer based on the input risk factors, as in Eq. (8):

$$P(y=1 \mid x) = \sigma(F(x;\ \Theta)) = \frac{1}{1+e^{-F(x;\ \Theta)}}, \quad (8)$$

where $\sigma(z)$ is the Sigmoid activation function.

Consequently, CatBoost uses an adaptive technique during the iterative training phase to automatically modify the model parameters. The algorithm is more successful because of its adaptive nature. It has several benefits, including the capacity to model with high accuracy, resilience to data noise, and less susceptibility to over-fitting, which may be troublesome for other algorithms. It also facilitates multi-class categorization and makes assessing feature significance possible[34].

We begin by importing the necessary modules and loading data from a "CSV" file into a Pandas Data Frame, separating features and target columns. Then, we divided the data into 80% for training, and 20% for testing, as the data were unbalanced. We used the SMOTE algorithm on the training data. After that, we used the CatBoost classifier with depth, iterations, and learning rate as hyper-parameters to adjust within the algorithm. Grid Search CV' is the proposed strategy for adjusting the hyper-parameters and thus evaluating the model using both grid search and k-fold cross-validation using 5-folds. The grid search algorithm is one of the methods used to find the optimal parameters for the performance of the proposed model, as it searches in a specific grid. In advance, systematically, in hyper-parameters, k-fold cross-validation divides the dataset into a specified number of folds. We used five folds here, one for validation and the other for training, to improve flexibility in evaluating the model. This process is repeated k times. Each time, a different fold is used to verify the validity. Ultimately, the average performance is determined from the iterations used to enhance the model's generalization ability. This accurately assesses the system's performance and creates an effective classification system by adjusting these parameters. Figure 8 shows the architecture of the proposed CatBoost model.

3.3 CNN for Malhari Dataset

This section clarifies the representation of a specific subset of machine learning models designed to manage structured grid data, such as audio or picture files. CNN model can extract hierarchical patterns and features from incoming data automatically, which has transformed the field of computer vision. CNN networks consist of a group of convolutional and pooling layers that are fully connected between the

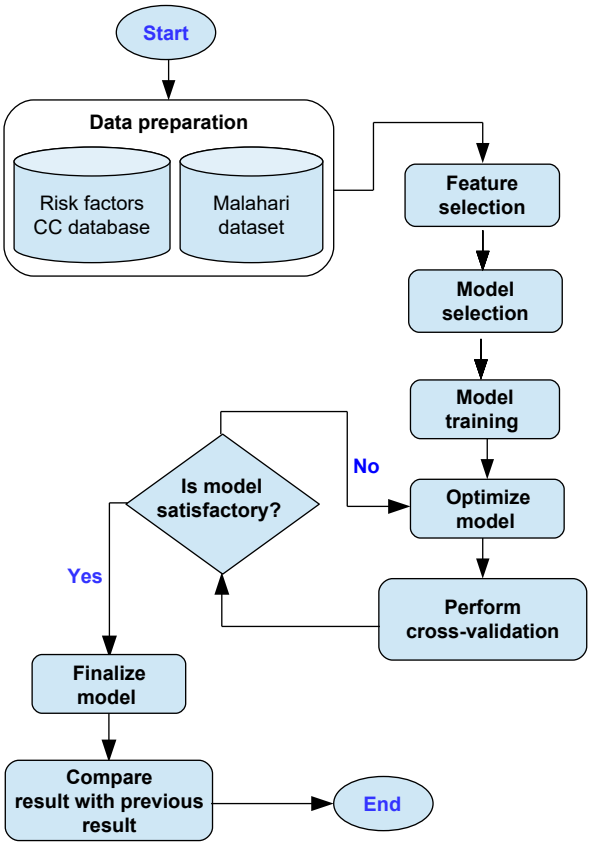

Fig. 8 The flowchart of the proposed CatBoost model for the risk factors cervical cancer dataset.

algorithm layers[35]. The task of the convolutional layer is to extract features by merging kernels across inputs. This is done by applying filters across the data, while the pooling layers work on simplifying calculations and preserving important information. CNNs can accurately classify objects, identify patterns, and segment images through a continuous learning and optimization process. They are an essential component of contemporary artificial intelligence and machine learning systems because of their extensive uses in image identification, object detection, medical imaging, and autonomous cars, among other fields.

Convolutional and pooling layers comprise most of the layers in CNNs, while batch normalization layers are included in certain deep learning architectures. Since there are three dimensions for each of the layers, which are width, height, and depth, or $(m \times m \times r)$, the width must be equal to the height, and it is expressed by m, since the depth is the same as the channel number and expressed by r. In RGB color images, the channel is equal to 3. Each layer contains a set of kernels denoted by k, and it also has three dimensions $(n \times n \times q)$, where n must be less than m and q must be less than or equal to r as variable dimension[35]. In the following are the layers of the CNN model:

1) Convolution Layer: The initial stage in obtaining valuable data from photos is the convolution layer. Convolution uses tiny squares of input data to learn image features, which aids in maintaining the relationship between the image's pixels. This is accomplished mathematically using two inputs: a kernel or filter and a portion of the image as a matrix[36]. It uses Eq. (9) to calculate the input-weight dot product[35]. Meanwhile, the foundation for feature maps is coupled with the input is the kernel:

$$h^k = (W^k \times x + b^k), \tag{9}$$

while generating k feature maps, h^k convoluted with input and each having a size of $(m-n-1)$. The W^k and b^k represent weights and bias, respectively.

2) Pooling Layer: The main goal of this layer is to reduce the number of parameters while retaining the most relevant data for very large images. Spatial pooling, also referred to as down-sampling or sub-sampling, minimizes the number of dimensions in every feature map. The three main forms of pooling layers are typically average pooling, max pooling, and sum pooling[36]. The max pooling layer is computed according to Eq. (10).

$$p_k^{(l)} = \max_{i,j \in \text{pooling window}} a_k^{(l)}(i,\ j) \tag{10}$$

3) Flatten Layer: When working with multi-dimensional inputs, such as image datasets, the flattening layer class is crucial for modeling your input sub-caste, creating a neural network model, and efficiently transferring the data to each neuron in the model. Eq. (11) illustrates the mathematical formula of Flatten Layer.

$$f = \boldsymbol{W}_{\text{fc}} \cdot \boldsymbol{p} + \boldsymbol{b}_{\text{fc}}, \tag{11}$$

where $\boldsymbol{p}$ is the Flattened feature vector. $\boldsymbol{W}_{\text{fc}}$ is the Weight matrix of the fully connected layer. $\boldsymbol{b}_{\text{fc}}$: is the Bias vector.

4) Dense Layer: It is based on the output of the convolutional layers. The images are classified using a dense layer. Each neuron in every layer computes a weighted average of its inputs, as in Eq. (12). This processed input is then passed through a non-linear function known as the activation function. Therefore, the output of the neuron is determined by the result of this activation function. Then, the final output layer uses the Softmax function to create a probability distribution over the classes:

$$\hat{y}_i = P(y = c \mid x_i) = \frac{\exp(f_c)}{\sum_{j=1}^{C} \exp(f_j)} \tag{12}$$

where c is the total number of classes (e.g., risk groups for cervical cancer), and f_c is the output for class c.

We use a powerful, well-known, and widely used methodology for analyzing images to perform classification and detection operations to detect cervical cancer in women. We used the CNN algorithm with several layers, which are 10 layers, one for inputs with 32 filters, each with a 3×3 kernel and a ReLU activation function. The ReLU function is the most commonly used activation function in the construction of CNN algorithm. It is the main function for converting all complete values coming from the inputs into positive numbers. It has many advantages, but the most important of them is the low computational load that is represented in Eq. (13).

$$f(x)ReLU = \max(0,\ x) \tag{13}$$

There are 8 layers for the hidden layer, then 3 layers for MaxPooling2D with $2 - layer$ of Conv2D layers, and (Flatten, Dropout, and Dense) layers. Then, there is one dense output layer with 1 unit and a Sigmoid

activation function for binary classification.

Sigmoid: The main inputs to the Sigmoid activation function are real numbers, while the outputs are the intervals between zero and one. The curve of the Sigmoid function ($\sigma(x)$) can be represented by the following mathematical Eq. (14):

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{14}$$

Consequently, we used two activation functions to deploy a powerful and effective learning framework using ReLU and the ability to make decisions via Sigmoid, where ReLU is characterized by both the following: first, computational intelligence, and second, its ability to address the vanishing gradient problem by maintaining gradients for values that are positive. Hidden layers often utilize non-linearity to ensure faster convergence during training. We used the Sigmoid activation function, which is ideal for output layers when classifying. However, we avoided using it in the hidden layers because it can lead to vanishing gradients during backpropagation. We used the ReLU function in the hidden layers of the CNN network, where ReLU facilitates effective gradient propagation by adding non-linearity to the feature representations. Then, we used the Sigmoid function for the final classification in the CNN's output layer.

The model takes images of the input data and begins applying the Conv2D layer. After completing, it moves to the MaxPooling2D layer, which reduces spatial dimensions. Then, it is followed by two layers of Conv2D, increasing the number of filters (32, 64, and 128) that have been applied and combined with the MaxPooling2D layers. The flattened layer converts the 2D output into a 1D vector, which will be suitable for dense layers. A dropout layer with a 50% dropout rate helps prevent over-fitting. Two dense layers follow, featuring 512 units with ReLU activation in the first layer, and a single unit with Sigmoid activation in the last layer, which is ideal for binary classification. The model is compiled using the "Adam" optimizer and binary cross-entropy loss, while accuracy acts as the monitoring metric during training. The training process involves feeding the training data and labels, and then it specifies the number of epochs. Validation data can evaluate performance during training, and the model's overall performance can be assessed using evaluation methods after training. Figure 9 shows the Keras layer models in sequence, and Figure 10 shows the main architecture of the CNN algorithm, which is adapted from[37, 38].

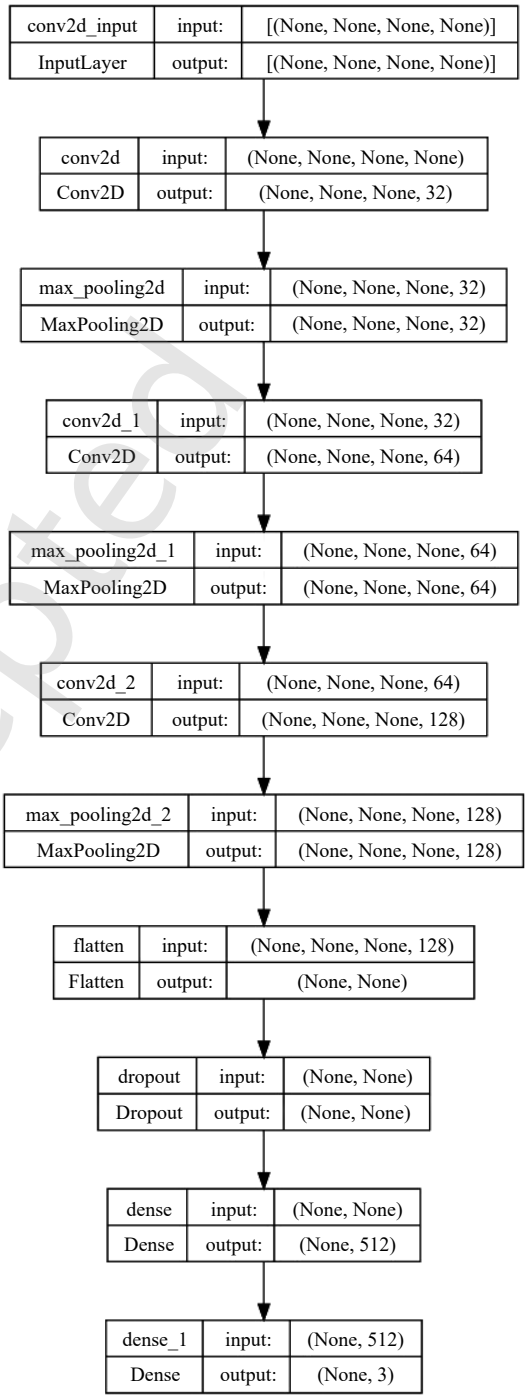

Fig. 9 The Keras layer models in sequence.

Algorithm 2 presents the main outlines of the CNN process. It begins by loading the data, and then dividing this data into two parts: one for training with a size of 80% of the data, and the other part for testing 20%. This will determine the size of the images and the patch size. Then, it performs data preprocessing and augmentation based on the given values. We then build

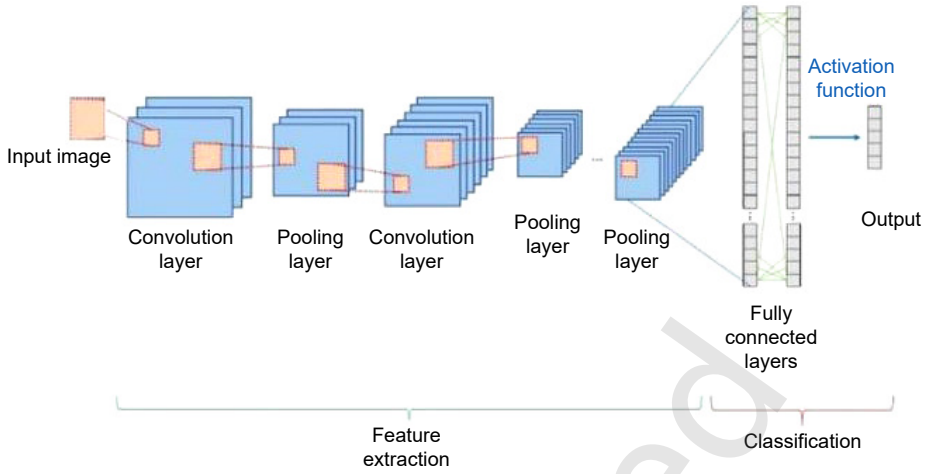

Fig. 10 The architecture of the CNN algorithm, adapted from[37, 38].

the model for the CNN algorithm and compile the model for fitting. In the end, we evaluated this model based on accuracy and calculated the evaluation metrics (Accuracy, Precision, Recall, and F1_score) for each predicted class. Figure 11 shows the flow chart of our developed CNN algorithm.

Algorithm 2: CNN Algorithm

Main steps of the proposed CNN algorithm:

1) Pre-processing

- Load Dataset.

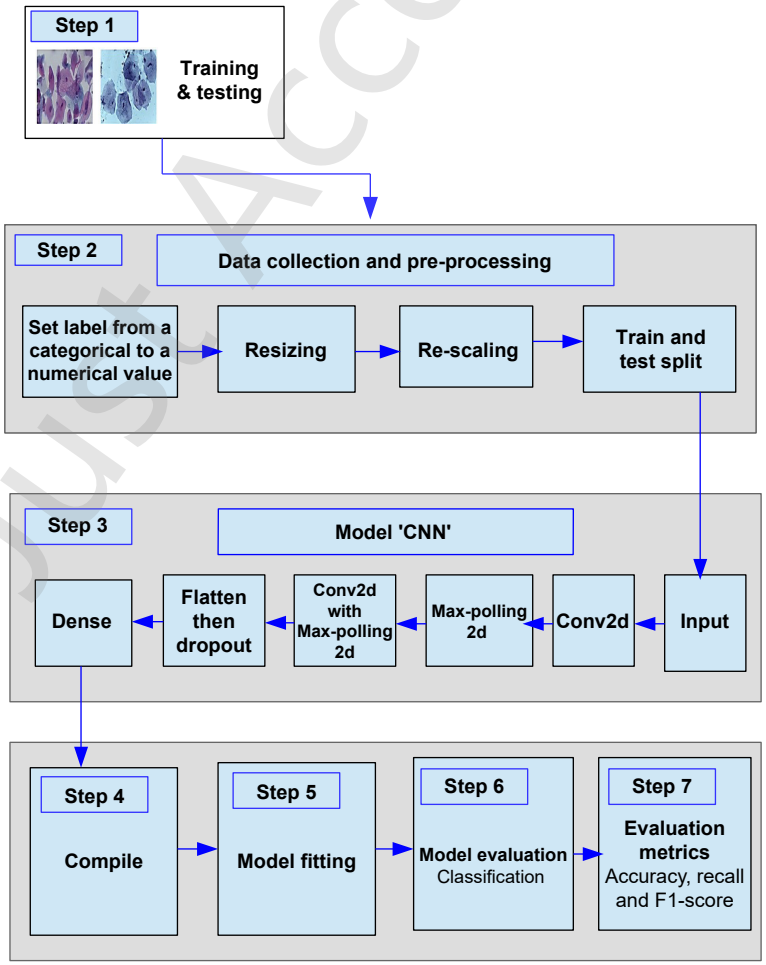

Fig. 11 The flowchart for processing the Malhari dataset based on developed CNN model.

● Split Dataset: into training with 80%, and testing sets using an 20% split.

● Define Image Dimensions and Batch Size: Set the image dimensions to 300×300 and the batch size to 32.

● Data Pre-processing and Augmentation.

2) Build CNN Model

● Import Modules.

● Freeze Layers.

● Add Custom Layers.

● Compile Model: The compilation process uses "Adam" optimizer, a defined loss function, and evaluation metrics (like Accuracy).

● PrInt. Model Summary.

3) Fitting a Model.

● Train model: Train the compiled model using training data.

4) Model Evaluation.

● Evaluate the model on the test set to obtain performance scores (like Accuracy).

● Generate Predictions.

● Convert Probabilities to Predictions.

● Get Class Names: Retrieve the class names from the training set.

5) Calculate Evaluation Metrics:

● Calculate Metrics for Predictions: Calculate various evaluation metrics (Accuracy, Precision, Recall, F1_score) for each predicted class.

4 Experimental Results

4.1 Experimental Settings

We trained and tested our proposed models on publicly available datasets. The CNN model demonstrated reliable performance in detecting and classifying images, through which we classified the seven categories of cervical cancer. We then calculated the loss by implementing a function called the loss function using the sparse categorical cross-entropy function. The lower the loss value, the greater the improvement in the training and testing results. Eq. (15) shows the mathematical formula for the loss function:

$$loss = \sum_{i=0}^{k} q_i \log p_i, \tag{15}$$

where q_i is the probability related to the i^{th} class, and p_i represents the true label. We used the "Adam" optimizer for prediction to improve the weights of the inputs.

We set the weights in the layers of the CNN model as mentioned previously in subsection 3.3. Then, we used 32 and 150 for the batch size and number of epochs for training, respectively, to ensure convergence.

4.2 Experimental Environments

The models have been implemented using Python deep learning with the Keras package, which is based on the TensorFlow 2.8.2 version. The network is trained at Google Colaboratory, which offers free access to the NVIDIA Tesla T4 GPU, version 460.32.03 of the graphic driver, and version 11.2 of CUDA programming.

4.3 Evaluation Metrics

We implemented CatBoost and CNN algorithms to evaluate the proposed classifiers from the cervical cancer datasets. We measured this model using Precision, Recall, F1_score, and Accuracy. Therefore, these are the most suitable metrics used to examine models for multi-category classification. The following formulas are for each one, respectively:

Precision: This parameter indicates the percentage of patients with cervical cancer and how accurately the algorithm predicts their diagnosis. As in Eq. (16), Precision calculates the proportion of individuals with cervical cancer that the model accurately predicts.

$$Precision = \frac{TP}{TP + FP} \tag{16}$$

Recall: The model's ability to diagnose cervical cancer is reflected in Eq. (17), where a recall of 1 signifies that every patient with cervical cancer is correctly identified.

$$Recall = \frac{TP}{TP + FN} \tag{17}$$

F-Measure (F1_score): The harmonic mean of the model's sensitivity and precision represents the combination of the two parameters, as seen in Eq. (18). A higher F-measure indicates fewer individuals who are incorrectly categorized as having cervical cancer or not.

$$\text{F1\_score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{18}$$

Accuracy: The number of accurate predictions made by the model out of all the predictions made, as seen in Eq. (19).

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{19}$$

4.4 Finding Results

4.4.1 Risk Factors Cervical Cancer Dataset

Figure 12 shows the percentage of accurate and inaccurate predictions the model produced, broken down by class. The class labels predicted by the model are represented by the "Predicted Labels" on the X-axis of this matrix. In contrast, the "True Labels" on the Y-axis represent the actual class labels from the dataset. The number 160 in the upper left cell indicates that the model accurately predicted the positive or negative class based on the label assignment 160 times. Five, which reflect accurate predictions for the opposite class, are displayed in the bottom right cell. There are neither false positives nor false negatives, indicating that the model did not mistakenly categorize any cases in these two classes. The top-right and bottom-left cells are both blank. This shows that the model is performing very well, with no reported prediction mistakes.

A Receiver Operating Characteristic (ROC) curve in Figure 13 is a graphical tool for assessing how well binary classification models work. As depicted in Figure 13, with an Area Under the Curve (AUC) of 1.00, this particular ROC curve shows flawless classifier performance. With a True Positive Rate (TPR) of 1 and a False Positive Rate (FPR) of 0, the curve begins at the origin $(0,0)$, climbs quickly to the top-left corner $(0,1)$, and then stretches horizontally to the right corner at $(1,1)$. The AUC value of 1.00 indicates that the classifier can accurately distinguish

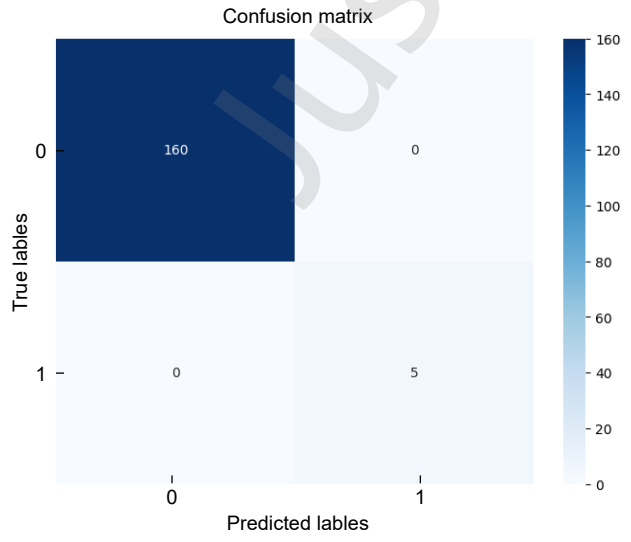

Fig. 12 Confusion Matrix (CM) for the risk factor cervical cancer dataset.

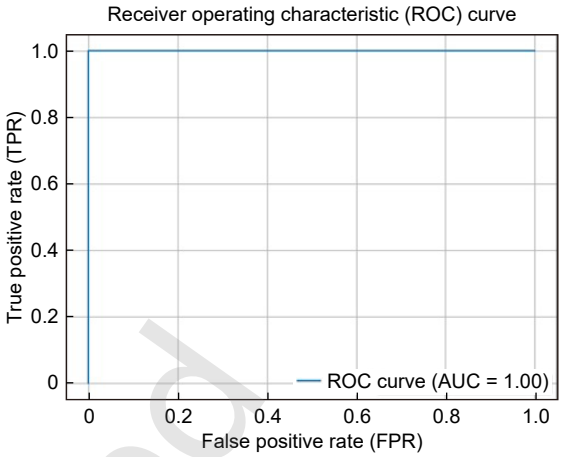

Fig. 13 ROC curve for cervical cancer.

between the two classes, according to the ideal ROC curve.

Figure 14 encapsulates the performance metrics, namely Precision, Recall, and F1_score, for a classification task. Each metric represents a different aspect of the model's effectiveness in classifying instances. Precision gauges the accuracy of positive predictions; Recall measures the model's ability to identify all positive instances; and the F1_score serves as a balanced metric, considering both Precision and Recall. With all metrics achieving a perfect 100% in both rows, the model demonstrates flawless performance. This indicates that it accurately identifies all positive instances while minimizing false positives. Such exemplary performance underscores the model's robustness for classification tasks.

Table 6 compares the presented research in the same field using the same dataset. This is essential for evaluating and comparing their strengths and weaknesses against our proposed model, which we

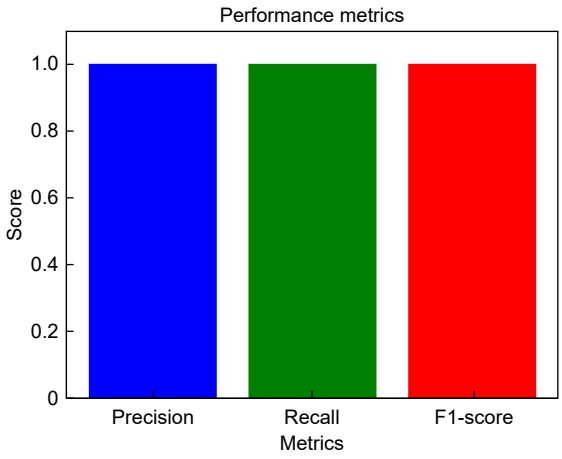

Fig. 14 Performance metrics for both classes.

developed based on different evaluation criteria. After applying the new enhanced CatBoost algorithm, we obtained a higher accuracy with a value of 100% compared to the research that was produced previously, where the highest accuracy by using the LMT algorithm is 99.31%, and the lowest accuracy is using Random Forest at 96.4%. Consequently, our proposed model achieved the best results in other measurements like Precision, Recall, and F1_score.

Table 6 compares evaluation metrics for four machine learning models: Decision Tree[5], Random Forest[16], LMT[7], and the proposed enhanced Catboost model. CatBoost has an outstanding accuracy rate of 100%, while Decision Tree comes close behind at 98.82%. Random Forest has a lower Accuracy of 96.4%, whereas LMT is somewhat better at 99.31%. When it comes to precision, CatBoost once again leads the way with 100% Precision, showing no false positives, while Decision Tree has 87.51%. Recall measures show that both Decision Tree and CatBoost have 100% Recall, proving their ability to catch all relevant instances. Random Forest, on the other hand, is trailing at 96.4%, while LMT is at 99.6%. Finally, in the F1_score category, CatBoost leads with a flawless score, followed by LMT with 99.6%. The Decision Tree is at 93.335%; however, the Random Forest does not provide a value of F1_score. Overall, CatBoost surpasses the other models on all criteria, demonstrating its efficacy in the evaluation setting.

4.4.2 Malhari Dataset

To our knowledge, we use the "Malhari" dataset for the first time, presenting groundbreaking discoveries. As a result, "*Pap Smear*" images are represented by a matrix (A) in Figure 15, and "*Colposcopy*" images are represented by a matrix (B) in Figure 16. The degree to which the suggested algorithm classifies instances of the data collection is represented by this matrix.

For categorization results, the "*Pap Smear*" images display a confusion matrix in Figure 15. NILM, SSC, LSIL, and HSIL are four different classes. The matrix shows that 11 times HSIL is properly predicted; however, 2 cases are incorrectly classified as LSIL, 1 case for SSC, and 0 case for NILM. Three examples are incorrectly classified as SSC and one as NILM, out of the 17 accurate predictions for LSIL. There is just one case of SSC incorrectly labeled as LSIL out of 44 correct predictions. Finally, the NILM prediction is accurate nine times, and no additional categories are misclassified. According to the matrix, most of the forecasts came true, with SSC having the highest

Table 6 Comparison performance of models based on the Cervical Cancer Dataset.

Model	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1_Score (%)
Decision Tree [5]	Risk Factors Dataset	98.82	87.51	**100.00**	93.34
Random Forest [16]	Cervical Cancer Risk Factors	96.40	96.50	96.40	---
LMT [7]	Cervical Cancer Dataset	99.31	99.60	99.60	99.60
CatBoost (Our Proposed Model)	Risk Factors Cervical Cancer	**100.00**	**100.00**	**100.00**	**100.00**

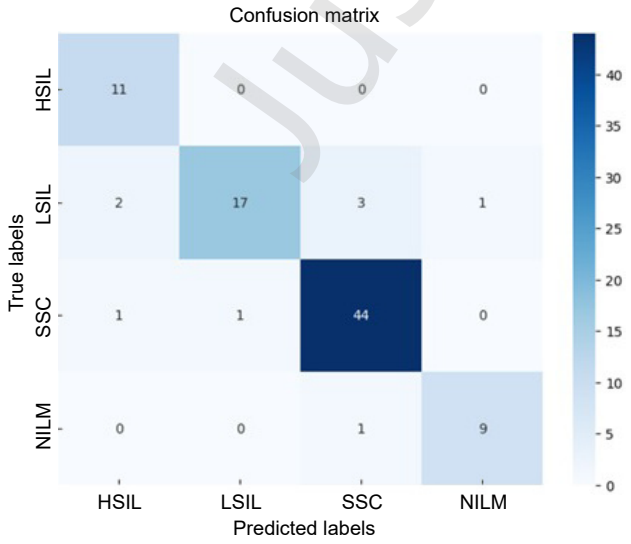

Fig. 15 Confusion Matrix (A) for "*Pap Smear*" images.

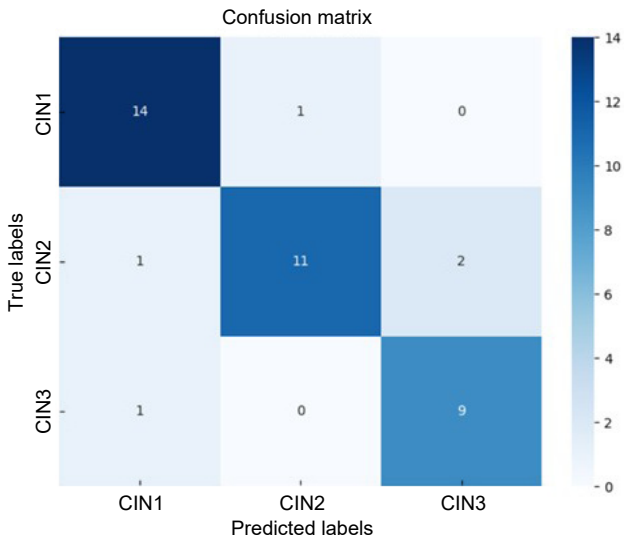

Fig. 16 Confusion Matrix (B) for "*Colposcopy*" images.

percentage of correct predictions.

Figure 16 shows a confusion matrix for the three expected labels: CIN1, CIN2, and CIN3 from the "*Colposcopy*" dataset. The matrix is a 3×3 grid where each cell displays the number of predictions. There are 14 accurate predictions for CIN1; however, there is just one misdiagnosis for CIN2 and CIN3. There are 11 accurate predictions for CIN2, with 0 cases incorrectly labeled as CIN3, and 1 instances correctly predicted as CIN1. In conclusion, for CIN3, 9 of the predictions are accurate, and the 2 case is incorrectly identified as CIN2. No CIN1 is mistakenly classified as a CIN2. This matrix can be used to assess how well a classification model is performing.

Based on Figure 17, the precision figure on the left presents the true positive percentage of the positive predictive ratios provided by the model, where the highest rate in CIN2, where the precision is 92%, and then CIN1, followed by CIN3, with percentages of (88% and 82%), respectively. The Recall represents the extent of the system's ability to sort positive cases from among the actual positive cases shown by the data, where CIN1 had the highest rate with a rate of 93%, followed by CIN3, then CIN2 with a rate of (90% and 79%), respectively. The F1_score in Figure 17 (on the right side) provides a fair assessment of a model's performance and is the harmonic mean of Precision and Recall. The highest percentage for CIN1 is 90%, followed by CIN3, then CIN2, with percentages of (86% and 85%), respectively.

Therefore, analyzing the three subfigures in Figure 17 reveals that CIN1 consistently outperforms in terms of Precision, Recall, and F1_Score, indicating the model's effectiveness. However, the performance of CIN2 is somewhat worse, particularly in recall, which indicates that the model struggles more to recognize true positive cases for CIN2 accurately. Regarding performance indicators, CIN3 is between CIN1 and CIN2, showing balanced results that are a little lower than those of CIN1. Overall, these numbers provide a clear assessment of the model's classification strengths and weaknesses for each class, offering valuable insights for future iterations and performance evaluations.

Based on Figure 18, the Precision shows that LSIL has the best precision rate (94%), suggesting a high degree of accurate positive predictions. With a precision of 92%, SSC trails closely behind, demonstrating good accuracy in positive predictions. NILM has a precision rate of 90%, indicating high precision in the identification of NILM instances. With a precision of 79%, HSIL does less well than the other classes in the positive prediction precision. According to the recall figure with a 100% recall rate, HSIL has

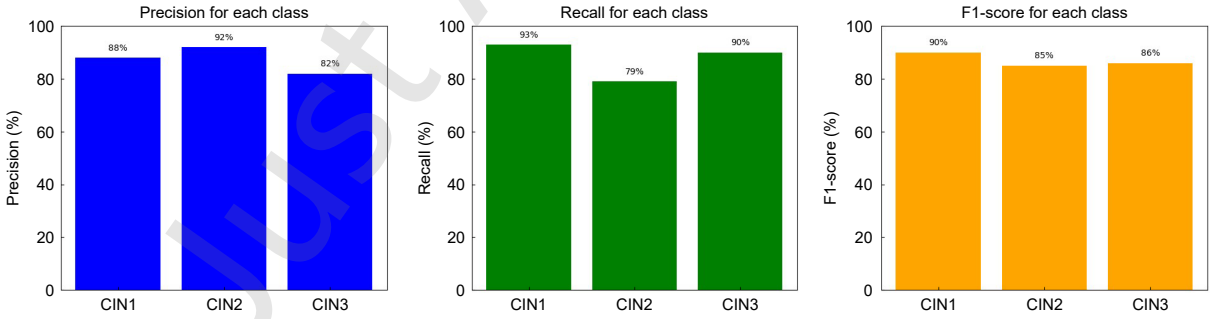

Fig. 17 The Precision, Recall, and F1_score for each class of "Colposcopy" type.

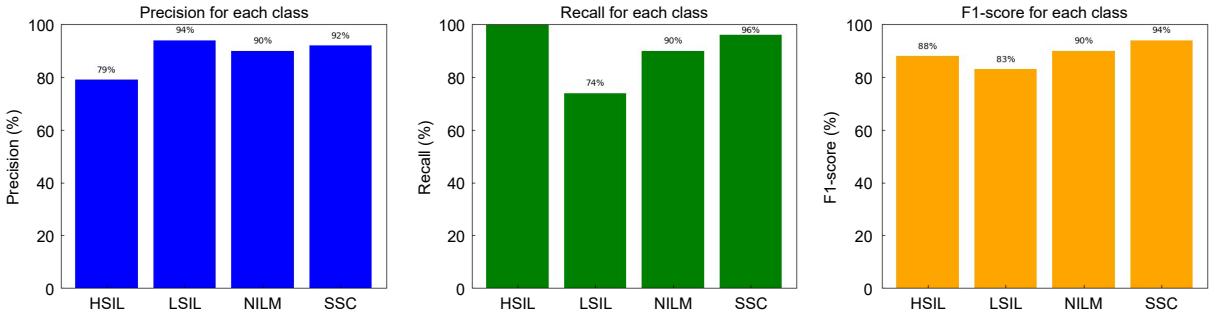

Fig. 18 The Precision, Recall, F1_score for each class of "*Pap Smear*" type.

the best recall, demonstrating a strong capacity to collect genuine positive cases for that class. Additionally, SSC has a high Recall of 96%, demonstrating a good capacity for accurate SSC instance identification. NILM has a 90% recall rate, indicating that many NILM occurrences are correctly identified. But LSIL's recall is just 74%, indicating that the model would have greater trouble accurately detecting genuine positive cases for LSIL.

Also, in Figure 18, the SSC achieves the highest F1_score for that class (94%), indicating a strong balance between recall and precision. Similarly, NILM demonstrates a high F1_score of 90%, suggesting balanced performance. Despite having less accuracy, HSIL's F1_score of 88% shows that it has a fair balance between Recall and Precision. LSIL, on the other hand, has the lowest F1_score (83%), showing difficulties in keeping memory and precision in check.

Upon comparing the three subfigures in Figure 18, HSIL performs well in Recall and F1_score, demonstrating efficient HSIL instance detection and categorization. However, compared to other classes, HSIL's Precision is somewhat poorer. SSC shows good performance in terms of Precision, Recall, and F1_score, pointing to efficient SSC instance detection and classification with a decent Recall-to-Precision ratio. Additionally, NILM does well across the board. LSIL has difficulties balancing recall and precision, lowering the F1_score even when Precision is good. Together, these numbers offer a thorough evaluation of the model's advantages and disadvantages for categorizing each class, which can help with further research or improvement of the model's functionality.

We applied statistical analysis and error analysis for a fair comparison with state-of-the-art models[39]. These two powerful analyses are important. First, the statistical analysis evaluates the viability of our proposed model concerning the use of the SMOTE technique, which is performed to create a statistically balanced dataset that enhances the final overall classification performance. Second, the error analysis presents examples from the dataset demonstrating instances where the base classifiers made incorrect predictions, while our proposed enhanced predictive model, which combines CatBoost and CNN algorithms, delivered correct predictions. Therefore, we performed the statistical analysis and error analysis as follows.

4.5 Statistical Analysis

(1) Only 2% of the Risk Factors Cervical Cancer dataset represented positive instances, indicating a notable class imbalance. This imbalance is corrected by using SMOTE technique, which produces a balanced dataset with better classification performance.

(2) Two types of images ("*Pap Smear*" and "*Colposcopy*") with several categories are included in the "Malhari" dataset. Techniques for data augmentation made sure there was enough data for efficient training.

4.6 Error Analysis

(1) Risk Factors Cervical Cancer Dataset:

- Although the CatBoost model produced flawless measurements, possible problems can occur as a result of addressing class imbalances and data preprocessing. The SMOTE technique and hyperparameter tuning are employed to reduce and adjust class imbalances.

- Overfitting Risks: Achieving ideal metrics, especially with a balanced dataset, can indicate overfitting. Therefore, Five-fold cross-validation is used to make sure the results are generalizable.

(2) Malhari Dataset:

a) Pap Smear images:

. Misclassification of LSIL as HSIL or SSC is the most common error found in its classification. This issue likely arises from the apparent similarity between these categories in low-quality images.

. The model's susceptibility to both normalized and augmented data quality suggests possible preprocessing biases.

b) Colposcopy images:

. The primary area of misclassification is differentiating between CIN2 and CIN3, likely due to the subtle visual differences between moderate and severe dysplasia.

. Performance may have been impacted by noise generated by changes in illumination and image resolution.

The enhanced performance of both models stemmed from high-quality inputs and well-balanced datasets. Nevertheless, accuracy is constrained by factors such as image resolution, class overlaps, and small datasets, particularly in the Malhari dataset.

5 Conclusion & Future Works

In conclusion, combining CNN with CatBoost

algorithms has proven to be a very successful method for classifying and predicting cervical cancer. The findings of this research indicate that our newly developed model displayed respectable accuracy rates in the Malhari image dataset based on the CNN algorithm, with 91.1% accuracy for "*Colposcopy*" photos, and 87.2% accuracy for "*Pap Smear*" images. Moreover, our proposed model based on the new enhanced CatBoost method attained a flawless accuracy rate 100% in the structured "*Risk Factors Cervical Cancer*" dataset. These findings highlight the potential of our proposed method for improving diagnostic accuracy and dependability in medical applications by integrating machine learning with deep learning approaches.

This study not only establishes a foundation for future research but also advances the creation of more accurate and reliable diagnostic instruments for cervical cancer. The high accuracy and resilience of the suggested algorithms demonstrate their potential to improve the early identification and treatment outcomes of cervical cancer patients, which will ultimately lead to improved healthcare and lower death rates. In the future, this paper suggests the potential for more AI-driven diagnostic tools in various medical fields, potentially enhancing patient sample collection through a mobile application for hospital physicians.

Acknowledgment

We appreciate the efforts of all authors in developing and implementing this research, and then in preparing, reviewing, and refining the writing process of this paper. Also, the authors would like to thank the anonymous reviewers for their helpful suggestions to improve the quality of this paper.

References

[1] W. H. Organization *et al.*, "Comprehensive cervical cancer prevention and control: a healthier future for girls and women (2013)," *Tersedia di: http://apps.who.int/iris/bitstream/10665/78128/3/9789241505147eng.pdf*, 2015.

[2] M. Marvan and E. L x opez-V x azquez, "The anthropocene: x Politik–economics–society–science: Preventing health and environmental risks in latin america," 2017.

[3] B. Nithya and V. Ilango, Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction, *SN Applied Sciences*, vol. 1, pp. 1–16, 2019.

[4] A. S. Gunnell, *Risk factors for cervical cancer development*. Karolinska Institutet (Sweden), 2007.

[5] J. Tanimu, M. Hamada, M. Hassan, H. Kakudi, and J. Abiodun, "Machine learning method for classification of cervical cancer," 2022.

[6] S. Mandal and I. Banerjee, Cancer classification using neural network, *International Journal*, vol. 172, pp. 18–49, 2015.

[7] S. H. Ripon, N. Q. Bhuiyan, et al, Cervical cancer risk factors: classification and mining associations, *APTIKOM Journal on Computer Science and Information Technologies*, vol. 4, no. 1, pp. 8–18, 2019.

[8] W. H. Organization *et al.*, "Cancer resolution: cancer prevention and control in the context of an integrated approach," *17th World Heal Assem.[Internet]. Retrospectivo descritivo. Fatores de risco para encaminhamento para hospital foram: cancer, problemas ˆ cardiovasculares, digestivos e metabolicos. Fatores x protetores: AD de enfermagem*, 2017.

[9] A. Alaiad, A. Migdady, R. M. Al-Khatib, O. Alzoubi, R. A. Zitar, and L. Abualigah, Autokeras approach: A robust automated deep learning network for diagnosis disease cases in medical images, *Journal of Imaging*, vol. 9, no. 3, p. 64, 2023.

[10] K. Fernandes, J. S. Cardoso, and J. Fernandes, "Transfer learning with partial observability applied to cervical cancer screening," in *Pattern Recognition and Image Analysis: 8th Iberian Conference, IbPRIA 2017, Faro, Portugal, June 20-23, 2017, Proceedings 8*, pp. 243–250, Springer, 2017.

[11] M. F. Unlersen, K. Sabanci, and M. Ozcan, Determining ¨ cervical cancer possibility by using machine learning methods, *Int. J. Latest Res. Eng. Technol*, vol. 3, no. 12, pp. 65–71, 2017.

[12] S. D. Singh, "Surveillance for cancer incidence and mortality—united states, 2013," *MMWR. Surveillance Summaries*, vol. 66, 2017.

[13] H. K. Fatlawi, Enhanced classification model for cervical cancer dataset based on cost sensitive classifier, *International Journal of Computer Techniques*, vol. 4, no. 4, pp. 115–20, 2017.

[14] W. TEAM, "Cervical cancer," 2024.

[15] F. Alkhateeb, R. M. Al-Khatib, and I. A. Doush, A survey for recent applications and variants of natureinspired immune search algorithm, *International Journal of Computer Applications in Technology*, vol. 63, no. 4, pp. 354–370, 2020.

[16] N. Razali, S. A. Mostafa, A. Mustapha, M. H. Abd Wahab, and N. A. Ibrahim, "Risk factors of cervical cancer using classification in data mining," in *Journal of Physics: Conference Series*, vol. 1529, p. 022102, IOP Publishing, 2020.

[17] C. Zhao, R. Shuai, L. Ma, W. Liu, and M. Wu, Improving cervical cancer classification with imbalanced datasets combining taming transformers with t2t-vit, *Multimedia tools and applications*, vol. 81, no. 17, pp. 24265–24300, 2022.

[18] R. M. Al-Khatib, N. E. A. Al-qudah, M. S. Jawarneh, and A. Al-Khateeb, A novel improved lemurs optimization algorithm for feature selection problems, *Journal of King*

Saud University-Computer and Information Sciences, vol. 35, no. 8, p. 101704, 2023.

[19] K. M. Nahar, M. Ra'ed, M. Al-Shannaq, M. Daradkeh, and R. Malkawi, Direct text classifier for thematic arabic discourse documents., *Int. Arab J. Inf. Technol.*, vol. 17, no. 3, pp. 394–403, 2020.

[20] S. Fekri-Ershad and M. F. Alsaffar, Developing a tuned three-layer perceptron fed with trained deep convolutional neural networks for cervical cancer diagnosis, *Diagnostics*, vol. 13, no. 4, p. 686, 2023.

[21] R. M. Al-Khatib, L. Heilat, W. Qudah, S. Alhatamleh, and A. Al-Khateeb, "A novel improved deep learning model based on bi-lstm algorithm for intrusion detection in wsn.," *Networks & Heterogeneous Media*, vol. 20, no. 2, 2025.

[22] A. Migdady, A. Alaiad, and R. M. Al-Khatib, Efficientnet deep learning model for pneumothorax disease detection in chest x-ray images, *International Journal of Business Information Systems*, vol. 50, no. 1, pp. 1–21, 2025.

[23] S. Fekri-Ershad and S. Ramakrishnan, Cervical cancer diagnosis based on modified uniform local ternary patterns and feed forward multilayer network optimized by genetic algorithm, *Computers in Biology and Medicine*, vol. 144, p. 105392, 2022.

[24] Q. Abuein, M. Ra'ed, A. Migdady, M. S. Jawarneh, and A. Al-Khateeb, "Arsa-tweets: A novel arabic sarcasm detection system based on deep learning model," *Heliyon*, vol. 10, no. 17, 2024.

[25] M. S. Jawarneh, S. M. Shah, M. M. Aljawarneh, R. M. Al-Khatib, and M. G. Al-Bashayreh, "Rib bone extraction towards liver isolating in ct scans using active contour segmentation methods," *International Journal of Advanced Computer Science and Applications*, vol. 16, no. 4, 2025.

[26] M. Kalbhor, S. Shinde, S. Lahade, and T. Choudhury, "Deepcervicancer-deep learning-based cervical image classification using colposcopy and cytology images," *EAI Endorsed Transactions on Pervasive Health and Technology*, vol. 9, 2023.

[27] L. M. Shoohi and J. H. Saud, "Dcgan for handling imbalanced malaria dataset based on over-sampling technique and using cnn.," *Medico-Legal Update*, vol. 20, no. 1, 2020.

[28] R. Hariprasad, T. Navamani, T. R. Rote, and I. Chauhan, "Design and development of an efficient risk prediction model for cervical cancer," *IEEE Access*, 2023.

[29] R. M. Al-Khatib, T. Zerrouki, M. M. Abu Shquier, and A. Balla, Tashaphyne0.4: a new arabic light stemmer based on rhyzome modeling approach, *Information Retrieval Journal*, vol. 26, no. 1, p. 14, 2023.

[30] R. M. Al-Khatib, T. Zerrouki, M. M. A. Shquier, A. Balla, and A. Al-Khateeb, "A new enhanced arabic light stemmer for ir in medical documents.," *Computers, Materials & Continua*, vol. 68, no. 1, 2021.

[31] P. Chatterjee, S. Siddiqui, and R. S. A. Kareem, "Vlr net-an ensemble model to enhance the accuracy of classification of colposcopy images," *Preprint*, 2024.

[32] I. Pacal, "Investigating deep learning approaches for cervical cancer diagnosis: a focus on modern image-based models.," *European Journal of Gynaecological Oncology*, vol. 46, no. 1, 2025.

[33] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," *Advances in neural information processing systems*, vol. 31, 2018.

[34] R. Sanjeetha, A. Raj, K. Saivenu, M. I. Ahmed, B. Sathvik, and A. Kanavalli, Detection and mitigation of botnet based ddos attacks using catboost machine learning algorithm in sdn environment, *International Journal of Advanced Technology and Engineering Exploration*, vol. 8, no. 76, p. 445, 2021.

[35] S. Ghosh, S. J. Sunny, and R. M. Roney, "Accident detection using convolutional neural networks," *2019 International Conference on Data Science and Communication* (*IconDSC*), pp. 1–6, 2019.

[36] H. Chen, J. Liu, Q.-M. Wen, Z.-Q. Zuo, J.-S. Liu, J. Feng, B.-C. Pang, and D. Xiao, Cytobrain: cervical cancer screening system based on deep learning technology, *Journal of Computer Science and Technology*, vol. 36, pp. 347–360, 2021.

[37] N. Moustafa, "Architecture of convolutional neural networks (cnn)," 2023.

[38] S. Khan, H. Rahmani, S. A. A. Shah, and M. Bennamoun, A guide to convolutional neural networks for computer vision, *Synthesis Lectures on Computer Vision*, vol. 8, no. 1, pp. 1–207, 2018.

[39] A. Manna, R. Kundu, D. Kaplun, A. Sinitca, and R. Sarkar, A fuzzy rank-based ensemble of cnn models for classification of cervical cytology, *Scientific Reports*, vol. 11, no. 1, p. 14538, 2021.

Ra'ed M. Al-Khatib received the PhD degree from Universiti Sains Malaysia (USM), Malaysia in 2012. He is currently an Associate Professor in the Department of Computer Sciences, Faculty of Information Technology and Computer Sciences, Yarmouk University. His research interests include Artificial Intelligence, Machine Learning, Natural Language Processing (NLP), Speech Processing, Game Development, Text-to-Speech (TTS), IoT, and Wireless Sensor Network (WSN) domain. He can be contacted at email: raed.m.alkhatib@yu.edu.jo.

Qasem Al-Radaideh received the PhD degree from Universiti Putra Malaysia (UPM), Malaysia in 2005. He is currently a Full Professor in the Department of Information Systems and the Dean of the Faculty of Information Technology and Computer Sciences, Yarmouk University. His research interests include Artificial Intelligence, Data Mining, Information Retrieval, Arabic Language Processing, Text Mining, and Educational Data Mining. He can be contacted at email: qasemr@yu.edu.jo.

Sally Haddad received the Master's degree in Computer Science from Yarmouk University, Jordan, in 2024. She was an Adjunct Lecturer at Al-Balqa Applied University. She is currently working as a Project Engineer in Abu Dhabi at Target Engineering Construction. Her research interests include computer science and related fields. She can be contacted at email: sallyhaddad56@yahoo.com.

Ahmad T. Al-Taani received his PhD in Computer Vision from the University of Dundee, UK in 1994. He obtained a Master's degree in Software Engineering from National University, USA, in 1988 and a Bachelor of Science degree in Computer Science from Yarmouk University, Jordan in 1985. Dr. Al-Taani worked at the Department of Computer Sciences at Yarmouk University since 1995 until 2025. He worked at several universities during his sabbatical and absence leave since 2000. Currently, Dr. Al-Taani is a Professor of AI at the Arab Open University, Amman, Jordan Branch. Dr. Al-Taani's research interests lie in the fields of AI, machine learning and deep learning, computer vision and image processing, and knowledge-based systems.

Heba A. Maabreh received the Master's degree in Computer Science from Yarmouk University, Jordan, in 2024. She has worked as a BTEC teacher, with professional experience in education and computer science–related subjects. Her research interests include data science, artificial intelligence, and related fields in computer science. She can be contacted at email: heba.maabreh198@gmail.com.