IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 5, May 2022

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.118

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 5, May 2022||

| DOI:10.15680/IJIRSET.2022.1105322|

Breast Cancer Detection using Logistic Regression Algorithm

Prof. Ajit N.Gedam[1], Kajol B. Deshmane[2], Nishigandha N.Jadhav[3], Ritul M.Adhav[4], Akanksha N.Ghodake[5]

Dept. of Computer Science, AISSMS Polytechnic, Pune, MH, India

ABSTRACT: Breast cancer is one of the common diseases specifically in women now days. It has become the second main reason of cancer death in females. Every year 4.5-5% new cancer cases are recorded and increasing the morbidity at worldwide. It has proved that early detection of any has proved that early detection of any cancer when followed up with appropriate diagnosis and treatment can increase the survival rate of the patients. Breast cancer is diagnosed by mammography. Mammograms are films generated by radiologist with a device. These mammograms are observed and diagnosed by the oncologist for further treatment. Since all general hospitals do not have the specialist and patients used to wait for their report. So, waiting for diagnosing a breast cancer may take time. This delay may be responsible for cancer spreading and reducing the survival rate of the patient. This does not mean to replace expert or physician by computer but it means that computer can assist the expert for better understanding the particular case and the results can be produced early. This paper presents a brief summary on breast cancer diagnosis using Logistic Regression machine learning algorithms used to increase the efficiency and effectiveness of predicting cancer. The correct diagnosis and accurate classification of breast cancer detection are the main objective of this paper.
diagnosis and accurate classification of breast cancer detection are the main objective of this survey paper.

KEYWORDS: Breast Cancer, Mammograms, machine learning algorithms, Logistic Regression.

I. INTRODUCTION

Over the years, a continuous evolution related to cancer research has been performed. Scientists used various methods, like early-stage screening, so that they could find different types of cancer before it could do any damage. With this research, they were able to develop new strategies to help predict early cancer treatment outcome. With the arrival of new technology in the medical field, huge amount of data related to cancer has been collected and is available for medical research. But physicians find the accurate prediction of the cancer outcome as the most.

In this project, I will be using different machine learning techniques to analyse the data given in the datasets. This analysis will help us predict whether the cancer is benign or malignant. Benign cancer is the cancer which doesn't spread whereas malignant cancer cells spread across the body making it very dangerous.

Machine learning algorithms will help with this analysis of the datasets. These techniques will be used to predict the outcome. The outcome can be either that the cancer is benign or malignant. Benign cancer is the cancer which doesn't spread whereas malignant cancer cells spread across the body making it very dangerous.

II. PROBLEM STATEMENT

Over the years, a continuous evolution related to cancer research has been performed. Scientists used various methods, like early-stage screening, so that they could find different types of cancer before it could do any damage. With this research, they were able to develop new strategies to help predict early cancer treatment outcome.

But physicians find the accurate prediction of the cancer outcome as the most interesting yet challenging part. For this reason, machine learning techniques have become popular among researchers. Patients have to spend a lot of money on different tests and treatments to check whether they have breast cancer or not. These tests can take a long time and the results can be delayed. Also, after confirmation that the patient has cancer, more tests need to be done to check whether the cancer is benign or malignant.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 5, May 2022||

| DOI:10.15680/IJIRSET.2022.1105322|

In this project, we will be using machine learning techniques to analyse the data given in the datasets. This analysis will help us predict whether the cancer is benign or malignant. Benign cancer is the cancer which doesn't spread whereas malignant cancer cells spread across the body making it very dangerous.

Logistic Regression Machine learning algorithm will help with this analysis of the datasets. These techniques will be used to predict the outcome. The outcome can be either that the cancer is benign or malignant.

III. MOTIVATION

Breast cancer is cancer that develops in breast cells. Typically, the cancer forms in either the lobules or the ducts of the breast. Cancer also can occur within the adipose tissue or the fibrous connective tissue within your breast. The uncontrolled cancer cells often invade other healthy breast tissue and may visit the lymph nodes under the arms. Doctors say that breast cancer happened due to abnormal growth of cells in the breast and these cells spread in size like Meta Size from breast to lymph nodes or the other parts of the body also. Hence it is necessary to detect and stop the growth of these unwanted cells as early as possible to avoid the next phase consequences.

IV. METHODOLOGY

- **Logistic model:**

Logistic regression was introduced by statistician DR Cox in 1958 and so predates the field of machine learning. It is a supervised machine learning technique, employed in classification jobs (for predictions based on training data). Logistic Regression uses an equation likeLinear Regression, but the outcome of logistic regression is a categorical variable whereas it is a value for other regression models. Binary outcomes can be predicted from the independent variables.

The general workflow is:

(1) get a dataset
(2) train a classifier
(3) make a prediction using such classifier

Main working of model:

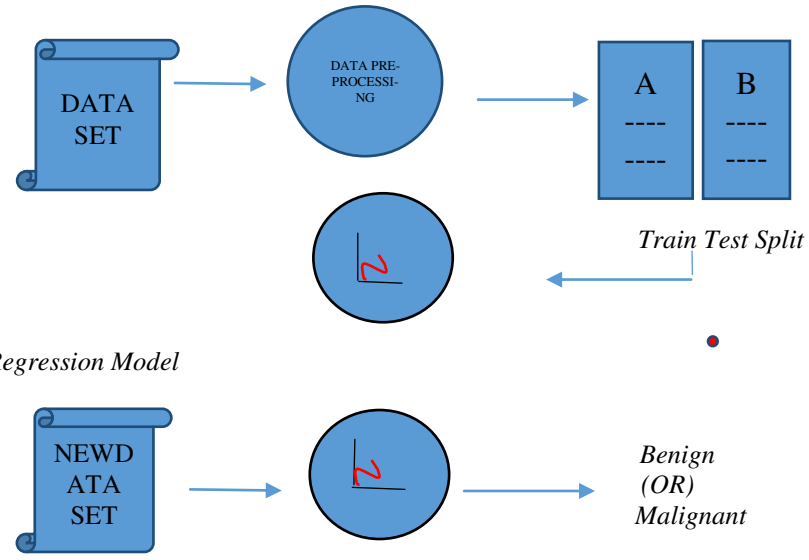

Fig: 1.1Work flow diagram for breast cancer detection

Step 1:Data Set

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 5, May 2022||

| DOI:10.15680/IJIRSET.2022.1105322|

Collecting the Data Set of breast cancer patients.

Data Set Characteristics:

Number of Instances
569

Number of Attributes
30 numeric, predictive attributes and the class

Attribute Information

- radius (mean of distances from centre to points on the perimeter)
- texture (standard deviation of Gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness (perimeter^2 / area - 1.0)
- concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- symmetry
- fractal dimension ("coastline approximation" - 1)

The mean, standard error, and "worst" or largest (mean of the three worst/largest values) of these features were computed for each image, resulting in 30 features. For instance, field 0 is Mean Radius, field 10 is Radius SE, field 20 is Worst Radius.

- **class:**
  - WDBC-Malignant
  - WDBC-Benign

Step 2: Data Pre-processing

Move the dataset to Data Pre-processing Stage.

Where, Data pre-processing is a data mining technique which is used to transform the raw data in a useful and efficient format. The dataset provided is already clean and does not have any missing values. As a part of the pre-processing stage, the data is standardized. It transforms the attributes to normal distributionand in this process the data set will be compatible for python language.

Step 3:Train-Test Split
Move the data towards Train-Test split Stage.Here, we split our data into training data and testing data for comparing the accuracy of the algorithm.

We have many records. If we use the entire data for model building, we will not be left with any data for testing. So generally, we split the entire data set into two parts, say 70/30 percentage. We use 70% of the data for model building and the rest for testing the accuracy in prediction of our created model.

Step 4: Use Logistic Regression Model.
Step 5: Train the logistic regression model for using it with dataset.
Step 6: Prediction
Insert the new data into Trained Logistic regression model.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 5, May 2022||

| DOI:10.15680/IJIRSET.2022.1105322|

Here, we can choose any new random value from the data set which will be tested in the trained logistic regression algorithm, where it will help us to predict whether the patient has Benign or Malignant type of cancer.

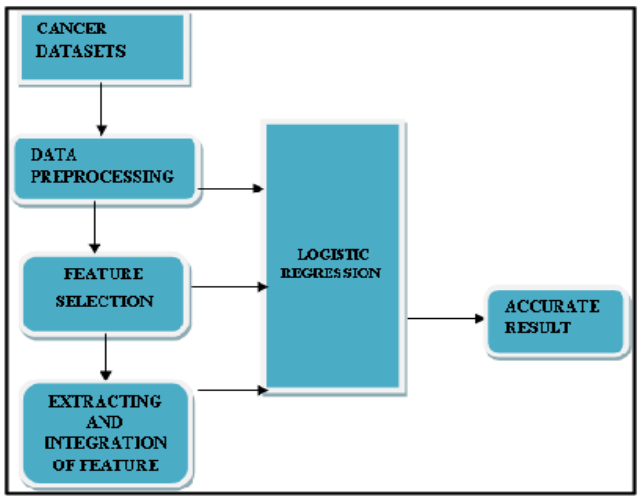

Fig2. Architecture Diagram of Logistic Regression

So according to the about work flow diagram first one data set will be used, after that the data set will move towards data pre-processing (Data pre-processing is a data mining technique which is used to transform the raw data in a useful and efficient format) and in this process the data set will be compatible for python language, and the data set will be trained.

After this data is pre-processed, we can use the data set in the logistic regression algorithm which is the machine learning algorithm. Then we can choose any new random value from the data set which will be tested in the trained logistic regression algorithm where it will help us to detect whether the patient has Benign or Malignant type of cancer.

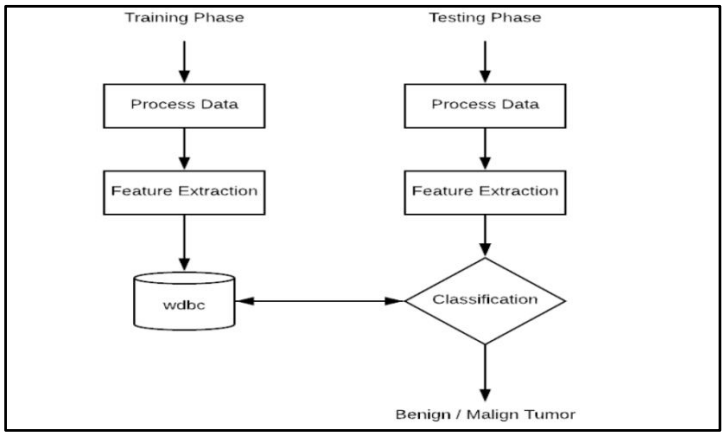

Fig3. Flow diagram for breast cancer detection.

V. LIBRARIES

1. Scikit-learn in Python:

Scikit-learn (formerly scikits-learn and also known as sklearn) is a free software machine learning library for the Python programming language.[3] It featuresvarious classification, regression and clustering algorithms including support-vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy. Scikit-learn is a NumFOCUS fiscally sponsored project.

2. NumPy:

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 5, May 2022||

| DOI:10.15680/IJIRSET.2022.1105322|

NumPy is an abbreviation of **Numerical Python**. It is one of the most fundamental and powerful python libraries to create and manipulate numerical objects. The basic purpose of designing the NumPy library was to support large multi-dimensional matrices. It helps to perform high-level mathematical functions and complex computations using single and multi-dimensional arrays. NumPy provides innumerable features that reduce the complicated tasks of data analytics, data scientists, researchers, etc.

3. Pandas:

Pandas is an abbreviation for Python Data Analysis Library. It is an open-source library specially designed for data analysis and data manipulation in Python. Pandas is built on the top of the NumPy package and hence it fundamentally relies on NumPy.

Pandas enable us to read from multiple sources such as Excel, CSV, SQL, and many more. Basically, Pandas possess two types of data objects:

1. Pandas Data Frame: It is a mutable two-dimensional data structure with labelled rows and columns which are generally compared with excel and SQL sheets.
2. Pandas Series: It is a One-dimensional labelled array to store the heterogeneous data elements generally compared with the columns in MS Excel.

VI. RESULT

Fig3.Output of code

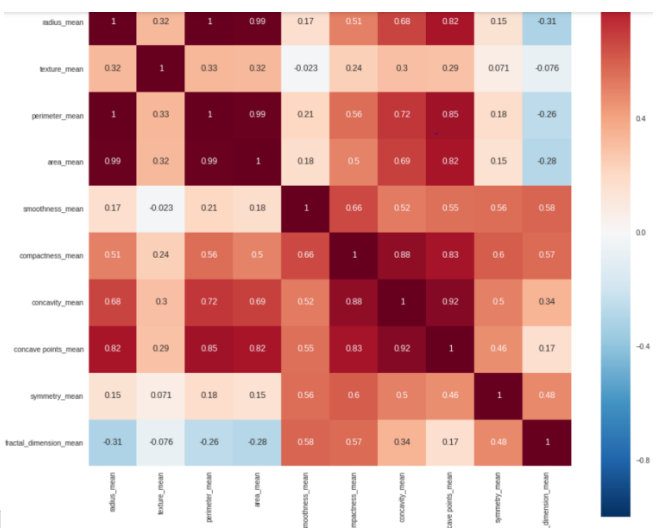

Fig4. Output in heatmap form

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 5, May 2022||

| DOI:10.15680/IJIRSET.2022.1105322|

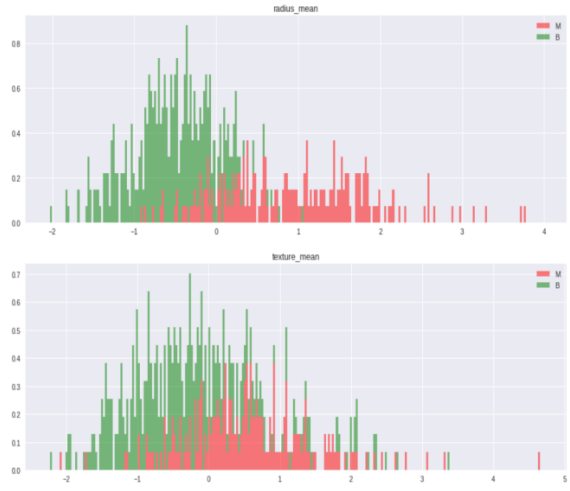

Fig 5. Graph of radius_mean and texture_mean.

Fig6. Graph of perimeter_mean and area_mean

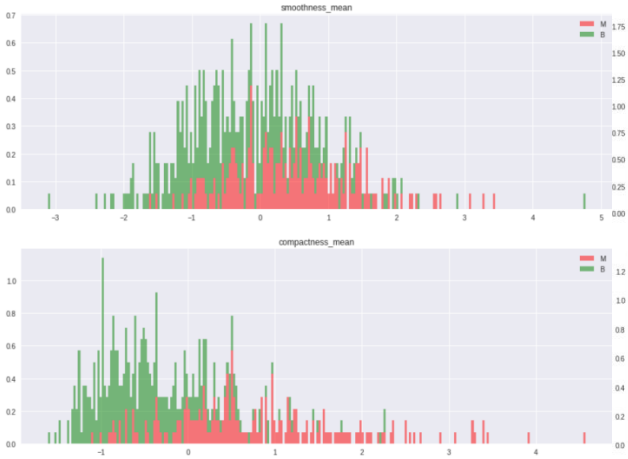

Fig7. Graph on smoothness_mean and compactness_mean

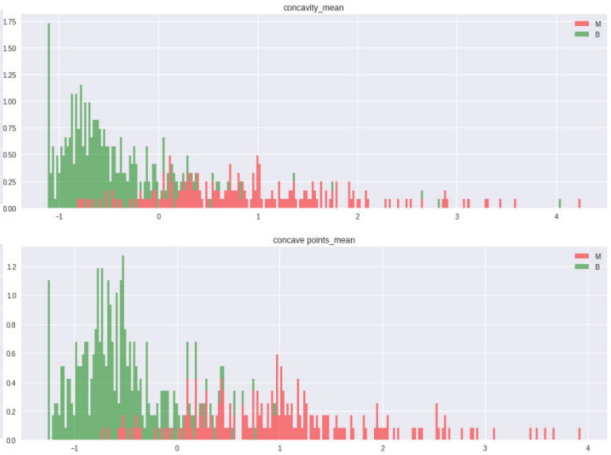

Fig8. Graph on concavity_mean and concave point_mean

Fig9. Graph on ssymmetry_mean and fractal_dimension_mean

VII. FUTURE SCOPE

The analysis of the results signifies that the integration of multidimensional data along with different classification, feature selection and dimensionality reduction techniques can provide auspicious tools for inference in this domain.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 5, May 2022||

| DOI:10.15680/IJIRSET.2022.1105322|

This topic will play an important role in the research and the medical advancement that is happening out there in the world; The analysis of the results signifies that the integration of multidimensional data along with different classification, feature selection and dimensionality reduction techniques can provide auspicious tools for inference in this domain. Further research in this field should be carried out for the better performance of the classification techniques so that it can predict on more variables. We are intending how to parameterize our classification techniques hence to achieve high accuracy. We are looking into many datasets and how further Machine Learning algorithms can be used to characterize Breast Cancer. We want to reduce the error rates with maximum accuracy.

VIII. CONCLUSION

The most frequently occurring type of across cancer is breast cancer. There is a chance of twelve percent for a women picked randomly to be diagnosed with the disease. Thus, early detection of breast cancer can save a lot of valuable life. The proposed work flow in our project is the study of machine learning algorithms, for the detection of breast cancer. It has been observed that logistic regression algorithm had an accuracy of more than 94%, to determine benign cancer or malignant cancer. Thus, supervised machine learning techniques will be very supportive in early diagnosis and prognosis of a cancer type in cancer research.

REFERENCES

1. Sameer Hamed1; Abdelwadood Mesleh2; Abdullah Arabiyyat3, Vol. 10, Issue. 11, November 2021, pg.4 – 11 Breast Cancer Detection Using Machine Learning Algorithms
2. Shubham Sharma[1], Archit Aggarwal[2], Tanupriya Choudhury[3], Breast Cancer Detection Using Machine Learning Algorithms
3. International Journal of Multidisciplinary Engineering in Current Research Volume 6, Issue 12, December 2021 BREAST CANCER DETECTION WITH MACHINE LEARNING
4. Breast Cancer Detection and Prediction using Machine Learning Anoy Chowdhury University of Engineering & Management, Kolkata.
5. 2020 International Conference on Emerging Trends in Communication, Control and Computing (ICONC3) Mody University of Science and Technology, Lakshmangarh, Feb 21-22, 2020 978-1-7281-1420-0/20/$31.00 ©2020 IEEE Diagnosis of Breast Cancer using Machine Learning Algorithms-A Review
6. M. R. Basunia, I. A. Pervin, M. A. Mahmud, S. Saha, and M. Arifuzzaman, "On Predicting and Analyzing Breast Cancer using Data Mining Approach," in 2020 IEEE Region 10 Symposium (TENSYMP), 2020, pp. 1257-1260.
7. P. Mekha and N. Teeyasuksaet, "Deep Learning Algorithms for Predicting Breast Cancer Based on Tumour Cells," in 2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON), 2019, pp. 343-346.
8. S. Kabir and M. Ahad, "ANN Classification of Female Breast Tumour Type Prediction Using EIM Parameters," in 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE), 2020, pp. 890-893.
9. H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, et al., "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," CA Cancer J Clin, vol. 71, pp. 209-249, May 2021.
10. M. Masud, A. E. Eldin Rashed, and M. S. Hossain, "Convolutional neural network-based models for diagnosis of breast cancer," Neural Comput Appl, pp. 1-12, Oct 9 2020.
11. H. Abdel-Razeq, A. Mansour, and D. Jaddan, "Breast Cancer Care in Jordan," JCO global oncology, vol. 6, pp. 260-268, 2020.
12. "Breast Cancer Screening and Diagnosis," Annals of Internal Medicine, vol. 172, pp. 839-840, 2020.
13. G. G. Hungilo, G. Emmanuel, and A. W. R. Emanuel, "Performance Evaluation of Ensembles Algorithms in Prediction of Breast Cancer," in 2019 International Biomedical Instrumentation and Technology Conference (IBITeC), 2019, pp. 74-79.
14. V. A. Telsang and K. Hegde, "Breast Cancer Prediction Analysis using Machine Learning Algorithms," in 2020 International Conference on Communication, Computing and Industry 4.0 (C2I4), 2020, pp. 1-5.
15. K. Sharma, B. Muktha, A. Rani, and C. Thirumalai, "Prediction of benign and malignant tumor," in 2017 International Conference on Trends in Electronics and Informatics (ICEI), 2017, pp. 1057-1060.
16. F. S. Ahadi, M. R. Desai, C. Lei, Y. Li, and R. Jia, "Feature-Based classification and diagnosis of breast cancer using fuzzy inference system," in 2017 IEEE International Conference on Information and Automation (ICIA), 2017, pp. 517-522.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 5, May 2022||

| DOI:10.15680/IJIRSET.2022.1105322|

17. P. Sivakumar, T. U. Lakshmi, N. S. Reddy, R. Pavani, and V. Chaitanya, "Breast Cancer Prediction System: A novel approach to predict the accuracy using Majority-Voting Based Hybrid Classifier (MBHC)," in 2020 IEEE India Council International Subsections Conference (INDISCON), 2020, pp. 57-62.
18. N. A. Mashudi, S. A. Rossli, N. Ahmad, and N. M. Noor, "Comparison on Some Machine Learning Techniques in Breast Cancer Classification," in 2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES), 2021, pp. 499-504.
19. S. Ara, A. Das, and A. Dey, "Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms," in 2021 International Conference on Artificial Intelligence (ICAI), 2021, pp. 97-101.
20. M. R. Basunia, I. A. Pervin, M. A. Mahmud, S. Saha, and M. Arifuzzaman, "On Predicting and Analyzing Breast Cancer using Data Mining Approach," in 2020 IEEE Region 10 Symposium (TENSYMP), 2020, pp. 1257-1260.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details