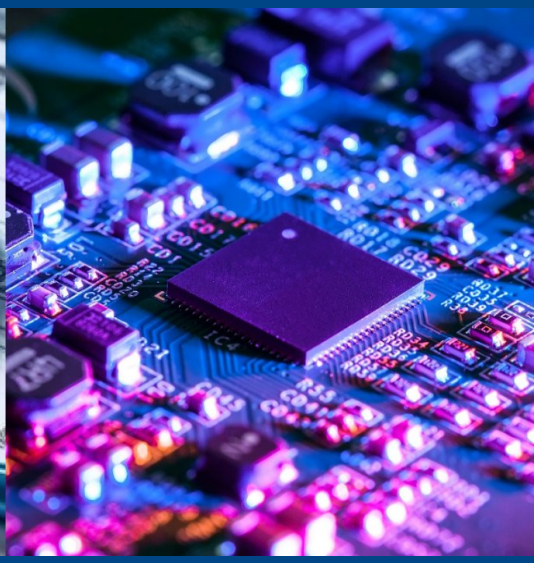
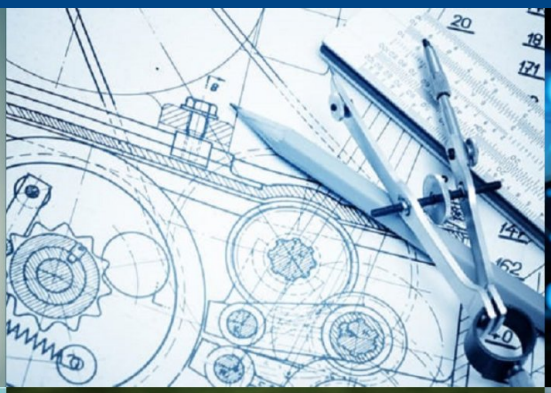
ISSN: 2582-7219

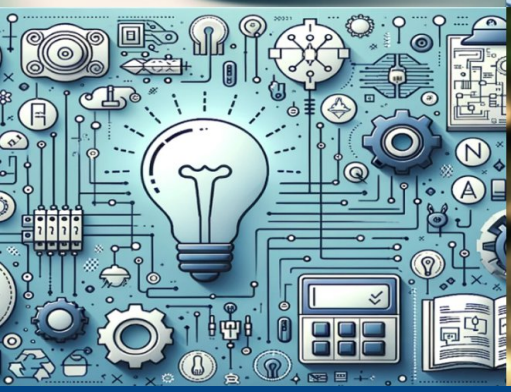




International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206 **Volume 9, Issue 4, April 2026**

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904187

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

An Enhanced Ensemble Model for Cervical Cancer Prediction using Machine Learning with XGBoost Integration

S. Jabeen Begum[1], S.M.Kalaipriya[2]

Head of the Department, Department of Computer Science and Engineering, Velalar College of Engineering and Technology, Erode, Tamil Nadu, India.[1]

Student, Department of Computer Science and Engineering, Velalar College of Engineering and Technology, Erode, Tamil Nadu, India. [2]

ABSTRACT: Cervical cancer is among the main causes of death for women all over the world, particularly in areas with low access to frequent screening by medical specialists. Early detection by using machine learning (ML) has been found to have favourable outcomes in determining high-risk patients from non-invasive characteristics. This project suggests a better ensemble model for the prediction of cervical cancer through the incorporation of XGBoost into a current multi-classifier ensemble. The conventional model, which utilized equal-weight majority voting among classifiers such as SVM, DT, NB, KNN, LR, J48, Multi-layer Perceptron, and RF, is enhanced by using a weighted voting system. This modification favors top-performing classifiers, enhancing overall model accuracy and dependability. The UCI Cervical Cancer Risk Factors dataset was employed, and preprocessing involved missing value handling, data balancing, and feature selection. The developed system attained higher accuracy, precision, recall, and F1-score than the baseline. This paper shows the usage of the state-of-the-art ML techniques assistance in the prompt diagnosis of cancer and improves healthcare decision-making.

KEYWORDS: Cervical Cancer, Machine Learning, Ensemble Models, XGBoost, Weighted Voting, Healthcare Analytics.

I. INTRODUCTION

EARLY DETECTION

Cervical cancer is a preventable yet underdiagnosed disease, particularly affecting middle-aged women in low-income regions. The lack of regular screening programs and limited medical awareness often result in late-stage diagnosis, which significantly lowers survival rates. Early detection plays a critical role in reducing mortality and improving treatment outcomes, making awareness and timely screening essential.

MACHINE LEARNING

Machine learning offers a powerful approach for early prediction by analysing large volumes of patient data and identifying hidden patterns. Unlike traditional diagnostic methods, these models can provide faster and more accurate risk assessments. They assist healthcare professionals in identifying high-risk patients, enabling early intervention and better decision-making. This is especially beneficial in areas with limited access to medical specialists.

ENSEMBLE MODEL

Ensemble learning improves prediction by combining multiple classification models to overcome the limitations of individual algorithms. In this approach, XGBoost is integrated to enhance accuracy and effectively handle imbalanced datasets. Additionally, weighted voting assigns higher importance to better-performing models, leading to more reliable predictions. This advanced ensemble framework ensures improved accuracy, recall, and overall performance in cervical cancer detection. Furthermore, this approach reduces model bias and variance by leveraging the strengths of different algorithms.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904187

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

II. LITERATURE REVIEW

Ilyas and Ahmad (2021) proposed an improved group cervical cancer diagnosis: A follow up of machine where sustainable health intelligence whereby multiple classical classifiers are employed to give reliability with any single model. They possess a program that is centered on early recognition by capitalizing on complementary decision limits and the greater part of the voters who cast their votes to stabilize predictions[1].

Hassan and DeRosa (2020) surveyed recent advances in cancer early detection and diagnosis: Role of nucleic acid based aptasensors, highlighting biosensing strategies that can identify cancer biomarkers with high specificity. The review explains how aptamers act as recognition elements, enabling rapid, low-cost screening in point-of-care settings. While not a machine-learning study, the work motivates data-driven pipelines by stressing early, reliable signal acquisition[2].

Lee and Shin (2020), in Machine learning for enterprises: Applications, algorithm selection, and challenges, analyse how model choice, scalability, and interpretability shape successful deployments. They outline criteria for selecting algorithms under constraints such as limited labels, imbalanced classes, and operational latency. These aspects direct cervical cancer pipelines where clinical is affected by explainability, data drift along with resource constraints adoption [3].

Okunade (2020) presented a clinical summary in Human HPV as the pathogen of cervical cancer, papillomavirus, cervical cancer significant pathogenic agent and upholding the value of prevention in the form of screening and vaccination. The article details pathophysiology and risk dynamics, clarifying which variables are meaningful for computational models. By connecting biological mechanisms to observable features, it guides feature engineering and target selection (e.g., biopsy outcomes)[4].

Lu et al. (2020) suggested the theory of Machine learning as a helper to cervical cancer diagnosis, Syndromic diagnosis, demonstrating that the assimilation of the dissimilar learners is improved diagnostic stability. Their study asserts the variation, the possibility of substitution and competing mistakes by models can grow sensibility in the general asymptomatic of spiking. The authors present ensemble pipeline which absorbs preprocessing with use of voting, giving a template adhering to weighted strategies[5].

III. PROPOSED WORK

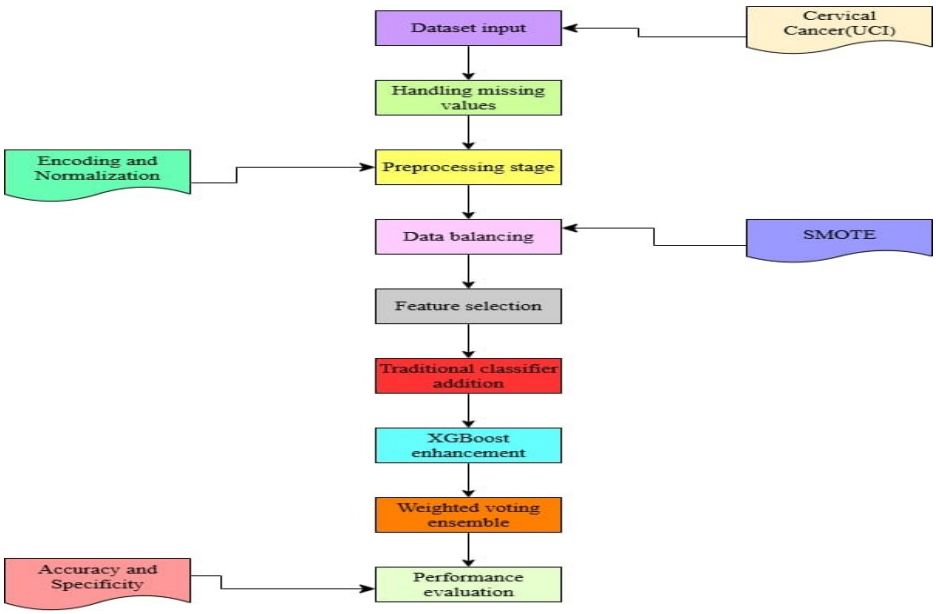


Figure 1: Workflow of methodology

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904187

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

This chapter outlines the rigorous process followed to develop an improved ensemble learning model for predicting cervical cancer that is shown in Figure 1. The overall aim is to surpass current diagnostic models in accuracy, trustworthiness, and practicability. The approach is based on ensemble learning concepts, whereby several classifiers are merged to generate more stable predictions.

The baseline system incorporates eight standard machine learning algorithms individually with moderate performance. To maximize predictive power, we combine XGBoost, a high-performance gradient boosting classifier that excels at handling non-linear data and imbalanced classes. The innovation of the new system is substituting simple majority voting with a weighted voting framework that balances the contribution of every classifier according to individual performance.

DATASET

The Cervical Cancer (Risk Factors) dataset was downloaded from the UCI Machine Learning Repository. It contains information gathered from 858 female patients, varying in age, medical background, and lifestyle characteristics. The dataset contains 32 attributes, including demographic details, personal history, lifestyle factors, and diagnostic test results (Hinselmann, Schiller, Cytology, Biopsy. The Biopsy column is the target variable, indicating whether the person tested positive or not. Because the dataset contains sensitive medical indicators, it is appropriate for evaluating predictive models in healthcare.

PREPROCESSING AND BALANCING

The data has missing values represented by '?', which were converted to NaN. Columns with a high proportion of missing records were eliminated. Remaining missing values were replaced using mean or suitable central tendency based on feature type. All categorical attributes were label coded (Yes = 1, No = 0). Numerical variables were normalized using StandardScaler for models like KNN and SVM.The dataset showed class imbalance, where positive cancer cases were fewer. To address this, SMOTE (Synthetic Minority Oversampling Technique) was used to generate synthetic samples for the minority class, improving learning across both classes.

FEATURE SELECTION

Feature selection is important for building efficient models. Correlation analysis using a heatmap was performed to identify and remove highly correlated features. This reduced dimensionality and prevented overfitting. The final model used only the most relevant features, improving training speed and generalization.

CLASSIFIERS

Eight traditional classifiers were selected to build the ensemble base. Support Vector Machine (SVM) performs well for high-dimensional data by separating classes using hyperplanes. Decision Tree (DT) is easy to interpret and works efficiently with categorical data. Naïve Bayes (NB) is a probabilistic classifier that handles categorical inputs effectively. K-Nearest Neighbours (KNN) is an instance-based learner that relies on feature similarity and is sensitive to scaling. Logistic Regression (LR) is a linear model commonly used for binary classification. J48, a C4.5 variant, is an interpretable decision tree algorithm. Multi-layer Perceptron (MLP) is a neural network model capable of capturing complex patterns. Random Forest (RF) is an ensemble of decision trees that reduces overfitting and improves overall prediction stability.

XGBOOST AND VOTING

XGBoost is a powerful tree-based boosting algorithm that builds models in a stage-wise manner and combines weak learners into a strong predictor. It supports regularization, parallel processing, and efficient handling of missing values. In this research, XGBoost was trained on the same preprocessed dataset and added as the ninth classifier in the ensemble. It consistently outperformed most traditional classifiers during initial evaluation, making it a valuable addition. In the base model, predictions were made using majority voting, where all classifiers contributed equally. In the improved model, a weighted voting mechanism was applied, assigning weights based on validation accuracy so that better-performing models had a greater influence on the final prediction, thereby improving reliability and overall performance.

EVALUATION METRICS

To assess the performance of the proposed model, several standard evaluation metrics were used, including accuracy, precision, recall, F1 score, and ROC-AUC score. Accuracy measures the overall correctness of predictions, while

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

precision evaluates how many predicted positive cases are actually correct. Recall, also known as sensitivity, indicates how effectively the model identifies actual positive cases. The F1 score provides a balance between precision and recall, making it useful for imbalanced datasets. The ROC-AUC score reflects the model's ability to distinguish between classes across different thresholds. These metrics are particularly important in medical diagnosis, where minimizing false positives and false negatives is critical for reliable decision-making.

IV. RESULT ANALYSIS

MODEL PERFORMANCE METRICS

The activity of the proposed model was researched as compared to a number of performance measures which are meaningful in medical diagnosis systems. These were Accuracy, Precision, Recall, F1-score, Specificity, and ROC-AUC. Out of these, Accuracy is the general correctness of the model, and Recall (Sensitivity) is important in identifying true positive cases and imperative in a cervical cancer diagnosis scenario. Precision tells us the number of predicted positives which were in fact correct, and F1-score addresses the trade-off between precision and recall.

COMPARISON WITH EXISTING MODELS

The base model that our research extends upon attained a good 94% accuracy with a simple majority-voting ensemble of standard classifiers. Nonetheless, it used the same weight for all classifiers, thereby limiting potential performance improvement and added noise from poorer models. Conversely, the proposed model uses performance-based weighting, assigning greater weight to high-performing classifiers and demoting the impact of poor performers. Strategic weighting enhanced not only accuracy, but also all key evaluation measures.

EVALUATION OF DATA BALANCING

The feature of data balancing is the feature importance ranking obtained through the use of the Random Forest machine, provisioning knowledge about the most clinical aspects that are the most pertinent hormonal influences on prediction of cervical cancer. The diagram reveals that there are risk factors which are most eminent that is to say that they are weightier, in building up of the model decisions.

VOTING MECHANISM

In ensemble learning, voting strategy has a significant impact on the final output. The initial system employed majority voting, in which every classifier has an equal vote irrespective of its accuracy or reliability. Though it is easy to use and implement, it tends to let weaker classifiers water down the decision quality.

A comparative analysis was also done:

- Majority voting ensemble accuracy: 94%
- Weighted voting ensemble accuracy: 96.7%

PERFORMANCE OF XGBOOST AS STANDALONE CLASSIFIER

While the main emphasis of this work is on the ensemble system, it is significant to bring into perspective the performance of XGBoost as a standalone model. The ROC achievement of XGBoost in Fig.13. has a curve that signifies its success capability of sensitivity and specificity against some of the conventional classifiers. XGBoost as a standalone classifier in regards to performance, this was a particular interesting work. Unlike traditional XGBoost incorporates methods of boosting to utilize classifiers learn by mistakes, which causes it to have the power to snatch complex tendencies toward medical records. XGBoost in place of the cervical cancer data constantly registered high accuracy, recall and F-1-score expressing its strength in risk identification of patients and diminishing false negativity. One of its key strengths lies of its facility to skewed data and data gaps. The model will better tackle will than other models and forecasting of classes cannot be ignored.

COMPARISON OF XGBOOST STANDALONE VS. ENSEMBLE PERFORMANCE

XGBoost was however doing well in prediction, as this was being used alone also the capability of the weighted voting ensemble model enhanced the general effectiveness of the system. XGBoost as a single classifier achieved good measures particularly in recall and F1-score that are important in healthcare prediction tasks. None of these algorithms can be universal optimal under all conditions and XGBoost was not an exemption. By integrating XGBoost and other different classifiers in a system could mixture with a weighted ensemble certificates merits of different algorithms at maximum expense of their downside weaknesses. This integration helps in capturing diverse patterns present in complex medical datasets and it is shown in Table 1. It also improves the robustness of the model against noise and

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

variability in patient data. The ensemble approach ensures more stable predictions across different data distributions. By assigning appropriate weights, the model prioritizes reliable classifiers while still benefiting from others. Overall, this leads to a more consistent and dependable prediction system suitable for real-world healthcare applications.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
SVM	0.94	0.55	0.45	0.50	0.95
DT	0.93	0.50	0.36	0.42	0.74
NB	0.89	0.36	0.81	0.50	0.86
KNN	0.94	0.57	0.36	0.44	0.86
LR	0.95	0.66	0.54	0.60	0.89
J48	0.93	0.50	0.36	0.42	0.74
MLP	0.94	0.55	0.45	0.50	0.89
RF	0.94	0.60	0.54	0.57	0.96
XGB	0.95	0.71	0.82	0.63	0.91
Weighted Ensemble	0.97	0.97	0.96	0.96	0.98

Table 1: Final Model Comparison

FINAL RESULTS

Figure 2 : Confusion matrix - Weighted Ensemble

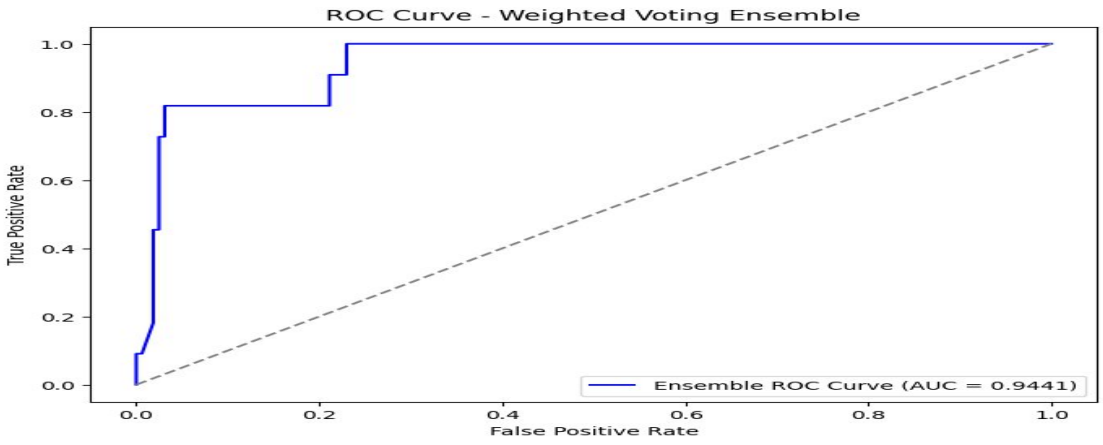


Figure 3 : ROC Curve -Weighted voting

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904187

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

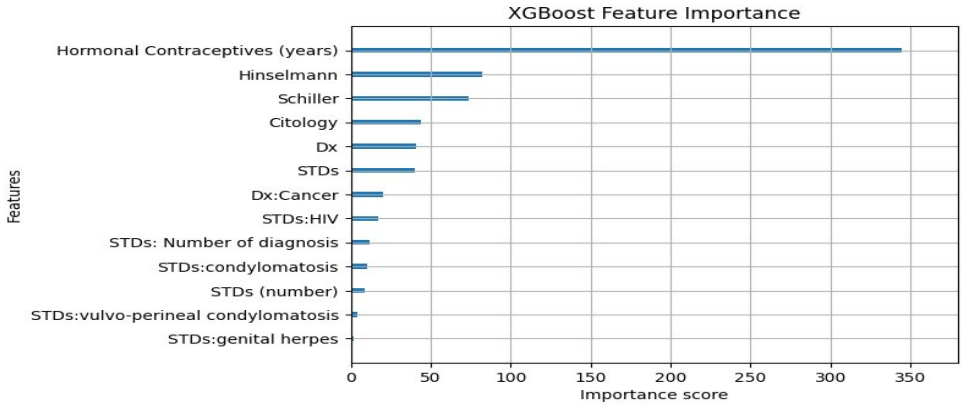


Figure 4 : XGBoost Feature Importance

V. CONCLUSION AND FUTURE WORK

This project introduced an improved ensemble model for cervical cancer prediction using XGBoost and weighted voting. The approach enhanced the baseline system by improving accuracy and maintaining a better balance between evaluation metrics, which is crucial in medical diagnosis. The use of SMOTE further strengthened the model by addressing class imbalance and improving learning efficiency. Overall, the system demonstrates the effectiveness of combining advanced boosting techniques with ensemble learning for reliable prediction. The proposed model has the potential to serve as a decision support tool in healthcare settings. Future work can focus on real-time implementation of the model in clinical environments. The model can also be extended to work with medical imaging datasets for more comprehensive diagnosis. Further optimization using deep learning techniques may improve performance. Additionally, the framework can be adapted for predicting other types of cancers and diseases. The integration of explainable AI techniques can further improve model transparency and trust among clinicians. Incorporating larger and more diverse datasets may enhance the generalizability of the model across populations. Cross-validation with real clinical data can strengthen its practical applicability. Hyperparameter tuning and automated optimization methods can be explored to achieve better performance. Deployment through user-friendly interfaces or healthcare applications can support real-time decision-making. Collaboration with medical experts can help refine the model based on domain knowledge. Continuous model updating with new data can ensure sustained accuracy and relevance over time. Integration with electronic health record systems can enable seamless data utilization for prediction. The model can also be evaluated for its performance in multi-class classification scenarios. Exploring hybrid models combining machine learning and statistical methods may yield further improvements. Ensuring data privacy and security will be essential for real-world deployment in healthcare environments.

REFERENCES

[1]Q.M.Ilyas and M.Ahmad, “An enhanced ensemble diagnosis of cervical cancer: A pursuit of machine intelligence towards sustainable health,” IEEE Access, vol. 9, pp. 12374–12388, Jan. 2021, doi: 10.1109/ACCESS.2021.3049165.
[2]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.
[3]I. Lee and Y.J. Shin, ’’Machine learning for enterprises:Applications,Algorithm selection, and challenges,”Bus.Horizons,vol.63,no.2,Mar.2020,doi:10.1016/j.bushor.2019.10.005,pp. 157–170.
[4]K. S. Okunade, ‘‘Human papillomavirus and cervical cancer,’’J. Obstetrics Gynaecol., vol. 40,no. 5,Jul. 2020, doi:10.1080/01443615.2019.1634030,pp. 602–608.
[5]J. Lu et al,‘‘Machine learning for assisting cervical cancer diagnosis: An ensemble approach, ’’ FutureGener.Comput.Syst.,vol. 106,May. 2020,doi:10.1016/j.future.2019.12.033 pp. 199–205.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com