Efficient Machine Learning Technique for Prediction of Cervical Cancer

M. Lavanya[1] Dr. B. Rama Ganesh[2]
[1]M..Tech Student, [2]Professor,
Department of Computer Science and Engineering,
Sri Venkatesa Perumal College of Engineering & Technology, Puttur, AP.

Abstract:

Health professionals identify cervical cancer as a potentially fatal condition. Patients' valuable lives are at risk due to the difficult late diagnosis and treatment. The formal screening for disease identification suffers in both developed and developing countries because of high medical costs, a lack of healthcare facilities, social norms, and the late onset of symptoms. Early diagnosis of various different diseases, including cervical cancer, is possible because to machine intelligence. It is also cost-effective and computationally cheap. The reliance on a single classifier's prediction accuracy is the issue with the present machine classification approaches for illness identification. Due to bias, over-fitting, improper treatment of noisy data, and outliers, the use of a single classification method does not guarantee the best prediction. In order to provide an appropriate diagnosis that addresses the patient's symptoms or problems, this research study presents an ensemble classification method based on majority voting. A wide variety of classifiers, including Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), Extra Tree (EX), and Cat Boost (CB) classifiers, are experimented with in this work. The study shows an increase in prediction accuracy of 94%, which is significantly higher than the prediction accuracies of individual classification methods tested on the same benchmarked datasets. As a result, the suggested paradigm grants health professionals a second opinion to aid in the early diagnosis and prompt treatment of diseases.

Keywords: News, efficiently, performance, interest.

1. INTRODUCTION

Today's deadliest diseases include cancer. The disease exhibits specific aberrant behaviour in the afflicted cells. Normal tissues suffer damage from cancerous cells, which alters how well they normally operate. The likelihood of cancer spreading to different bodily areas is likewise very high. Consequently, in many situations, the inability to identify cancer at an early stage can result in mortality. If cancer is discovered early, there is typically a higher chance of survival depending on the type of cancer. One of the most prevalent cancers in women is cervical cancer, which is primarily brought on by HPV(Human Papilloma Virus). Numerous studies have demonstrated that cervical cancer therapy and recovery can be considerably impacted by early detection. Pap smear (Papanicolaou test), HPV DNA genotyping, HCII (Hybrid Capture II), hybrid capture, and Southern blot hybridization assay are frequently employed methods for the identification of cervical cancer. The most recent innovation in this area is the biosensor, which has enormous potential due to its low cost, quick findings, and simplicity of use. Each of these methods has its drawbacks. Significant drawbacks of the pap smear test include a high false negative rate, low sensitivity, and high cost. Southern blot hybridization assay testing is laborious and insensitive. The test for HPV DNA genotyping is time-consuming and rather expensive. One of the most cutting-edge methods for the early identification of cervical cancer is HCII hybrid capture. Its biggest drawback is that it can only identify 13 different HPV strains. Several operational limitations, including the inability to reuse electrodes, operational instability, and short lifetime, affect the performance of biomedical sensors. These shortcomings of the current approaches highlight the demand for more effective methods for cervical cancer early detection.

"Computational approaches employing experience to improve performance or to produce accurate predictions" is how machine learning is defined in Reference. A wide range of industries, including the personal, financial, business, government, military, and even space science, can benefit from machine learning approaches. The medical field has effectively employed machine learning techniques. Cervical cancer is not an exception when it comes to the appeal of various machine learning technologies. The methods currently in use for cervical cancer detection mainly concentrate on a single classifier. This article suggests using ensemble techniques for cervical cancer diagnosis from two publically available datasets because they can typically give better outcomes in specific problems. Support Vector Machine, Random Forest, and Logistic Regression, Extra Tree, and Cat Boost classifiers are among the classifiers used in the work strategy. The target attribute is categorised using the majority voting method. The remainder of the essay is structured as follows. A critical evaluation of recent efforts on cervical cancer detection is presented in Section 2. Section 3 goes into further detail about the research process. Section 4 provides the findings, and Section 5 brings the study to a close.

2. RELATED WORKS

Application of cyclodextrins in cancer treatment:

One of the main causes of death is cancer. Chemotherapy is a popular therapeutic approach that uses a combination of drugs to either eliminate or slow the growth of cancer cells. The majority of cytotoxic chemotherapy drugs, however, are insoluble in water, which makes formulation difficult. Utilizing cyclodextrins (CDs), which are frequently utilised to increase drug solubility, bioavailability, stability, and safety by creating non-covalent inclusion complexes, is an intriguing strategy. This review's objective is to explain how CDs are utilised in various cancer treatment plans. Particularly intriguing is the fact that CDs have been shown to have an anticancer effect both in vitro and in vivo. The use of CDs as anticancer agents is examined, as is the most plausible method for preventing the multiplication of cancer cells. Future applications as well as CDs/anti-neoplastic-drug combinations with improved solubility, stability, and anti-cancer activity are described. The application and advantages of CDs in various drug delivery systems, including liposomes, conjugates, nanoparticles, and siRNA carriers for cancer treatment, are covered in this work.

The Warburg effect and cancer's telltale signs Whether cancer is a single illness or a group of quite dissimilar illnesses has been a topic of discussion for a very long time. The goal of this research is to provide compelling evidence that the Warburg effect is to blame for the majority, if not all, of the characteristics of cancer. The metabolic impairment of oxidative phosphorylation results in a decrease in ATP concentration. Massive glucose uptake, anaerobic glycolysis, and an increase in biosynthesis through the Pentose Phosphate Pathway, which results in improved cell division and local pressure, are used to make up for the lower energy yield. The tumor's fractal structure, fibroblast collagen release, and metastatic spread are all caused by this increased pressure. The significant release of lactic acid raises extracellular acidity and stimulates the immune system. Reduced oxidative phosphorylation affects CO_2 levels both within and outside the cell, increasing internal alkalosis and causing extracellular acidosis, which is mediated by at least two carbonic anhydrase isoforms linked to cancer.

Recent developments in cancer early diagnosis and detection: The role of aptasensors based on nucleic acids: People all across the world are affected by the complicated illness known as cancer. Early and sensitive cancer detection, together with cancer biomarkers, can improve outcomes and aid in the research of cancer progression and the creation of effective treatment methods. The recognition and identification of cancer cells and biomarkers have shown significant promise when using aptasensors, one of the several detection methods currently accessible. In this review, we address recent advances in aptamer-based platforms for cancer exosome detection as well as recent advances in optical and micro-fluidic aptasensors for early cancer diagnosis.

First, an overview of the most recent optical aptasensors for identifying and detecting cancer cells is given, followed by an introduction to the aptamer fabrication method and characterizations. The methods for sensitive biomarker and cancer cell detection based on microfluidic aptamers are discussed in the following section. Then, new fluorescence, colorimetric, and other techniques are used to further develop the detection of cancer exosomes. Finally, we share our opinions on aptasensors' potential for use in cancer diagnosis in the future.

Human papillomavirus e6 biosensing: Current developments in cervical cancer early detection methods: One of the biggest problems facing the medical system now is the early detection of cancer. Medical disagreements between specialised physicians and researchers on early screening for cervical cancer, the interaction of risk factors, portable devices, and a speedy and free labelling system have been a major topic of concern. The electrical biosensing-based system shown improved specificity and selectivity due to the hybridization of DNA duplexes between analyte target and DNA probes. Numerous researchers are interested in the electrical DNA sensor for cervix cancer because of its high performance, simplicity, speed, and potential for miniaturisation. The development of HPV E6 genobiosensing and its potential effects on cervical cancer early detection technologies are examined in this review.

Biomarkers for detecting cervical cancer: Cervical cancer incidence has been markedly reduced as a result of population-wide cytological screening programmes using the Pap test in developed countries. Pap-stained cytological screening programmes have a number of disadvantages despite their clear success. First, a proportion of ambiguous or slightly abnormal test results require pricey follow-up treatments, such as additional testing or direct colposcopy and biopsy, because they may conceal a small percentage of high grade lesions that need urgent treatment. The workup of cytological tests that are only slightly abnormal or unclear accounts for a sizeable amount of the entire cost of cervical cancer screening. Improved sample triage could result in significant cost savings. Cervical cancer is brought on by recurrent infections with the oncogenic human papillomavirus (HPV). Although HPV infection is required, it is not enough to cause cancer. Less than 10% of acute HPV infections proceed to high grade lesions or invasive malignancies, while more than 90% cause low grade precursor lesions that spontaneously disappear within a few months. Progression is characterised by the uncontrolled production of the viral oncogenes E6 and E7 in infected basal and parabasal cells. The detection of lesions with a high risk of progression in both primary screening and triage settings is anticipated to be improved by the use of novel biomarkers that enable the monitoring of these crucial molecular processes in histological or cytological specimens. In this review, we will discuss potential biomarkers for cervical cancer screening with a focus on the strength of the clinical data that underpins their usage as cutting-edge indicators in sophisticated cervical cancer screening programmes.

2. Methodology:

Proposed system:

In the suggested approach, an ensemble classification method based on majority voting is used to make a precise diagnosis that takes into account the patient's symptoms or circumstances. Cat Boost (CB), Support Vector Machine (SVM), Random Forest (RF), Extra Tree (EX), and Logistic Regression (LR) classifiers are only a few of the many classifiers that are tested in this study. The study shows an increase in prediction accuracy of 94%, which is significantly higher than the prediction accuracies of individual classification methods tested on the same benchmarked datasets.

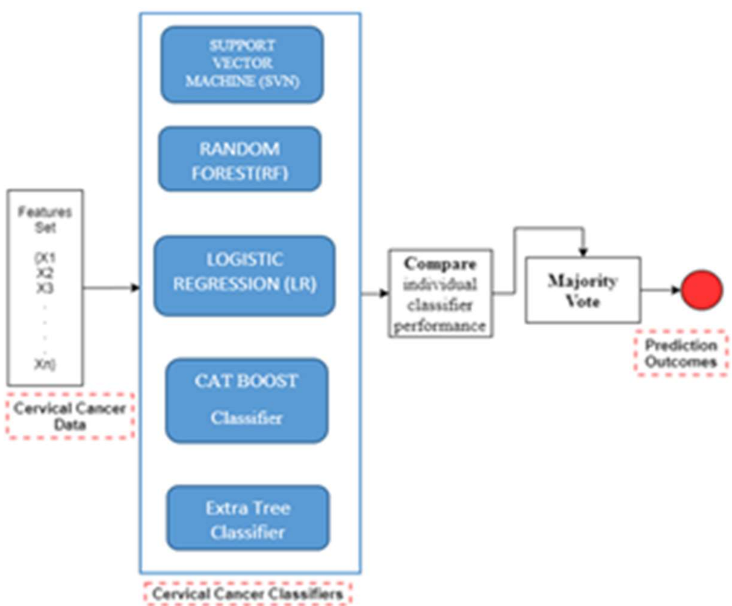


Figure 1: Block diagram of proposed method

3. IMPLEMENTATION

The project was carried out using the algorithm given below.

Random Forest:

A random forest is a machine learning method for tackling classification and regression issues. It makes use of ensemble learning, a method for solving complicated issues by combining a number of classifiers. In a random forest algorithm, there are many different decision trees. The random forest algorithm creates a "forest" that is trained via bagging or bootstrap aggregation. The accuracy of machine learning algorithms is improved by the ensemble meta-algorithm known as bagging. Based on the predictions of the decision trees, the (random forest) algorithm determines the result. It makes predictions by averaging or averaging out the results from different trees. The accuracy of the result grows as the number of trees increases. The decision tree algorithm's shortcomings are eliminated with a random forest. It improves precision and decreases dataset overfitting. Without requiring numerous configurations in packages, it generates forecasts (like Scikit-learn).

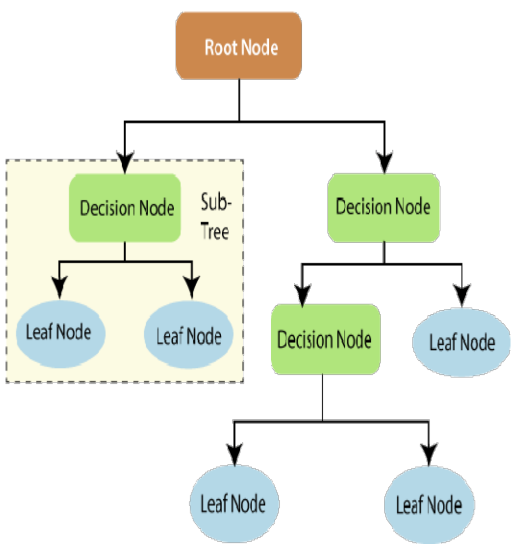
Characteristics of a Random Forest Algorithm:

• It offers a useful method of handling missing data; • It is more accurate than the decision tree algorithm.
• Without hyper-parameter adjustment, it can generate a fair prediction.
• It addresses the issue of decision trees' overfitting.
• At the node's splitting point in every random forest tree, a subset of features is chosen at random.

A random forest algorithm's building components are decision trees. A decision support method that has a tree-like structure is called a decision tree. We will learn about decision trees and how random forest methods function.

Decision nodes, leaf nodes, and a root node are the three parts of a decision tree. A training dataset is divided into branches by a decision tree algorithm, which then separates those branches further. This process keeps going until a leaf node is reached. It is impossible to further separate the leaf node.

The attributes that are utilised to forecast the outcome are represented by the nodes in the decision tree. Links to the leaves are provided by decision nodes. The three different sorts of nodes in a decision tree are depicted in the diagram below.

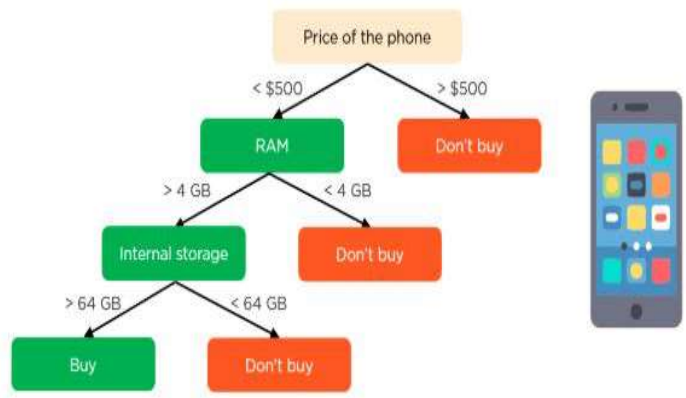


Information theory can shed further light on decision trees' operation. The foundation of a decision tree is information gain and entropy. An review of these key ideas will help us better comprehend the construction of decision trees.
Uncertainty can be measured using entropy. Given a set of independent variables, information gain measures the degree to which uncertainty in the target variable is minimised.

Using independent variables (features) to learn more about a target variable is the idea behind information gain (class). The information gain is calculated using the entropy of the target variable (Y) and the conditional entropy of Y (given X). In this instance, the entropy of Y is reduced by the conditional entropy.

Decision trees are trained using information acquisition. It helps to make these trees' uneasiness lessened. A significant information gain denotes the removal of a large amount of uncertainty (information entropy). Splitting branches, a crucial step in the creation of decision trees, depends on entropy and information gain. Take a look at a straightforward decision tree example. Let's say we want to forecast whether or not a customer will buy a mobile phone. His decision is based on the phone's features. A decision tree diagram can be used to display this study.

The above-mentioned phone features are represented by the decision's root node and decision nodes. The leaf node represents the outcome, whether a purchase is made or not. The pricing, internal storage, and Random Access Memory are the primary criteria for selection (RAM). The following is how the decision tree will look.



Applying decision trees in random forest:

The fundamental distinction between the random forest method and the decision tree algorithm is that the latter randomly selects the root nodes and groups the nodes. To produce the necessary forecast, the random forest uses the bagging approach.vBagging entails using multiple samples of data (training data) as opposed to a single sample. Predictions are made using features and observations from a training dataset. Depending on the training data that the random forest algorithm receives, the decision trees generate a variety of results. The highest ranking of these outputs will be chosen as the final output. The operation of random forests can still be explained using our initial example. The random forest will have numerous decision trees rather than just one. Assume that there are only four decision trees in all. In this instance, four root nodes will be created from the training data, which consists of the phone's observations and features.

The four features that potentially affect the customer's choice are represented by the root nodes (price, internal storage, camera, and RAM). By randomly choosing features, the random forest will divide the nodes. The results of the four trees will be used to choose the final prediction.
The majority of decision trees will select the final result. The final forecast will be buying if three trees predict buying and one tree predicts not buying. In this instance, it is anticipated that the client will purchase the phone.

1. LOGISTIC REGRESSION

Logistic regression was used in the biological sciences at the turn of the twentieth century. Later, it was used in many different social science contexts. Logistic regression is used when the dependent variable (target) is categorical. For instance, Whether a tumour is malignant (1) or not (0) To assess whether an email is spam (1) or not (0) (0)

Think about the following example: Finding out if an email is spam is necessary. In order to accomplish classification if we use linear regression to tackle this issue, a threshold must be established. The data point will be classified as nonmalignant if the actual class is malignant, the projected continuous value is 0.4, and the threshold value is 0.5. This classification can have significant consequences in real time.

This illustration demonstrates why categorization issues do not lend themselves to linear regression. Logistic regression comes into play because linear regression has no bounds. Their value is exactly in the range of 0 and 1.

Goals and illustrations of logistic regression

One of the most popular machine learning techniques for binary classification problems, or issues with two class values, includes logistic regression. These include predictions like "this or that," "yes or no," and "A or B."

The goal of logistic regression is to estimate event probabilities, including establishing a correlation between feature frequencies and specific outcome probabilities.

Predicting whether a student will pass or fail an exam using the number of study hours as a feature and pass or fail values for the response variables is one example of this.

Organizations can improve their business strategies to meet their objectives by using the insights from logistic regression outputs, such as decreasing costs or losses and boosting ROI in marketing efforts, as examples.

An online retailer that distributes pricey promotional offers to consumers would like to know whether or not a specific customer will take advantage of the offers. They can inquire as to whether the customer will "reply" or "not respond," for instance. Propensity to respond modelling is the term used for this in marketing.

Similar to this, a credit card company will create a model to determine whether or not to issue a customer with a credit card and will attempt to predict whether the customer will default on the credit card based on factors like annual income, monthly credit card payments, and the number of defaults. This is referred to as default propensity modelling in the banking industry.

Uses of logistic regression:

Online advertising has benefited greatly from the growing popularity of logistic regression since it allows advertisers to forecast the likelihood, expressed as a percentage, of individual website users clicking on particular advertisements.

Additionally, logistic regression can be utilised in:

- Medical care to detect illness risk factors and devise preventive strategies.
- Apps for weather forecasting that can forecast weather and snowfall.
- Voting apps that predict if voters will support a specific candidate.

Insurance that calculates the likelihood that a policyholder will pass away before the policy's term has run out based on factors including gender, age, and physical examination.

Using annual income, prior defaults, and historical debts, banking can forecast whether a loan applicant would default or not.

Logistic regression vs. linear regression:

The primary distinction between logistic regression and linear regression is that the output of logistic regression is constant whereas the output of linear regression is continuous.

A dependent variable's result in logistic regression has a constrained range of possible values. However, the output of a linear regression is continuous, which means that there are an endless number of possible values for it.

When the response variable is categorical, such as yes/no, true/false, and pass/fail, logistic regression is utilised. When the response variable is continuous, such hours worked, height, and weight, linear regression is utilised.

Logistic regression and linear regression, for instance, can predict various outcomes depending on the information about the amount of time a student spent studying and the results of their exams.

Predictions using logistic regression are limited to certain values or categories. Logistic regression can therefore forecast whether a pupil will succeed or fail. The student's test score on a scale of 0 to 100 can be predicted by linear regression because its predictions are continuous, such as numbers in a range.

CatBoost:

An open source decision tree gradient boosting library with high performance is called CatBoost. A gradient boosting method for decision trees is called CatBoost. It was developed by Yandex researchers and engineers and is used by Yandex and other companies including CERN, Cloudflare, and Careem taxi for search, recommendation systems, personal assistants, self-driving cars, weather forecasting, and a range of other tasks. It is open-source and available to everyone. With just a little over a year in existence, Catboost, the new kid on the block, is already posing a threat to XGBoost and LightGBM. On the benchmark, Catboost earns the best results, which is fantastic.

However, this improvement becomes considerable and obvious when you look at datasets where categorical variables are heavily weighted.

Figure 2: Features of CatBoost Algorithm

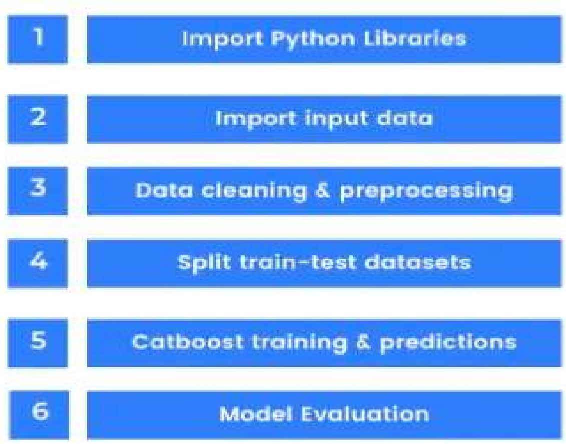


Figure-3: The sequence of steps in methodology

Implementation:

According to the Yandex benchmark, prediction time is 13–16 times faster than the other libraries, even if training time can be greater than with other GBDT implementations. For beginners who want a plug-and-play model to start using tree ensembles or Kaggle contests, Catboost's default parameters are a better starting point than in other GBDT techniques. The object importance, feature interactions, and snapshot support are some of Catboost's other important innovations. Catboost offers ranking out of the box in addition to classification and regression.

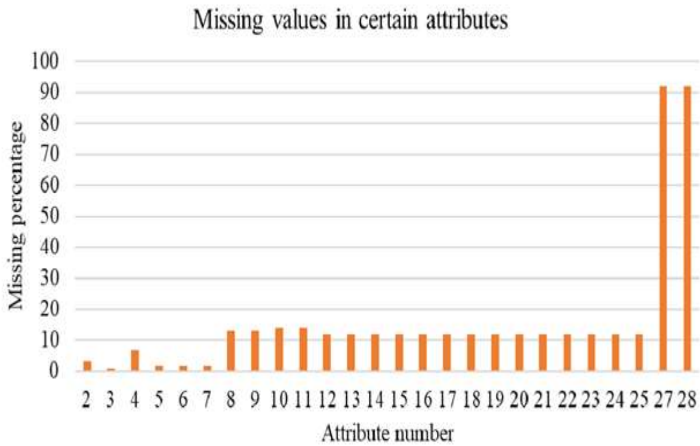


Figure-4 percentage of missing data in specific attributes of our dataset

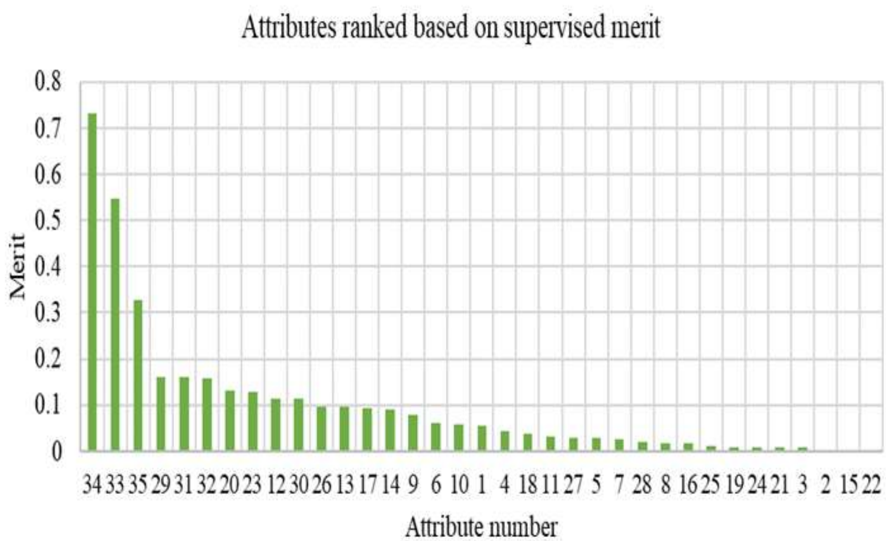


Figure-:5 Ranking attributes based on correlated and supervise ranking

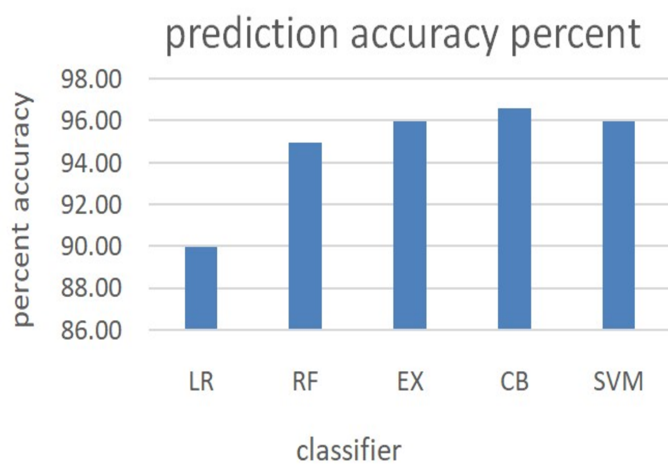


Figure-6: output performance of cat boost classifier prediction accuracy

Measurements:

*Precision

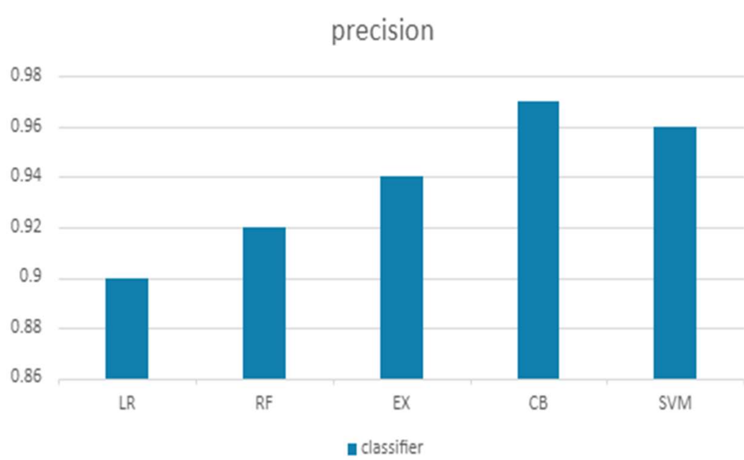


Figure-7 performance measure of terms in precision

*Recall

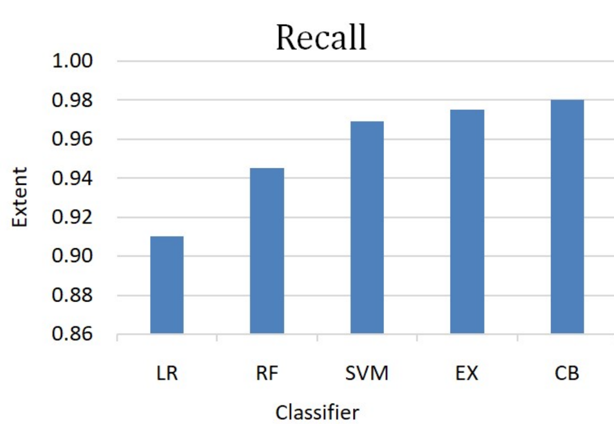


Figure-8 performance measure of terms in recall

*F-measure

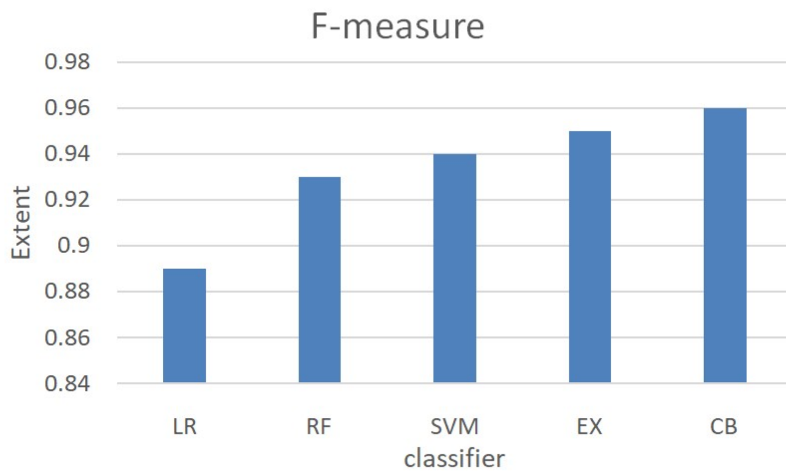


Figure-9 performance evaluation in context of F-measure

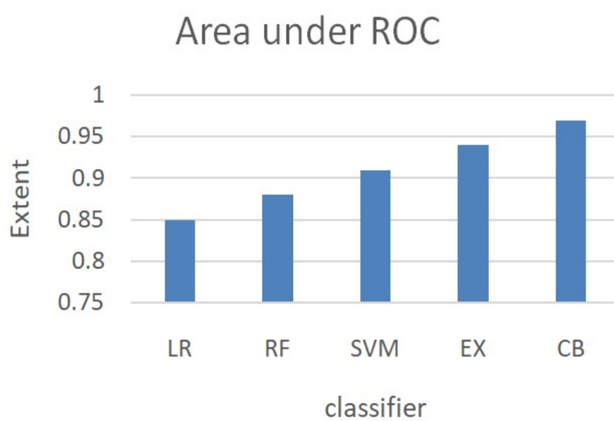


Figure-10 performance evaluationin terms areas of classifiers under ROC curve

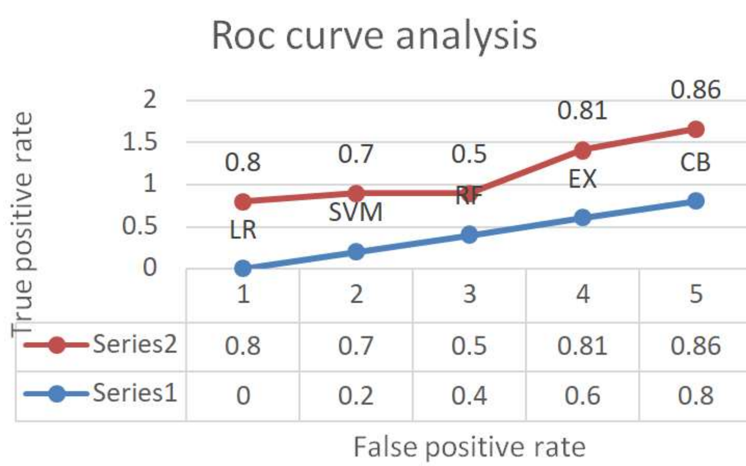
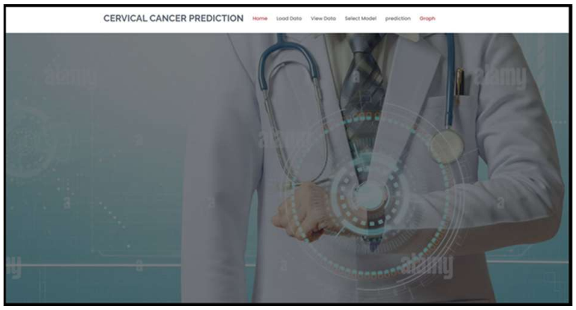


Figure-11 ROC curve analysis of different classifiers

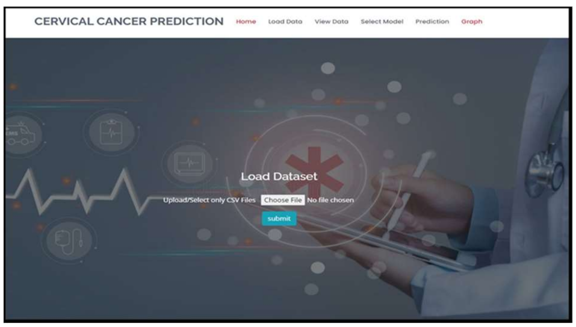
2. Results and Discussion:

The following screen photos show how this revolutionary treatment for cervical cancer utilising machine learning approaches actually works.

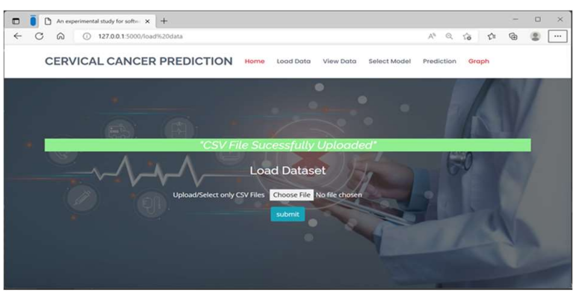
Home page: This is the home page of a novel approach for cervical cancer using machine learning Techniques.

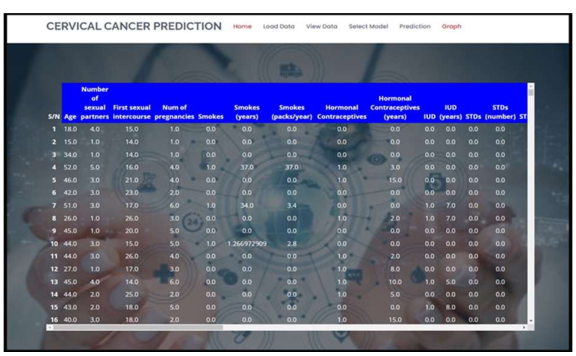


Upload data: This is the upload page where the data is being uploaded.



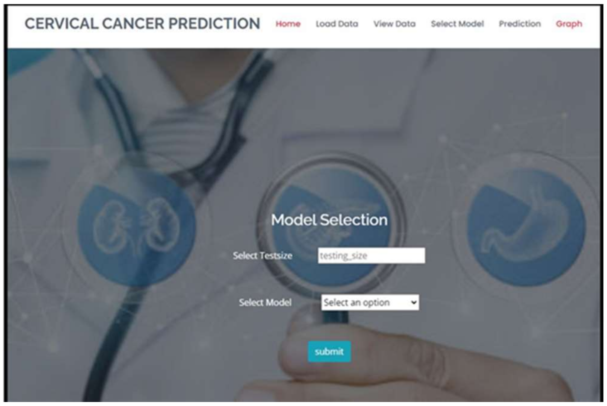
Loaded data: This page displays the uploaded data.



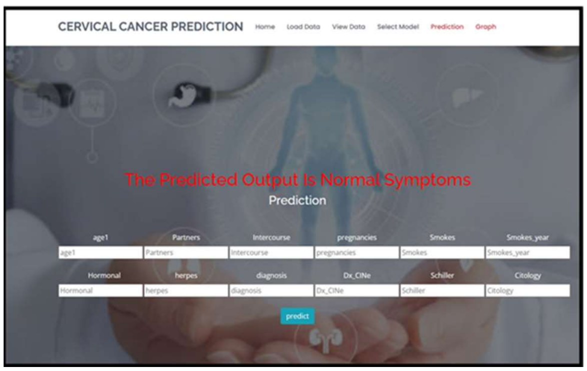


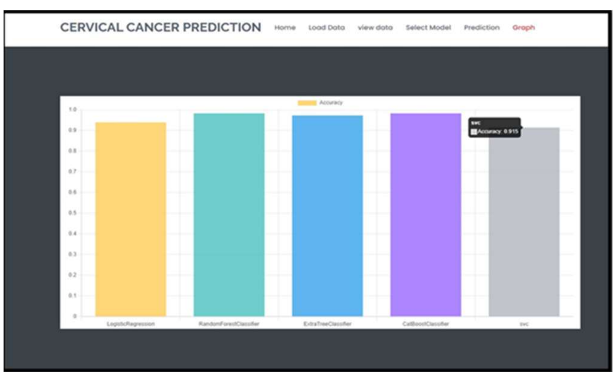
View data: Here we can see all the data.

Select model: In this page we can select the model.



Prediction: Once after model selection prediction will be done here.





Graph page: This page will displays the graphical results.

5. Conclusion:

cervical cancer detection to preserve the subjects' priceless life. Despite the difficulties and problems, machine intelligence-based solutions now available are regarded as reliable. However, the effectiveness of various categorization systems for the detection of cervical cancer is also a problem. Different classifiers produce distinctly different results when used to the same datasets because the specific classification algorithms are sensitive to the nature of the data. Based on a major voting procedure to accept the best classification outcomes for cervical cancer prediction, this research study proposed an ensemble classification approach for cervical cancer diagnosis.

REFERENCES:

[1] N. Qiu, X. Li, and J. Liu, ''Application of cyclodextrins in cancer treatment,'' J. Inclusion Phenomena Macrocyclic Chem., vol. 89, nos. 3–4, pp. 229–246, Dec. 2017, doi: 10.1007/s10847-017-0752-2.

[2] L. Schwartz, C. Supuran, and K. Alfarouk, ''The warburg effect and the hallmarks of cancer,'' Anti-Cancer Agents Med. Chem., vol. 17, no. 2, pp. 164–170, Jan. 2017, doi: 10.2174/1871520616666161031143301.

[3] E. M. Hassan and M. C. DeRosa, ''Recent advances in cancer early detection and diagnosis: Role of nucleic acid based aptasensors,'' TrAC Trends Anal. Chem., vol. 124, Mar. 2020, Art. no. 115806, doi: 10.1016/j.trac.2020.115806.

[4] N. A. Parmin, U. Hashim, S. C. B. Gopinath, S. Nadzirah, Z. Rejali, A. Afzan, and M. N. A. Uda, ''Human papillomavirus e6 biosensing: Current progression on early detection strategies for cervical cancer,'' Int. J. Biol. Macromolecules, vol. 126, pp. 877–890, Apr. 2019, doi: 10.1016/j.ijbiomac.2018.12.235.

[5] T. A. Kessler, ''Cervical cancer: Prevention and early detection,'' Seminars Oncol. Nursing, vol. 33, no. 2, pp. 172–183, May 2017, doi: 10.1016/j.soncn.2017.02.005.

[6] N. Wentzensen and M. von KnebelDoeberitz, ''Biomarkers in cervical cancer screening,'' Disease Markers, vol. 23, no. 4, pp. 315–330, 2007, doi: 10.1155/2007/678793.

[7] C. J. Meijer, P. J. Snijders, and P. E. Castle, ''Clinical utility of HPV genotyping,'' Gynecol. Oncol., vol. 103, no. 1, pp. 12–17, Oct. 2006, doi: 10.1016/j.ygyno.2006.07.031.

[8] H. N. Luu, K. R. Dahlstrom, P. D. Mullen, H. M. Vonville, and M. E. Scheurer, ''Comparison of the accuracy of hybrid capture II and polymerase chain reaction in detecting clinically important cervical dysplasia: A systematic review and meta-analysis,'' Cancer Med., vol. 2, no. 3, pp. 367–390, Jun. 2013, doi: 10.1002/cam4.83.

[9] H. N. Luu, K. Adler-Storthz, L. M. Dillon, M. Follen, and M. E. Scheurer, ''Comparing the performance of hybrid capture II and polymerase chain reaction (PCR) for the identification of cervical dysplasia in the screening and diagnostic settings,'' Clin. Med. Insights, Oncol., vol. 7, Jan. 2013, Art. no. CMO.S12811, doi: 10.4137/CMO.S12811.

[10] T. A. Brown, ''Southern blotting and related DNA detection techniques,'' in Encyclopedia of Life Sciences. Hoboken, NJ, USA: Wiley, 2001.

[11] Q. Wang, B. Zhang, X. Lin, and W. Weng, ''Hybridization biosensor based on the covalent immobilization of probe DNA on chitosan–mutiwalled carbon nanotubes nanocomposite by using glutaraldehyde as an arm linker,'' Sens. Actuators B, Chem., vol. 156, no. 2, pp. 599–605, Aug. 2011, doi: 10.1016/j.snb.2011.02.004. 12386 VOLUME 9, 2021 Q. M. Ilyas, M. Ahmad: Enhanced Ensemble Diagnosis of Cervical Cancer: A Pursuit of Machine Intelligence.