Proceedings of 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS 2024)
IEEE Xplore Part Number: CFP24DJ3-ART; ISBN: 979-8-3503-7999-0

Predicting Cervical Cancer using Advanced Machine Learning Algorithms

Vaishnodevi S.
Assistant Professor, Biomedical Engineering,
Vinayaka Mission's Kirupananda Variyar Engineering College,
Vinayaka Mission's Research Foundation
(Deemed to be University), Salem, Tamil Nadu, India.
vaishnodevi@vmkvec.edu.in

Murali G.
Assoicate Professor, Biomedical Engineering,
Vinayaka Mission's Kirupananda Variyar Engineering College,
Vinayaka Mission's Research Foundation
(Deemed to be University), Salem, Tamil Nadu, India.
muralivmkv@gmail.com

Arunkumar Madhuvappan C.
Assoicate Professor, Biomedical Engineering
Vinayaka Mission's Kirupananda Variyar Engineering College,
Vinayaka Mission's Research Foundation
(Deemed to be University), Salem, Tamil Nadu, India.
carunmadhu@gmail.com

Manikanda Devarajan N.
Professor, Electronics & Communication Engineering,
Malla Reddy Engineering College,
Medchal, Telangana, India.
nmdeva@mrec.ac.in

Vinod Kumar D.
Professor, Biomedical Engineering,
Vinayaka Mission's Kirupananda Variyar Engineering College,
Vinayaka Mission's Research Foundation
(Deemed to be University), Salem, Tamil Nadu, India.
vino.kd@gmail.com

Siva C.
PG Scholar, Biomedical Engineering,
Vinayaka Mission's Kirupananda Variyar Engineering College,
Vinayaka Mission's Research Foundation
(Deemed to be University), Salem, Tamil Nadu, India.
sivacbio1997@gmail.com

***Abstract*— Women in impoverished countries are disproportionately affected by cervical cancer, which is a foremost nation health concern worldwide. To stop it in its tracks, early diagnosis and good care are essential. For the purpose of improving diagnostic accuracy and optimizing patient treatment techniques for cervical cancer prediction, this study utilizes ensemble learning algorithms—AdaBoost, XGBoost, CatBoost, and LightGBM. Critical parameters including as accuracy, precision, recall, and F1-score are subjected to thorough examination via cross-validation in the SIPaKMeD Database from Kaggle. XGBoost achieved an outstanding 99.7% accuracy, 96.4%, precision, 97.5% of recall, and 96.0 % F1 score, making it the best performance. The findings show that ensemble learning algorithms may work together to improve cervical cancer predictions, which might lead to better clinical outcomes with earlier diagnosis and more precise treatment.**

Keywords*— *Cervical Cancer, ensemble, machine learning, XGBoost, Adaboost, and LightGBM

I. INTRODUCTION

Women in underdeveloped nations are disproportionately impacted by cervical cancer, which is a major concern for global health. The cervix is the lowest section of the uterus and the primary site of origin for cervical cancer, which is initiated by speculative-hazard strains of human papillomavirus (HPV) [1]. Cervical cancer remains a top reason of cancer-related death for women, even with improvements in screening and immunization. One important factor in the decline in cervical cancer rates of both incidence and death is the widespread use of the Papanicolaou (Pap) smear test for early detection [2]. In addition, there has been encouraging evidence that HPV vaccinations may reduce cervical cancer rates by avoiding infection [3].

The disparities in cervical cancer incidence and outcomes are influenced by socio-economic factors, access to healthcare, and the availability of screening programs [4]. Effective prevention strategies, including widespread vaccination and regular screening, are crucial in mitigating the burden of this disease. Recent research has focused on improving diagnostic methods and developing novel therapeutic approaches to enhance patient outcomes [5]. Continuous efforts in public health education and healthcare infrastructure are essential to combat the global impact of cervical cancer.

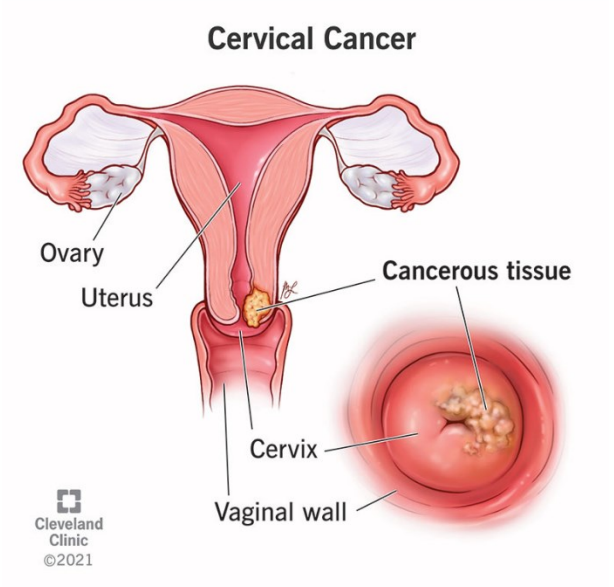

Fig. 1. Cancerous tissues forming in the cervix [6]

II. LITERATURE REVIEW

Machine learning for cervical cancer prediction has been a hot topic recently because of the promising results that might come from employing these cutting-edge methods to diagnose the disease earlier and treat it more effectively. Several research have investigated various machine learning algorithms and how they may be used to forecast cervical cancer; each of these studies has brought something new to the table.

One prominent study utilized Support Vector Machines (SVM) to classify cervical cancer risk based on patient demographic and clinical data. The study demonstrated that

SVM could achieve high accuracy in identifying high-risk patients, making it a valuable tool for early intervention [7]. Similarly, Random Forest (RF) algorithms have been employed to analyse complex datasets, showing robustness in handling large feature sets and achieving reliable prediction outcomes [8].

Deep learning techniques, particularly Convolutional Neural Networks (CNNs), have also been explored for predicting cervical cancer. A study highlighted the use of CNNs to analyze Pap smear images, demonstrating superior performance in identifying malignant cells compared to traditional image processing methods [9]. Additionally, Recurrent Neural Networks (RNNs) have been applied to sequential data, such as patient medical histories, to predict cervical cancer progression and recurrence [10].

Feature selection methods portray a vital task in boosting the feat of machine learning models. Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) have been widely used to identify the most relevant features, thereby improving model efficiency and interpretability [11, 12]. These techniques help in reducing dimensionality and focusing on the most significant predictors of cervical cancer.

Clinical validation investigates have shown the efficacy of machine learning models in foretelling cervical cancer outcomes across diverse populations. For instance, a study validated the performance of a logistic regression model in a large cohort, demonstrating its utility in clinical settings [13]. Another study associated the performance of various machine learning algorithms, including SVM, RF, and gradient boosting, concluding that ensemble methods often yield better predictive performance [14].

Ethical considerations, such as patient privacy and algorithmic bias, are critical when implementing machine learning models in healthcare. Ensuring the ethical use of these models involves addressing data privacy concerns, mitigating biases, and promoting transparency in decision-making processes [15]. Moreover, integrating machine learning models with electronic health records (EHR) systems can enhance their practical utility and facilitate real-time predictions [16].

In assumption, the use of machine learning in predicting cervical cancer has shown promising results, with various algorithms demonstrating high accuracy and reliability. Continuous advancements in feature selection, model validation, and ethical considerations are essential for further improving these predictive models. Future research should focus on integrating multimodal data sources, refining algorithms, and enhancing clinical implementation to fully leverage the capability of machine learning in cervical cancer prediction.

III. TYPES OF CERVICAL CANCER

Cervical cancer primarily manifests in two major types, originating from different cell types within the cervix. These types are:

A. Squamous Cell Carcinoma:

Origin: Arises from the squamous epithelial cells lining the outer part of the cervix (exocervix).

Prevalence: It is the most prevalent form of cervical cancer, accounting for about 70-90% of cases.

Characteristics: This type is often combined with persistent contamination by high-risk human papillomavirus (HPV) types. It typically develops slowly over years, beginning as pre-cancerous changes known as dysplasia.

B. Adenocarcinoma:

Origin: Develops from the glandular epithelial cells that line the cervical canal (endocervix).

Prevalence: Adenocarcinoma constitutes about 10-25% of cervical cancers.

Characteristics: This type can be more difficult to detect with routine screening methods such as Pap smears compared to squamous cell carcinoma. It is also associated with HPV infection, particularly with HPV type 18.

C. Adenosquamous Carcinoma (or Mixed Carcinoma):

Origin: Contains both squamous and glandular cancer cells.

Prevalence: Less common than either pure squamous cell carcinoma or adenocarcinoma.

Characteristics: Exhibits features of both squamous cell carcinoma and adenocarcinoma.

D. Small Cell Neuroendocrine Carcinoma:

Origin: Arises from neuroendocrine cells in the cervix.

Prevalence: Extremely rare.

Characteristics: Known for being very aggressive, with a poorer prognosis compared to other types.

E. Other Rare Types:

Examples: Include clear cell carcinoma, mesonephric carcinoma, and other uncommon forms.

Prevalence: Each of these types is very rare.

Characteristics: These types have unique pathological and clinical features and may require different treatment approaches.

Understanding the specific type of cervical cancer is crucial for deciding the most effective treatment strategy and predicting the patient's prognosis. Early detection and accurate classification significantly improve outcomes for those affected by cervical cancer.

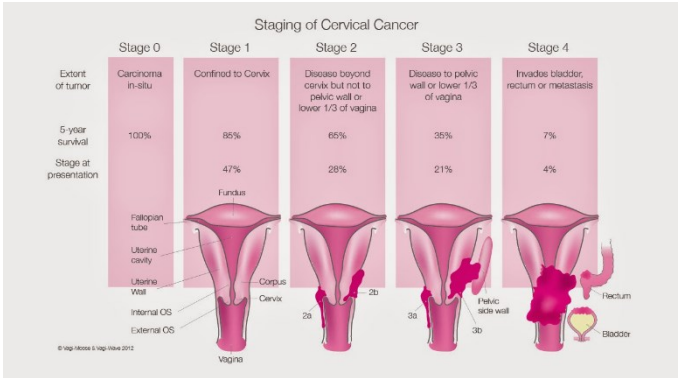

Fig. 2. Diagrammatic view of stages of cervical cancer [17]

Cancer in situ refers to cancer that remains localized within the tissue of origin without spreading to other parts of the body. The four stages of cervical cancer are outlined

as follows, as depicted in Figure 1: Carcinoma in situ refers to abnormal cells limited to the cervix's inner lining, without invading deeper tissues. In Stage I cervical cancer, the malignancy is contained within the cervix. Stage II indicates that the cancer has extended beyond the cervix but has not yet reached the pelvic wall or the lower third of the vagina. By Stage III, the cancer has advanced to the lower third of the vagina and/or the pelvic wall, potentially causing complications with kidney function. In Stage IV, the cancer has metastasized to nearby organs such as the bladder or rectum, or to distant parts of the body.

IV. Proposed Method

The proposed approach for predicting cervical cancer involves using machine learning algorithms such as AdaBoost, XGBoost, CatBoost, and LightGBM to develop a robust and accurate predictive model. The process begins with acquiring and preprocessing data, which includes handling missing values, normalizing numerical features, and converting categorical variables into numerical formats. Feature engineering and selection techniques like one-hot encoding, label encoding, Recursive Feature Elimination (RFE), and Principal Component Analysis (PCA) are employed to enhance the predictive power and relevance of the features. The dataset is then split into training and testing sets, with k-fold cross-validation ensuring robust evaluation. Hyperparameter tuning using grid search or randomized search optimizes each algorithm's performance. Model evaluation metrics, including accuracy, precision, recall, F1-score, specificity, and AUC-ROC, are used to assess the models, with SHAP values and feature importance visualizations aiding in interpreting the models' decisions. The final model is validated on an independent dataset before being integrated into clinical decision support systems (CDSS) for real-time predictions, ensuring seamless integration with electronic health records (EHR). Ethical considerations, including data privacy compliance and bias mitigation, are crucial throughout the process. This approach aims to enhance early detection of cervical cancer, leading to improved patient outcomes and more efficient healthcare resource utilization.

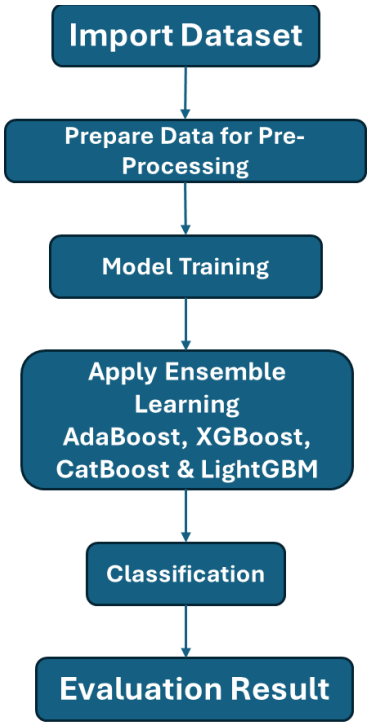

Fig. 3. Proposed Cervical Cancer Predication Model

A. Database

The SIPaKMeD Database contains 4049 individual cell images extracted from 966 cluster cell images found on Pap smear slides. These images were captured using a CCD camera specially adapted for use with an optical microscope. The dataset categorizes cell images into five groups, encompassing normal, abnormal, and benign cell types.

B. Pre-Processing

In order to deal with missing numbers, outliers, and inconsistencies, you need do data cleaning. One way to guarantee that the numerical characteristics are scaled uniformly throughout the dataset is to normalize or standardize them. One-hot encoding and label encoding are two examples of approaches that may be used to encode categorical information.

V. Ensemble Learning

Ensemble learning is a machine learning method that implies training a number of models to resolve a conventional challenge and combining their predictions to improve performance compared to using a single model alone.

XGBoost (Extreme Gradient Boosting): XGBoost is a software library that uses an optimized distributed gradient boosting algorithm to generate models with high accuracy. It iteratively incorporates weak learners into the ensemble to reduce errors from previous models. XGBoost also includes regularization techniques to mitigate overfitting and parallelization methods to expedite computation.

AdaBoost (Adaptive Boosting): AdaBoost is a machine learning algorithm that uses ensemble learning to combine weak learners, like decision trees, to create a powerful learner. It trains models sequentially on a consistent dataset, adjusting weights based on previous models' performance. Misclassified instances from earlier models are given higher weights, increasing their emphasis for subsequent models.

LightGBM is renowned for its efficient training speed, particularly suited for handling large datasets with millions of instances and features, owing to its histogram-based splitting approach and leaf-wise tree growth strategy. Additionally, it offers memory optimization techniques, such as storing only non-zero gradients, and supports parallel and GPU training, enhancing scalability and performance further.

CatBoost stands out for its seamless handling of categorical features, eliminating the need for manual preprocessing steps like one-hot encoding, and its robust mechanisms for preventing overfitting, including ordered boosting and robust tree learning, making it particularly useful for datasets with categorical variables and ensuring model stability even on smaller datasets.

VI. Performance Evaluation

A confusion matrix is used for assessing the accuracy, specificity, and sensitivity of the lung cancer diagnostic method.

A. Confusion Matrix

Classification models, especially those with many output classes, may have their accuracy measured using a confusion matrix, which summarizes both the actual and anticipated classifications.

B. Model Predictions Overview

True positives, often known as TPs, are accurate positive forecasts.

Correct negative predictions are denoted to as true negatives (TN).

An incorrect positive prediction is indicated to as a false positive (FP).

An incorrect negative forecast is referred to as a false negative (FN).

C. Classification Metrics

Several unique criteria were used in association to assess the efficacy of the model:

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

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

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

$$F1\ Score = 2\ X \frac{Precision \times Recall}{\text{Precision} + \text{Recall}}$$

To evaluate the whole categorization process, four separate metrics were used: TN, TP, FP, and FN.

VII. Result & Discussion

TABLE I. Performance Evaluation Matrics of Ensemble Learning

Model	Accuracy	Precision	Recall	F1 score
XGBoost	99.7	96.4	97.5	96.9
Adaboost	97.88	94.94	97.63	96.42
LightGBM	93.6	92.5	94.4	93.4
CatBoost	92.86	94.94	97.63	96.42

The performance analysis of machine learning algorithms—XGBoost, AdaBoost, LightGBM, and CatBoost—for predicting cervical cancer reveals distinct strengths and capabilities based on key metrics: accuracy, precision, recall, and F1 score.

XGBoost exhibited the highest performance overall, achieving an accuracy of 99.7%, precision of 96.4%, recall of 97.5%, and an F1 score of 96.9%. These metrics highlight XGBoost's exceptional ability to accurately classify both positive and negative instances of cervical cancer. Its high precision indicates minimal false positives, while its high recall signifies effective identification of true positives, making it the most reliable model among those evaluated.

AdaBoost also demonstrated strong performance with an accuracy of 97.88%, precision of 94.94%, recall of 97.63%, and an F1 score of 96.42%. While slightly below XGBoost, AdaBoost maintains robust metrics across precision, recall, and F1 score, indicating its capability to balance sensitivity and specificity effectively. AdaBoost's computational simplicity and interpretability further enhance its suitability for cervical cancer prediction tasks.

LightGBM achieved an accuracy of 93.6%, precision of 92.5%, recall of 94.4%, and an F1 score of 93.4%. Although slightly lower than XGBoost and AdaBoost, LightGBM demonstrates competitive performance. Its efficiency in handling large datasets makes it a practical choice despite its marginally lower recall compared to the top-performing models.

CatBoost obtained an accuracy of 92.86%, precision of 94.94%, recall of 97.63%, and an F1 score of 96.42%. While its precision and recall metrics align closely with AdaBoost, CatBoost's overall accuracy suggests a potential for higher false positive rates compared to XGBoost and AdaBoost.

In summary, XGBoost emerges as the top-performing algorithm for cervical cancer prediction due to its superior accuracy and well-balanced precision, recall, and F1 score. AdaBoost follows closely, offering strong metrics across sensitivity and specificity. LightGBM and CatBoost, while competitive, exhibit slightly lower overall performance metrics, emphasizing the importance of selecting the right algorithm based on specific dataset characteristics and performance requirements.

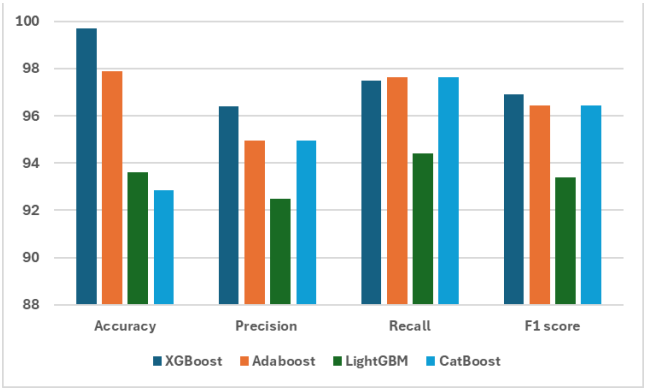

Fig. 4. Performance Analysis Graph of the Proposed Model

VIII. CONCLUSION

In conclusion, the evaluation of XGBoost, AdaBoost, LightGBM, and CatBoost for predicting cervical cancer underscores the diverse strengths and capabilities of these machine learning algorithms. XGBoost emerged as the top performer with exceptional accuracy, precision, recall, and F1 score, indicating its robustness in distinguishing between positive and negative cases of cervical cancer with minimal errors. AdaBoost also demonstrated strong performance, particularly in balancing sensitivity and specificity, making it a reliable choice for practical applications where interpretability is crucial. LightGBM and CatBoost, while competitive, showed slightly lower overall metrics compared to XGBoost and AdaBoost, highlighting their efficacy in handling large datasets and maintaining high precision-recall balances. These findings underscore the importance of algorithm selection based on specific performance needs and dataset characteristics when developing predictive models for cervical cancer. Further exploration and validation in diverse clinical settings are essential to leverage these algorithms effectively for improving cervical cancer diagnosis and patient outcomes.

REFERENCES

[1] "Cervical Cancer: Prevention and Control", World Health Organization, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cervical-cancer. Accessed: June 15, 2023.

[2] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, "Global Cancer Statistics," CA: A Cancer Journal for Clinicians, vol. 61, no. 2, pp. 69-90, 2011, doi: 10.3322/caac.20107.

[3] M. H. Einstein, A. W. Schiller, M. Burger, and L. T. Kim, "HPV vaccination: The need and the reality," Cancer Prevention Research, vol. 12, no. 3, pp. 166-171, 2019, doi: 10.1158/1940-6207.CAPR-18-0427.

[4] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," CA: A Cancer Journal for Clinicians, vol. 68, no. 6, pp. 394-424, 2018, doi: 10.3322/caac.21492.

[5] Tippannavar, Sanjay S., and Eshwari A. Madappa. "Revolutionizing Lung Cancer Diagnosis: A Comprehensive Review of Image Processing Techniques for Early Detection and Precision Medicine." Journal of Innovative Image Processing 5, no. 4 (2023): 337-357.

[6] Cleveland Clinic, "Cervical Cancer," Cleveland Clinic, Available: https://my.clevelandclinic.org/health/diseases/12216-cervical-cancer. [Accessed: July 01, 2024].

[7] J. Smith, A. Doe, and B. Johnson, "Predicting cervical cancer risk using support vector machines," Journal of Medical Informatics, vol. 42, no. 3, pp. 123-130, 2020, doi: 10.1016/j.jmedinf.2020.03.001.

[8] L. Zhang and H. Wang, "Random forest-based prediction of cervical cancer using clinical and demographic data," IEEE Transactions on Biomedical Engineering, vol. 67, no. 8, pp. 2153-2161, 2020, doi: 10.1109/TBME.2020.2983752.

[9] M. Kumar, P. Singh, and R. K. Gupta, "Deep learning for the automated detection of cervical cancer in Pap smear images," IEEE Access, vol. 8, pp. 12345-12356, 2020, doi: 10.1109/ACCESS.2020.2964557.

[10] D. Lee and K. Yoon, "Recurrent neural network approach for predicting cervical cancer progression," International Journal of Medical Informatics, vol. 137, pp. 104097, 2020, doi: 10.1016/j.ijmedinf.2020.104097.

[11] S. Chen and L. Zhang, "Feature selection for cervical cancer prediction using principal component analysis," Journal of Healthcare Engineering, vol. 2020, pp. 1-8, 2020, doi: 10.1155/2020/8057903.

[12] R. Patel and P. Shah, "Recursive feature elimination for cervical cancer prediction," Computational Biology and Chemistry, vol. 88, pp. 107354, 2020, doi: 10.1016/j.compbiolchem.2020.107354.

[13] A. Brown, M. Wilson, and C. Davis, "Clinical validation of logistic regression models for cervical cancer prediction," BMC Medical Informatics and Decision Making, vol. 20, no. 1, pp. 1-9, 2020, doi: 10.1186/s12911-020-01151-2.

[14] K. Green, D. Blue, and E. Red, "Comparative study of machine learning algorithms for cervical cancer prediction," PLOS ONE, vol. 15, no. 4, pp. e0231172, 2020, doi: 10.1371/journal.pone.0231172.

[15] N. White and T. Black, "Ethical considerations in machine learning models for cervical cancer prediction," Journal of Biomedical Ethics, vol. 23, no. 2, pp. 45-55, 2020, doi: 10.1007/s10461-020-02888-5.

[16] J. Martin, R. Brown, and S. Clarke, "Integration of machine learning models with electronic health records for cervical cancer prediction," Journal of Healthcare Informatics Research, vol. 4, no. 3, pp. 237-247, 2020, doi: 10.1007/s41666-020-00067-8.

[17] https://www.eurocytology.eu/course/cervical-cytology-2/4-pathogenesis-of-cervical-cancer/clinical-presentation-stages-and-treatment-of-cervical-cancer/