scientific reports

Check for updates

OPEN

Integrating cat boost algorithm with triangulating feature importance to predict survival outcome in recurrent cervical cancer

S. Geeitha[1], K. Ravishankar[2], Jaehyuk Cho[3]✉ & Sathishkumar Veerappampalayam Easwaramoorthy[4]

Cervical cancer is one of the most dangerous malignancies in women. Prolonged survival times are made possible by breakthroughs in early recognition and efficient treatment of a disease.The existing methods are lagging on finding the important attributes to predict the survival outcome. The main objective of this study is to find individuals with cervical cancer who are at greater risk of death from recurrence by predicting the survival.A novel approach in a proposed technique is Triangulating feature importance to find the important risk factors through which the treatment may vary to improve the survival outcome.Five algorithms Support vector machine, Naive Bayes, supervised logistic regression, decision tree algorithm, Gradient boosting, and random forest are used to build the concept. Conventional attribute selection methods like information gain (IG), FCBF, and ReliefFare employed. The recommended classifier is evaluated for Precision, Recall, F1, Mathews Correlation Coefficient (MCC), Classification Accuracy (CA), and Area under curve (AUC) using various methods. Gradient boosting algorithm (CAT BOOST) attains the highest accuracy value of 0.99 to predict survival outcome of recurrence cervical cancer patients. The proposed outcome of the research is to identify the important risk factors through which the survival outcome of the patients improved.

Keywords Recurrence, Gradient boosting, Cervival cancer, ReliefF, Attribute selection

The models based on machine learning linked to classification methods have been compared and analyzed. Machine learning ideas have gained substantial recognition in the study fields by utilizing a variety of categorization algorithms and their methodologies. This study evaluates the effectiveness of several classifiers and artificial intelligence models by comparing their outcomes using metrics like recall, precision, and accuracy[1]. The PSO-based mixed method offers the classification of the sick gene from the normal gene with higher precision. The naive Bayes model creates a successful analysis of classification to identify a characteristic of gene expression. Additionally, the amount of noise in the gene information has been reduced. The PSO system is the greatest method for getting the greatest accuracy out of the testing data that the Naïve Bayes algorithm sends. Then, this technique offers an accurate answer with cutting-edge technology and identifies the discriminative gene data[2]. For patients with cervical cancer, tree-based data mining techniques offer reliable prognoses. The SMOTE method was utilized to solve the problem of imbalanced data gathering, in which malignant sufferers were represented compared to healthy persons. The enhanced decision tree outperforms choice forests and decision jungles regarding prediction accuracy as evaluated by the AUROC value. Random forest algorithms and decision jungle techniques were not the best prognostic classification models due to their poor AUROC values[3]. A CNN-based approach for cervical carcinoma identification and categorization was suggested. When the CNN

[1]Department of Information Technology, M. Kumarasamy College of Engineering, Thalavapalayam, Karur, Tamil Nadu, India. [2]Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, Tamil Nadu, India. [3]Department of Software Engineering and Division of Electronics and Information Engineering, Jeonbuk National University, Jeonju-Si, Republic of Korea. [4]Department of Computing and Information Systems, Sunway University, Selangor Darul Ehsan, 47500 Petaling Jaya, Malaysia. ✉email: chojh@jbnu.ac.kr

framework was included, the ELM-based classification was used. Shallow and deep CNN models were proposed. The suggested system used the ELM classification technique and obtained 99.7% accuracy. These accuracy levels exceeded the previously reported accuracy level[4].Patients are more likely to die and have a higher chance of recurrence if they have bigger tumours, a higher FIGO stage, and lymph node metastases. The chance of recurrence increases depending on tumour size but becomes less significant as the individual lives longer without advancing the illness. Gynaecologists might discover this information useful for determining the tumour danger for cervical cancer patients and during patient assessment[5].Long non coding RNA gene signatures are used to predict the survival outcome of recurrence.cervical cancer recurrence genes are rendered using Internet of Thing. Recurrent neural network algorithm is used for recognising recurrence samples.High risk and low risk recurrence genes are identified through risk score.By using these values cervical cancer survival is predicted as Long term survival and short term survival[6].This study looks at ways to back up a doctor's judgement in prediction. Analysis Pathologic staging of profoundly invasive tumours and Pathologic T were distinct risk variables.Our data corroborate this since Pathologic Level and Pathologic T are significant and unique forecasting factors[7]. The ability of machine learning tools to identify essential characteristics from complicated data is crucial. Artificial neural network techniques, support vector machine models, and tree approaches are some technologies that are rapidly utilized to predict malignancy. Machine learning techniques that can be utilized to determine a significant level of validity are necessary to utilize these procedures frequently in medical practice, regardless of how illness evolves[8].A two-step medical incident forecasting framework is suggested in this research. GRU to a neural network is constructed in the second step to determine if the clinical occurrence series K-means approach is used in the first step[9].Multivariate Cox, principal component analysis, and K-means clustering analyses were once utilized to examine the factors' ability to forecast recurrence. Eight machine learning algorithms' prediction capabilities were contrasted with the logistic or Cox model. Using Machine learning algorithms, a unique online forecast calculator was created[10].

A good method to forecast a short lifespan in women with recurrent cervical carcinoma is to use a deep learning method to evaluate specific clinic laboratory criteria. This method may subsequently guide patients towards comfort treatment[11]. The choice of treatments for individuals with recurrent cervical carcinoma ought to take their prior radiation status and the location of the recurrence into account to enhance the chance of survival since these characteristics have a substantial impact on the efficacy of treatments[12]. Data on cancer treatments administered for a period exceeding three months following the initial finding were evaluated as recurrences, and this information was then cross-referenced with patient records. Women at FIGO level 1A1 are not required for specialized hospital monitoring, but FIGO stage 1B1 needs monitoring[13]. Surgery, radiation, chemotherapy, or a mixture of these techniques may be utilized for treating cervical carcinoma recurrences, depending on what type of therapy is being used, the site of the recurrence, the length of time since the last disease flare-up, the symptoms experienced by the individual, their level of effectiveness, and the likelihood that the particular therapy will be helpful[14].The most effective way for feature selection and balancing data are key components of the cervical cancer prognosis. In the suggested work, the issue of majority and minority class samples is overcome. The work aims to address the uneven distribution of information and validate risk factors for cervical carcinoma prognosis[15].Thus in an excisting method of survival **prediction** in cervical cancer the digital hunt of several database are explored through which 85 to 14,946 patients are analysed through different machine learning algorithm. Accuracy obtained is varied from 0.40 to 0.99 but there is no stable variation and not used the feature importance techniques are not used efficiently[16].This research aimed to predict the survival through miRNA by cox proportional and saying that it is tough to predict through the clinical features. And it distinguishes into high, moderate and low survival is mentioned and accuracy is about 90%[17].Cervical cancer has the possibility to reoccur for 26% of patients after surgery. Aim is to find the important prognostic factor of disease specific survival after recurrence. The accuracy were obtained as 81.8% for 5 years survival[18].Thus in a proposed approaches accuracy is low in prediction. And also efficient feature importance techniques are lagging. And also in existing technique[17]based on miRNA data,but practically getting miRNA data is difficult and highly expensive.

The main objective of this study is to find individuals with cervical cancer who are at greater risk of death from recurrence by predicting the survival. A novel approach in a proposed technique is Triangulating feature importance to find the important risk factors through which the treatment may vary to improve the survival outcome.The initial step of a research is to deal with the issue of choosing features by utilizing methods for preprocessing and industry-standard three feature selection methods, such as ReliefF, information gain, and FCBF algorithm for selecting the particular set of attributes. These characteristics then get used to successful training and testing of the ML techniques to determine which attribute selection method and the classification algorithm give satisfactory outcomes in terms of accuracy. Determine the dataset's weak characteristics that impact the outcome of the algorithms. At last, research is to identify the important risk factors through which the survival outcome of the patients improved.

By following the Introduction, Section "Literature survey" illustrates the Literature survey that is briefly explained about existing challenging methods. Section "Materials and methods" is the Materials and methods under which discussed about the Data set collection, Pre-processing techniques ,Three Feature importance techniques are compared in which highly performed model is selected through ranking the features by Orange data mining tool. Section "Proposed classifier models" is the Proposed six classifier models Logistic regression, Decision tree, Gradient boosting, Support vector machine, Random forest, and NaïveBayes are discussed detail in which Gradient cat boost algorithm has attained highest accuracy of 99.1%. Section "Results" is the analysis of Result. Section "Performance evaluation" is the conclusion stated.

Literature survey

Women are at risk of Dangerous reproductive illnesses globally, such as cervical carcinoma. The discovery of abnormal genes in cervical carcinoma samples from patients may open the door to creating more accurate indicators of prognosis and treatment strategies. We gathered 10 samples, seven from patients with cervical carcinoma and three belonging to healthy women who had the procedure, to determine the dysregulated genes (DEGs). The human ClariomD Affymetrix system was used to do microarray analysis on all RNA extracted from biopsies[19]. CCPM is the proposed technique of the study. The CCPM can assist people in the early detection of cervical carcinoma.Chi-square was an attributes extraction method employed by CCPM. Ten features are extracted.The mean equation is used for missing values[20]. Cox regression approaches to examine the connection between lncRNAs and CC recurrence. Using the strongly related long non-coding RNA RRS system was built. Bioinformatics analysis was carried out to evaluate the essential lncRNAs' possible contribution to cervical cancer recurrence[21]. An increase in the quantity of recurrence rate was calculated using Kaplan–Meier analysis of survival, and the Cox proportional hazards framework was employed to determine risk variables related to outcome. Results might provide crucial information for developing tailored therapies for the medical management of cervical carcinoma[22].The decision tree is an appropriate approach, according to the outcomes. It retains all the benefits of traditional decision trees. However, working with medical professionals to validate the device's construction is necessary for medical assessment. Perhaps by analyzing already-existing measurable data on an individual, we can come up with some conclusions that will help a doctor caring for the patient determine when to start the therapy[23]. In computerized methods, fuzzy-based logistic regression was particularly applied for selecting features for tumour genomic information, and a logistic regression technique was applied. The approach also used a unique approach called LASSO Logistic Regression, which was developed from logistic regression to determine the target genomes in categorizing tumour data and to acquire the gene pattern accurately[24]. Incorporating human papillomavirus DNA within a person's genome is crucial in developing cervical cancer. Results give knowledge about HPV integration, their oncogenic development in cervical carcinoma, the spatial connection of HPV, its numerous structural variants, and also its functional ramifications[25]. Recognizing risk factors for cervical carcinoma may help increase the number of patients who are cured as well as the proportion of female patients who survive cancer. Approaches of data mining have a huge potential for creating Apps with diagnostic and prognostic indicators that can aid in the right start of therapy for serious conditions. It is crucial to boost medical care standards and patient recovery rates using various methods for data mining. The results of the data minimization also kept the approaches' effectiveness from declining[26]. Discovering info regarding screening for cervical carcinoma in the shape of standards using a constrained-rule development method. On actual data, experiments are run to demonstrate the effects of limitations and the removal of pointless rules, with validation on several test sets. The regulations' impact on support, confidence, and lift is assessed in terms of medicine. The guidelines offer useful information that can assist the medical community in developing better cervical cancer screening programmes so that individuals can obtain appropriate care at an earlier stage when signs have not yet been shown[27]. Exclusively examined recurrent cervical carcinoma patients after CCRT from a single institution.After definitive CCRT, TFI had a prognostic impact on outcome and patient opinion in cases of recurrent cervical carcinoma[28].A two-step medical occurrence forecasting framework is suggested in this research. The standardized diagnostic and treatment model is mined using the bisecting K-means approach in the initial step to determine if the medical event series under test is standard. To build a forecasting model, we perform GRU on a neural network system in the second[9]. Given that the occurrences mentioned in the free text were discovered to be the most significant factor among the variables considered for the diagnostic prediction, it demonstrates the potential relevance of integrating free text in forecasting models using medical information. Examined methods to visualize electronic health record (EHR) information to forecast cervical carcinoma, a dangerous condition where early diagnosis improves the prognosis of therapy[29]. A unique and effective supplementary framework for the forecasting of cervical carcinoma by utilizing a sequence of gene components was presented after a thorough investigation of various cervical cancer treatment methods.The provided technique is more scalable and practical.The findings show that machine learning techniques have a limitless chance in the study of medicine[30]. Machine learning may improve cervical carcinoma forecasting. According to the research's findings, decision tree algorithms have the potential to be used to find the best-suited forecasters. Additionally, research appears that elevating individuals' socio cultural status as well as their wellness might contribute to avoiding the occurrence of cervical carcinoma[31]. Globally, metastatic Skin Cutaneous Melanoma (SKCM) has been linked to low survivability and high fatality rates. Genes ESM1, NFATC3, C7orf4, CDK14, ZNF827, and ZSWIM7 have been proposed to be new potential indicators for skin melanoma metastasis in this work[32]. This study's main goal is to rank the top-notch investigations into omics-related computational intelligence techniques. Based on omics data challenges are discussed in AI based on preprocessing, feature extraction, model validation and then its application.Models developed using deep learning are integrated using omics data while overcoming a number of difficulties[33].The goal of the initiative's is to decrease the makespan of the cloud and fog networks lung process activities while meeting security and time limitations. We discuss sophisticated security, where assaults using realtime malware have even been detected in fog and cloud networks. The models findings show that, while meeting the given criteria, BDFS executes better in pipeline execution than any current BIoMT design. All things considered, the BDFS algorithm included in the BIoMT architecture offers a secure and effective way to fuse data related with the process of lung cancer in fog cloud networks, advancing digital medical facilities in an universal setting[34].

Materials and methods

The background information regarding the research tools and methods is covered in the sections that follow.

Dataset collection

This historical investigation was conducted in a tertiary teaching medical centre. Between March 2012 and April 2018, thorough epidemiology, therapeutic, and information were acquired from the case reports. On March 1 2020, the investigation concluded. The dataset was built using an online database[30].Women with cervical carcinoma who had recurrences totalled 260 of a total of 4913 patients who are affected with cervical cancer. The dataset has been broadly classified into 27 features relevant to treatments and staging of a disease. Individuals receiving a particular form of recurrence therapy are categorized as pelvic cavity Recurrence, Recurrence beyond the pelvic cavity, and pelvic cavity Recurrence in both cases.

Average/most frequent imputation method

Women with cervical carcinoma who had been affected with recurrences is 260 of a total of 4913 patients who are affected with cervical cancer. Of the 260 parents with recurrence, the data set consists of many missing values. If missing values exceed 80%, they are deleted from the dataset. After deletion, we get the data of 160 patients, and 27 features are considered to proceed with the research. Out of 160 patients and 27 features, some values are missing; those missing values are imputed by the Average/Most frequent imputation method. Average and most frequent imputation substitutes the values not present in columns with the mean value for numerical attributes and the most commonly available values for categorical features. It is a straightforward replacement technique that substitutes the data at hand for the features instead of the information not present[31]. However, it does not portray a standard result. The following procedures allow inputting values absent from a data set employing the "Average" approach for continuous characteristics in a dataset. If a data collection has the continuous feature "α" but specific values are missing (NaN). Generate the average mean value by using the values that are present in the feature column "α" of the dataset. Values that are missing are to be replaced in attributes"α "with the Average value that is calculated.

$$Average\,\alpha = \frac{Sum\ of\ all\ values\ that\ are\ present\ in\ attribute\,\alpha}{no.\ of\ .\ values\ that\ are\ present\ in\ attribute\,\alpha} \tag{1}$$

Use the methods below to impute the values that are absent for categorical characteristics in a collection of data using the "Most Frequent" imputation method: Let's say that dataset with an attribute "β " that is categorical and has some values that aren't present (NaN) in it. Find the category that exists the most frequently in attribute "β ". Change the missing values in attribute "β" to the most common category found.

Mode_ β = attribute's most commonly used category β.

If an attribute β has a missing value (NaN), then:

$$\beta_i = Mode\beta \tag{2}$$

β_i—ith value missing in attribute β_i, $Mode\beta$—most frequently used categorised attribute.

Advanced paradigms for feature elicitation—exploring ReliefF algorithm, FCBF, and information gain

Following the pre-processing of information, feature selection is the most crucial step for research progress. Generally, Feature selection is an essential stage in creating a classification system. It functions by decreasing the features in a given dataset and selecting the most essential features required for an accurate prediction of a model[32]. In this research, we discussed four feature selection algorithms and how they worked in our given dataset and selected one of the best algorithms for our model predictions, which attains the highest accuracy.

The relief technique gives each data set's attributes weights, and the values assigned to them are continuously updated. Strong weight characteristics should be chosen, while low-weight ones should be ignored. The technique relief is looped through n random training samples (T), with no replacement choice chosen as 'n'.The algorithm for Relief feature selection is given below is the step-by-step procedure for relief algorithm for feature selection[33]

Load the Recurrence cervical cancer dataset into the Data Frame of pandas. It comprises 26 features; the last column should be a target variable. Define –Relief represents the Input of 'n' random samples. Initialize the weights (W) for all attributes (X) to zeros. To perform the relief F algorithm, Iterate over 'n' randomly chosen samples for training. The target is chosen randomly. Consider, for example, Target is (T) find the closest hit (I) and closest miss (M). For every attribute (A), determine the difference between the attribute values of T and I as well as the difference between T and M.Sum the difference between T and M and subtract the difference between T and I, both divided by m, to change the weights (W) of each Attribute (A). For each of the 'n' random samples for training, restart the procedure.Normalize the attribute values. Name the weight of each attribute. Greater weights denote qualities that are more crucial for separating classes.

The Best-ranked 5 features are selected from the given dataset of 26 attributes. Accuracy value has been increased gradually.

Information Gain is a crucial metric for categorizing characteristics. We can develop a metric representing new data about β provided by α that indicates the amount by which the overall entropy of β lowers since the entropy is a criterion of impurities in the training data set S.This is known as Information gain[34];and it is represented as,

It is formulated as,

$$Corr(\alpha, \beta) = \frac{P(\alpha U \beta)}{P(\alpha)P(\beta)} \tag{3}$$

It can aid in locating the most pertinent characteristics that considerably aid forecasting. The algorithm for information gain is described as the Load Recurrence cervical cancer dataset. It consists of 26 features. Predicting the survival, whether the patient dies or is alive, should be a target variable. Calculating the entropy of a target feature "STATUS" in a given dataset. Table 1 is the important feature selected by ReliefF technique. Calculate the target attributes entropy to understand how uncertain the group distribution is.

The mathematical representation of entropy is,

$$Entropy(\beta) - \sum (P(X_i) * \log 2(P(X_i))) \tag{4}$$

In which $P(X_i)$ is the percentage of cases in the dataset that correspond to class X_i.

To calculate information gain, determine the information gain for every attribute by considering its capacity to forecast the target value. For each feature, PR-OS, Date of first recurrence diagnosis, PR-PFS, Therapeutic effect, FIGO 2009 staging, etc. Information gain is calculated as,

$$Information\ Gain(N, \alpha) = Entropy\ (\beta) - \sum \left(\left(\frac{|N_{i|}}{|N|} \right) * Entropy(\beta | \alpha = W_i) \right) \tag{5}$$

$|N_{i|}$—no. of. instances in the dataset, *Feature value*$\alpha = W_i, |N|$—Total no. of. instances in a given recurrent cervical cancer dataset.

Information gain will calculate the most important and uncertain features when the target feature is fixed. The highest Information gain score feature is selected. To determine the target variable's entropy "Status" for multiple features present in the dataset (PR-OS, Date of first recurrence diagnosis, PR-PFS, Therapeutic effect, FIGO 2009 staging, etc.). The Information Gain of every attribute is then calculated by adding the numbers for each distinct attribute and deducting it from the entropy of the target variable. Attributes withthe highest information gain score are selected[35]. Best The 5 features are selected to forecast the prognostic fate of individuals with recurrent cervical carcinoma.

The Best-ranked 5 features are selected from the given dataset of 26 attributes using information gain as a feature selection algorithm is listed in Table 2. Scored features have been selected to increase the model's accuracy.

Fast correlation-based feature selection[36] chooses a feature's efficiency for classification. The attribute is chosen only if it is considered good, applicable to the field of study, and does not duplicate any additional applicable characteristics. The relationship between the two characteristics is calculated. Important attributes from the dataset are chosen to be strongly associated with every other type. Fast correlation-based attribute chosen techniques can solve dimensionality problems. The algorithm for the fast correlation-based attribute chosen is given as a technique for FCBF.Load Recurrence cervical cancer dataset. It consists of 26 features. Predicting the survival, that is, the patient's death or life, should be a target variable.Target variable to be fixed.The target variable is fixed as "status", as we predict the survival outcome after a disease recurrence.For every attribute, determine the Symmetrical Uncertainty (SU). As we have 26 features in our Pre-processed dataset (such as PR-OS, Date of first recurrence diagnosis, PR-PFS, Therapeutic effect, FIGO 2009 staging,disease progression once again, etc.), we must select the crucial attributes based on the score value.To calculate the symmetrical uncertainty for every 26 features and select the best features using the formula,

$$To\ calculate\ SU\left(X, Target\ variable\right)$$

S. no	Features	Score
1	Date of first recurrence	0.053
2	PR-OS (months)	0.059
3	PR-PFS (months)	0.093
4	Therapeutic effect	0.218
5	Disease progression	0.404

Table 1. Important features selected by ReliefF.

S. no	Features	Score
1	FIFO 2009 staging	0.082
2	Therapeutic effect	0.280
3	Disease progression once again	0.377
4	PR-OS (months)	0.409
5	PR-PFS (months)	0.427

Table 2. Important characteristics chosen by Information gain.

$$SU(\alpha, T) = 2 * \left(\frac{MI(\alpha, T)}{H(\alpha) + H(T)} \right) \quad (6)$$

here $MI(\alpha, T)$ is the Mutual Information between features α in a given dataset. Here α denotes the 26 features of recurrent cervical cancer, and the target variable is "status", $H(\alpha)$—Entropy of feature α, and $H(T)$—Entropy of target 'status'.

Set the threshold for FCBF. Select a threshold value for symmetrical uncertainty. Features with SU exceeding this cut-offare picked for additional investigation and will be considered meaningful. Choosing pertinent characteristics based on SU. To Calculate the Joint mutual information for a relevant feature, which is selected by the step set the threshold for FCBF. For any pertinent characteristic α,

$$JMI\left(\alpha, Target\prime T\prime\right) = MI(\alpha, T) - 1/((|F| - 1) * \sum(MI(\alpha, \alpha_i))) \quad (7)$$

$MI\ (\alpha, T)$ is the Mutual Information between attribute α and the target variable "status", $MI(\alpha, T)$—Mutual Information between α and the target variable 'status', $|F|$—Total no. of. Features, Summation of feature α and all other relevant feature α_i, $\sum(MI(\alpha, \alpha_i$ —Summation of feature α and all other relevant feature α_i, α_i—features excluded by α.

Choose the attribute with the highest JMI. The features with the highest JMI are the most useful for predicting the "Survival outcome after recurrence". The ranked 5 features are selected from the given dataset of 26 attributes using FCBF as a feature selection algorithm is listed in Table 3. Scored features have been selected to increase the model's accuracy.

Triangulating feature importance

By analyzing the above 3 algorithms, ReliefF, Information gain, and FCBF detail, we selected the essential features based on a score available.By selecting the essentialattribute from the given data, the accuracy and performance of a model is increased.A total of 8 features are selected from the 26 features of the given dataset by correlating all 3 algorithms.This is a Novel Approach to Feature Selection Algorithm. From the above-referred table, the essential features which are selected are mentioned below in Table 4,

Proposed classifier models

For predicting survival outcomes of recurrent cervical cancer patients, six classifiers are used in this paper. The classifier models used in this paper are Logistic regression, Decision tree, Gradient boosting, Support vector machine, Random forest, and NaïveBayes.

Epitome of prognostic advancement using gradient boosting algorithm

To predict the survival outcome in recurrent cervical carcinoma patients, a proposed method has been developed. For features selection techniques, the standard state-of-the-art algorithms are ReliefF, Information gain and FCBF techniques are used. Proposed a model by combining these three algorithms to select the most essential features to increase the model's precision. Figure 1 represents the working procedure of the proposed system.

S. no	Features	Score
1	Histological subtype	0.052
2	Definite adjuvant after primary treatment	0.065
3	Therapeutic effect	0.314
4	PR-OS (months)	0.375
5	Disease progression once again	0.657

Table 3. Important features selected by FCBF.

S. no	Features
1	Disease progression once again
2	PR-PFS (months)
3	PR-OS (months)
4	Therapeutic effect
5	FIGO 2009 staging
6	Definite adjuvant after primary treatment
7	Histological subtype
8	Date of first recurrence diagnosis

Table 4. Triangulating feature importance from ReliefF,IG,FCBF.

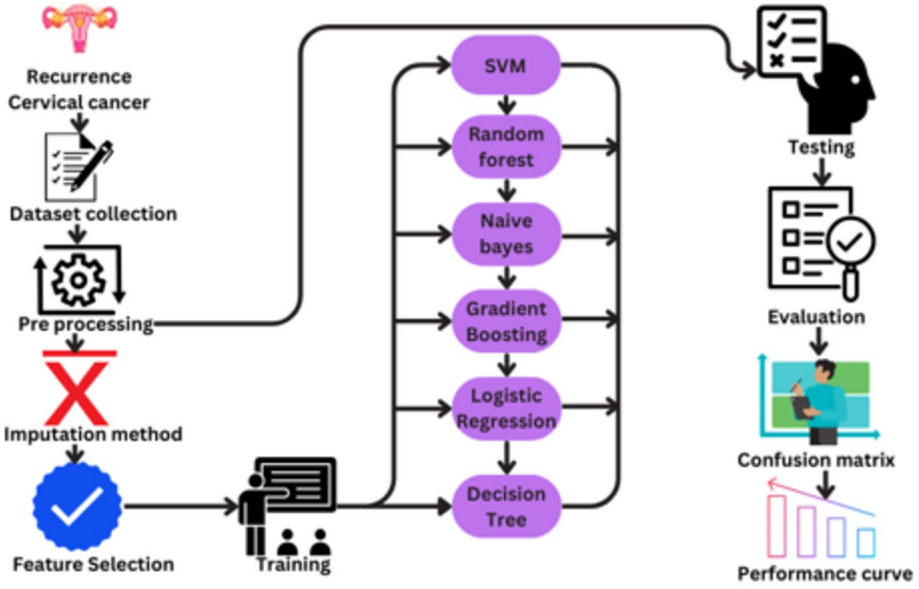

Figure 1. Working procedure of the proposed system.

Stratified tenfold cross-validation techniques are also used to assess the best models. Classification Accuracy, Area under the Receiver Operating Characteristic Curve, F1-Score, Precision, Recall and MCC are indicators of the model's performance[37]. The initial step is the pre-processing of the recurrent cervical cancer dataset, and the second step is choosing the methods for selecting features. The dataset consist of 26 features of a cervical cancer patients. we have to identify which all are the important factors to identify the survival outcome. If the important factors are identified then in a real time treatment procedure may varies to improve the survival of patients. Triangulating feature importance is a novel technique followed to find the important risk factors they are Disease progression once again, PR-PFS(months), PR-OS(months), Therapeutic effect, FIGO 2009 staging, Definite adjuvant after primary treatment, Histological subtype, Date of first recurrence diagnosis. These are the most dangerous features related to survival outcome of recurrence cervical cancer patients.The third step is a stratified tenfold cross-validation approach.The fourth step is comparing different machine learning classification algorithms and listed them as Support vector machine,Naïve Bayes,Gradient boosting,Random forest, Logistic regression and decision tress. These six different machine learning algorithms are trained and compared to describe which model would attain highest accuracy. By comparison Gradient boosting has highest accuracy. And finally, different metrics for performance measurement is calculated. The Gradient boosting algorithm consists of four methods: cat boost, extreme Gradient boosting random forest, sci-kit learn Gradient boosting and extreme gradient boosting. Among these, the cat boost algorithm performs well with an accuracy of 99.1%. Finally the important attributes selected by triangulating feature importance is accurate to detect the survival outcome of recurrence cervical cancer patients.

CAT boost

Extreme gradient Boost is a Gradient-boosting decision-tree modification created to be highly effective, adaptable, and portable for solving classification and regression issues. It is built on the enhancing method and uses the gradient descent algorithm to reduce losses while incorporating novel algorithms into boosting. After dividing the trees, it uses an enticing route to group the unconditional characteristics, which can be merged with the categorical attributes in the dataset to create a distinct novel characteristic that can subsequently be translated into a numerical value. Categorical characteristic non-combination is part of the initial partition. Integrating all of the categorical characteristics in the dataset results in a novel attribute, starting with the following split and moving forward. It is beneficial to use the Cat Boost methodology to forecast survival outcomes of patients with recurrent cervical carcinoma.The dataset used in this study contains the maximum of categorical features. Categorical characteristics are particularly intended to be handled effectively by catBoost. Datasets include intricate relationships between features and highly dimensional. Catboost technique will reduce the overfitting of the model and class imbalance challenge is handled effectively. It works well with the default parameters and not in a need of hyper parameter tuning. Resulting in its perfect fit for the complex and multidimensional medical information compared to the other algorithms.The category characteristic is converted to a numerical value using the target statistic approach. Ordered boosting represents the technique of category feature unbiased boosting. The processing of categorical features, regularisation, and loss function are combined in Cat Boost's goal value.

$$L(x, F(y)) - \textit{loss function for prediction error} \tag{8}$$

$$\Omega(\mathrm{F}) - \text{Regularisation term} \tag{9}$$

$$c(data) - categorical\ attributes \tag{10}$$

CatBoost uses ordered boosting to handle categorical attributes, which entails encoding the categorical values and discovering the best order. The method distributes scores to various stages of categorical traits in training based on how they affect the loss function. Calculating gradients for categorical features. Gradient and hessian for categorical attributes y_j and level M is calculatedas:

$$Gradient = \frac{\partial L}{\partial F(y_i)} * \frac{\partial F(y_i)}{\partial U_{ij}} \tag{11}$$

$$Hessian = \frac{\partial^2 L}{\partial F(y_i)^2} * \left(\frac{\partial F(y_i)}{\partial U_{ij}}\right)^2 + \frac{\partial F(y_i)}{\partial U_{ij}} * \frac{\partial^2 F(y_i)}{\partial U^2{}_{ij}} \tag{12}$$

here U_{ij} symbolizes the given weight to the M th level of the categorical attribute y_j

By iteratively locating divides that minimize the goal function, CatBoost builds decision trees. The split search minimizes the following for each node t.

$$obj = \frac{\left(\sum j \epsilon leaf\,(u) \frac{\partial L(x_i, F(y_i))}{\partial F(x_i)}\right)^2}{\left(\sum j \epsilon leaf\,(u) \frac{\partial^2 L(x_i, F(y_i))}{\partial F(x_i)^2}\right) + ^{\alpha}} \tag{13}$$

The above formulas give a fundamental knowledge of how CatBoost handles categorical information, adds Gradient boosting, and optimizes the ensemble of trees.

After collecting the recurrent cervical cancer dataset, data has been split into testing and training. 80% of the data has been trained,and 20% is tested. Data are trained in machine learning models and the catboost Gradient boosting algorithm. Figure 2 represents the flow of working procedure in catboost gradient boosting algorithm Compared to other algorithms, catboost Gradient boosting achieves the highest accuracy.

Structure of CatBOOST algorithm

The ability of CatBoost, a potent gradient-boosting algorithm, to handle category data naturally makes it stand out. It uses an ordered boosting approach that arranges categorical data according to how they relate to the target variable, allowing for wise tree-split decisions. Figure 3 represents the structure of catboost classifier. The approach successfully creates an ensemble of decision trees using gradient descent optimization to minimize a user-defined loss function. Overfitting may be avoided by using regularisation techniques like L2 regularisation on leaf values. The effect of each tree's contribution is adjusted according to its learning rate. Early halting is also supported by CatBoost for better generalization and effective hyperparameter adjustment. CatBoost excels at managing mixed data types thanks to parallelization and GPU support, making it a reliable option for many predictive modelling workloads.

Figure 2. Working procedure of catboost Gradient boosting classifier.

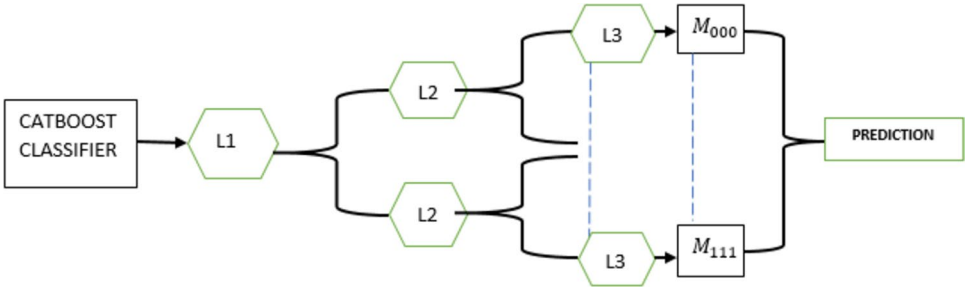

Figure 3. Catboost classifier structure.

Results

Various statistical operations such as average/most frequent imputation of missing values, min, max,mean,median, standard deviation,skewness, and kurtosis have been applied to the dataset.Women with cervical carcinoma who had been affected with recurrences is 260 of a total of 4913 patients who are affected with cervical cancer. Of the 260 parents with recurrence, the data set consists of many missing values. If missing values exceed 80%, they are deleted from the dataset. After deletion, we get the data of 160 patients, and 27 features are considered to proceed with the research. Out of 160 patients and 27 features, some values are missing, and those values are imputed by the Average/Most frequent imputation method. The graphical representation of information is known as data visualization. Figure 4 represents the dataset's numerical features min, mean,median,max, standard deviation, skewness and kurtosis, which condense and display a vast amount of data straightforwardly and understandably. It aids in the practical and clear communication of details and lets individuals realize the relevance of data. In the data set's histogram displays the regularity of occurrence of particular values that fall within the specified range of values and are arranged in regular, predetermined intervals.

We use three algorithms to select appropriate features to predict the survival outcome of recurrence cervical cancer patients: Information gain, ReliefF and Fast correlation-based feature selection.The most important attributes selected by information gain, ReliefF and fast correlation-based feature selection, and "triangulation of feature selection algorithm" are listed along with their scores in Tables 1, 2, and 3.Disease progression, Therapeutic effect, PR-PFS(months), PR-OS (months) and date of first recurrence diagnosis are the essential features for the Prediction of survival outcome, as per the analysis obtained from the ReliefF algorithm.Other feature selection algorithms,such as information gain and FCBF, also select these four features. Disease progression, therapeutic effect, PR-PFS (months) and PR-OS (months) are similar.The selected featuresare the most important for predicting the survival outcome in recurrent cervical cancer patients.Every feature selection algorithm commonly chooses a few features.

These eight features 'Date of first recurrence', 'PR-OS(months)', 'PR-PFS(months)', 'Therapeutic effect', 'Disease progression', 'FIFO 2009 staging', 'Histological subtype', 'Definite adjuvant after primary treatment' has been

Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis
Age of initial diagnosis	27	47.1813	46.5	74	9.7625	0.3468	-0.5683
DFS	3.1	27.3587	18.2	174.5	27.1989	2.4266	7.1614
PR-PFS (months)	0.0	17.9738	7.8	94	22.1754	1.7344	2.5317
PR-OS (months)	1.8	34.5381	26.6	149.1	24.9311	1.4048	2.447

Figure 4. Statistical representation of numerical data.

selected to train the model for the best performance measure. The algorithms' results for classification based on these chosen characteristics are excellent.

Figure 5 represents the graphical representation of feaures vs score for ReliefF,Information gain and FCBF. In Fig. 6 which represents the graphical representation of all the 8 features.

Figure 7 represents Important feature selection. Recurrent cervical cancer patients dataset consist of diverse clinical histories, treatment responses and influenced by complex interaction between data. Survival prediction involves the imbalance dataset. By boosting the validity, consistency, and comprehensibility of feature importance information, feature significance triangulation assists to overcome the challenges associated with forecasting survival outcomes in recurrent cervical cancer and finally facilitates improved clinical decision-making.After a feature selection technique, the most essential 8 features are selected,and then the model is trained using a six-classifier model.Gradient boosting is the best algorithm to predict from all six classifier models.The AUC of the model is 99%.

Figure 8 represents the Four different model parameters.The Gradient boosting algorithm consists of four methods: catboost,extreme Gradient boosting random forest, sci-kit learn Gradient boosting and extreme gradient boosting.Among these, the catboost algorithm performs well. The tabular column is mentioned below,

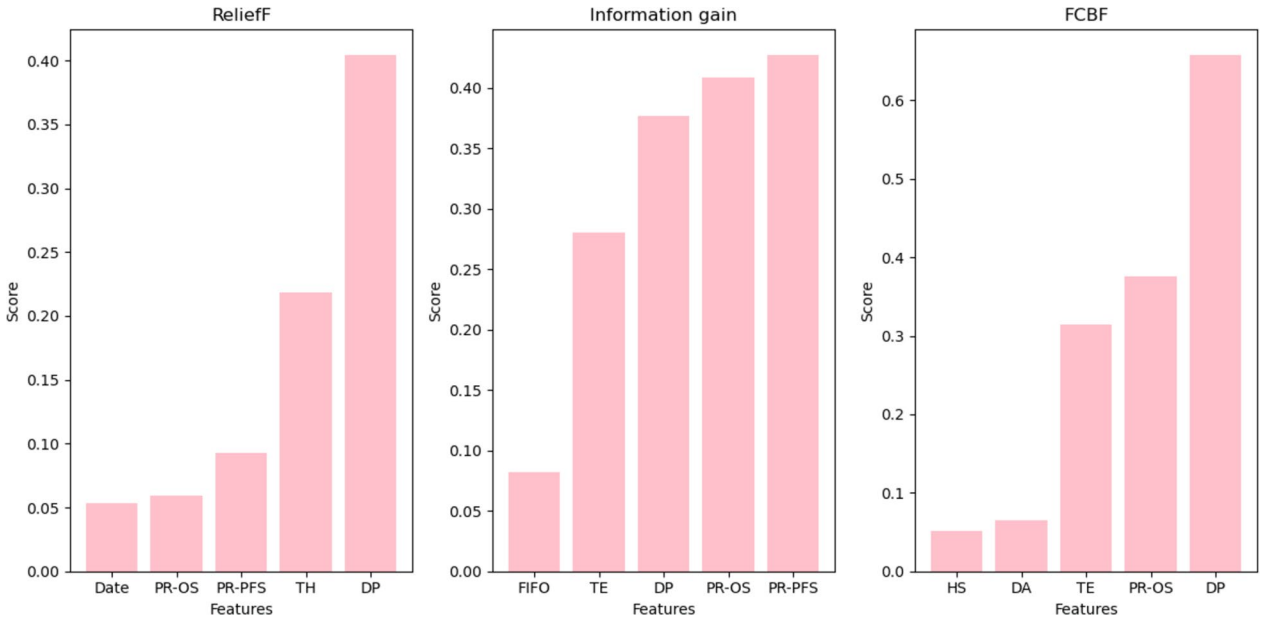

Figure 5. Feature score of ReliefF, Information gain, FCBF.

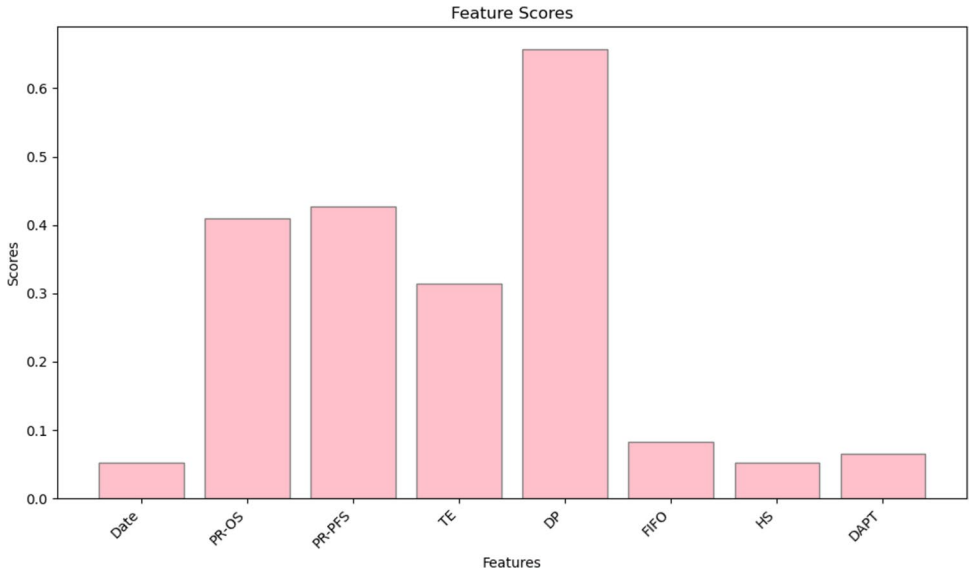

Figure 6. Feature scores of 8 Features.

		#	Info. gain	ReliefF	FCBF
1	C Definite adjuvant after primary treatment	2	0.061	-0.010	0.065
2	C Histological subtypes	3	0.047	-0.002	0.052
3	T Date of first recurrence diagnosis		0.039	0.076	0.000
4	N PR-PFS (months)		0.427	0.077	0.000
5	N PR-OS (months)		0.409	0.090	0.375
6	C FIGO 2009 staging	9	0.082	0.154	0.000
7	C Therapeutic effect	4	0.280	0.300	0.314
8	C Disease progression once again	2	0.377	0.312	0.657

Figure 7. Important feature selection.

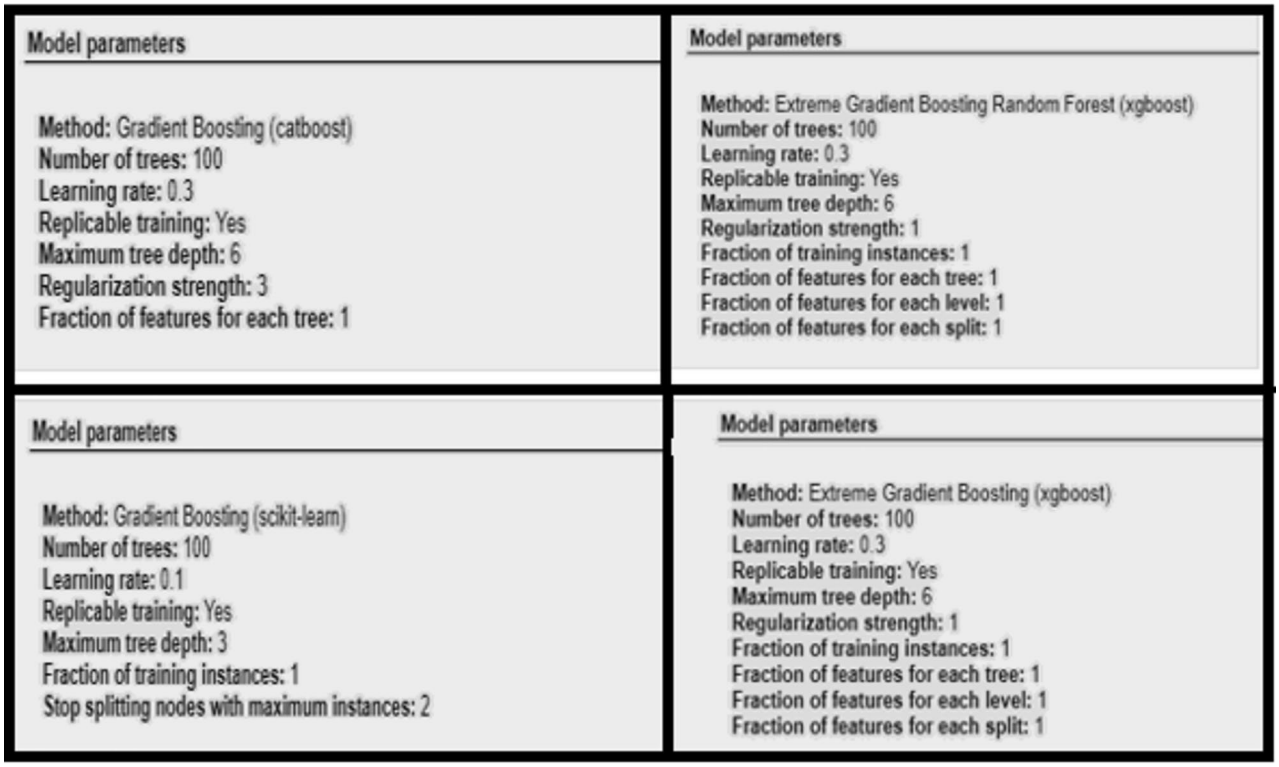

Figure 8. Four different model parameters.

Table 5 represents the Metrics evaluation of 4 Different gradient boosting algorithms Gradient boosting(Scikit learn), Extreme Gradient Boosting, Extreme Gradient boosting Random forest, Gradient boosting(catboost).The values in the table mentioned below are Area under curve, classification accuracy,F1,precision,Recall and MCC values of six different types of algorithms,andthe overall Gradient boosting algorithm(CAT BOOST) attains the highest accuracy value of 0.99 to predict survival of recurrence cervical cancer patients. Table 6 represents the metrics evaluation of different classifiers.

MODEL	AUC	CA	F1	Precision	Recall	MCC
GB (Scikit learn)	0.990	0.936	0.936	0.936	0.936	0.898
Extreme GB	0.990	0.944	0.944	0.944	0.944	0.888
Extreme GBRF	0.946	0.912	0.912	0.915	0.912	0.827
GB (Cat boost)	0.991	0.956	0.956	0.956	0.956	0.913

Table 5. Metrics evaluation of 4 different gradient boosting algorithms.

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.931	0.856	0.856	0.859	0.856	0.715
Gradient Boosting	0.991	0.956	0.956	0.957	0.956	0.913
Random forest	0.971	0.906	0.906	0.906	0.906	0.813
Decision Tree	0.954	0.944	0.944	0.944	0.944	0.888
SVM	0.485	0.487	0.487	0.487	0.487	−0.025
Naive Bayes	0.947	0.850	0.850	0.852	0.850	0.702

Table 6. Metrics evaluation of different classifiers.

Performance evaluation

Various metrics for assessing performance have been employed to determine the different classifier performances. Different performance metrics used in this research are Area Under the Receiver Operating Characteristic Curve, Classification Accuracy, F1-Score, Precision (Positive Predictive Value),Recall, and Matthews Correlation Coefficient. The confusion matrix is used for calculating these measures.

ROC analysis

We utilize several Evaluation Metrics in Machine learning to analyze and confirm how excellent our predicted machine-learning techniques are.Such an assessment measure is the AUC-ROC curve,which may be used to demonstrate the efficacy of an approach for classification.The receiver operating characteristic graph illustrates a Probability to show the degree to which an algorithm works at different threshold levels. The curve forms between the two variables listed below. The true positive and false positive rates.

$$True\ positive\ rate = Recall = \frac{TP}{TP + FN} \tag{14}$$

$$False\ positive\ rate = \frac{FP}{FP + TN} \tag{15}$$

AUC assesses the precision the binary classifier provides over different thresholds and provides an overall statistic in the ROC curve region. AUC ranges from 0 to 1; hence, a good model will have an AUC close to 1, indicating a significant level of separability.Hence, in our model, gradient boosting has a value of 0.992, which shows the model's high performance.The threshold gradient boosting value is 0.592.The ROC curve enables us to comprehend better the compromise between sensitivity and specificity at various threshold locations.Fig. 9 is the Mean TP and FP at the thresholdcalculated for the Gradient boosting cat boost algorithm.The graph is drawn between sensitivity and specificity.ROC Analysis curve has been drawn between merge prediction from folds, mean TP Rate, mean TP and FP at the threshold and the individual curves between all six different algorithms.

Figure 10 shows an ROC analysis curve of all six machine learning algorithms. The curve is drawn between the merge prediction from folds. False positive cost is maintained as 500, and False negative cost is maintained as 500. All the actions have been carried out through the orange data mining tool.The default threshold point is maintained as 0.5. The figure below mentioned the different colours of a classifier to be plotted.

Evaluation metrics

We can calculate various evaluation metrics from the confusion matrix to assess the classifier's performance in predicting survival status.

Accuracy: This measure represents how accurate the classifier predicts.

$$Accuracy = \frac{TN + TP}{TN + FP + FN + TP} \tag{16}$$

Precision (Positive Predictive Value): This metric measures the accuracy of the positive class predictions.In this case, it representspredicting "Death" correctly.

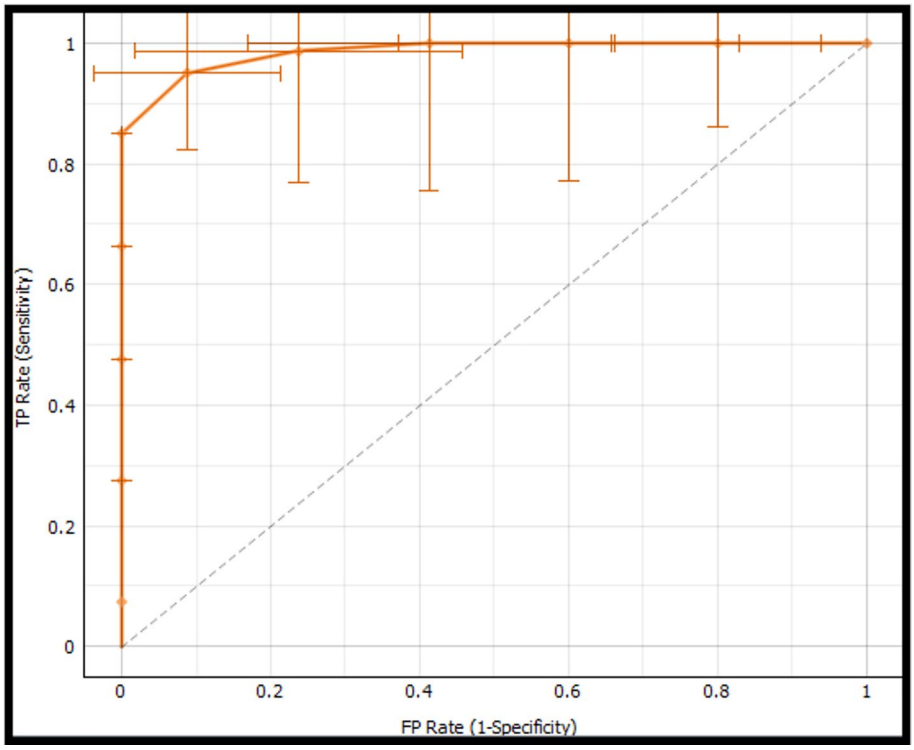

Figure 9. Mean TP and FP at the threshold for GB.

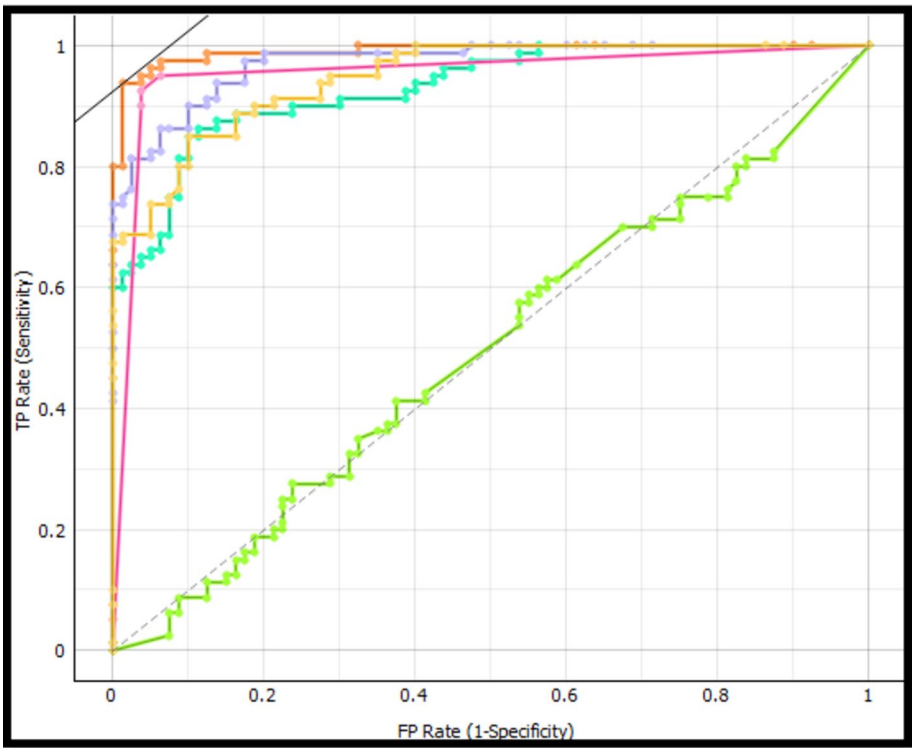

Figure 10. ROC Curves of six classifiers.

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

Recall (Sensitivity, True Positive Rate): This metric measures the classifier's sensitivity in correctly identifying positive class instances ."Death" is mentioned in this case.

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

Specificity (True Negative Rate): This metric measures the specificity of the classifier in correctly identifying negative class instances ("Alive" in this case)

$$Specifity = \frac{TN}{TN + FP} \tag{19}$$

F1-Score: This metric compromises accuracy and recall, offering a unified assessment of the classifier's efficiency.

$$F1 - Score = \frac{2 * (Precision * recall)}{Precision + Recall} \tag{20}$$

The confusion matrix and assessment metrics evaluate how well the classifier predicts whether an individual survives or passes away. For the classification of survival status, a high accuracy, precision, recall, specificity, and F1-score point to a well-performing classifier with predictive solid powers.

Performance curve

Concerning the classifier's threshold calculation, the Performance Curve displays curves for measuring the fraction of true positive data occurrences. The performance curve is of 3 types: lift curve, cumulative gain, and Precision recall curve.

Our classification output model is to predict the survival outcome of recurrent cervical cancer. Figure 11 represents the performance of the lift curve.Lift is defined in this case as the ratio of the percentage of patients diagnosed to the percentage of patients' prediction of survival outcome.

$$Lift = \frac{Percentage\ of\ patient\ diagnosed}{Percentage\ of\ patients\ predicted} \tag{21}$$

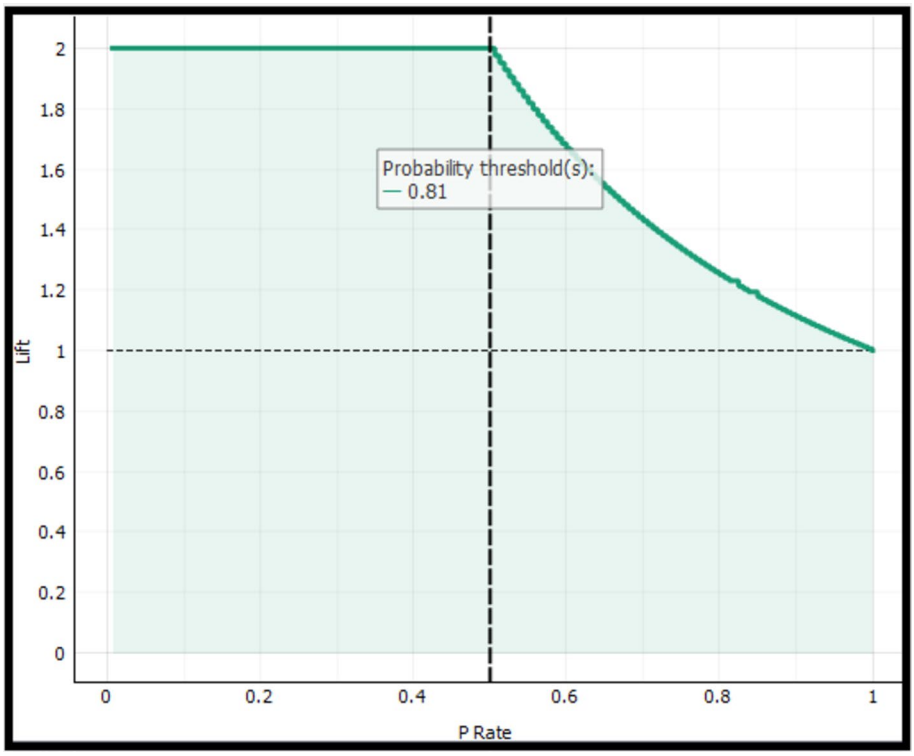

Figure 11. Performance of lift curve.

Cumulative gain curve, which shows the ratio of people affected with recurrent cervical cancer to predict the survival outcome of recurrence patients. Figure 12 represents the performance curve of cumulative gain. The area between the curve and base is greater, and the model performance is good.

The precision and recall curve is the ratio between precision and recall that is true positive in prediction and true positive in their respective class. The target value here is chosen as alive. Viewing the forecast accuracy of the probability classes in a graph using the calibration plot. The calibration curve compares the classifier forecasted probability to the actual group probability. The desired target value is to be mentioned. Status is our target variable. Status is subdivided into death and alive. Here, for the calibration plot, the target is mentioned as alive. There are different metrics of the classifiers calibration curve, classification accuracy,F1,Precision and recall, sensitivity and specifity, positive and negative predictive values and True and false positive rate. Alive is mentioned as a target variable for a calibration plot. The graph between threshold probability to classify as positive and CA is drawn in classification accuracy metrics. In an information window, the value is Threshold $p = 0.5$, Gradient boosting = 1.00. In metrics F1, the graph is drawn between F1 and threshold probability to classify as positive.

In sensitivity and specificity metrics, the graph is drawn between sensitivity and specificity and threshold probability to classify as positive. Figure 13 represents the precision curve of precision and recall.In a precision and recall, positive and negative predicted value, and true and false positive rate, the graph is drawn between their respective values and threshold probability to classify them as positive.

Calibration curve

A crucial component in evaluating the effectiveness of prediction models is calibration. Determine how closely the proposed model's projected survival probability matches the actual survival results using a calibration plot-designed mainly for survival forecasting. The process for making a calibration plot for forecasting survival is as follows. The calibration plot consists of a different metrics calibration curve,F1,Classification accuracy, sensitivity and specificity, precision and recall, True and false favourable rates, and pos and neg predictive value. Figure 14 below represents the calibration curve and a sigmoid calibration.The sigmoid calibration curve is A visual illustration of anticipated probabilities. Figure 15 illustrates the classification accuracy of the cat boost model.The equation below represents the true positive for forecasting probabilities,

$$P_{true}P_{pred} = \frac{\sum_{t=1}^{n} I\left(x \le P_{pred}(t) < y\right) z_t}{\sum_{t=1}^{n} I(x \le P_{pred}(t) < y)} \tag{22}$$

$P_{pred}(t)$—predicted probability for t th sample, z_t—Label of t th sample, n —no. of. samples, x, y—no. of. samples, boundaries of the curve.

According to the calibration curve analysis, the prediction model for survival prediction in recurrent cervical carcinoma showed relatively excellent calibration, with most of the information's points following a diagonal line. The efficiency and calibration of the model may be improved by using calibration procedures.

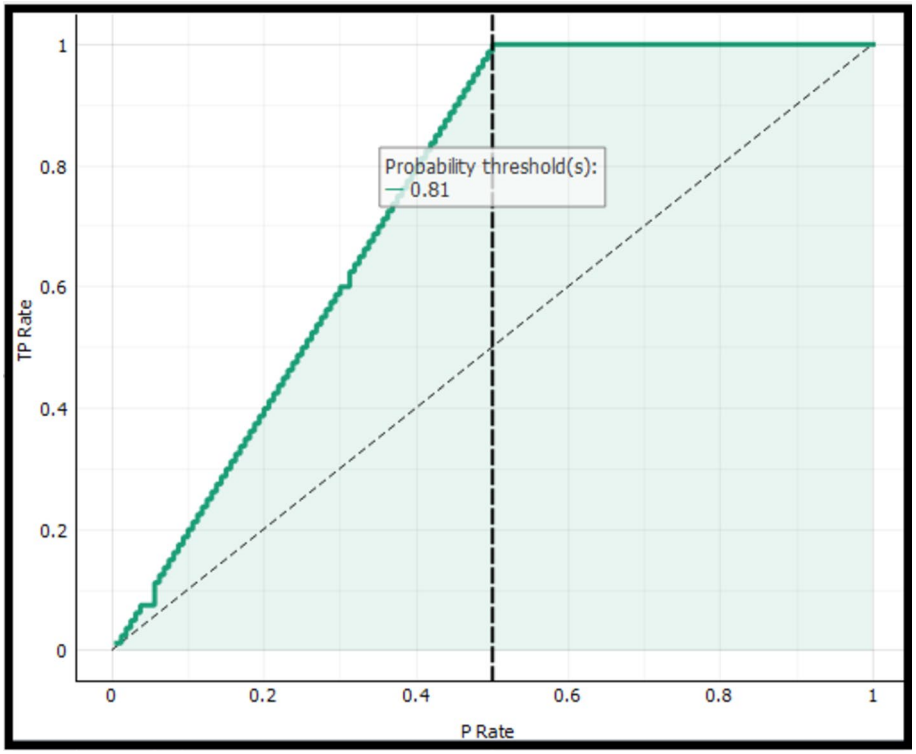

Figure 12. Performance curve of cumulative gain.

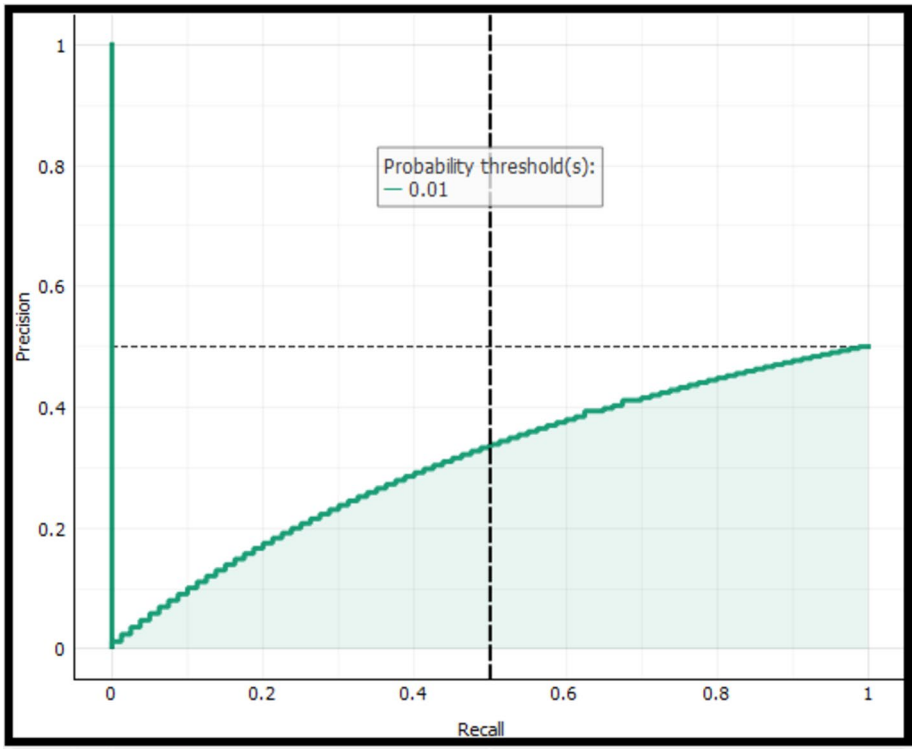

Figure 13. Performance curve of precision and recall.

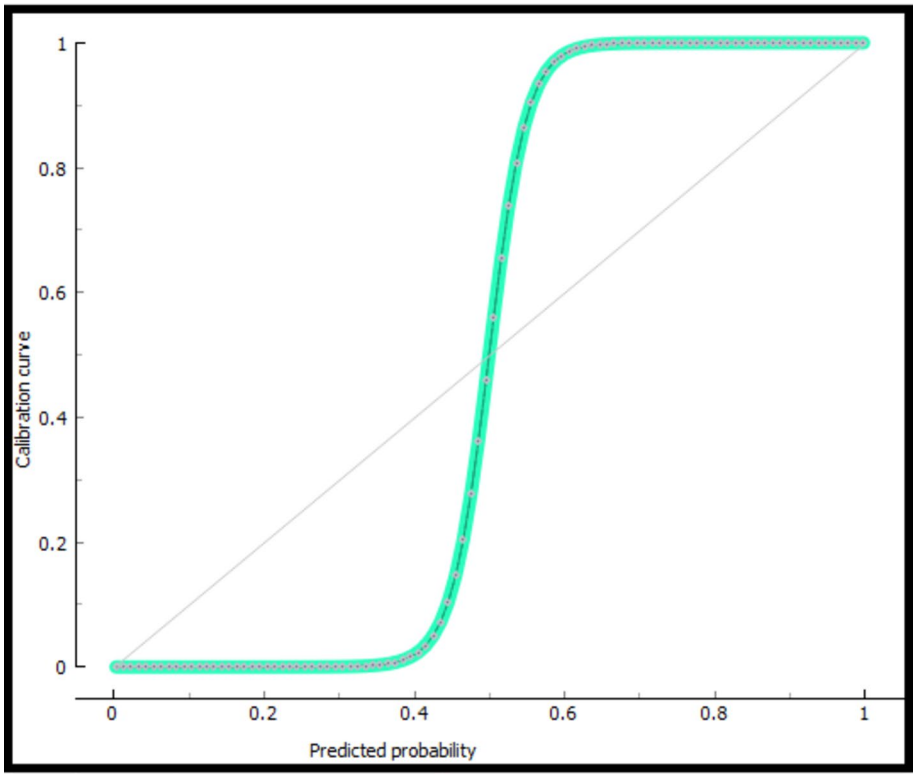

Figure 14. Calibration curve of a Cat boost model(Sigmoid calibration).

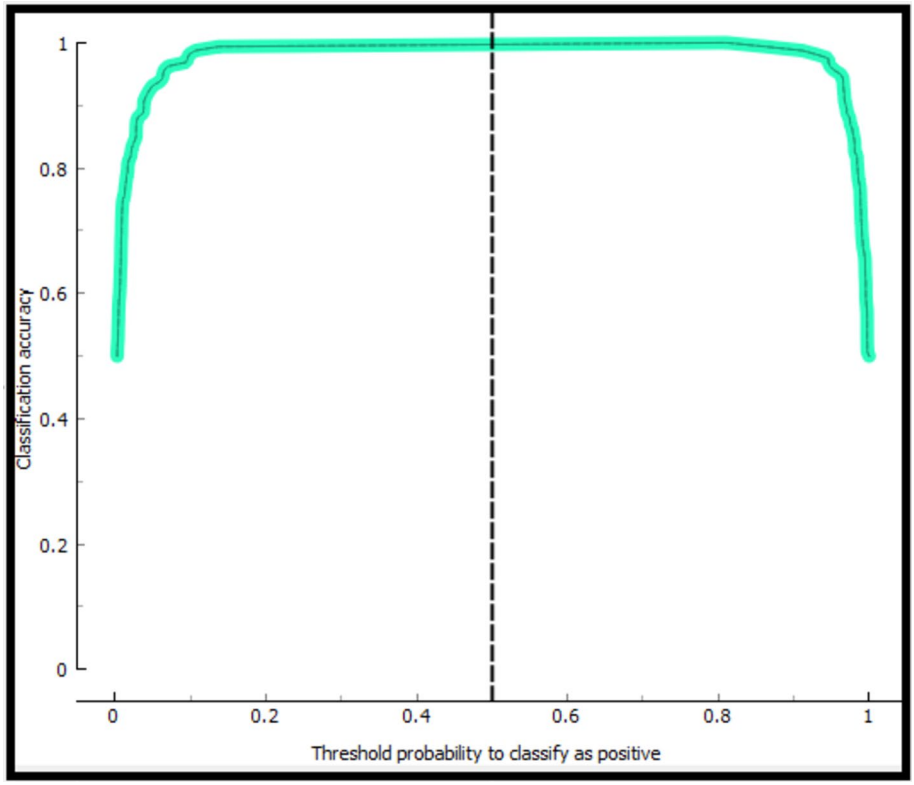

Figure 15. Classification accuracy of Catboost model.

Limitations of the proposed work is the threshold gradient boosting value is 0.592.It can be improved in future enhancement.The default threshold point is maintained as 0.5 for ROC curve. The area between the curve and base is greater, and the model performance is good in the performance curve of cumulative gain but still it can be improved.

Conclusion

The issue of predicting the recurrence of cervical carcinoma remains difficult.For predicting survival outcomes of recurrent cervical cancer patients, six classifiers are used in this paper. The classifier models used in this paper are Logistic regression,Decisiontree,gradientboosting,Support vector machine,Randomforest,and Naïve Bayes. Supervised classification tests have been carried out to assess the classifier's performance.Common feature selection techniques, including ReliefF, Information gain, and FCBF, are used to select the relevant attributes.The necessary features are chosen by correlating the proposed Feature selection algorithms, and then the outcome of the classifiers is assessed using a subset of the features. After a feature selection technique, the most important 8 features are selected, and then the model is trained using the six-classifier model. From all the six classifier models, gradient boosting is the best algorithm to predict. The AUC of the model is 99%. The gradient boosting algorithm consists of four methods: catboost, extreme gradient boosting random forest, sci-kit learn Gradient boosting and extreme gradient boosting. Among these catboost algorithm performs well. The classification accuracy of the catboost model is 95.6%. As a result, in recurrence cervical carcinoma, survival outcome prediction employs precision, recall, f-measure, AUC, MCC and accuracy.A stratified tenfold cross-validation technique is used for every classifier.

The theoretical application of a research is by using different algorithm and feature selection techniques the survival outcome is predicted for recurrence cervical cancer, This research contains the novel technique of triangulating feature importance using this the risk factors are analysed and steps are taken by clinicians to predict the survival outcome .As a practical applications it is to be implemented in healthcare centres to easily predict the recurrence cervical cancer which is very tough now a days. The practical advantage hospitals are lagging in storing the data of the patients and easily predict through already trained machine learning techniques. If this technique is trained and efficiently used in hospital it is a good welfare for recurrence cervical cancer patients. The data shown imply that the Gradient-boosting cat boost is an effective model. Numerous performance evaluation measures have been calculated to evaluate the algorithms' performances.Each investigation was carried out in an orange datamining toolkit.Orange is used as a Python library for manipulating data.Orange employs popular libraries that are open-source in Python for scientific computation, including numpy, scipy, and sci-kit-learn, and its GUI is built on the cross-platform. The proposed study predicts survival outcomes in recurrent cervical cancer based on the different features. Important features determine the survival of an individual, andthis offers a reliable computational justification for actual cervical tumour therapy. Limitation of a research to train the efficient machine learning model need more and more data but in a online repository data is limited. Hospital

need to be used cloud based storage system to use machine learning techniques for the welfare of patients.The future enhancement of a research is offering the prospect of a future treatment, the objective is to provide the individual an opportunity to live everyday life easily.Shortly, individuals who feel doubtful regarding their signs and symptoms can verify them through a web portal, and after being clinically evaluated, they can confirm their condition.

Data availability

The datasets used and/or analysed during the current study available from publicly accessible dataset from Xiaopei Chao et al. [31].

Received: 21 March 2024; Accepted: 12 July 2024
Published online: 27 August 2024

References

1. Geeitha, S. & Thangamani, M. A cognizant study of machine learning in predicting cervical cancer at various levels-a data mining concept. *Int. J. Emerg. Technol.* **11**(1), 23–28 (2020).
2. Geeitha, S. & Thangamani, M. A Hybrid Model for Mining and Classification of Gene Expression Pattern for Detecting Neuro-degenerative Disorder. In *Progress in Advanced Computing and Intelligent Engineering* (Springer, 2019).
3. Alam, T. M., Khan, M. M. A., Iqbal, M. A., Wahab, A. & Mushtaq, M. Cervical cancer prediction through different screening methods using data mining. *Int. J. Adv. Comput. Sci. Appl.* **10**(2), 9 (2019).
4. Ghoneim, A., Muhammad, G. & Hossain, M. S. Cervical cancer classification using convolutional neural networks and extreme learning machines. *Future Gener. Comput. Syst.* **102**, 643–649 (2020).
5. Chang, C., Chen, J., Chang, W.-Y. & Chiang, A. J. Tumor size has a time-varying effect on recurrence in cervical cancer. *J. Lower Genital Tract Dis.* **20**(4), 317–320 (2016).
6. Senthilkumar, G. *et al.* Incorporating artificial fish swarm in ensemble classification framework for recurrence prediction of cervical cancer. *IEEE Access* **9**, 83876 (2021).
7. Tseng, C.-J., Chi-Jie, L., Chang, C.-C. & Chen, G.-D. Application of machine learning to predict the recurrence-proneness for cervical cancer. *Neural Comput Appl* **24**, 1311 (2013).
8. Shaikh, F. J. & Rao, D. S. Predication of cancer disease using machine learning approach. *Mater. Today Proc.* **50**, 40–47 (2021).
9. Yan, Y., Zhao, K., Cao, J. & Ma, H. Prediction research of cervical cancer clinical events based on recurrent neural network. *Proc. Comput. Sci.* **183**, 221–229 (2021).
10. Guo, C. *et al.* Novel artificial intelligence machine learning approaches to precisely predict survival and site-specific recurrence in cervical cancer: A multi-institutional study. *Transl. Oncol.* **14**, 101032 (2021).
11. Matsuo, K. A pilot study in using deep learning to predict limited life expectancy in women with recurrent cervical cancer. *Am J Obst Gynecol* **217**, 703–705 (2017).
12. Chao, X. *et al.* Selection of treatment regimens for recurrent cervical cancer. *Front. Oncol.* **11**, 618485 (2021).
13 Taarnhoetal, G. A. Risk of recurrence, prognosis, and follow-up for danish women with cervical Cancer in 2005–2013: A National Cohort Study. *Cancer Am. Cancer Soc.* **124**, 943–951 (2017).
14. Peiretti, M. *et al.* Management of recurrent cervical cancer: A review of the literature. *Surg. Oncol.* **21**, e59–e66 (2012).
15 Geetha, S. & Thangamani, M. Integrating HSICBFO and FWSMOTE algorithm-prediction through risk factors in cervical cancer. *J. Amb. Intell. Humaniz. Comput.* **12**, 3213–3225 (2020).
16. Rahimi, M., Akbari, A., Asadi, F. & Emami, H. Cervical cancer survival prediction by machine learning algorithms: A systematic review. *BMC Cancer* **23**(1), 341. https://doi.org/10.1186/s12885-023-10808-3.PMID:37055741;PMCID:PMC10103471 (2023).
17. Ding, D. *et al.* Machine learning-based prediction of survival prognosis in cervical cancer. *BMC Bioinf.* **22**(1), 331. https://doi.org/10.1186/s12859-021-04261-x.PMID:34134623;PMCID:PMC8207793 (2021).
18. Cibula, D. *et al.* Post-recurrence survival in patients with cervical cancer. *Gynecol Oncol.* **164**(2), 362–369. https://doi.org/10.1016/j.ygyno.2021.12.018 (2022).
19 Annapurna, S. D. *et al.* Identification of differentially expressed genes in cervical cancer patients by comparative transcriptome analysis. *BioMed Res Int* **2021**, 1–13 (2021).
20. Ijaz, M. F., Attique, M. & Son, Y. Data-driven cervical cancer prediction model with outlier detection and over-sampling methods. *Sensors* **20**, 2809 (2020).
21 Zhang, Y. *et al.* Identification of potential prognostic long non-coding RNA biomarkers for predicting recurrence in patients with cervical cancer. *Cancer Manag. Res.* **12**, 719–730 (2020).
22. Li, J. *et al.* Cervical cancer prognosis and related risk factors for patients with cervical cancer: A long-term retrospective cohort study. *Sci. Rep. Nat. Portfolio* **12**, 13994 (2022).
23. Chang, C.-C., Cheng, S.-L., Chi-Jie, L. & Liao, K.-H. Prediction of recurrence in patients with cervical cancer using MARS and classification. *Int. J. Mach. Learn. Comput.* **3**, 75–78 (2013).
24. Nandagopal, V. *et al.* Feasible analysis of gene expression–a computationally based classification for breast cancer. *Measurement* **140**, 120–125 (2019).
25. Zhou, L. *et al.* Long-read sequencing unveils high-resolution HPV integration and its oncogenic progression in cervical cancer. *Nat. Commun.* **13**, 2563 (2022).
26 Mehmood, M., Rizwan, M., Gregus ml, M. & Abbas, S. Machine learning assisted cervical cancer detection. *Front. Public Health* **9**, 788376 (2021).
27. Lee, C. K. H. *et al.* Uncovering insights from healthcare archives to improve operations: An association analysis for cervical cancer screening. *Technol. Forecast. Soc. Change* **162**, 120375 (2021).
28. Kozaki, M. *et al.* Therapy-free interval has prognostic value in patients with recurrent cervical cancer treated with chemotherapy following definitive concurrent chemoradiotherapy. *ArchGynecol. Obstet.* **296**, 997–1003 (2017).
29. Weegar, R. & Sundstrom, K. Using machine learning for predicting cervical cancer from Swedish electronic health records by mining hierarchical representations. *PLOS ONE* **15**(8), e0237911 (2020).
30. Lu, J. *et al.* Machine learning for assisting cervical cancer diagnosis: An ensemble approach. *Future Gener. Comput. Syst.* **106**, 199–205 (2020).
31. Asadi, F., Salehnasab, C. & Ajori, L. Supervised algorithms of machine learning for the prediction of cervical cancer. *J. Biomed. Phys. Eng.* **10**(4), 513–522 (2020).
32. Bhalla, S. *et al.* Prediction and analysis of skin cancer progression using genomics profiles of patients. *Sci. Rep.* **9**, 15790 (2019).
33 Ali, A. M. & Mohammed, M. A. A comprehensive review of artificial intelligence approaches in omics data processing: evaluating progress and challenges. *Int. J. Math. Statist. Comput. Sci.* **2**, 114–167. https://doi.org/10.59543/ijmscs.v2i.8703 (2023).
34. Abdullah, L. *et al.* Secure blockchain assisted Internet of Medical Things architecture for data fusion enabled cancer workflow. *Internet Things* **24**, 100928. https://doi.org/10.1016/j.iot.2023.100928 (2023).

35. Chao, X. *et al.* Diagnostic strategies for recurrent cervical cancer: A cohort study. *Front. Oncol.* **10**, 591253. https://doi.org/10.3389/fonc.2020.591253 (2020).
36 Priya, S., Karthikeyan, N. K. & Palanikkumar, D. Pre screening of cervical cancer through gradient boosting ensemble learning method. *IASC* **35**(3), 2673–2685 (2023).
37. Tamane, S. *et al.* Applying gini importance and RFE methods for feature selection in shallow learning models for implementing effective intrusion detection system, ICAMIDA 2022. *ACSR* **105**, 214–234 (2023).
38. Urbanowicz, R. J. *et al.* Relief-based feature selection: Introduction and review. *J. Biomed. Inf.* **85**, 189–203 (2018).
39. S. Lei, A feature selection method based on information gain and genetic algorithm, IEEE (2012).
40. M. .Jeyanthi, C. Velayutham, Analysis of information gain ranking feature selection algorithm using uci machine learning datasets, *Proc. JETIR*. Vol 6, (2019).
41. N.Gopika, A. M. Kowshalaya, Correlation feature selection algorithm for machine learning, *Proc. International Conference on Communication and Electronics Systems* (ICCES 2018).
42 Hariprasad, R. *et al.* Design and development of an efficient risk prediction model for cervical cancer. *IEEE Access* **11**, 74290 (2023).

Acknowledgements

This work was supported the Korea Environmental Industry & Technology Institute (KEITI), with a grant funded by the Korea government, Ministry of Environment (The development of IoT-based technology for collecting and managing big data on environmental hazards and health effects), under Grant RE202101551.

Author contributions

S.G.—Writing—original draft, Validation, Methodology, Investigation, Data curation; K.R.—Writing—review & editing; J.C. -Methodology, Investigation, Data curation, Conceptualization; S.V.E.—Writing—original draft, Resources.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).
Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.
These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.
We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.
While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.
These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.
To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.
Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.
If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com