INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

A Survey on Non-Payment Risk Automation Using Machine Learning

Shashank Jagtap[1], Parth Apamarjane[2], Sourav Chitnis[3], Akshay Patil[4],

Prof. Naved Raza Q. Ali[5].

[1,2,3,4,5]Department of Computer Engineering, Smt. Kashibai Navale College of Engineering, Pune, India

ABSTRACT: With the advancement in technology, there are so many enhancements in the banking sector. The number of applications is increasing every day for loan approval. Some bank policies that are made, they also have to consider it while preparing a policy. Based on some parameters, the bank has to decide which one is best for approval. A credit scoring model is a tool that is typically used in decision making process of accepting or rejecting a loan. A credit scoring model is the result of a statistical model which, based on information about the borrower (e.g. age, salary, etc.), allows one to distinguish between "good" and "bad" loans and give an estimate of the probability of default using machine learning. This work's primary objective is to predict whether the loan approval to a specific individual is safe or not with the help of machine learning algorithms and identify some potential future research directions in the field.

KEYWORDS: Loan Repayment, Machine Learning, Random Forest algorithm.

I. INTRODUCTION

Machine learning has become an important in our life. We may or may not be aware that machine learning is used in various application. It works by analysing the data, relating patterns, and incorporating minimum mortal intervention. Nearly any work that can be done using a data description pattern or set of rules can be done using a machine learning machine. This allows companies to modify processes that preliminarily only humans could make hypotheticals for client service calls, accounts, and reviews. Loan distribution is the middle enterprise of virtually every bank. Almost all of the bank's means come directly from the gains it makes from the loans it makes. The main thing of the banking terrain is to put wealth in a safe place. Currently, many banks / financial institutions approve loans after a verification and validation redemption process, but it is not yet clear if the selected applicant is correct among all applicants. Through non- payment risk system, it is possible to predict whether a particular applicant is safe and the entire process of verifying the characteristics will be automated by machine learning technology. Credit forecasting is very useful for both bank employees and applicants. The goal of this system is to provide a quick and immediate and easy way to select good applicants. It can offer banks special benefits.

Thus, financial institution uses machine learning to reduce risk and frauds and to find the loan defaulters. Defaulter is an applicant who failed to repay the loan. With millions of transactions happening every day, it is impossible for a human to monitor each one. Automated techniques based on AI are the only solution when dealing with such volume.

Credit risk has always been a challenging area for banking sectors, given the multiple factors that go into forming an individual's risk profile. For business borrowers, the process is even more complicated as data across a variety of parameters and time periods.
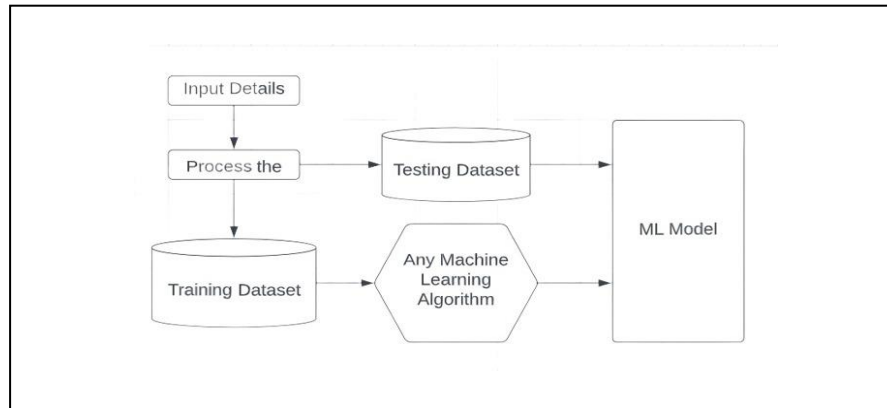
Figure 1: General Architecture of any Machine Learning Model building.

Mainly the loan prediction system is a software that checks the eligibility of the customer who is capable of paying loan or not. This process can be applied by many customers and the dataset can be obtained. Loan prediction system is used for prediction according to the attributes considered from the dataset. The purpose of study is to test and train the dataset using different machine learning algorithms and understanding how the data is extracted through the features. The results against a particular loan Id can be helpful by sending it to other departments so that an appropriate action can be taken on the applicant. Briefly having study on how to train the model and predict the features from the model. Understanding the concepts of loan prediction and having knowledge on validation of accuracy and model loss while training and testing data. Then selection of the model is much important according to the accuracy obtained from each model and the automation of this process will surely benefit.

II. MOTIVATION

The banking sector is about to undergo a revolution thanks to machine learning and artificial intelligence, which will use massive amounts of data to create models that will enhance decision-making, customise services, and enhance risk management. Banking, finance, and insurance all benefit from the "automatic" component of automated machine learning, often known as Auto ML. In order to determine whether someone is qualified to obtain a loan from the bank or not, automated non-payment risk automation is a key instrument for bank security. So, the primary goal of this project was to make it simple to handle such a method and cross-verify the client's information without too much difficulty.

Also, we employ a variety of algorithms in this project, which greatly improves accuracy in comparison to previous models. The prior model had many problems, including a dearth of metrics that made it difficult for lenders to identify victims who were not repaying their loans and decreased accuracy. Due to the lack of background information, lenders are unable to determine whether a client is qualified for a loan or not. But in this case, we employ a lot of parameters to determine whether or not a person is eligible for a loan based on those parameters. And because of different parameters and techniques, our model's speed greatly increases.

III. LITERATURE SURVEY

Various researchers proposed different algorithms and techniques to detect loan defaulters. Some of the latest techniques as well as old methods are discussed. Different algorithms used by researchers provide different results. The review research on various algorithms such as:

A. Survey on Decision Tree, Random Forest and Support Vector Machine
B. Survey on techniques which do not use any machine learning algorithm
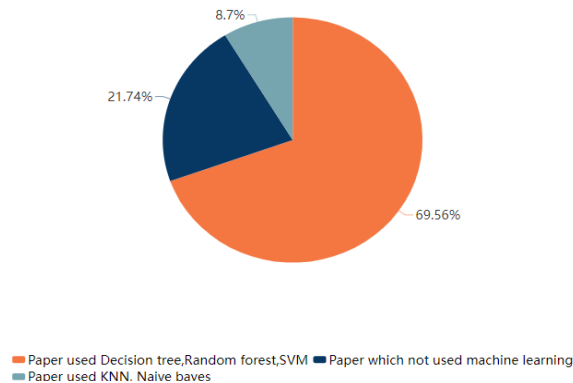C. Survey on KNN and Naïve Bayes.

Figure 2: Division of surveyed papers based on the following sections. (in %)

A. SURVEY ON DECISION TREE, RANDOM FOREST AND SUPPORT VECTOR MACHINE.

Anant Shinde et al. [1] used algorithms like Logistic regression using stratified k-folds cross validation and Random Forest. Various features that are taken into the dataset are compared with each other to know their relation with each other. From the comparison done, Credit_History is the most major point. The dataset consists of 600 sample data points. This model obtains a maximum accuracy of about 82 percent.

Ritika Purswani et al. [2] proposed a methodology on loan approval prediction method used by banks. Initially the data is collected then EDA is performed on the data. Outliers are removed and the data is structured. Then model is built using various classification algorithm like logistic regression, Decision Tress, Random Forest, Support Vector Classification. The accuracy obtained by logistic regression was 81 percent whereas the accuracy by Decision tree was 82 percent. Using random forest there is less chances of overfitting and higher the precision.

Lin Zhua et al. [3] elaborated a system on Loan Repayment using various ML models. This paper was presented in the 7th International Conference on Information Technology and Quantitative Management in year 2019. The algorithm used for building model is Random Forest. Almost 115,000 original loan data of users with 102 attributes is used for training model. Finally, 15 attributes are shortlisted which are considered more beneficial than any other attributes. As the input dataset from training is imbalanced, SMOTE (Synthetic Minority Oversampling Technique) is used to balance it by oversampling. Various performance evaluation metrics are used to evaluate the performance of the model. Random forest performs best among other algorithms like decision tree, SVM , Logistic regression.

Sudhamathy G. [4] used random forest algorithm for credit risk analysis. Dataset is taken from UCL Repository having 1000 records and 21 attributes. Initially data normalization is performed on the dataset. Dataset is divided into 4:1 ratio. 80% training dataset and 20% test dataset. The training dataset is balanced using SMOTE function. A threshold value is fixed for selecting number of features based on their rankings.

Kumar Arun et al. [5] proposed a model to reduce the risk factor for selecting the right person to get the loan from the banks. Data is collected first and then comparison is done between different machine learning models. Training is done on the most appropriate model and then testing is performed. Every time a new user fills the data it is treated like testing data. Algorithms used for building model are Decision tree, Random Forest, SVM, Linear model, Neural network, Adaboost.

Pidikiti Supriya et al. [6] proposed a model which reduces bank's efforts for selecting a candidate for loan who will not default on loan. Data mining technique is used to understand the underlying trends and patterns. Decision tree is used to build model. The data collected is divided into the ratio 4:1 of training set and testing set. The data collected is cleaned and correlation among the attributes is considered. Decision tree brings out an accuracy of 81%.

Ms. Kathe Rutika Pramod et al. [7] used decision tree to build a model used for predicting whether a customer will

default or not on loan. This will help the lending club to approve loans requested. Decision tree model works by selecting best attribute as the tree's root. The training set is divided into subsets, such that each subset comprises similar value for an attribute. Applicants with high income sanctioning low amount is more likely to get approved which make sense, more likely to pay back their loans.

Saqib Aziz et al. [8] proposed a system based on SVM and Decision tree algorithm with neural network. However, it explored how artificial intelligence (AI) and machine learning solutions are transforming risk management. This paper explained the application to credit risk, application to market risk, application to operational risk, application to Reg Tech. Moreover, after taking the dataset for reducing to appropriate dimensions, they used Principal component analysis (PCA).

Majid Bazarbash et al. [9] elaborated what is Machine Learning. Machine learning models as nonparametric models, Prominent Machine Learning Models and proposed decision tree and random forest algorithm. They Achieving widespread financial inclusion is a major step in achieving sustainable development goals in many countries, and FinTech credit appears as a promising solution and a potential leapfrog for countries with low financial inclusion.

Dr.C K Gomathy et al. [10] has studied decision tree algorithm on Oct 2021. Now-a-days banks face a huge loss as the customers who take loan are not able to repay the loan on time or totally not able to repay the loan. As a result, now banks check the history of the customer taking loan and various other parameters manually. The model built is based on old training dataset in future it should also be tested on new dataset. The problem is not completely solved here but it is made more predictive.

C N Sujatha et al. [11] has studied on Decision Tree and Random Forest Algorithms. The paper has been published by IEEE and presented in 2021 Innovations in Power and Advanced Computing Technologies (i-PACT) conference. Kaggle open-source dataset was used to train the models. There is a comparative study between Logistic Regression, SVC, KNN, Decision Tree and Random Forest Algorithms where Random Forest Algorithm provides more accuracy and so is chosen as the preferred ML model. User Interface is created using Heroku app and Flask is used for API calls.

Mehul Madaan et al. [12] has studied on Decision tree random forest algorithm. The paper automates the process of borrower's likelihood of repaying the loan. Machine Learning concepts are used to solve this problem. Both the algorithms have been used on the same dataset and the conclusions have been made with results showing that the Random Forest algorithm outperformed the Decision Tree algorithm with much higher accuracy. Dataset used was used from Lending Club in Kaggle. Random Forest showed accuracy of 80% while Decision Tree provided 73% accuracy. The paper showed comparative study and did not consist of an actual user interface.

Anshika Gupta et al. [13] has studied on Logistic Regression and Random Forest. The paper was published and presented in 9th International Conference on System Modelling & Advancement in Research Trends organized by IEEE. Date of presentation was 4th-5th December,2020. to train the models was taken from Kaggle competition which belong to different age group and gender of the applicants having 13 attributes in all. Moreover, the proposed system has user interface developed in HTML and CSS and connected to local server via Django. The paper lacked in amount of accuracy and user interface.

Shamsa Khalid et al. [14] researched on decision trees and random forests was conducted in the year 2022. The goal of this research study was to analyze, assess, and create machine learning algorithms that could precisely and error-free estimate risk for nonfinancial enterprises listed on PSX. These kinds of models, particularly those that focus on quantitative investment, could be used in nonfinancial organization's investment management since they provide a thorough description of the most recent technology, their potential uses, and the possibility that they would be successfully applied.

Peter Martey Addo et al. [15] published an article in "Maison des Sciences Économiques," they focus on a variety of algorithms, including the logistic algorithm. Random forest Models for gradient decent enhancing Moreover, deep learning was employed in these 2018 models for Credit Risk Analysis. They have observed that artificial neural network-based algorithms do not always offer the optimum performance and that authorities must also assure the transparency of decision-making algorithms to prevent prejudice and a potential adverse effect on the industry.

Sivasree M. S. et al. [16] assisted underperforming banks improve their operations and turn a profit. Two key goals

are to identify pertinent attributes and choose the appropriate models to assess credit risk. There are 17 attributes and 4520 records in the collection. The remaining 34% of the dataset is used to test the model, and 66% of it is utilized to train the model. 'Job' is the most significant attribute taken into account here.

B. SURVEY ON TECHNIQUES WHICH DO NOT USE ANY MACHINE LEARNING TECHNIQUES.

Nancy Gathoni Kiliswa et al. [17] researched on small-scale businesses have features that make the cost of managing credit significantly higher than the return on loans. Their paper was published in "Management Studies and Economic Systems (MSES)" in 2019. Small-scale businesses typically have shallow management, little experience, and inadequate training. They are also frequently start-ups with little history, poor financial records, and shallow management. Small-scale businesses may have new, unproven products, little to offer as security to lenders, and a reluctance to raise outside equity capital due to costs, control loss, and increased disclosure requirements. Small-scale business characteristics like these don't really encourage any aggressive debt recovery strategies.

Dr. V. Ramanujam et al. [18] released an article in the International Research Journal of Business and Management (IRJBM) in the year 2019 that made the case for better understanding the connection between demographic traits and MSME credit repayment performance. This study reveals that borrowers' demographic features are a significant factor that banks take into account when approving loans, contrary to conventional wisdom.

Michal Sterbak et al. [19] focused on delivering security as a crucial component of every firm. The digitization and transformation of data and information into the cyberspace has also increased and transferred their risks, dangers, and vulnerabilities, which have an impact on the market and on business. For the majority of firms, dealing with connected cybersecurity risks is a significant burden. There are various research and techniques that can help solve these issues, and it also emphasizes automating facility Our study demonstrates that enterprises are compelled to utilise a variety of techniques to evaluate information security and put protective measures in place to reduce risks and safeguard their information assets. These tools automate several information security risk management subprocesses, at least in part.

KaoutarErramy et al. [20] proposed an international trade strategy. The following article appeared in the peer-reviewed journal "Journal of Advanced Research in Accounting & Financial Management." This paper's main goal is to outline the acceptable strategies used by a multinational corporation to address scenarios including the risk of international non-payment. This study shows that a number of variables affect the international non-payment risk management strategy, which makes it necessary to assess the experiences of all Moroccan exporting enterprises.

X.FrancisJency et al. [21] assisted in understanding the nature of customer who applies for a loan. EDA is done on a given dataset. The dataset initially undergoes the process of normalization, missing values are taken care, essential columns are filtered, deriving new columns, identifying the target variables and visualizing the data in graphical format. Pandas' library available in python is used to derive information from given dataset. Various graphs are used to visualize and from the graphs it can be concluded that short term loan was preferred by the majority of the loan applicants and customers mainly apply loan for debt consolidation.

C. SURVEY ON KNN AND NAÏVE BAYES

Aditi Kacheria et al. [22] suggested using Naive Bayesian classification, K-NN, and Binning algorithms to keep the bank safe. The following article appeared in the 2016 issue of the International Journal of Soft Computing and Engineering (IJSCE). Using K-NN and Binning algorithms, the quality of the data is improved prior to classification in order to increase classification accuracy. Using pre-processing methods is necessary since there are many situations when the data set is inconsistent as a result of missing values and abnormalities. They utilise the K-NN method to handle missing values. K-NN is a straightforward algorithm that categorizes new data based on a similarity metric and stores all accessible data.

Aboobyda Jafar Hamid et al. [23] developed a classification system that employs BayseNet, NaiveBayes, and J48 to forecast loan defaulters. The used dataset is split into 20% of testing data and 80% of training data. naveBayes accuracy is 73%, bayesNet accuracy is 77%, and j48 accuracy is 78%. Weka application is used to put the model into action. As a result of its high accuracy and low mean absolute error, the J48 algorithm is determined to be the best.

IV.   OVERVIEW

Table 1: Evaluation of Algorithms in surveyed papers.

Sr. No.	Author Name and Paper Title.	Techniques used.	Performance Measures	Limitations
1.	Anant Shinde et.al [1] Loan prediction system using machine learning	Logistic Regression, Random Forest.	Accuracy: 82%	1. Low accuracy score.
2.	Ritika Purswani et.al [2] Loan approval prediction using machine learning	Logistic Regression, Decision tree, Random Forest.	Accuracy: 81%,	1. Low accuracy score.
3.	Sudhamathy G. [4] Credit risk analysis and prediction modelling of bank loans using R	Decision Tree.	Accuracy: 94%, Precision: 83%	1. Low precision score. 2. No user-friendly GUI.
4.	Pidikiti Supriya et.al [6] Loan prediction by using machine learning models	Decision tree, SVM, K-neighbors, Gradient Boosting.	Accuracies respectively: 81%, 62%, 76%, 79%	1. Less scores of performance measures. 2. No in-depth analysis of dataset.
5.	Mehul Madaan et.al [12] Loan default prediction using decision trees and random forest: A comparative study	Decision tree, Random Forest.	Accuracies respectively: 73%,80%	1. Less accuracies delivered by both the algorithms.
6.	Shamsa Khalid et.al [14] Predicting risk through artificial intelligence based on machine learning algorithms: A Case of Pakistani Nonfinancial Firms	Decision tree.	RMSE: 0.08093, MAE: 0.03049	1. No user-friendly GUI.
7.	Peter Martey Addo et.al [15] Credit risk analysis using machine and deep learning models	Decision tree, Random Forest.	RMSE: 0.2014	1. No user-friendly GUI.
8.	Aboobyda Jafar Hamid et.al [23] Developing prediction model of loan risk in banks using data mining machine learning and applications	Bayse-Net and Naïve Bayes, J48 classification.	Accuracies respectively: 73%,77%,78%	1. Less accuracies delivered.

.

Table 2:Analysis of Algorithms.

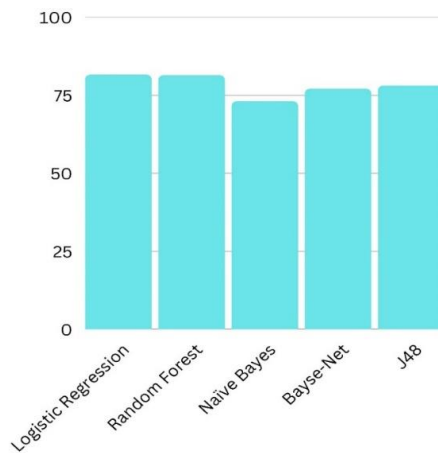| Algorithms used in following papers.	Mean Accuracy.
Logistic Regression	81.5%
Random Forest	81.33%
Naïve Bayes	73%
Bayse-Net	77%
J48	78%

Figure 3: Graphical comparisons of Algorithms Accuracy

V. CONCLUSION

According to the following survey we noticed that use of Machine Learning in Loan prediction has significant effects. The use of machine learning to predict loan repayment is a rapidly growing field. Machine learning can be a useful tool for predicting loan repayment, according to several studies. As we have surveyed several algorithms used for prediction include logistic regression with accuracy 81.5%, Random Forest (81.33%), Naïve Bayes (73%), etc. However, there is a need to increase the accuracy, precision, and recall of these models to get desirable results quickly. Hence, by offering a secure and effective solution to the banks and customers, the web-based application has potential to transform the way customers have been lending loan from banks. The possibility of loss to banks is reduced by significant margin. All customers along with bank stakeholders and employees will be satisfied with the system as long as it is implemented using specified methodology and approach. The system provides a user-friendly interface making it easier to use by customers.

REFERENCES

[1] A. Shinde, Y. Patil, I. Kotian, A. Shinde, and R. Gulwani, "Loan prediction system using machine learning," in ITM Web of Conferences, vol. 44, 2022, pp. 1-5, doi: 10.1051/itmconf/20224400001.
[2] R. Purswani, S. Verma, Y. Jaiswal and Prof. S. M., "Loan approval prediction using machine learning," *International Research Journal of Engineering and Technology (IRJET),* vol. 08, no. 06, pp. 01-05, Jun. 2021.
[3] L. Zhua, D. Qiua, D. Ergua, C. Yinga and K. Liub, "A study on predicting loan default based on the random forest," *Innovations in Power and Advanced Computing Technologies (i-PACT),* pp. 236-239, Jun. 2019.
[4] S. G., "Credit risk analysis and prediction modelling of bank loans using R," *International Journal of Engineering and Technology (IJET)*, vol. 8, no. 6, pp. 2631-2635, 2016.
[5] K. Arun, G. Ishan and S. Kaur, "Loan approval prediction based on machine learning approach," *IOSR Journal of Computer Engineering (IOSR-JCE),* vol. 18, no. 4, pp. 56-61, 2016.
[6] P. Supriya, M. Pavani, N. Saisushma, N. V. Kumari and K. Vikas, "Loan prediction by using machine learning models," *International Journal of Engineering and Techniques*, vol. 5, no. 2, pp. 32-37, 2019.
[7] Ms. K. R. Pramod, Ms. P. S. Dattatray and Ms. A. P. Prakash, "An approach for prediction of loan approval using machine learning algorithm," *International Journal of Creative Research Thoughts (IJCRT)*, vol. 9, no. 2, pp. 2357-2362, 2021.
[8] S. Aziz and M. Dowling, "AI and machine learning for risk management," SSRN Electronic Journal, Jan. 2018, doi: 10.2139/ssrn.3101253.
[9] M. Bazarbash, "FinTech in financial inclusion: machine learning applications in assessing credit risk," *International Monetary Fund*, 2019, doi: 10.5089/9781484393044.086.
[10] C. K. Gomathy et al., "The loan prediction using machine learning*," Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya University*, 2021.
[11] C. N. Sujatha et al., "Loan prediction using machine learning and its deployment on web application," *Innovations in Power and Advanced Computing Technologies (i-PACT)*, 2021.

[12] M. Madaan et al., "Loan default prediction using decision trees and random forest: A comparative study," *IOP Conf. Series: Materials Science and Engineering*, 2021.

[13] A. Gupta et al., "Bank loan predicting system using machine learning," *9th International Conference System Modeling and Advancement in Research Trends (SMART)*, 2020.

[14] S. Khalid et al., "Predicting risk through artificial intelligence based on machine learning algorithms: *A Case of Pakistani Nonfinancial Firms*," 2022.

[15] P. M. Addo, D. Guegan, and B. Hassani, "Credit risk analysis using machine and deep learning models," *Maison des Sciences Économiques*, 2018.

[16] S. M. S and R. S. T, "Loan credibility prediction system based on decision tree algorithm," *International Journal of Engineering Research & Technology (IJERT)*, 2015.

[17] N. G. Kiliswa and M. S. Bayat, "Determinants of loan repayment in small scale enterprises in developing countries," *Management Studies and Economic Systems (MSES)*, 2014.

[18] V. Ramanujam and K. A. Vidya, "A study on the credit repayment behaviour of borrowers," *International Research Journal of Business and Management (IRJBM)*, 2017.

[19] M. Sterbak, P. Segec, and J. Jurc, "Automation of risk management processes," *University of Zilina, Faculty of Management Science and Informatics,* Žilina, Slovakia, 2021.

[20] S. Ahrouch and K. Erramy, "Exportations non-payment risk management," *Journal of Advanced Research in Accounting & Finance Management Peer Reviewed Journal*, 2018.

[21] X. F. Jency, V. P. Sumathi, and J. S. Sr, "An exploratory data analysis for loan prediction based on nature of the clients," *International Journal of Recent Technology and Engineering (IJRTE),* 2018.

[22] A. Kacheria, N. Shivakumar, S. Sawkar, and A. Gupta, "Loan sanctioning prediction system," *International Journal of Soft Computing and Engineering (IJSCE),* 2016.

[23] A. J. Hamid and T. M. Ahmed, "Developing prediction model of loan risk in banks using data mining machine learning and applications," *An International Journal (MLAIJ)*, vol. 6, no. 2, pp. 55-63, 2016.

INTERNATIONAL JOURNAL
OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING