DETECTING ABERRANT CERVICAL CANCER USING MACHINE LEARNING

PBS ADITYA KUMAR[1], SHAIK HANISHMA[2]

[1&2]Assistant Professor, Dep of CSE, Narasimha Reddy Engineering college, Hyderabad.

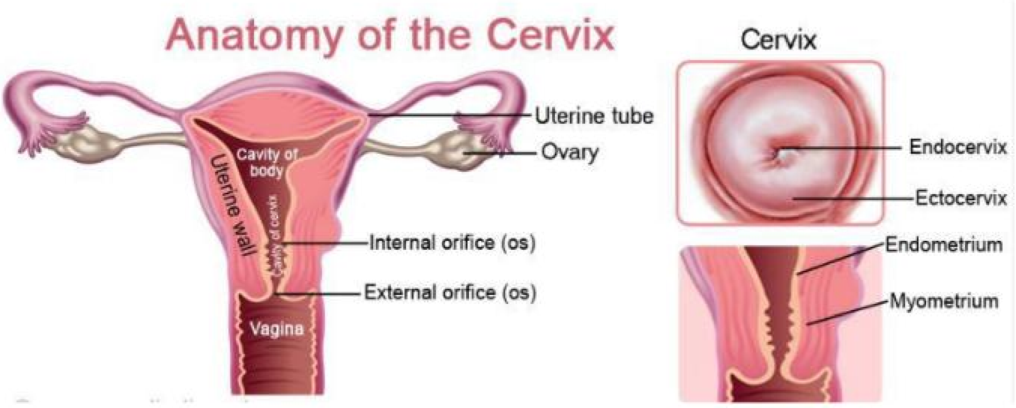
Abstract: Atypical cervical cancer is one of the main causes of mortality, hence early detection is crucial. Atypical cervical cancer, which originates within women's cervixes, may be found via vaginal examination. The HPV virus may cause cervical tissue abnormalities in women. Female patients with cervical infections have a systemic viral infection. As the leading cause of maternal mortality in developing countries, it is catastrophic. Atypical cervical carcinoma ranks third in female cancers globally. One of India's top 10 murders. India has 2-2.5 million cancer sufferers. Over seven lakh new cancer cases are anticipated in India annually. Two innovative methods for cervical cancer detection from aberrant Pap smear pictures are presented in this thesis. The first module detects atypical cervical cancer using a creative decision tree. Deciding tree creates a prediction model that uses an input variable to predict the feature vector. The picture must be divided and analyzed using a decision tree to pinpoint the affected areas. This research examines alternative segmentation methods for abnormal cervical cancer detection. This study examines single-cell Pap smear pictures. Smear tests reveal blood abnormalities. Image processing is a popular data extraction approach. This treatment measures uterine length and cervical carcinoma size. Martin's open-source database can validate and analyze findings.

Keywords: Machine Learning, Cervical cancer, HPV and Decision Tree,

I. INTRODUCTION

Cancer, the leading cause of death on a worldwide scale, places a significant burden on healthcare systems, families, and people impacted by the disease. A hereditary component is found in 5-10% of all cancer cases. One of the most distinguishing hallmarks of cancer is uncontrolled cell proliferation, which may be triggered by several internal factors. Some of these causes include somatic and germ line mutations, cigarette smoke, pathogenic bacteria, a poor diet, and other internal variables. When cancer starts to develop in cells, it causes some parts of the body to expand in an unforeseen way. Cancer may emerge in several ways, but it always starts with the uncontrolled proliferation of certain cells. To distinguish themselves from other cells, cancer cells have the potential to go mad and attack other tissues. Cancer cells are considered to have metastasized when they spread to other parts of the body. This broad category of diseases includes a variety of conditions characterized by abnormal growth of ordinarily sized cells throughout the body. These issues may occur anywhere in the body and can even infect and spread to other organs. Cancers that start in squamous cells and epithelial cells are two examples of cancer categorization based on the cells from which they arise. Certain cell types are the source of certain cancers.

Cervical cancer (CC) [1] is a dangerous tumor that develops from cells in the cervix uteri, the lower section of the womb's neck in the female reproductive system. Squamous and glandular cells surround the cervix. Human Papillomavirus (HPV) [3] is the primary cause of cervical cancer, which begins in the cervix's core layer, at the vaginal junction, and in the uterus. It is a slowly advancing illness that starts with dysplasia, a precancerous condition. Cancer is one of India's ten major causes of death, with an estimated 2-2.5 million people affected [2]. Cervical cancer is commonly regarded as one of the most devastating malignancies on the globe. Women in developing countries have a greater frequency of cancer identification at later stages, most likely owing to fewer treatment choices, which leads in a worse prognosis and a higher mortality rate. Most cervical cancer cases are caused by high-risk human papillomaviruses (HPV), which are very infectious (99 percent). Women aged 15 to 44 are more likely to acquire cervical cancer. Every year, more than 200000 women die from cervical cancer, with approximately 90% dying in developing countries due to a lack of medical resources and experience. Around 80% of cervical cancer cases occur in developing countries, where it is the most prevalent malignancy in many regions. In recent decades, the widespread use of cytologic screening programs has shown to be a successful way to reducing cervical cancer incidence and mortality rates.



(Source: https://www.medindia.net/patients/patientinfo/cervix-function.htm)

Figure 1 Cervical cancer

II. METHODS OF MACHINE LEARNING

Machine learning methods are classified in to 4 types, they are

Supervised machine learning

Unsupervised machine learning

Semi-supervised machine learning

Reinforcement machine learning algorithms

Supervised machine learning makes use of labeled examples to generalize learnt concepts to new data to make predictions about the future. To forecast the output values, the learning algorithm analyses a certain training dataset and then generates an inferred function. The system will provide goals for every new input after enough training is done. Another way the learning algorithm might find errors and fix the model is by comparing its results to the target one.

Without an adult supervisor: When there are no labels or classifications in the training data, an unsupervised learning approach is used. Computers can learn to explain hidden structures from unlabeled data using unsupervised learning. The system finds hidden structures using unlabeled datasets instead of deciding what output to provide.

A hybrid approach that blends supervised and unsupervised learning is known as semi-supervised learning. Training in semi-supervised learning makes use of both labelled and unlabeled data. System learning accuracy may be significantly enhanced using this approach.

When training using labeled data that involves the employment of competent and enough resources becomes common, semi-supervised learning becomes the method of choice. But unlabeled data often doesn't need any extra resources.

Methods for enhancing machine learning: This algorithm takes cues from its surroundings to produce actions and determines whether they were successful or not. Delay in rewards and trial-and-error search are hallmarks of reinforcement learning. With this innovation, software agents and robots may increase their productivity by deciding for themselves what actions to take in different scenarios. The agent needs a simple reward, or reinforcement signal, to learn the correct course of action.

Predicting the result of a medical image as a dependent variable, logistic regression is the simplest basic classifier in machine learning. It is a probability value that falls somewhere in the medium, rather than two extremes like 0 or 1, like Yes, No, 0, 1, true, false, etc. Its ability to accurately predict sickness based on medical images is lacking. The issue was addressed by developing the Naive Bayes classifier. A straightforward and efficient categorization method, it facilitates the development of fast machine learning models that can identify illnesses with minimal input data and provide predictions quickly. Predicting multiple classes is another area where it excels. The results are better when contrasted with logistic aggression. If a categorical variable not included in the training set is present in the test set, Naive Bayes will be rendered incapable of producing any predictions. But it's not without its flaws. If there is a categorical variable in the test set that wasn't in the training set, the Naive Bayes model will fail to make any predictions. In medical imaging, for more accurate illness detection We created a K-nearest-neighbor classifier, and it new data may be added without compromising the accuracy of the picture, and training is not necessary before making predictions. Using KNN is a breeze. On the other hand, it is very vulnerable to noisy data and cannot handle big datasets. When applied to high-dimensional spaces, support vector machine (SVM) outperforms k-nearest neighbour (KNN) algorithms. This approach is effective when there are more dimensions than samples. If there are more features per data point than there are training data samples, SVM will underperform. This is to avoid problems with noise and data set overlaps. The suggested decision tree can manage both continuous and categorical inputs, shorten training time, and more. One issue with this classifier is that it becomes overfit with the data, which causes it to make inaccurate predictions. It's also susceptible to noise and doesn't deal with missing values.

III. ABNORMAL CERVICAL CANCER DETECTION

According to [9,10], cervical cancer is one of the most common types of cancer that is diagnosed in women all over the globe because of its prevalence. This results in women passing away at a young age and being incapacitated for the rest of their lives, squandering time that might have been used on more useful endeavors. People do not have sufficient access to medical treatment and screening, which is the primary source of the problem. Additionally, there is a lack of public awareness on the issue. According to [6,7], the government of India has initiated a multitude of cancer control programs; however, these efforts have not been successful in reaching out to rural regions because to problems with availability to trained staff, infrastructure, transportation, regulatory compliance, and scanning frequencies. According to the current statistics, cervical cancer is the second most common kind of cancer that affects women throughout the world. It is the cervix, also known as the entrance to the uterus, that is the place of genesis for this cancer. The nucleus of cancer cells is not always possible to detect with the naked eye, which makes the process of diagnosis more difficult that is being performed. When compared to the nuclei of abnormal cells, the nuclei of normal cells are much thicker and thinner. Being able to visually differentiate between the phases of cervical cancer is particularly challenging since abnormal nuclei tend to be larger. This is due to the fact that there is no system that is widely approved for categorizing cancer stages based on nucleus monitoring, and every specialist has their own distinct point of view. According to the most recent figures, one woman passes away around every seven minutes because of cervical cancer. According to [8], by the year 2025, that amount will have increased to 4.6 minutes each hour. Even though India has been working tirelessly to combat cervical cancer for the last three decades, the nation is still ranked fourth in the world. This is even though India has made very little progress in decreasing the mortality and morbidity rates associated with cervical cancer. There is a substantial likelihood of survival for cervical cancer if it is discovered at an early stage. Pap smears have the capability of detecting tumors, which are precancerous growths that may be treated with early genetic testing. This has the potential to be a useful diagnostic tool. Consequently, this demonstrates the urgent need for a system that can facilitate sample analysis. Byriel, Martin, and Norup concentrated their efforts primarily on the classification of Pap smear pictures. The function of cervical cell segmentation and classification was accomplished by Martin and Norup using the use of the champ software that was made available by Dimac Imaging. Ustafson-kessel clustering, fuzzy c-means [13] (fcm), and hard c-means (hcm) were the primary classification techniques (gk) that Martin applied in his research. With a neuro-fuzzy classification system, Byriel was able to create the categorization of cervical cells.

Adaptive network based fuzzy inference system (anfis), kessel clustering (gk), logistic model [12], and clustering algorithm (fcm) were the methods that Byriel used to classify the cells that were found in the cervix. The classification of cervical cells was accomplished by Byriel via the use of fuzzy c-means (fcm), gustafsonkessel clustering (gk), and an autonomous network-based fuzzy inference system (anfis). The study that was conducted by Nithya and colleagues and released in the year 2020 indicates that the nucleus boundary has been segmented by the use of watershed transform-based techniques. Using the findings of Martin, Byriel, and Norup's study as a foundation, we provide an innovative approach to the development of a diagnostic and detection software for cervical cancer. The single-cell photos that were acquired using Pap smear are the primary areas of concentration in this study [11]. Researchers have access to the open-source database that Martin controls and utilizes for analysis and validation. This database is also available for usage by researchers. The bottom part of the uterus is referred to as the cervix, which is also occasionally referred to as the cervix uteri. The significant medical issue of cervical cancer, which affects an estimated 14,000 women in less developed countries, is directly attributable to the insufficient vaccination efforts that have been made. The fatality rate of this condition is much higher among females. When it comes to radiation therapy, this situation involves both volumetric and anthropomorphic techniques. Detecting abnormalities in the blood may be accomplished via the use of smear testing. It is possible that the extraction of data will be successful if image processing is used. According to [10], this medical method has the capability of determining both the length of the uterus as well as the quantity of cervical cancer already present.

IV. CLASSIFICATION TYPES OF ABNORMAL CERVICAL CANCER

When cells lining the cervix begin to change improperly, abnormal cervical cancer occurs. Malignant transformation or eventual return to normalcy is possible for these abnormal cells. Among atypical cervical cancers, melanoma and denocarcinoma predominate. The cellular makeup of each one is what distinguishes them when seen via a light microscope. Squamous cell carcinoma begins in the thin, smooth cells that line the cervix's bottom. About 80% of cervical malignancies are this kind of cancer, according to the National Cancer Institute. The seminiferous tubules that cover the surface of the cervix are the origin of cervical adenocarcinomas. About 20% of all instances of atypical cervical cancer are cervix adenocarcinomas. When cervical cancer is abnormal, it might affect both cell types. Abnormal On the other hand, some types of cervical malignancies are very uncommon. For instance, the cervix is the origin of metastatic atypical cervical carcinoma.

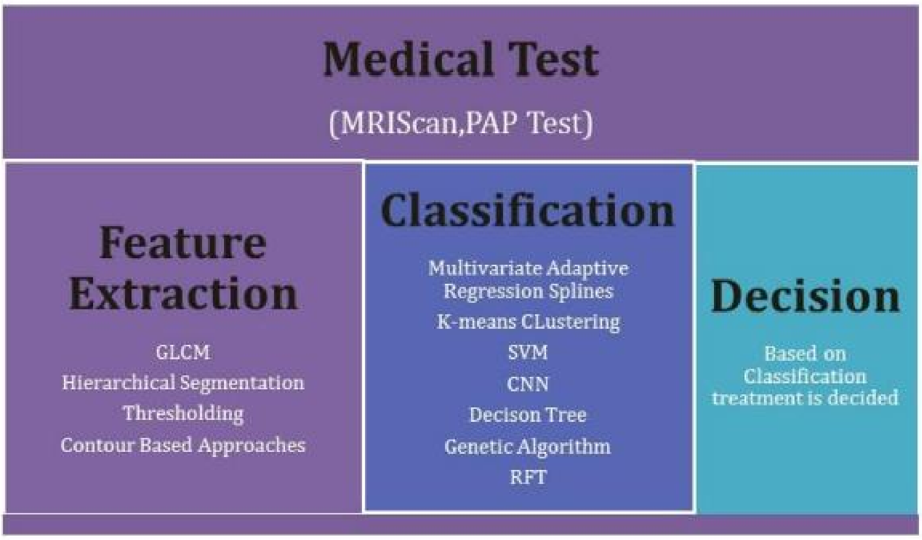


Figure 2 Classification types of abnormal cervical cancer

It is a screening procedure for the cervical region that may identify cancerous or precancerous cells. This identical medical procedure is also called the Papanicolaou test or the Pap smear. A Pap smear is a photographic analysis of cells that have been infected with the HPV [8] virus. Obtaining cervix cells raises the possibility that all of them are cancerous, even if only one is aberrant. Regardless, cervical cancer that is not typical has emerged. The first step in getting a slide ready for a Pap smear analysis is to fix and stain the cells on it. Next, the prepared slide is examined using a microscope lens, and digital images are captured. After this, the digital photos are blown up while keeping the pixel intensity correct. The next step is to use various deep learning and machine learning algorithms to divide up the photographs into their respective categories. In the segmentation process, the nuclei and cytoplasm are separated from one another. The segmented image is the output of this step; it is sent into the classification phase, which oversees deciding whether the cell is normal or aberrant.

V. METHODS AND ALGORITHMS

Adaptive threshold metric is one of the heuristic tools used in strategic thinking. Most of the time, it's utilized to find out which data splitting standard works best for the available tests. Common metrics include supervised learning, entropy, and gain value. The study introduces a new decision tree approach for classifying atypical cervical cancers using characteristics extracted from the preprocessed image. Based on the traits that were studied, this method was developed.

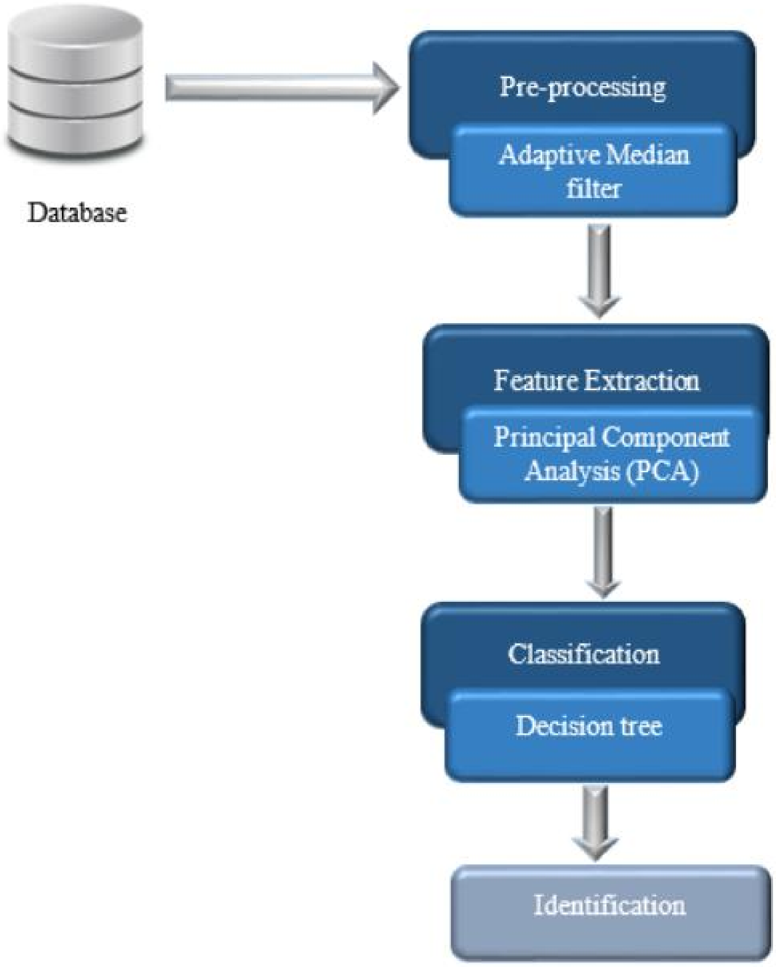


Figure 3 Proposed methodologies

Figure depicts the overall process of using the Novel Decision tree algorithm technique to diagnosis the disease in pap smear image more efficiently and in less time. In the above figure, for pre-processing the image Adaptive filter is used. By extracting the feature by the image Principle Component Analysis is used. For classification Decision tree is used.

VI. RESULTS AND DISCUSSION

On a Windows 7 Intel Core i7 CPU (Central processing unit) with a 2.53GHz operating frequency and 8GB of accessible memory, the implementation is carried out in MATLAB (MATrix LABoratory) R2018. In past findings, single cells did the most of the job, while the Pap smear slides received many and overlapping cells. We obtained cytological pictures from pathology labs in Jaipur. We have 50 shots in total; all of them are in JPEG format. There are 15 normal cells, 20 CIN1 cells, and the rest are CIN2/CIN3 [7]. The image is made up of numerous cells that are overlapping. The attribute is labeled on a node, branches are formed for each value of the attribute and the data are subdivided properly and the predictor variables are tabulated in Table 4

Predictor Variables	Total Number of Instances	Correctly Classified Instances	Percentage	Incorrectly Classified Instances	Percentage	Detailed Accuracy by Class
Decision tree (J48)	28	24	85.7143 %	4	14.2857 %	85
Random tree Size of the tree: 07	28	20	71.4286 %	8	28.5714 %	71
Random tree Size of the tree: 13	28	16	57.1429 %	12	42.8571 %	57

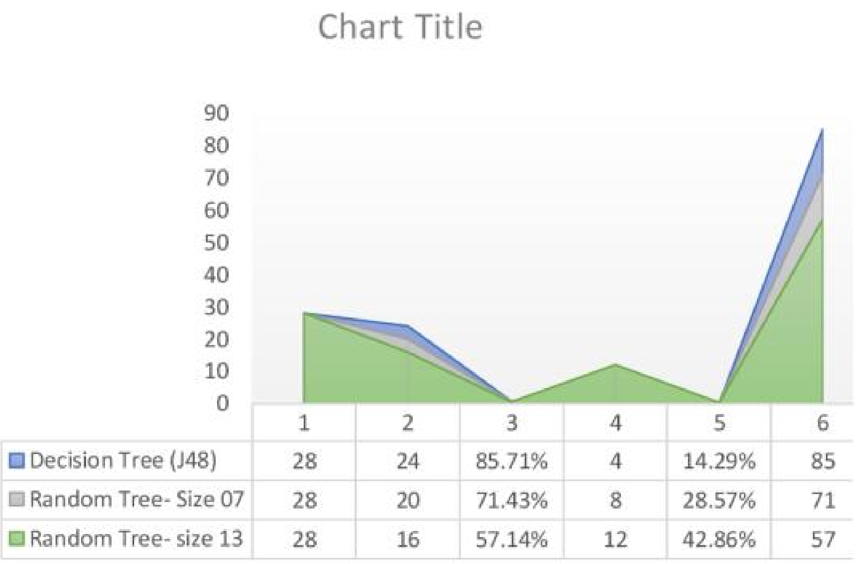


Figure 5: Comparison of different phases of processing

Techniques	Sensitivity	Specificity	Accuracy
Decision Tree	80	40	72
Random Tree Size (7)	77.5	30	68
Random Tree Size (13)	72.5	20	62

Fig 6: Comparison of Machine learning Technology

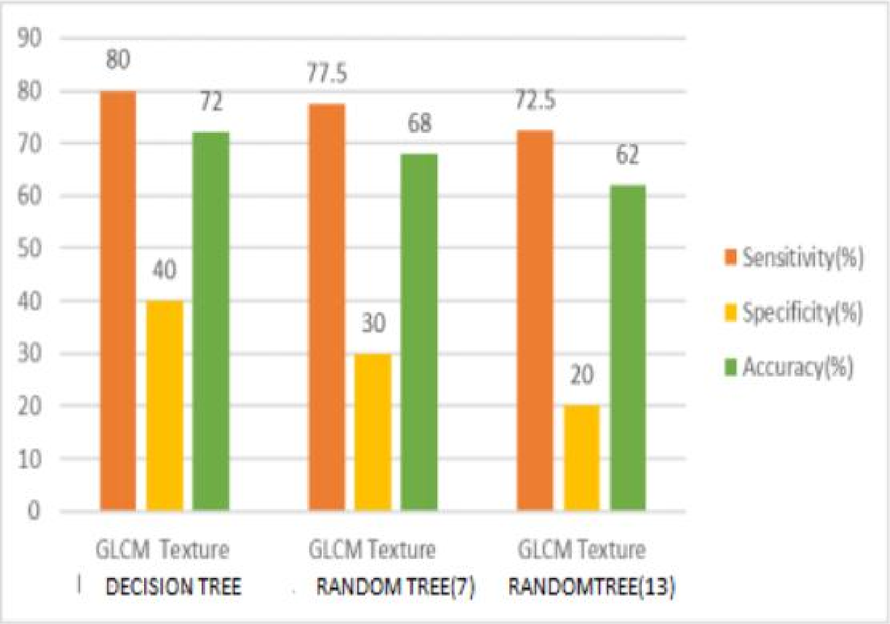


Figure 7 Graphical representation of texture for various techniques

Figure 4 illustrates the error rate that occurs when using the choice tree technique and the random tree approach with sizes of seven and thirteen, respectively. When using the decision tree approach, there is very little chance for error to potentially occur. In Figure 7, the data that correspond to the performance of the texturing function are visually shown. With regard to the detection of aberrant cervical cancer by the use of GLCM [4] characteristics, the suggested technique unequivocally exhibits a significant improvement in accuracy. The results of the accuracy tests that were performed on the various categorization techniques are graphically shown in the figure. The purpose of this study is to investigate the current state of automated cell segmentation and categorization in cervical cytopathology. During the process of compiling this exhaustive research, a number of databases, such as PubMed, arXiv, Google Scholar, ACM, IEEE, Springer, and Elsevier, were investigated. In addition to that, we made certain that the references in each of the articles were checked.

CONCLUSION

This study looked at a variety of non-standard cervical cancer screening approaches. A combination of image processing and rigorous computer approaches produced more reliable results for atypical cervical cancer diagnosis than medical diagnostics. To achieve reliable findings, advanced technology is used at all stages of the discovery process, from preprocessing to highlight extraction and arrangement. According to the findings, medical pictures that utilize an adaptive filter for preprocessing, Principal Component Analysis for feature extraction (with shade as a component), an island evacuation post-handling technique, and a decision tree as a classifier provide the best results with high accuracy. Combining these approaches and applying them to a preset clinical imaging enables one to determine the cancer's underlying stage. A variety of information mining approaches may be used to forecast cervical cancer. To predict cervical cancer, this research initially employed three categorization methods and then assessed the outcomes. Furthermore, the article discusses the trait choice measure utilized by numerous choice tree calculations, including the J48 calculation's entropy, the choice tree calculation's most significant increase value, and the irregular tree's data gain. The report also includes the methodologies used to calculate these property choice measures. We conclude that various calculations for choice tree enrollment should be employed at different times depending on the scenario, since they need less computing time overall. Using the test dataset, the findings showed that the decision tree is the best classifier indicator. Other study should be undertaken to improve the execution of these order strategies by including other components and selecting a longer future period.

REFERENCES

1. *Abdoh, SF, Rizka, MA & Maghraby, FA 2018, 'Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques', IEEE Access, vol. 6, pp. 59475-59485.*
2. *Ali, MM, Ahmed, K, Bui, FM, Paul, BK, Ibrahim, SM, Quinn, JM & Moni, MA 2021, 'Machine learning-based statistical analysis for early stage detection of cervical cancer', Computers in Biology and Medicine, vol. 139, p.104985.*

3. *Bhatt, AR, Ganatra, A & Kotecha, K 2021, 'Cervical cancer detection in pap smear whole slide images using convNet with transfer learning and progressive resizing', PeerJ Computer Science, vol. 7, p.e348.*

4. *Deng, X, Luo, Y & Wang, C 2018, 'Analysis of risk factors for cervical cancer based on machine learning methods', 5 th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS), pp. 631-635. IEEE.*

5. *Prasadu Peddi (2023). AI-Driven Multi-Factor Authentication and Dynamic Trust Management for Securing Massive Machine Type Communication in 6G Networks. International Journal of Intelligent Systems and Applications in Engineering, 12(1s), 361–374.*

6. *Singh, SK & Goyal, A 2020, 'Performance analysis of machine learning algorithms for cervical Cancer detection', International Journal of Healthcare Information Systems and Informatics (IJHISI), vol. 15, no. 2, pp.1-21.*

7. *Uday Chandrakant Patkar, Sushas Haribabu Patil and Prasad Peddi, "Translation of English to Ahirani Language", International Research Journal of Engineering and Technology(IRJET), vol. 07, no. 06, June 2020.*

8. *Soni, VD & Soni, AN 2021, 'Cervical cancer diagnosis using convolution neural network with conditional random field', Third International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 1749-1754. IEEE.*

9. *Taha, B, Dias, J & Werghi, N 2017, 'Classification of cervical-cancer using papsmear images: a convolutional neural network approach', Annual Conference on Medical Image Understanding and Analysis, pp. 261-272. Springer, Cham.*

10. *Unlersen, MF, Sabanci, K & Özcan, M 2017, 'Determining cervical cancer possibility by using machine learning methods', International Journal of Latest Research in Engineering and Technology, vol. 3, no. 12, pp.65-71.*

11. *William, W, Ware, A, Basaza-Ejiri, AH & Obungoloch, J 2019, 'A pap-smear analysis tool (PAT) for detection of cervical cancer from pap-smear images', Biomedical Engineering Online, vol. 18, no. 1, pp.1-22.*

12. *Yang, L & Huang, X 2014, 'A novel piezoelectric immunosensor for early cervical cancer detection', In Proceedings of the Symposium on Piezoelectricity, Acoustic Waves, and Device Applications, pp. 453-456. IEEE.*

13. *Zhang, L, Kong, H, Chin, CT, Liu, S & Fan, X 2014, 'Automation-assisted cervical cancer screening in manual liquid-based cytology with hematoxylin and eosin staining', Cytometry Part A, vol. 85, no. 3, pp. 214-30.*