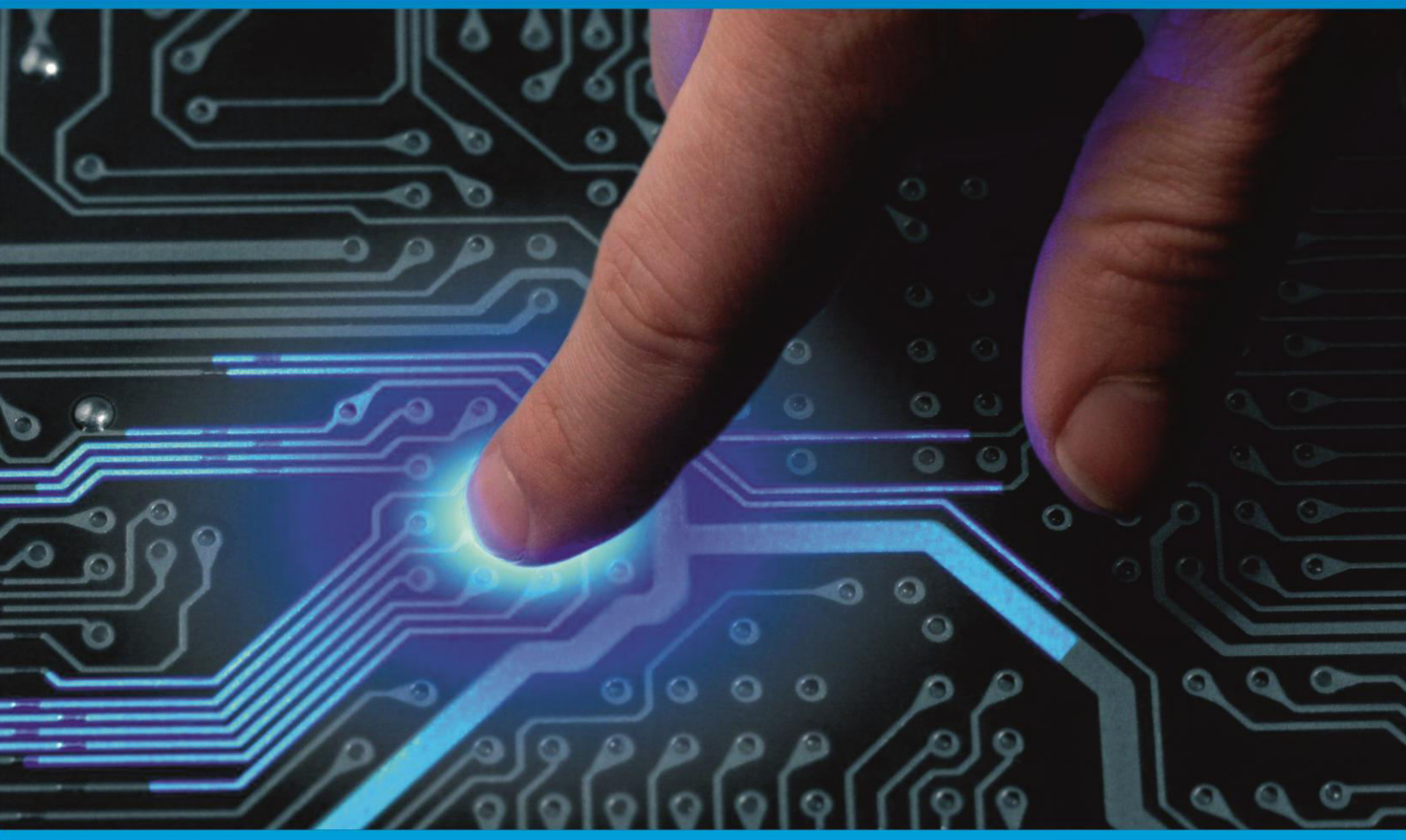
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 5, May 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.488

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905264 |

Annoyed Turn out Ability Transmittal Using Heap – Structure

B.Manjubashini, Abitha.S, Meena.K, Abinaiya.B, Aaisa Siddiqa.A

Assistant Professor, Department of Computer Science and Engineering, Mahendra Institute of Technology Autonomous Namakkal, Tamilnadu, India

Department of Computer Science and Engineering, Mahendra Institute of Technology Autonomous Namakkal, Tamilnadu, India

ABSTRACT: Cancer is a dominant cancer in women worldwide and is increasing in developing countries where the majority of cases are diagnosed in late stages. The projects that have already been proposed show a comparison of machine learning algorithms with the help of different techniques like the ensemble methods, data mining algorithms or using blood analysis etc. This paper proposed now presents a comparison of six machine learning (ML) algorithms: This paper is a relative study on the implementation of models using Logistic Regression, Support Vector Machine (SVM) and K Nearest Neighbor (KNN) is done on the dataset taken from the UCI repository. With respect to the results of accuracy, precision, sensitivity, specificity and False Positive Rate the efficiency of each algorithm is measured and compared. These techniques are coded in python and executed in Spyder, the Scientific Python Development Environment. Our experiments have shown that SVM is the best for predictive analysis with an accuracy of 92.7%.We infer from our study that SVM is the well suited algorithm for prediction and on the whole KNN presented well next to SVM.

KEYWORDS: Classification, Logistic Regression, KNN, SVM.

I. INTRODUCTION

Cancer is the prime reason for demise of women. It is the second dangerous cancer after lung cancer. In the year 2018 according to the statistics provided by World Cancer Research Fund it is estimated that over 2 million new cases were recorded out of which 626,679 deaths were approximated. Of all the cancers, breast cancer constitutes of 11.6% in new cancer cases and come up with 24.2% of cancers among women. In case of any sign or symptom, usually people visit doctor immediately, who may refer to an oncologist, if required. The oncologist can diagnose breast cancer by: Undertaking thorough medical history, Physical examination of both the breasts and also check for swelling or hardening of any lymph nodes in the armpit.

This research paper has gathered information from ten different papers based on breast cancer using machine learning and other techniques such as ultrasonography, blood analysis etc. The project by S. Gokhale, is using the ultrasonography(USG) technique which is a powerful method in detecting details about the breast mass that usually cannot be detected even by mammography. Another project by Pragya Chauhan and Amit Swami, which is based on the ensemble method usually used to increase the prediction accuracy of breast cancer. A Genetic algorithm based weighted average method that includes crossover and mutation is used for the prediction of multiple models. Further more, a project by Abien Fred M. Agarap uses different methods like GRU-SVM, NN, multilayer perceptron (MLP), softmax regression to classify the dataset into benign or malignant. A project by Priyanka Gupta shows the comparison of the lesser invasive techniques such as Classification and Regression Trees (CART), random forest, nearest neighbour and boosted trees. These four classification models are chosen to extract the most accurate model for predicting cancer survivability rate.

Further more, a project by Yixuan Li and Zixuan Chen shows a performance evaluation using three indicators including prediction accuracy values, F-measure metric and AUC values are used to compare the performance of these five classification models. Other experiments show that random forest model can achieve better performance and adaptation than other four methods. A project by Mumine Kaya Keles, which is a comparative study of data mining classification algorithms. Another project by Sang Won Yoon and Haifeng Wang that uses four data mining models are applied in this paper, i.e., support vector machine (SVM), artificial neural network (ANN), Naive Bayes classifier,

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905264 |

AdaBoost tree. Furthermore, feature space is highly deliberated in this paper due to its high imapct on the efficiency and effectiveness of the learning process. Lastly a project by Wenbin Yue and Zidong Wang that shows the algorithms that helped them with the diagnosis and prognosis of their dataset.

As the diagnosis of this disease manually takes long hours and the lesser availability of systems, there is a need to develop the automatic diagnosis system for early detection of cancer. Data mining techniques contribute a lot in the development of such system. For the classification of benign and malignant tumor we have used classification techniques of machine learning in which the machine is learned from the past data and can predict the category of new input. Keywords - Breast cancer classification, Breast cancer prediction, benign, malignant, Naïve Bayes, KNN, Support Vector Machine, Artificial Neural Network, Random Forest, Decision tree, SQLAlchemy.

II. RELATED WORK

AlirezaOsarech, BitaShadgar used SVM classification technique on two different benchmark datasets for breast cancer which got 98.80% and 96.63% accuracies[2]. MandeepRana, PoojaChandorkar, AlishibaDsouza worked on the diagnosis and the prediction of recurrence of breast cancer by applying KNN, SVM, Naïve Bayes and Logistic Regression techniques, programmed in MATLAB. The classification techniques are applied on two datasets taken from UCI depository. A dataset of them is used for identification of disease(WDBC) and the next one is used for recurrence prediction (WPBC)[3].Vikas Chaurasia, BB Tiwari and Saurabh Pal used three famous algorithms such as J48, Naive bayes, RBF, to build predictive models on breast cancer prediction and compared their accuracy. The results had shown that Naive Bayes predicted well among them with an accuracyof97.36% [4]. Haifeng Wang and Sang Won Yoon compared Naive Bayes Classifier, Support Vector Machine (SVM), Ada Boost tree, Artificial Neural Networks (ANN), to find a powerful model for breast cancer prediction. They implemented PCA for dimensionality reduction[5]. S.Kharya worked on breast cancer prediction and stated that artificial neural networks are widely used. The paper featured about the advantages and short comings of using machine learning methods like SVM, Naive Bayes, Neural network and Decision trees[6]. Naresh Khuriwal,Nidhi Mishra took data from Wisconsin Breast Cancer database and worked on breast cancer diagnosis..The results of their experiments proved that ANN and Logistic Algorithm worked better and provided a good solution. It achieved an accuracy of 98.50% [7].

Ultrasound characterisation of breast masses by S. Gokhale written by proposed a system where they found that doctors have known and experienced that breast cancer occurs when some breast cells begin to grow abnormally. These cells divide more briskly and disperse faster than healthy cells do and continue to accumulate, form- ing a lump or mass that the may start causing pain. Cells may spread rapidly through your breast to your lymph nodes or to other parts of your body. Some women can be at a higher risk for breast cancer because of their family history, lifestyle, obesity, radiation, and reproductive factors. In the case of cancer, if the diagnosis occurs quickly, the patient can be saved as there have been advances in cancer treatment. In this study we use four machine learning classifiers which are Naive Bayesian Classifier, k-Nearest Neighbour, Support Vector Machine, Artificial Neural Network and random forest.

Harmonic imaging and real-time compounding has been shown to enhance image resolution and lesion characterisation. More recently, USG elastography seems to be quite ncouraging. Initial results show that it can improve the specificity and positive predictive value of USG within the characterisation of breast masses. The reason why any lesion is visible on mammography or USG is that the relative difference within the density and acoustic resistance of the lesion, respectively, as compared to the encompassing breast tissue. [1]

Breast Cancer Prediction Using Genetic Algorithm Based Ensemble Approach written by Pragya Chauhan and Amit Swami proposed a system where they found that Breast cancer prediction is an open area of research. In this paper dierent machine learning algorithms are used for detection of Breast Cancer Prediction. Decision tree, random forest, support vector machine, neural network, linear model, adabost, naive bayes methods are used for prediction. An ensemble method is used to increase the prediction accuracy of breast cancer. New technique is implemented which is GA based weighted average ensemble method of classification dataset which over- came the limitations of the classical weighted average method. Genetic algorithm based weighted average method is used forthe prediction of multiple models. The comparison between Particle swarm optimisation(PSO), Dierential evolution(DE) and Genetic algorithm(GA) and it is concluded that the genetic algorithm outperforms for weighted average methods. One more comparison between classical ensemble method and GA based weighted average method and it is concluded that GA based weighted average method outperforms. [2]

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905264 |

On Breast Cancer Detection: An Application of Machine Learn ing Algorithms on the Wisconsin Diagnostic Dataset by the Abien Fred M. Agarap. In this paper, six machine learning algorithms are used for detection of cancer . GRUSVMmodel is used for the diagnosis of breast cancer GRUSVM, Linear Regression, Multilayer Perceptron (MLP), Nearest Neighbour (NN) search, Softmax Regression, and Support Vector Ma- chine (SVM) on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset by measuring their classification test accuracy, and their sensitivity and specificity values. The said dataset consists of features which were computed from digitised images of FNA tests on a breast mass. For the implementation of the ML algorithms, the dataset was partitioned in the following fashion 70 percent for training phase, and 30 percent for the testing phase. Their results were that all presented ML algorithms exhibited high performance on the binary classification of carcinoma, i.e. determining whether benign tumour or malignant tumour. Therefore, the statistical measures on the classification problem were also satisfactory. To further corroborate the results of this study, a CV technique such as k-fold cross-validation should be used. The appliance of such a way won't only provide a more accurate measure of model prediction performance, but it'll also assist in determining the foremost optimal hyper-parameters for the ML algorithms. [3]

III. METHODOLOGY

We obtained the breast cancer dataset from UCI repository and used spyder as the platform for the purpose of coding. Our methodology involves use of classification techniques like Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), Logistic Regression, with Dimensionality Reduction technique i.e. Principal Component Analysis (PCA) .

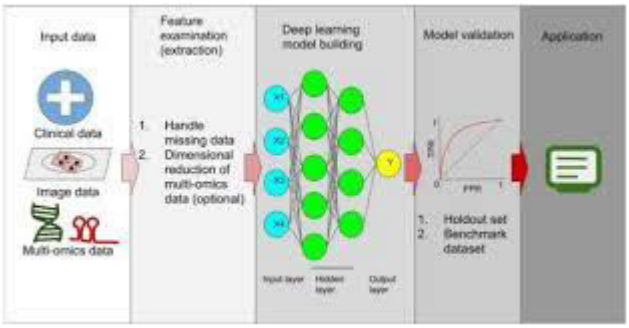


Fig1: Propose Methods

Dimensionality Reduction is a process in which the number of independent variables is reduced to a set of principle variables by removing those which are less significant in predicting the outcome. Dimensionality Reduction is used to get two dimensional data so that better visualization of machine learning models can be done by plotting the prediction regions and the prediction boundary for each model. Whatever may be the number of independent variables, we often end up with two independent variables by applying a suitable dimensionality reduction technique. There are two methods, namely Feature selection and Feature Extraction.

It is a linear technique which is used to compress lots of data into something which gives essence of the original data. Based on the variance of the data it plots the actual data into a dimensional space with less attributes such that the variance is maximized.PCA extracts p independent variables from n independent variables of our dataset (p<=n) that explain the most variance of our dataset, despite of the independent variables.With the help of covariance matrix of the dataset, the eigen vectors are calculated. The principal components are those eigen vectors which have the largest eigen values and these can be used to rebuild a huge portion of the variance of the actual data. These few eigen vectors (with most important variance) span a lesser space reducing the original space But this process may cause some data loss. So, we should make sure that they retain the remaining eigenvectors.All these individual principal components sum up to give total variance. Each individual principal component is the ratio to the variance of the principal component to the total variance.The result of applying PCA gives us two principal components PC1 (the first principal component) and PC2 (the second principal component).PC1 gives the most variance and PC2 gives the second most variance.Now, our dataset is ready and data mining techniques can be applied on it for classification of benign and malignant tumors.
The most exciting phase in building any machine learning model is selection of algorithm. We can use more than one kind of data mining techniques to large datasets. But, at high level all those different algorithms can be classified in two groups: supervised learning and unsupervised learning. Supervised learning is the method in which the machine is trained on the data which the input and output are well labeled. The model can learn on the training data and can

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905264 |

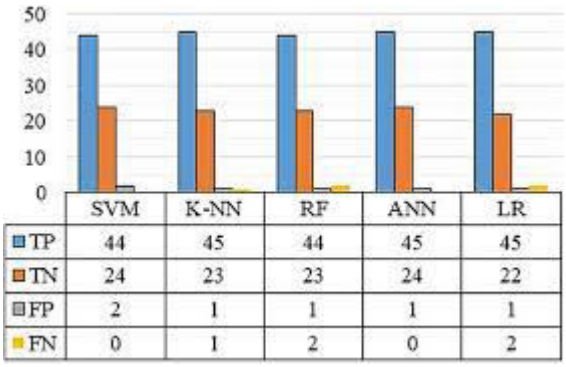
process the future data to predict outcome. They are grouped to Regression and Classification techniques. A regression problem is when the result is a real or continuous value, such as “salary” or “weight”. A classification problem is when the result is a category like filtering emails “spam” or “not spam”. Unsupervised Learning : Unsupervised learning is giving away information to the machine that is neither classified nor labeled and allowing the algorithm to analyze the given information without providing any directions. In unsupervised learning algorithm the machine is trained from the data which is not labeled or classified making the algorithm to work without proper instructions. In our dataset we have the outcome variable or Dependent variable i.e. Y having only two set of values, either M (Malign) or B(Benign). So Classification algorithm of supervised learning is applied on it. We have chosen three different types of classification algorithms in Machine Learning.

Logistic Regression Logistic Regression is a supervised machine learning technique, employed in classification jobs (for predictions based on training data).Logistic Regression uses an equation similar to Linear 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 outcome of dependent variable is discrete. Logistic Regression uses a simple equation which shows the linear relation between the independent variables. These independent variables along with their coefficients are united linearly to form a linear equation that is used to predict the output[8].

Support Vector machine Support Vector Machine is a supervised machine learning algorithm which is doing well in pattern recognition problems and it is used as a training algorithm for studying classification and regression rules from data.SVM is most precisely used when the number of features and number of instances are high. A binary classifier is built by the SVM algorithm [13]. This binary classifier is constructed using a hyper plane where it is a line in more than 3-dimensions.The hyper plane does the work of separating the members into one of the two classes.

Hyper plane of SVM is built on mathematical equations. The equation of hyper plane is WTX=0 which is similar to the line equation y= ax + b. Here W and X represent vectors where the vector W is always normal to the hyper plane.WTX represents the dot product of vectors. As SVM deals with the dataset when the number of features are more so, we need to use the equation WTX=0 in this case instead of using the line equation y= ax + b. If a set of training data is given to the machine, each data item will be assigned to one or the other categorical variables, a SVM training algorithm builds a model that plots new data item to one or the other category. In an SVM model, each data item is represented as points in an ndimensional space where n is the number of features where each feature is represented as the value of a particular coordinate in the n-dimensional space. Classification is carried out by finding a hyper-plane that divides the twoclasses proficiently. Later, new data item is mapped into the same space and its category is predicted based on the side of the hyper-plane they turn up.

IV. RESULTS AND DISCUSSION



	SVM	K-NN	RF	ANN	LR
TP	44	45	44	45	45
TN	24	23	23	24	22
FP	2	1	1	1	1
FN	0	1	2	0	2

Fig2: Analysis

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905264 |

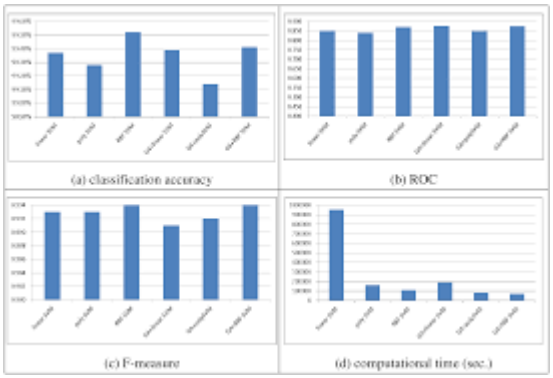


Fig3: Performance analysis

As our dataset contains 32 attributes dimensionality reduction contributes a lot in decreasing the multidimensional data to a few dimensions. Of all the three applied algorithms Support Vector Machine, k Nearest Neighbor and Logistic Regression, SVM gives the highest accuracy of 92.7% when compared to other two algorithms. So, we propose that SVM is the best suited algorithm for the prediction of Breast Cancer Occurrence with complex datasets.

IV. CONCLUSION

The research on nine papers has helped us gather the data for the project proposed by us. By using machine learning algorithms we will be able to classify and predict the cancer into being or malignant. Machine learning algorithms can be used for medical oriented research, it advances the system, reduces human errors and lowers manual mistakes. Our work mainly focused in the advancement of predictive models to achieve good accuracy in predicting valid disease outcomes using supervised machine learning methods. The analysis of the results signify 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.

REFERENCES

[1] ”Ultrasound characterisation of breast masses”, The Indian journal of radiology imaging by S. Gokhale., Vol. 19, pp. 242-249, 2009. K. Elissa, “Title of paper if known,” unpublished.
[2] “Breast Cancer Prediction Using Genetic Algorithm Based Ensemble Approach” by Pragya Chauhan and Amit Swami, 18 October 2018
[3] “On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset” by Abien Fred M. Agarap, 7 February 2019
[4] “Analysis of Machine Learning Techniques for Breast Cancer Prediction” by the Priyanka Gupta and Prof. Shalini L of VIT university, vellore, 5 May 2018.
[5] “Breast Cancer Diagnosis by Dierent Machine Learning Methods Using Blood Analysis Data“ by the Muhammet Fatih Aslan, Yunus Celik , Kadir Sabanci and Akif Durdu, 31 December, 2018
[6] ”Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction”, by Yixuan Li, Zixuan Chen October 18, 2018
[7] “Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms: A Comparative Study” by Mumine Kaya Keles, Feb 2019
[8] “Breast Cancer Prediction Using Data Mining Method ” by Haifeng Wang and Sang Won Yoon, Department of Systems Science and Industrial Engineering State University of New York at Binghamton Binghamton, May 2015.
[9] “Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis” by Wenbin Yue , Zidong Wang, 9 May 2018

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.488

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details