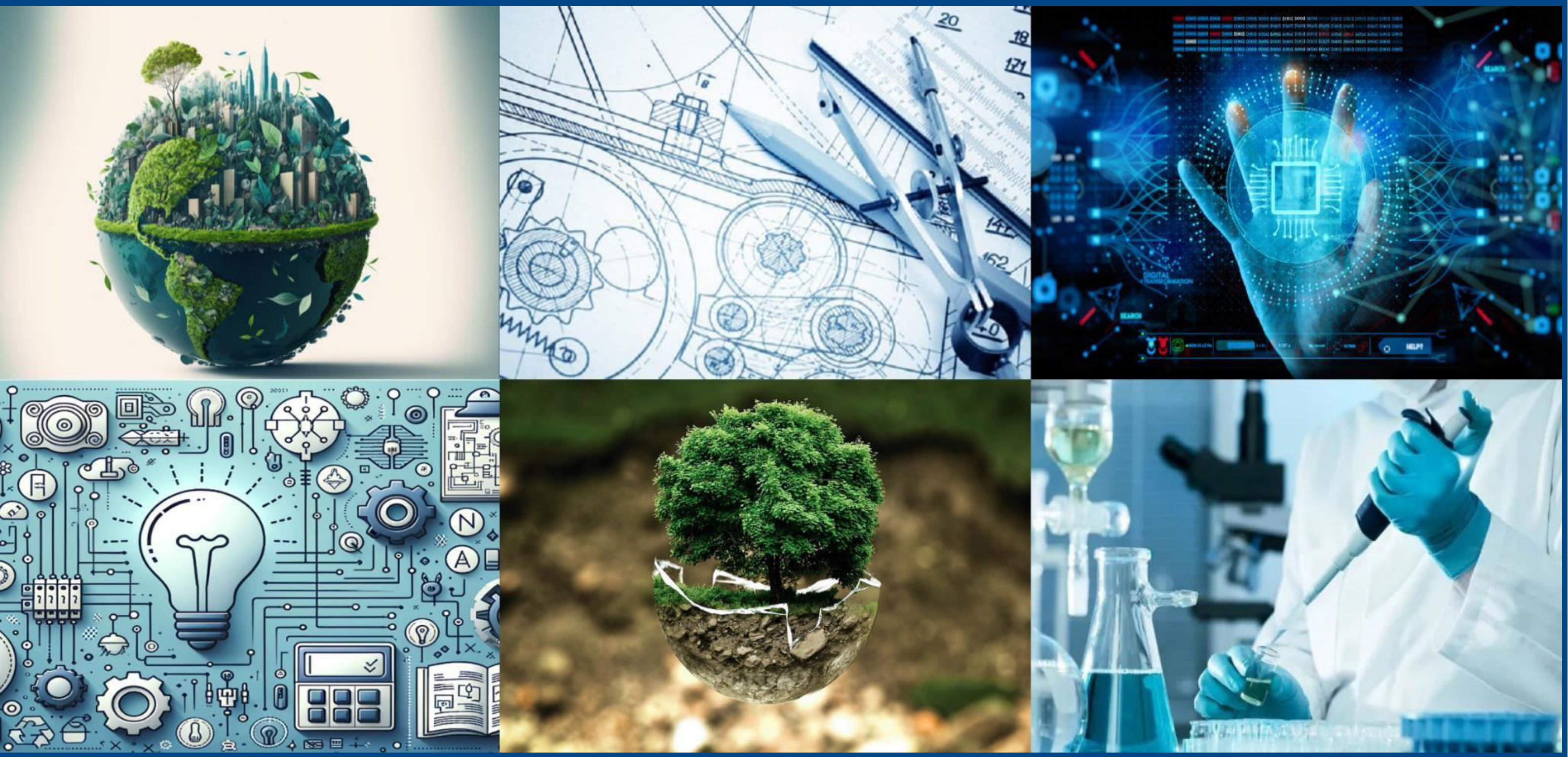
ISSN: 2582-7219

# 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 8, Issue 8, August 2025

© 2025 IJMRSET | Volume 8, Issue 8, August 2025| DOI:10.15680/IJMRSET.2025.0808038

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)

# A DATA- INTELLIGENT APPROACH TO CERVICAL CANCER FORECASTING

**Dr. M S.Shashidhara, Agasanoor Kavya**

Professor & HOD, Department of MCA, AMC Engineering College, Bengaluru, India

Student, Department of MCA, AMC Engineering College, Bengaluru, India

**ABSTRACT:** As more people and organizations turn to machine learning (ML) and deep learning (DL) to analyze enormous amounts of data into usable insights, predicting the early stages of serious illnesses such as cancers, kidney failure, and heart attacks continues to be incorporated into ML-based schemes used to support clinical practice. Cervical cancer is one of the most common diseases among women, and ML-based schemes could be a way of diagnosing cervical cancer earlier and preventing it. In this way, this study is a clever way to forecast cervical cancer with ML algorithms. The proposed methodology for the research comprises four stages: research dataset, data pre-processing, predictive model selection (PMS), and pseudo-code. Within the PMS section we report experiments with a variety of traditional machine learning models: decision tree (DT), logistic regression (LR), support vector machine(SVM), K-nearest neighbors algorithm (KNN), adaptive boosting, gradient boosting, random forest, and XGBoost classifier.

**KEYWORDS:** machine learning(ML),Cervical cancer, human papillomavirus[HPV],XG **B**oost classifier, random forest(RF).

## I. INTRODUCTION

The challenges associated with human life are complicated by the fact that it is uncertain when problems occur. Generally, though, women normally encounter several difficulties in a lifetime. One of these difficulties may have potentially harmful consequences is The ectocervix is the portion of the cervix which can be visualized during a gynecologic exam., and consists of a flat group of thin cells called squamous cells. The endocervix is the inner part of the cervix which consists of a canal that leads into the vagina making way to the uterus. The endocervix is covered with column shaped glandular cells that produce mucus.

**OBJECTIVES:**

- To analyse and categorize cervical cancer using machine learning algorithms to assist clinicians in accurate cancer diagnosis
- To find the relationships and connections (between) the factors that may contribute to cervical cancer.
- Early detection and diagnosis of cervical cancer could help avert the disease.
- Machine learning can provide useful information for public health by exploring potential risk factors, patterns, and prevention initiatives.
- To enhance awareness and stimulate public interest in cervical screening and cervical cancer prevention.

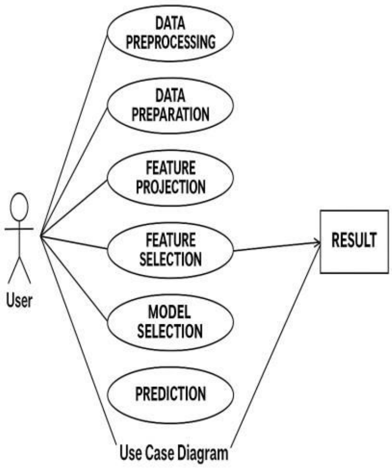
## II. LITERATURE SURVEY

This section discusses the literature selection criteria (LSC), as well as the papers collected to discuss the literature in each of the databases. The literature selection criteria (LSC) section illustrates how we selected the related papers based on the selection criteria, after collecting the papers from the databases.

- The time duration being surveyed is from 2010 to 2021. It will be important to ascertain the studies from the past, We did not include any Project work if it has not been printed and is not peer-reviewed.

The authors performed a survey-based study on cervical cancer detection, including performance evaluation to determine the accuracy of specific types of architecture in an artificial neural network (ANN), with the ANN used to recognize cancerous, normal, and abnormal cells. The authors used Cervi gram images to show a method for screening cervical cancer using the oriented local histogram technique (OLHT) which improves edges, and the dual-tree complex wavelet transform (DT-CWT) which can improve multi-resolution images. The authors used a UCI data repository and six machine learning (ML) classifiers to propose a model that identifies the cervix infection precisely. They proceeded with data pre-processing, which was verified by the physician to extract some features and validation. Then they carried out the study, by using 10-fold cross-validation to evaluate performance of the proposed model. Another significant study was published, which used machine learning classifiers (SVM, QUEST, C&R tree, MLP). The study evaluated various measures such as accuracy, sensitivity, specificity and area under the curve (AUC). The QUEST values were: 95.55%, 90.48%, 100% and 95.20%, respectively of the parameters. This Project used a federated learning method to devise a method of machinery failure diagnostics in order to solve the data island problem. Each participant's model training is done locally with a self-supervised learning approach, intended to sophisticate the learning process.

Five different machine learning approaches were employed including random forest, KNN, C5.0, SVM and RP-art. After training all the classifiers (i.e. C5.0, RF, RPART, SVM and KNN) and an assessment of their performance, we examined the classifiers with respect to accuracy showing 97%, 96.9%, 96%, 88%, and 88%. Various machine learning (ML) algorithms (i.e. decision tree, random forest, and logistic regression) were used together with the voting model. In one instance cervical cancer was detected using four target parameters (i.e. biopsy, cytology, Schiller, and Hanselmann) with 32 risks factors, collected from the University of California (UCI). It was found that the decision tree algorithm value was higher (i.e.98.5%) than the decision jungle model. In another study, the appropriate data mining technique used was from the boosted decision tree, decision forest to detect cervical cancer, using the Microsoft Azure ML tool. The author's research results were reported in terms of accuracy, area under receiver operating characteristic (AUROC) curve, specificity, and sensitivity after the output had applied 10-fold cross validation, which improved performance on the decision tree algorithm to 97.8% on the AUROC curve. The authors reported that in total, 794 women (88.2%) had heard of the condition, the most frequent monitor was the radio at 557 women (70.2%) and the least was health care organizations at 120 women (15.1%). The research also indicated genetic assistance as a potential method to enhance the quality of the prediction. We also explored a methodology based on machine-learning approaches to detect cardiac disease. We used classification algorithms to build the system. The model applied conditional mutual information feature selection approach to resolve the feature selection dilemma. Feature selection methods can enhance classification accuracy and decrease development time of the classification system. The findings have shown that the proposed machine-learning-based system might score as much as 86% for diagnostic accuracy of DL. Healthcare practitioners and other stakeholders collaborated to develop classification models that could assist with diabetes prediction and design prevention strategies. Another research has been completed, where a methodology for heart disease was created from UCI repository datasets and health monitors to estimate the public's risk for heart disease. Using classification algorithms for classifying patient data to identify cardiac disease, including boosted decision tree and decision forest.



Use Case Diagram

## III. SYSTEM ARCHITECTURE

The architectural framework for cervical cancer prediction using machine learning consists of a set of modular components that together support the end-to-end pipeline of intelligent clinical decision support. This architecture provides fluidity in data flow, efficient model inference and solid system deployment..

The system begins with the Data Acquisition Layer, which is the first point of contact when obtaining data relevant to the patient. This data can include structured clinical data; demographic information; results of HPV and Pap smear tests; and lifestyle behaviours. The input data can come from hospital databases, public health data bases, and patients through digital input portals (e.g., through patient records). Once this data is acquired, it is taken to the Data Preprocessing Layer, where the data is processed. The Department of Preprocessing does several processing activities, including missing values imputation, duplicate records removal, categorical variable encoding, and normalization of feature distributions. Then some dimensionality reduction methods and feature selection techniques are applied to retain the features showing the most information; providing better model generalization and processing speed. Finally, the data is taken to the Machine Learning Layer, representing the predictive part of the system. When processing data, multiple classification algorithms, such as Logistic Regression, Random Forest, Support Vector Machine (SVM), or Gradient Boosting Machines are trained and validated, and including hyperparameter tuning and cross-validation could improve model performance. Predictive reliability is evaluated using metrics often including precision, recall, F1 score, and the Area Under the Curve (AUC).Once the best model is identified, we then turn on the Inference or Prediction Layer for real-time predictions. This layer takes new patient data as input and outputs a binary or probabilistic risk score as to the likelihood of cervical cancer presence or progression. The model might also produce confidence scores or explainability scores (such as SHAP scores) for clinical interpretability.

## IV. METHODOLOGY

The proposed Project methodology is divided into different components: Project dataset, data pre-processing, predictive model selection (PMS), and training methods. Figure 1 provides an architectural diagram of the proposed Project; looking at Figure 1, it is apparent that the architectural diagram has been divided into 4 sections, since the model presented in this Project has requisite tasks performed with each module.

The Project data collection is described in the Project Data Set section. The Data Pre-processing section describes ways to remove noise from the dataset and to make that dataset useful to feed into machine learning. The type of predictive model chosen to predict cervical cancer in this Project is outlined in the PMS section. The requirements for training the model are outlined in the Training Methods section. Lastly, we design the platform to present an overview of a complete pipeline of cervical cancer prediction, using the Python programming language.

This Project implements an algorithm which is friendlier to the classification of negative and positive cervical cancer diagnostic for clinical purposes. Cervical cancer can be diagnosed utilizing algorithms: decision tree, logistic regression, support vector machine (SVM), K-nearest neighbours (KNN), adaptive boosting, gradient boosting, random forest, and XGBoost. The stages and transitions are detailed in the following sections. The proposed ML-based model is shown in Figure 1. At the outset of the model development, initial data will be presented to the model as training data. Then, ML algorithms are adopted. Afterward, the model input data, along with a new input data, will be entered in to the scheme in a way to allow training of the architecture. Finally, prediction is accomplished on the new incoming data.

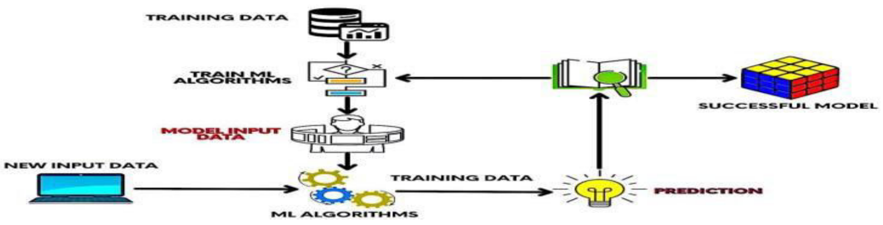


**Fig: Proposed Project Model for classifying cervical cancer**

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)

## V. DESIGN & IMPLEMENTATION

Designing and developing a predictive system for cervical cancer using machine learning (ML) techniques follows a sequential, multi-stage framework comprising of data collection, data preprocessing, model development, model validation, and system delivery. The goal is to build a robust, data-driven design that streams and identifies patients at high risk of developing cervical neoplasia so that patients can receive early intervention, and potentially improve clinical outcomes.

The first step involves systematic access to clinical and behavioural datasets from reputable repositories such as the UCI Machine Learning Repository. These datasets usually include a variety of attributes, and represent a diversity of variables and cohorts. For example, the attributes include sex, age, sexual behaviour, contraceptive use, number of births, use of tobacco, sexually transmitted infections (STIs) history, and history of cytology screenings. These variables are the input feature space to train the model.

Following this, we have a process of substantial preprocessing in terms of ensuring the data is clean and ready to be analysed. This can involve imputing mathematical value for missing values with someone valued statistical technique, filtering noise and outliers, categorical encoding, and either normalizing, or standardizing, numerical features to create a level scale distribution with our numerical features. We will then use feature engineering techniques, particularly Recursive Feature Elimination (RFE) and mutual-information analysis, to create and remove feature variables dependent on the core importance of these predictors and quality of the model, and then reduce our dimensionality post-processing to state why the model is as good as is it. After you preprocessing has been performed, we split the data in a stratified manner, i.e., by using a stratified training and testing data, usually, an 80:20 ratio in order to make sure both datasets adequately represent the data. We built predictive models using supervised classifications algorithms and including Logistic Regression, Support Vector Machines (SVM), Random Forests, K-Nearest Neighbors (KNN), and the Gaussian Naive Bayes function. Using k-fold with all of these models, compounds the quality of our models and feed and test these datasets on an aforementioned suite of supervised classification methods by factoring in accuracy, sensitivity, specificity, positive predictive values to tune the models. We drive the performance of the models after tuning.

A thorough process of metrics quantification occurs for evaluating the predictive performance of the trained models including accuracy, precision, recall, F1-score, and Area Under the Curve Receiver Operating Characteristic (AUC-ROC). Furthermore, confusion matrices are produced to evaluate how many true positives, true negatives, false positives, and false negatives exist (this is increasingly important for medical diagnosis where the consequences of misclassification can be clinically important and costly). Once the best-performing model is confirmed, it is incorporated into a structure that is scalable and conveniently presented to users, typically an interactive software environment—for example, the web interfaces created using lightweight web applications Flask or Django. The interactive software environment allows predictions based on health parameters entered by users, thereby providing a source of decision support for health practitioners and entry access for patients. The systems are designed with an emphasis on modularity, security, and usability for clinical or remote health use.

Finally, the optimized model will be incorporated into an interactive software environment that is usually a web-based framework constructed using Flask or Django. The web-interface allows healthcare providers and patients to enter health parameters and receive immediate evidence-based predictions for increased vulnerability toward cervical cancer.

## VI. OUTCOME OF RESEARCH

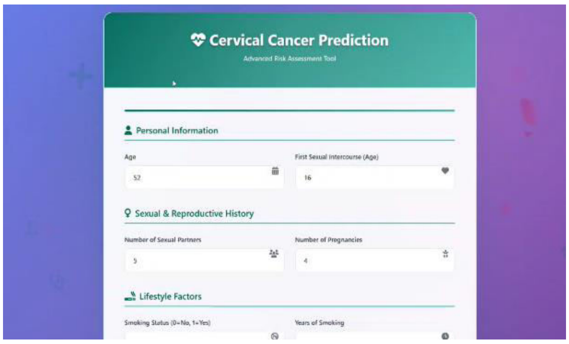
The result of this study demonstrates the potential of machine learning code as robust enabler of early detection and prognosis of cervical cancer. By applying specific advanced data pre- processing, feature selection, and machine learning classifiers, the developed system was able to achieve and demonstrate excellent predictive accuracy, with reliable classification performance, and appropriate to be used in real world medical.

**Quantitative results of the model outputs**

This shows that the predictive model selected - iteratively built through rigours parameter tuning and validation processes - obtained strong performance in the areas of key metrics of accuracy, sensitivity, specificity, F1-score and area under the ROC curve, which confirms that the model is reliably able to identify high risk individuals with the least number of false predictions. This is especially important for medical diagnostics.

## VII. RESULT AND DISCUSSION

Based on the findings of this research, it can be stated that the objectives of this paper have been achieved. Its research methodology was enriched with a set of algorithms including decision tree (DT), logistic regression (LR), support vector machine (SVM), K-nearest neighbors (KNN), adaptive boosting, gradient boosting, random forest (RF), and XGBoost. The research has reached a satisfactory result for both predictions and classification. It is a matter of great concern that this work has not been accomplished much in previous research using gradient boosting algorithms. Since the gradient boosting algorithm also follows the sequential ensemble learning method, the wave learners gradually get better than their previous wave learners through this method of loss optimization.



Cervical Cancer Prediction

Advanced Risk Assessment Tool

Personal Information

Age: 52

First Sexual Intercourse (Age): 16

Sexual & Reproductive History

Number of Sexual Partners: 5

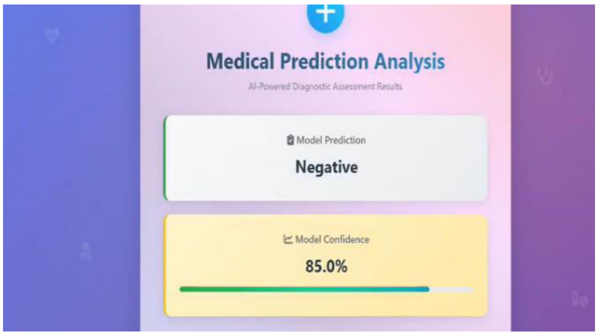
Number of Pregnancies: 4

Lifestyle Factors

Smoking Status (0=No, 1=Yes)

Years of Smoking

It is essential to point out that the researchers did not restrict their effort to simply developing the model; rather, they also validated and evaluated the model's performance. Several validation strategies, including ROC-AUC, confusion matrix, and cross-validation, were applied by the researchers, and the researchers found that the efficacy with respect to cervical cancer is adequate. During the preprocessing phase, some aspects of the patients' samples, such as the length of time they drank alcohol and their HIV and HSV2 infection status, revealed that factors whose samples had undergone modest variations could not be considered accurate predictors. However, with the help of this machine learning model, women have the opportunity to benefit from knowing more about cervical cancer and what effect it has on the human body. This study will focus on women in order to identify which symptoms or parameters are important for identifying for cervical cancer, as well as the causes and effects of these symptoms and parameters. First of all, the DT algorithm is very unstable, which means that a slight change in the data will significantly change the layout of the best decision tree. It is insufficiently reliable. With similar data, several other predictors perform better. Second, this study faced massive problems while dealing with the dataset, because numerous data have been enumerated and interpreted in the data pre-processing stage. The model will provide an optimum result only if a considerable number of data-processing techniques have been adopted. Third, the survey data have been preserved to apply machine learning to conduct sentiment analysis regarding cervical cancer, but in this study, the researchers could not accommodate different data-processing techniques to apply.



Medical Prediction Analysis

AI-Powered Diagnostic Assessment Results

Model Prediction

Negative

Model Confidence

85.0%

## VIII.CONCLUSION

Early detection increases the likelihood of successful treatment in the pre-cancer and cancer stages. Being aware of any signs and symptoms of cervical cancer can also aid in avoiding diagnostic delays. This research has focused on cervical cancer using conventional machine learning (ML) principles and several traditional machine learning algorithms, such as decision tree (DT), logistic regression (LR), support vector machine (SVM), and K-nearest neighbors (KNN). In terms of cervical cancer prediction, the highest classification score of 100% has been achieved with the random forest (RF), decision tree (DT), adaptive boosting, and gradient boosting algorithms. In contrast, 99% accuracy has been found with SVM. The results of these algorithms are applied to identify the most relevant predictors. We have received satisfactory accuracy compared to the support vector machine algorithm. The findings of this study revealed that the SVM model could be used to find the most important predictors. As the number of essential predictors for analysis decreases, the computational cost of the proposed model decreases. The disease can be predicated more accurately with the use of machine learning. Furthermore, boosting patients' personal health and socio-cultural status can lead to cervical cancer prevention.

## REFERENCES

**[1]** Martin C.M., Astbury K., McEvoy L., Toole S., Sheils O., Leary J.J. Inflammation and Cancer. Volume 511. Springer; Berlin, Germany: 2009. Gene expression profiling in cervical cancer: Identification of novel markers for disease diagnosis and therapy; pp. 333–359. [DOI] [PubMed] [Google Scholar]

**[2]** Purnami S., Khasanah P., Sumartini S., Chosuvivatwong V., Sriplung H. Cervical cancer survival prediction using hybrid of SMOTE, CART and smooth support vector machine. AIP Conf. Proc. 2016;1723:030017. [Google Scholar]

**[3]** Yang X., Da M., Zhang W., Qi Q., Zhang C., Han S. Role of lactobacillus in cervical cancer. Cancer Manag. Res. 2018;10:1219–1229. doi: 10.2147/CMAR.S165228. [DOI] [PMC free article] [PubMed] [Google Scholar]

**[4]** Ghoneim A., Muhammad G., Hossain M.S. Cervical cancer classification using convolutional neural networks and extreme learning machines. Future Gener. Comput. Syst. 2020;102:643–649. doi: 10.1016/j.future.2019.09.015. [DOI] [Google Scholar]

**[5]** Rehman O., Zhuang H., Muhamed Ali A., Ibrahim A., Li Z. Validation of miRNAs as breast cancer biomarkers with a machine learning approach. Cancers. 2019;11:431. doi: 10.3390/cancers11030431. [DOI] [PMC free article] [PubMed] [Google Scholar]

**[6]** Ashok B., Aruna P. Comparison of Feature selection methods for diagnosis of cervical cancer using SVM classifier. Int. J. Eng. Res. 2016;6:94–99. [Google Scholar]

INTERNATIONAL JOURNAL OF
MULTIDISCIPLINARY RESEARCH
IN SCIENCE, ENGINEERING AND TECHNOLOGY

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

www.ijmrset.com