

ADVANCED NEURAL NETWORK ARCHITECTURE FOR DETECTING FRAUD IN INTERNET LOAN APPLICATIONS

P.Anil Jawalkar¹, Ch.Aprana², D.Spandana Reddy², D.Sheetal Reddy²

¹Assistant Professor,²UG Students, Department of Information Technology

^{1,2}Malla Reddy Engineering College for Women, Maisammaguda, Dhulapally, Kompally, Secunderabad-500100, Telangana, India.

To Cite this Article

P.Anil Jawalkar, Ch.Aprana, D.Spandana Reddy, D.Sheetal Reddy, "ADVANCED NEURAL NETWORK ARCHITECTURE FOR DETECTING FRAUD IN INTERNET LOAN APPLICATIONS" *Journal of Science and Technology*, Vol. 08, Issue 12 - Dec 2023, pp94 -104

Article Info

Received: 21-11-2023

Revised: 01 -12-2023

Accepted: 11-12-2023

Published: 21-12-2023

ABSTRACT

The background of the modernized loan approval system lies in the inefficiencies and limitations of traditional loan approval processes. The history of modernizing loan approval systems using machine learning techniques can be traced back to the early 2000s when financial institutions started exploring data-driven approaches to assess credit risks. With the growth of the internet and digitalization, lenders began collecting vast amounts of data on borrowers, including transaction history, social media activities, and online behavior. This data became valuable for predicting creditworthiness and revolutionized the way loans were approved. Traditional loan approval systems typically involved manual paperwork, face-to-face interviews, and subjective judgment. Loan officers would assess applicants based on credit scores, income statements, and collateral. The process was time-intensive and often led to delays in loan approvals. Moreover, these methods were not always accurate in predicting repayment capabilities, leading to higher default rates. In addition, existing methods were often time-consuming, paper-based, and relied heavily on human judgment, making them prone to errors and biases. With the advent of technology and the availability of vast amounts of data, there was a need to develop a more efficient, accurate, and unbiased loan approval system. This need gave rise to the use of machine learning techniques to predict loan approvals based on various factors and data points. Therefore, this research work proposes a machine learning model to develop accurate predictive models that can assess a borrower's creditworthiness using diverse data sources. Further, the proposed model automates the loan approval process, which reduces the time taken for approval, enabling quicker disbursement of funds and it can analyze large datasets to make accurate predictions about a borrower's creditworthiness. This also reduces the operational costs associated with manual loan processing and it will reduce biases in loan approval decisions, promoting fairness and equal opportunities.

Keywords: Internet Loan, Neural Network, Machine Learning, Credit risk.

1.INTRODUCTION

The evolution of modernized loan approval systems stems from the inherent inefficiencies and limitations of traditional loan approval processes that were prevalent in the early 2000s. During this period, financial institutions began to explore data-driven approaches as a means to address the challenges associated with assessing credit risks. Traditionally, loan approval procedures were characterized by manual paperwork, in-person interviews, and subjective evaluations by loan officers. These evaluations were primarily based on factors such as credit scores, income statements, and collateral. Unfortunately, this conventional approach proved to be time-consuming, prone to errors, and often resulted in delays in the approval of loans. Moreover, the methods employed were not always accurate in predicting the repayment capabilities of borrowers, leading to higher default rates. The landscape began to change with the rise of the internet and the increasing trend of digitalization. Lenders started accumulating substantial volumes of data on borrowers, encompassing transaction histories, social media activities, and online behaviors. Recognizing the potential of this data to predict creditworthiness, financial institutions sought to revolutionize the loan approval process.

Recognizing the shortcomings of traditional methods, there emerged a pressing need for a more efficient, accurate, and unbiased loan approval system. The advent of technology and the availability of extensive data paved the way for the integration of machine learning techniques. This innovative approach aimed to leverage diverse data sources to develop predictive models capable of accurately assessing a borrower's creditworthiness. The proposed machine learning model not only seeks to enhance accuracy but also aims to automate the loan approval process. By doing so, the time required for approval is significantly reduced, facilitating quicker disbursement of funds. The model is designed to analyze large datasets, enabling precise predictions regarding a borrower's creditworthiness. This automation not only expedites the approval process but also mitigates the operational costs associated with manual loan processing. A crucial aspect of this evolution is the reduction of biases in loan approval decisions. Traditional methods, reliant on human judgment, were susceptible to errors and biases. The incorporation of machine learning technologies contributes to promoting fairness and equal opportunities in the loan approval landscape. In essence, the history of modernized loan approval systems reflects a progression from traditional, manual processes to data-driven, technologically advanced approaches. The integration of machine learning not only addresses the limitations of the past but also sets the stage for a more efficient, accurate, and equitable future in loan approvals. The genesis of the modernized loan approval system is rooted in the inherent inefficiencies and constraints of conventional loan approval methodologies. Traditional approaches, characterized by manual paperwork, face-to-face interviews, and subjective evaluations, were found wanting in terms of accuracy and efficiency. This inadequacy prompted a paradigm shift in the early 2000s, as financial institutions embarked on the exploration of data-driven strategies to evaluate credit risks. As the internet and digitalization gained momentum, lenders seized the opportunity to amass extensive datasets on borrowers, encompassing transaction histories, social media engagements, and online behaviors. This wealth of data emerged as a pivotal asset, facilitating the accurate prediction of creditworthiness and ushering in a transformative era in loan approval processes. Conventional loan approval practices, reliant on credit scores, income statements, and collateral assessments, proved to be labor-intensive, prone to delays, and, significantly, lacked precision in predicting repayment capabilities, thereby contributing to elevated default rates. Compounded by the time-consuming and paper-based nature of these approaches, coupled with their dependence on subjective human judgment, errors and biases were pervasive in the decision-making process. In response to these challenges, the advent of technology, coupled with the availability of vast datasets, necessitated the development of a more efficient, accurate, and impartial loan approval system. This imperative gave

rise to the adoption of machine learning techniques, leveraging various factors and data points to predict loan approvals. The crux of this research lies in the proposal of a machine learning model designed to construct precise predictive models, assessing a borrower's creditworthiness through the utilization of diverse data sources. Moreover, the proposed model seeks to automate the loan approval process, substantially curtailing the time required for approval and facilitating the expeditious disbursement of funds. Its capacity to analyze large datasets ensures the generation of accurate predictions regarding a borrower's creditworthiness. This automation not only addresses the operational inefficiencies associated with manual loan processing but also holds the promise of reducing biases in approval decisions, thereby fostering fairness and equal opportunities in the lending landscape.

2. LITERATURE SURVEY

[1] Ashwini Etal, The three main components of this study are data collection, data cleaning, and performance evaluation. As a result, it is safe to claim that when it comes to loan forecasting, the Naive Bayes model is superior to other models in terms of effectiveness and yields better outcomes. It functions properly and meets bankers' requirements. This technique accurately and precisely calculates the result. The accuracy of this study is 75%. It correctly predicts whether or not a loan application or customer will be accepted. Experimental results show that the Naive Bayes model performs more effectively.

In [2] Kshitiz Gautam Etal, a decision tree random forest-based machine learning model has been suggested. Finding out about the characteristics, history, and reliability of the loan applicant is the goal of this investigation. In order to address the issue of granting or denying the loan request, or more simply, loan prediction, this study applied exploratory data analysis approaches. This essay's major goal is to decide whether or not to authorize the loan granted to a specific person or organization.

In [3] SOURAV KUMAR, By analyzing the data with the aid of decision tree classifiers, which may produce an accurate result for the prediction, the study's main goal is to determine whether or not the person can obtain a loan. This study came to the conclusion that the Decision tree version is extremely effective and produces a higher final product. This created a model that can predict if someone will return their debt or not with ease. The bankers' work has been cut back thanks to this model. The accuracy of the study's findings is 80%.

In [4] J.Tejaswini 1 Etal used three machine learning methods to forecast a customer's loan approval: Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF). According to the experimental findings, the Decision Tree machine learning algorithm is more accurate than Logistic Regression and Random Forest machine learning methods.

In [5] Pidikiti Supriya1 Etal , investigates whether or not it is safe to give the loan to a specific person. In order to save the bank a lot of time and money, this project aims to lower the risk involved in choosing the safe individual. The Big Data of the individuals to whom the loan was previously issued is mined for this information, and the machine was trained using the machine learning model that produces the most accurate result based on these records and experiences. This study uses gradient boosting, decision trees, and logic regression to make predictions about loan data. It was shown that decision trees make predictions that are straightforward to understand and to interpret. It generates out of bag estimated error, which has been shown to be impartial in numerous testing. It is not too difficult to tune in. It provides the problem's greatest accuracy result, which is close to 90%.

In [6] E.Chandra Blessie Etal, It is inevitable that credit would be extended to businesses and individuals to ensure the smooth operation of developing economies like India. His study makes an

effort to reduce the risk associated with choosing the right borrower who may return the loan on time and keep the bank's non-performing assets (NPA) on hold. This is accomplished by feeding historical data on bank customers who have obtained loans into a trained machine learning model, which could produce reliable results. Determining whether or not it will be secure to distribute the loan to a specific person is the paper's main goal. The sections of this study are (i) Data Collection, (ii) Data Cleaning, and (iii) Performance Evaluation. According to this study, the Naive Bayes model is more effective and generates 80% accuracy.

In [7] Kumar Arun Etal, The major goal of this essay is to determine whether or not it will be safe to assign the loan to a specific individual. This essay is broken down into four sections. Data collection (i) Machine learning model comparison using the data collected (ii) (iii) Testing (iv) System training using the most promising model. With an accuracy of 82%, experimental tests have shown that the Naive Bayes model performs better.

In [8] Kumar employed Decision Tree (DT), Random Forest (RF), and naïve bayes as three machine learning methods to predict whether or not consumers would be approved for loans. According to the experimental findings, the Decision Tree machine learning algorithm is more accurate than naive bayes and Random Forest machine learning techniques.

In [9] Gautam, proposed a machine learning model using the Multilayer perceptron model. He worked to improve accuracy and get beyond the limitations of earlier research. He sought to address the issues with decision tree, naive bayes from earlier research. He came to the conclusion that MLP gives higher results and accuracy (85% accuracy)

In [10] T.Sunitha and colleagues developed a machine learning model based on Naive Bayes to forecast a customer's loan acceptance. The experimental results show that naive bayes machine learning algorithm has superior accuracy than earlier methods. He came to the conclusion that Naive Bayes has an accuracy of 85%, delivers better outcomes, and has a low error rate.

In [11] Suman developed a machine learning model to forecast a customer's loan acceptance utilizing ensemble learning algorithms like decision trees and random forests. Previous research only employed one machine learning technique, which led to less precise findings. He therefore made an effort to increase precision and overcome the restrictions of past study. He came to the conclusion that ensemble learning yields superior results and 90% accuracy

In [12] According to Belaid Bouikhalene and Safi, supervised learning can be used to predict the rank of a scientific research publication. Work is being done by Kumar Arun, Garg Ishan, and Kaur Sanmeet [1] on bank loan prediction on loan approval. With the use of SVM and neural network-like machine learning methods, they put forth a model. We are able to complete our job and make a solid bank loan prediction model with the assistance of this literature review.

3. PROPOSED SYSTEM

3.1 Overview

1) Dataset:

The first crucial step in our research is selecting a suitable dataset. The dataset should encompass a diverse range of software projects with labeled information on defects. This information might include code metrics, historical data, and other relevant features that can aid in predicting defects accurately.

2) Data Preprocessing: Once the dataset is secured, the next step involves data preprocessing. This entails cleaning the data by handling missing values, removing irrelevant features, and normalizing numerical values. Additionally, categorical variables may need encoding to make them compatible with the ensemble learning algorithms.

3) Data Splitting: To evaluate the performance of our model, it's essential to split the dataset into training and testing sets. The training set will be used to train the ensemble learning model, while the testing set will assess the model's ability to generalize to new, unseen data.

6) Performance Evaluation: To gauge the effectiveness of our ensemble learning model, we need robust performance evaluation metrics. Metrics such as precision, recall, F1 score, and area under the receiver operating characteristic (ROC) curve are pertinent for evaluating the model's ability to identify defects accurately. A comprehensive evaluation will provide insights into the strengths and weaknesses of our approach.

7) Prediction from Test Data: Using the trained ensemble learning model, we apply it to the test dataset to make predictions on unseen instances. This step allows us to observe how well the model generalizes to new data and identify potential areas for improvement.

3.2 Data Preprocessing

Handling Missing Values: Begin by identifying and addressing any missing values in the dataset. This might involve imputing missing values using mean or median for numerical features and mode for categorical features. Alternatively, you may choose to remove instances with missing values, depending on the impact on the dataset's integrity.

Removing Irrelevant Features: Analyze the dataset to identify features that do not contribute significantly to the prediction of defects. Removing these irrelevant features not only simplifies the model but also enhances its efficiency and reduces the risk of overfitting.

Normalization of Numerical Values: Normalize numerical features to ensure uniformity in their scales. This step is crucial, especially when using algorithms sensitive to the magnitude of values. Common normalization techniques include Min-Max scaling or Z-score normalization.

Encoding Categorical Variables: Ensemble learning algorithms typically require numerical input. Therefore, categorical variables need to be encoded. This can be achieved through techniques such as one-hot encoding, where categorical variables are converted into binary vectors, making them compatible with the algorithms.

Dealing with Imbalanced Data (if applicable): In software defect prediction, datasets may sometimes be imbalanced, with a disproportionate number of non-defective instances compared to defective ones. Addressing this imbalance can involve techniques like oversampling the minority class or undersampling the majority class to ensure the model is not biased towards predicting the majority class.

Feature Scaling (if applicable): Some ensemble learning algorithms, such as those based on distance metrics, benefit from feature scaling. Ensure that features are appropriately scaled to prevent certain features from dominating the learning process.

3.2 Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

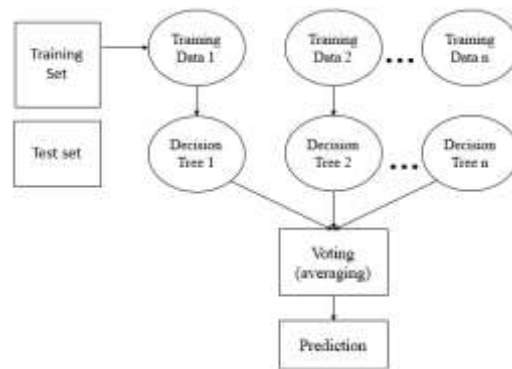


Figure 1: Random Forest algorithm.

3.2.1 Random Forest algorithm

Step 1: In Random Forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.

3.2.2 Important Features of Random Forest

- **Diversity**- Not all attributes/variables/features are considered while making an individual tree, each tree is different.
- **Immune to the curse of dimensionality**- Since each tree does not consider all the features, the feature space is reduced.
- **Parallelization**-Each tree is created independently out of different data and attributes. This means that we can make full use of the CPU to build random forests.
- **Train-Test split**- In a random forest we don't have to segregate the data for train and test as there will always be 30% of the data which is not seen by the decision tree.
- **Stability**- Stability arises because the result is based on majority voting/ averaging.

3.2.3 Assumptions for Random Forest

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random Forest classifier:

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
- The predictions from each tree must have very low correlations.

Below are some points that explain why we should use the Random Forest algorithm

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

3.2.4 Types of Ensembles

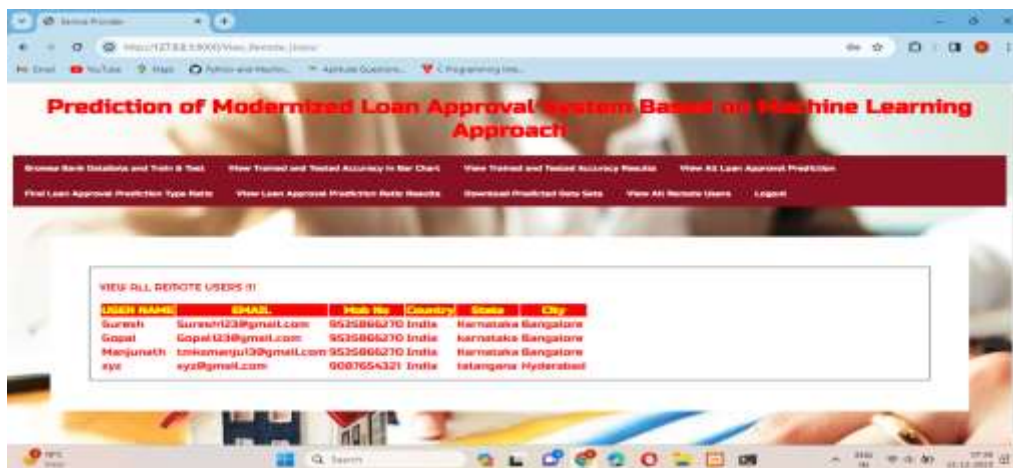
Before understanding the working of the random forest, we must look into the ensemble technique. Ensemble simply means combining multiple models. Thus, a collection of models is used to make predictions rather than an individual model. Ensemble uses two types of methods:

Bagging– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest. Bagging, also known as Bootstrap Aggregation is the ensemble technique used by random forest. Bagging chooses a random sample from the data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as row sampling. This step of row sampling with replacement is called bootstrap. Now each model is trained independently which generates results. The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting is known as aggregation.

Boosting– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST.

4. RESULTS AND DISCUSSION



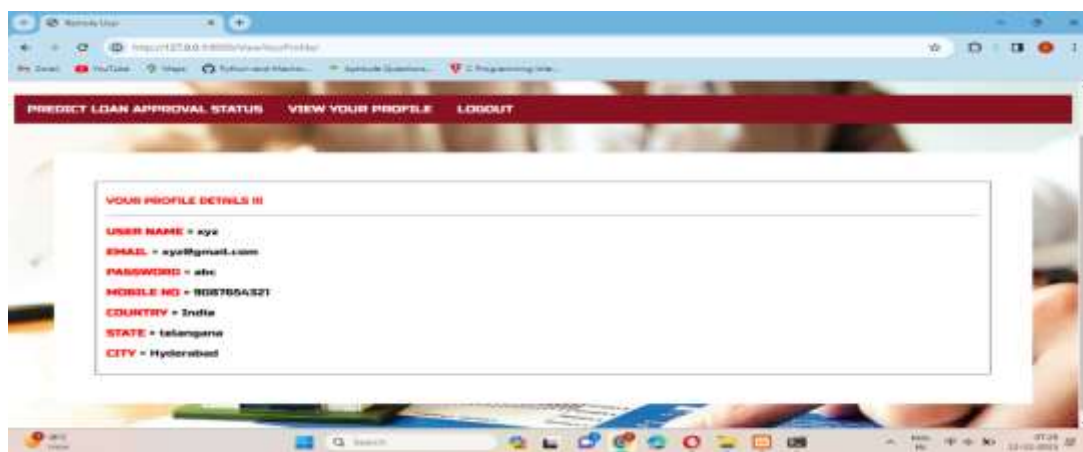




Year All Loan Approval Prediction Status II

Loan ID	Gender	Married Status	Dependents	Education	Employed Status	Applicant Income	Coapplicant Income	Loan Amount	Loan Term (Years)	Credit History	Property Area	Prediction
LP001460	Male	Yes	0	Graduate	No	6080	2569	182	360	0	Rural	Not Approved
LP001460	Male	No	0	Graduate	Yes	20168	0	650	480	1	Urban	Approved
LP001013	Male	Yes	0	Not Graduate	No	2333	1596	95	360	1	Urban	Approved
LP003630	Male	Yes	2	Graduate	No	1298	1066	17	120	1	Semiurban	Approved
LP001950	Male	Yes	1	Graduate	No	1911	139	360	360	0	Semiurban	Approved
LP001990	Male	No	0	Not Graduate	No	1442	0	35	360	1	Urban	Not Approved
LP001990	Male	No	0	Not Graduate	No	1442	0	35	360	1	Urban	Not Approved
LP001730	Male	Yes	0	Graduate	No	2221	0	60	360	0	Urban	Not Approved
LP001463	Male	Yes	2	Graduate	No	4069	1717	116	360	1	Semiurban	Approved
123456	Male	No	Other	Graduate	Yes	6080	5000	60000	3	0	Urban	Approved





5. CONCLUSION AND FUTURE SCOPE

In conclusion, the evolution of loan approval systems from traditional, manual processes to modern, machine learning-based approaches represents a significant leap forward in the financial industry. The limitations of the outdated systems, characterized by time-consuming procedures, subjectivity, and inefficiency, prompted the exploration of innovative solutions.

The integration of machine learning techniques into loan approval processes gained momentum in the early 2000s, aligning with the era of digitalization and the widespread collection of borrower data. The conventional methods, relying on credit scores, income statements, and collateral, often proved inadequate in accurately predicting creditworthiness, resulting in delays and higher default rates.

The proposed SMART LOANS model aims to address these shortcomings by leveraging predictive machine learning models. By incorporating diverse data sources such as transaction history, social media activities, and online behavior, the model enhances its ability to assess a borrower's creditworthiness more comprehensively. The automation of the loan approval process not only

expedites decision-making but also significantly reduces operational costs associated with manual processing.

Furthermore, the model's capacity to analyze large datasets enables precise predictions, contributing to a more accurate evaluation of borrowers' creditworthiness. Importantly, the implementation of this advanced system seeks to mitigate biases inherent in traditional approval methods, fostering fairness and equal opportunities in lending.

REFERENCES

- [1] Lakshman Narayana Vejendla and A Peda Gopi, (2019),” Avoiding Interoperability and Vol 11, Issue 4 , April/ 2020 ISSN NO: 0377-9254 www.jespublication.com Page No:530 Delay in Healthcare Monitoring System Using BlockChain Technology”, *Revue d'Intelligence Artificielle* , Vol. 33, No. 1,2019,pp.45-48.
- [2] Gopi, A.P., Jyothi, R.N.S., Narayana, V.L. et al. (2020), “Classification of tweets data based on polarity using the improved RBF kernel of SVM” . *Int. j. inf. technol.* (2020). [hps://doi.org/10.1007/s41870-019-00409-4](https://doi.org/10.1007/s41870-019-00409-4).
- [3] Lakshman Narayana Vejendla and A Peda Gopi, (2020),” Design and Analysis of CMOS LNA with Extended Bandwidth For RF Applications”, *Journal of Xi'an University of Architecture & Technology*, Vol. 12, Issue. 3,pp.3759-3765. [hps://doi.org/10.37896/JXAT12.03/319](https://doi.org/10.37896/JXAT12.03/319).
- [4] Lakshman Narayana Vejendla and Bharathi C R ,(2018),“Multi-mode Routing Algorithm with Cryptographic Techniques and Reduction of Packet Drop using 2ACK scheme in MANETs”, *Smart Intelligent Computing and Applications*, Vo1.1, pp.649-658. DOI:10.1007/978-981-13-1921-1_63
- [5] Lakshman Narayana Vejendla and Bharathi C R, (2018), “Effective multi-mode routing mechanism with master-slave technique and reduction of packet droppings using 2-ACK scheme in MANETS”, *Modelling, Measurement and Control A*, Vol.91, Issue.2,pp.73-76. DOI: 10.18280/mmc_a.910207
- [6] Lakshman Narayana Vejendla , A Peda Gopi and N.Ashok Kumar,(2018),“ Different techniques for hiding the text information using text steganography techniques: Asurvey”, *Ingénierie des Systèmes d'Information*, Vol.23, Issue.6,pp.115- 125.DOI: 10.3166/ISI.23.6.115-125
- [7] A Peda Gopi and Lakshman Narayana Vejendla (2018), “Dynamic load balancing for client server assignment in distributed system using genetic algorithm”, *Ingénierie des Systèmes d'Information*, Vol.23, Issue.6, pp. 87-98. DOI: 10.3166/ISI.23.6.87-98
- [8] Lakshman Narayana Vejendla and Bharathi C R,(2017),“Using customized Active Resource Routing and Tenable Association using Licentious Method Algorithm for secured mobile ad hoc network Management”, *Advances in Modeling and Analysis B*,Vol.60, Issue.1, pp.270-282. DOI:10.18280/ama_b.600117
- [9] Lakshman Narayana Vejendla and Bharathi C R,(2017),“Identity Based Cryptography for Mobile ad hoc Networks”, *Journal of Theoretical and Applied Information Technology*, Vol.95, Issue.5, pp.1173-1181. EID: 2-s2.0-85015373447
- [10] Lakshman Narayana Vejendla and A Peda Gopi, (2017),” Visual cryptography for gray scale images with enhanced security mechanisms”, *Traitement du Signal*,Vol.35, No.3-4,pp.197-208. DOI: 10.3166/ts.34.197-208
- [11] Cowell,R.G.,A.P.,Lauritez,S.L.,and Spiegelhalter,D.J.(1999). *Graphical models and Expert Systems*. Berlin: Springer. This is a good introduction to probabilistic graphical models.
- [12] Kumar Arun, Garg Ishan, Kaur Sanmeet, May-Jun. 2016. Loan Approval Prediction based on Machine Learning Approach, *IOSR Journal of Computer Engineering (IOSR-JCE)*