FMDB *Transactions on Sustainable Technoprise Letters*

Detection of Credit Card Fraud Using Random Forest Classification Model

M. A. Thinesh[1,*], S. S. Mukhil Varmann[2], S. Leoni Sharmila[3], Sonjoy Ranjon Das[4]

[1,2,]Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India.
[3]Department of Mathematics, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Thandalam, Chennai, Tamil Nadu, India.
[4]Department of Computing, Shipley College, Shipley, England, United Kingdom.
tm9045@srmist.edu.in[1], sm7225@srmist.edu.in[2], leonisharmilas.sse@saveetha.com[3], sanjoy.das@shipley.ac.uk[4]

Abstract: Credit card fraud significantly threatens financial institutions and consumers worldwide. To address this issue, this project leverages machine learning techniques, specifically a RandomForest Classifier, to detect fraudulent credit card transactions. The dataset is from Kaggle and contains transaction details, including transaction amounts and class labels indicating fraud or non-fraudulent transactions. The project begins with data exploration and visualization to gain insights into the dataset's characteristics. It uses various data visualization techniques, such as classification plots and correlation matrices, to understand the understood patterns. After preprocessing the data and dividing it into training and test sets, the random forest classifier is trained on training data. Learning curves visualize the model's performance as the training dataset size varies. A comprehensive set of metrics is utilized to evaluate the model's effectiveness. It includes accuracy, specificity, error rate, and a confusion matrix to assess the model's ability to classify fraudulent and non-fraudulent transactions. In addition, precision, recall, and F1-score are computed. Receiver Operating Characteristic (ROC) and Precision-Recall curves are generated to give a detailed understanding of the model's performance and to assess the power mean to discriminate between classes' precision-recall trade-offs. The project concludes with an evaluation of the model's performance, highlighting its strengths and areas for improvement. This project serves as a valuable example of the application of machine learning for fraud detection in financial transactions, benefiting financial institutions and consumers by reducing financial losses due to fraud and increasing security.

Keywords: Random Forest Classification Model; Accuracy and Precision; Receiver Operating Characteristic; F1-Score and Recall; Machine Learning; Confusion Matrix; Credit Card and Personal Information; Reducing Financial Losses.

Received on: 15/06/2023, **Revised on**: 10/09/2023, **Accepted on:** 11/10/2023, **Published on:** 24/12/2023

Cite as: M. A. Thinesh, S. S. Mukhil Varmann, S. Leoni Sharmila, and S. Ranjon Das, "Detection of Credit Card Fraud Using Random Forest Classification Model," *FMDB Transactions on Sustainable Technoprise Letters*, vol. 1, no. 4, pp. 181–194, 2023.

1. Introduction

Cashless transactions are becoming increasingly common worldwide as credit card usage rises. These credit cards enable simplified purchases both online and in-person, taking the place of paper bills. Credit cards are accepted online and in physical stores, making them the best replacement for cash. The convenience of digital transactions is a significant benefit of credit card usage. In the coming years, the number of credit cardholders globally is expected to grow by 2.79%, reaching 1.25 billion in 2023 from 1.1 billion in 2018. Credit cards are used in restaurants, petrol bunks, and online shopping [13]. Credit Cards give money when needed and can be returned later with no interest for a specific period, which is a big reason for the increase in the

[*]Corresponding author.

usage of credit cards [14]. Not everyone will have cash in hand when needed for very important situations. At those times, we cannot go and get cash; rather, credit cards are used to get money for whatever we need and pay it back later. There are many advancements in credit cards, too [15]. One of which is you can now tap the credit card on the credit card machine in physical stores like grocery shops, restaurants, and petrol bunks to pay the bill, which can save a lot of time before you have to scan the credit card in the credit card machine and type the pin. It has to be processed, which takes a lot of time [16].

Credit cards store people's credit card information. However, an individual or a party does not have the right to use the credit card of another party named without resorting to credit card fraud. These types of crimes have a serious impact on individuals and institutions [17]. One of the main causes of credit card fraud is the rapid advancement of technology. The Internet has grown and changed during the last decade and continues to evolve at a fast tempo. This includes e-commerce, tap price fashions, and online payments [18]. This has brought about the growth and substantial use of such features, leading to a boom in credit card fraud. With the great increase in digital transactions and online shopping, criminals discover new ways to commit such crimes daily [19]. Credit card information can be obtained from places so easily in practice. There are many methods to defend credit card transactions, including encryption and tokenizing credit card facts. While these methods are regularly effective, it does now not shield credit cards from fraud [20].

Credit card fraud is described as fraud by unauthorized humans using credit or debit cards. Credit card fraud is a common problem in the digital world and causes financial losses for individuals and businesses [21]. Credit card fraud can take many forms, such as identification robbery, cloning, forgery, skimming, and phishing [22]. In each of those instances, the fraudster attempts to reap the sufferer's credit score card records, either online or via bodily stealing a wallet or different item containing a credit card, and the fraudster obtains the facts after entirety, which may be used for purchases or withdrawals. Due to technological advancements, credit card fraud has become one of the biggest threats to people [23]. Scammers may send fake emails that look like they are from a bank or credit card issuer and may steal your personal information [24].

Hackers can even breach security and steal customers' personal information, including credit card information. The increased usage of online transactions where a physical credit card is unnecessary makes it riskier, as fraudsters may steal the card information and use it for their benefit [25]. Solving this problem will require proactive measures from both individuals and businesses. Businesses should invest in reliable security systems that can safeguard customers' data, and individuals must monitor their bank statements regularly and check for any suspicious activities; if they find some suspicious activities, they must report to the authorities immediately [26]. PwC's 2022 Global Crime Survey found that approximately 51% of organizations surveyed said they had encountered fraud in the past two years. According to the American Federal Trade Commission (FTC), credit score card fraud is the most unusual problem, accounting for 1,579 facts breaches and 179 million information factors [27]. Credit cards have become a major concern for consumers and financial institutions. As technology evolves, so do the techniques used to exploit physical weaknesses. To solve this problem, machine learning algorithms have become powerful tools for business fraud detection [28].

Machine Learning is an artificial intelligence department that allows computers to analyze from beyond records and increase unexplained predictive skills. Machine learning algorithms can adapt and enhance as new situations are encountered [29]. By reading consumer remarks and including additional statistics in the training records, the version will become more accurate at detecting fraud patterns [30]. One such algorithm is the random forest classifier, used in credit card fraud in this paper. The algorithm creates a series of decision trees and combines their predictions to make the correct classification. The algorithm can recognize the patterns of fraud and make new predictions accordingly [31]. The random forest method is especially useful when dealing with large data sets with many features and is ideal for credit card fraud.

One of the largest problems with using machines to resolve credit card fraud is that maximum reporting is impossible because credit card data are very personal [32]. Therefore, the statistics used to create gadgets to learn fashions for credit card evaluation have unknown abilities [33]. Credit card fraud is also a large trouble due to adjustments and constant modifications. Additionally, present device mastering fashions for credit card fraud detection are flawed and cannot cope with credit card fraud records nicely, so there is a want to expand higher trained models with excessive ratings that could capture credit card fraud [34].

2. Objective

- Using the Random Forest Classifier for credit score card fraud detection aims to pick out fraudulent transactions and accurately minimize false positives. This set of rules works by constructing more than one choice bushe based totally on random subsets of the data, after which they combine their predictions to make a very last decision.
- By utilizing the Random Forest Classifier, we aim to obtain an excessive level of accuracy in detecting fraudulent transactions. This will permit monetary institutions and credit score card organizations to take instantaneous movement while suspicious activities are detected, thereby preventing financial losses for individuals and businesses.

- Furthermore, by minimizing false positives, we can ensure that legitimate transactions aren't mistakenly flagged as fraudulent. This is important in preserving client pleasure and belief within the credit card machine.

3. Literature Survey

Alarfai et al. [1] discuss credit card fraud detection using state-of-the-art machine learning and deep learning algorithms. Various models, including CNN sequential models, were developed and compared for accuracy and performance. The CNN models were trained with different epoch sizes to optimize results.

Ghaleb et al. [2] discuss utilizing Ensemble Synthesized Minority Oversampling-Based Generative Adversarial Networks and the Random Forest Algorithm for credit card fraud detection. The authors aim to improve fraud detection through innovative techniques and interdisciplinary approaches.

Mienye and Sun [3] present a deep-learning ensemble approach for credit card fraud detection using LSTM and GRU neural networks. A multilayer perceptron (MLP) is the ensemble framework's meta-learner. The authors address the critical issue of fraud detection in disruptive technologies.

Kalid et al. [4] discuss credit card fraud detection challenges due to imbalanced class distribution and overlapping classes. The authors provide valuable insights for future research on credit card fraud and payment default detection.

Esenogho et al. [5] present a neural network ensemble with feature engineering for enhanced credit card fraud detection. The authors highlight the importance of advanced techniques in addressing the challenges of imbalanced datasets in fraud detection.

Alamri and Ykhlef [6] present a hybrid sampling method, BCBSMOTE, to balance imbalanced credit card transaction datasets. By combining Tomek links for undersampling and BIRCH clustering with Borderline-SMOTE for oversampling, the authors aim to improve fraud detection models.

Nguyen et al. [7] present an approach using CatBoost and neural networks to enhance fraud detection in the financial industry. The author aims to improve efficiency in real-time fraud detection scenarios by eliminating redundant features.

Ding et al. [8] discuss enhancing credit card fraud detection through an innovative oversampling method using VAEGAN. The training set is enriched by generating diverse and convincing synthetic fraud samples for improved classification accuracy.

Taha and Malebary [9] present an intelligent approach to credit card fraud detection using an optimized Light Gradient Boosting Machine.

Tingfei et al. [10] discuss using Variational Auto Encoding (VAE) in credit card fraud detection to address imbalanced datasets. The approach involves generating diverse synthetic cases from minority groups to enhance the training set.

Ghosh and Reilly [11] discuss developing and testing a neural network-based fraud detection system for credit card transactions. It highlights the significant improvements in fraud detection accuracy and reduction in false positives achieved by the neural network compared to traditional rule-based methods.

Alenzi and Nojood [12] suggested using a logistic regression model for credit card fraud detection. They have also used the K-nearest neighbors model and the voting classifier model. The logistic regression model was then compared with the other two models, showing better results in accuracy, sensitivity, and error rate.

4. Methodology

4.1. Existing model

Alenzi and Nojood [12] suggested using a logistic regression classifier, K-nearest neighbors classifier, and voting classifier for fraud detection in credit cards. Their dataset contained 492 fraudulent cases out of 284,807 data, which is highly unbalanced.

Logistic regression is a machine learning algorithm for binary and multiclass classification tasks. Even though its name is Logistic Regression, it is not a regression algorithm but a classification algorithm. For binary classification, logistic regression predicts whether it is of any of the targeted classes. For Multiclass, there are more than two target classes, which are achieved using one-vs-all or softmax regression. The logistic regression classifier got results with an accuracy of 97.2%, sensitivity of 97%, and error rate of 2.8%.

K-Nearest Neighbours classifier is a simple classification algorithm that is widely used. In KNN, the model predicts new data points based on the similarity to the neighboring data points in training data. KNN can also be used for regression tasks. This model performs well with both small and large data. The K-Nearest Neighbours classifier got results with an accuracy of 93%, a sensitivity of 94%, and an error rate of 7% [35].

A Voting Classifier, also known as an Ensemble Voting Classifier or Majority Voting Classifier, is a machine learning model that combines the predictions from multiple individual classifiers (or models) to make a final prediction [36]. The idea behind using a voting classifier is to leverage the strengths of different classifiers to improve overall prediction accuracy and robustness, especially when individual classifiers may have different biases or make errors in different ways [37]. The Voting Classifier got results with an accuracy of 90%, sensitivity of 88%, and error rate of 10%.

4.2. Proposed model

A classification algorithm that uses multiple decision trees is proposed to perform a classification task to determine whether the transaction is fraud. The classification algorithm used to classify whether the transaction is fraudulent or not is the RandomForestClassifier.

RandomForestClassifier is a popular machine learning algorithm generally used for classification tasks and can be adapted for regression. RandomForestClassifier can work with datasets with a mix of categorical and numerical features. RandomForestClassifier produces prediction with high accuracy. It uses multiple decision trees, reducing the risk of overfitting. RandomForestClassifier is best for handling large and unbalanced datasets [38]. The RandomForestClassifier is an efficient algorithm used in a wide range of applications. RandomForestClassifier can be used in many industries and applications for classification tasks, and some areas are marketing and advertising, Healthcare, Weather and climate, etc [39]. The model can make predictions about which one of the categories it comes under. RandomForestClassifier is from the sci-kit-learn library in Python. It can imported from the 'scklearn.ensemble' module [40].

The proposed RandomForestClassifier model is trained using a dataset containing a total of 30 columns, of which 28 columns are encoded personal data and two other columns contain the amount transferred and whether the transaction is fraud. The dataset contains 7231 data with 30 columns [41]. The column "Class" is the target variable as it contains the required data for classification, whether the transaction is fraud or not. The column Class contains data in the form of 1 and 0; 1 represents the transaction is fraud, and 0 represents the transaction is not fraud. The target and categorical columns are split into target and categorical variables [42]. Overfitting occurs when the model learns the training data too well and thus will have high training accuracy but low test accuracy [43]. To avoid overfitting, the dataset is split into as 80:20 ratio for training and testing. The training dataset contains 5784 data, and the testing dataset contains 1447 [44].

The training dataset is used to train the RandomForestClassifier Model, and the trained model will be validated using the testing dataset [45]. The model is trained with the encoded personal data, with the amount as one variable and the class as the target variable. The model is tested with the testing dataset with the encoded amount in one variable, and the model is used to classify whether the transaction is fraud or not [46]. The model is then validated using multiple parameters such as precision, accuracy, F1-score, Recall, Error rate, specificity, and ROC curve. The parameters are then visualized using multiple graphs from the Seaborn and matplotlib libraries [47]. These parameters show how well the model can classify the data and help us check if the model can be used in real-time [48].

The proposed model, the RandomForest Classifier, has performed better than the existing Logistic Regression Classifier method in performing credit card fraud detection. Both were trained on highly unbalanced datasets with low fraud and high non-fraud data. The models are compared based on the accuracy and error rate of the parameters.

Table 1: Parameters comparison between existing and proposed models

Classifier	**Metrics**	
	Accuracy	**Error Rate**
Logistic Regression	97.2%	2.8%
K-NN Classifier	93%	7%
VC Classifier	90%	10%
Random Forest	99.59%	0.41%

Table 1 compares the parameters obtained by the existing model and the proposed model. The existing models' Logistic regression has an accuracy of 97.2%, the K-NN Classifier has an accuracy of 93%, and the VC Classifier has an accuracy of 90%. In contrast, in the RandomForest Classifier, the proposed model achieved an accuracy of 99.59%, which makes the proposed model highly accurate while the existing model lagging. The other parameter used to compare is error rate, which represents classification errors. Random Forest Classifier is a better way to minimize classification errors as it has a very low error rate of 0.41%. In contrast, the LogisticRegression, K-NN Classifier, and VC Classifier have higher error rates of 2.8%, 7%, and 10%, respectively, showing poorer classifier performance than the random forest classifier.

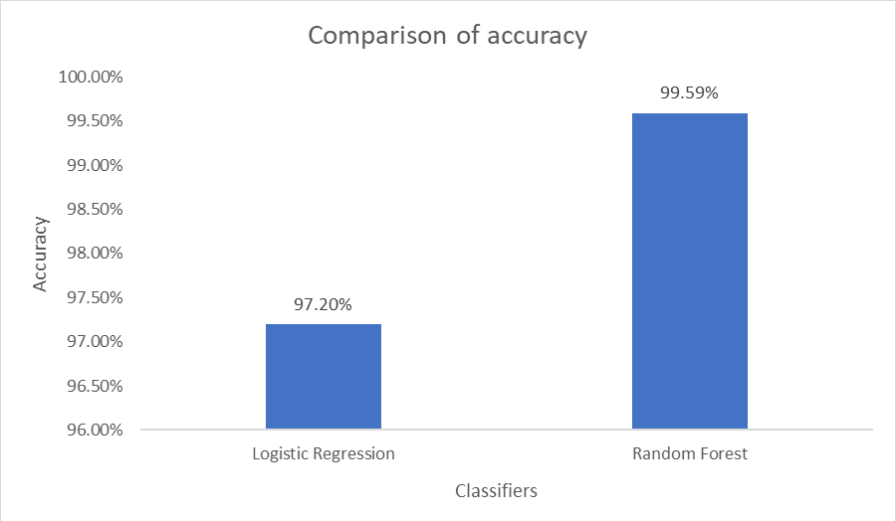

Figure 1: Comparison of accuracy between models

Figure 1 shows the visualization of the accuracy comparison between the best of the existing model Logistic Regression and the proposed model Random Forest Classifier. On the horizontal axis, we have two models, "Logistic Regression" and "Random Forest Classifier. The vertical axis represents the accuracy score, measured as a percentage (%). There are two bars on the chart. One bar represents the accuracy of the regression model representing the Logistic, and the other represents the accuracy of the random forest classification model. The Logistic Regression is 97.20%, and the Random Forest Classifier is 99.59%. This makes it very clear how the proposed model performs well and shows the huge difference in accuracy between the two models. Accuracy is the most important parameter when comparing models, and this figure shows how better the proposed model performs than the existing model.

4.3. Architecture

Figure 2 shows the architecture diagram of the model. First, the necessary libraries, such as numpy, pandas, matelot, seaborn, and sklearn, are imported. The sklearn library provides a RandomForestClassifier model and functions to calculate the result parameters such as accuracy, average precision, average recall, average f1-score, error rate, learning curve, and ROC Curve. The dataset is then loaded using the Pandas library, which uses Pandas functions for data exploration. The data are visualized using matplotlib and Seaborn Library. There are three graphs: a bar graph between fraud and not fraud counts, a Line Plot of Amount, and Heatmap Of Correlation between Columns. The target column and categorical columns are separated into two variables. The dataset is split into 80% and 20% training and testing data, respectively. This is done so that the model doesn't overlearn the data and gives poor performance when facing new data.

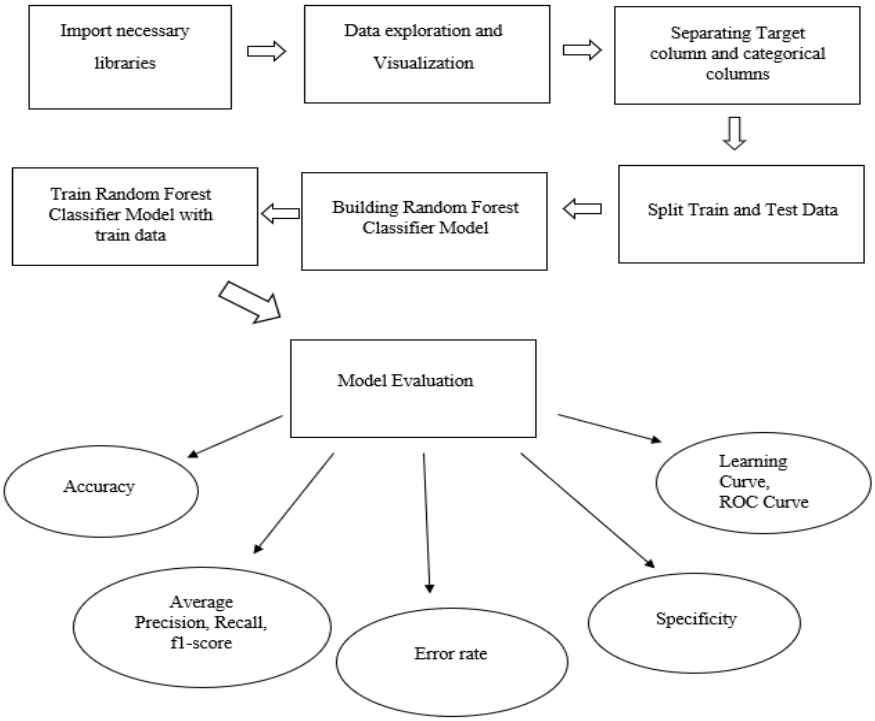

Figure 2: Architecture Diagram

The Random Forest Classifier is built using the RandomForestClassifier function. The Random Forest Classifier Model with the training data. The model is then evaluated using various parameters: Accuracy, Average Precision, Average Recall, Average f1-score, Error rate, and Specificity. The Learning and ROC curves are then plotted using the data to evaluate the model.

4.4. Algorithm : Credit Card Fraud Detection with RandomForest and Evaluation

Step 1: Import Necessary Libraries

- Import pandas, numpy, seaborn, matplotlib, pyplot, RandomForestClassifier, train_test_split, and relevant functions for metrics.

Step 2: Load Data

- Read a CSV file containing credit card fraud data in a Pandas Data Frame.

Step 3: Data Exploration

- Display basic information about the dataset:
- Display the first few rows of the data (data. head ()).
- Determine the shape of the data (data. shape).
- Generate descriptive statistics (data. describe ()).
- Display data type information (data.info ()).
- Display the last few rows of the data (data. tail ()).
- Check for missing values (data. isnull(). sum ()).

Step 4: Data Visualization

- Use seaborn and matplotlib to visualize the data:
- Create a bar graph of ‘Class’.
- Create a line plot of ‘Amount’.
- Create a Heatmap of all columns.

Step 5: Prepare Data for Prophet

- Rename Data Frame columns to ‘x’ (all columns except target variable) and ‘y’ (target variable).

Step 6: Split Data into Training and Testing Sets

- Define the training size (e.g., 80% of the data).
- Split the data into training and testing sets.

Step 7: Create and Configure RandomForestClassifier Model

- Create a RandomForestClassifier model with specific configuration settings (n_estimators, random_state, etc.).
- Fit the RandomForestClassifier model to the training data.

Step 8: Make Predictions with the RandomForest Classifier

- Use the trained model to make predictions on credit card fraud.

Step 9: Evaluate Model Performance

- Calculate various evaluation metrics such as Accuracy, Recall, F1-score, Specificity, Error rate, AUC Score, and Precision.
- Visualize the evaluation metrics.

Step 10: End

- End of the algorithm.

4.5. Execution

To implement the RandomForest Model, we need to import it. Sklearn library from Python provides a random forest classifier model that can be imported from the library. It can done using the following code: “rfc = RandomForestClassifier(n_estimators=100, max_depth=10, random_state=42)”.Then, we classify whether the transaction is fraud or not. The minimum requirements for this model are a Python interpreter and the necessary Python libraries.

5. Implementation

5.1. Data and pre-processing

First, we need to acquire credit score card fraud information on how to use it for the random forest model. The data is acquired online from Kaggle. After data is acquired, the dataset is processed by cleaning records. This can also include changing categorical variables to numeric variables and handling outliers and other inconsistencies in the statistics. The dataset used for the proposed model has no null values. It contains no categorical variables, reducing the time and effort of cleaning the records and encoding the data. The dataset is then split into training data and testing data. This is vital for evaluating the model’s overall performance on unobserved statistics. After splitting the information, you can train the Random Forest classifier model. Finally, we use an assessment method to evaluate version performance. This will help you determine if the model can detect credit score card fraud.

5.2. Data visualisation

The data seen in numerical terms can be better for machines but not for humans. Visualizing the data helps to understand how the data is in the dataset.

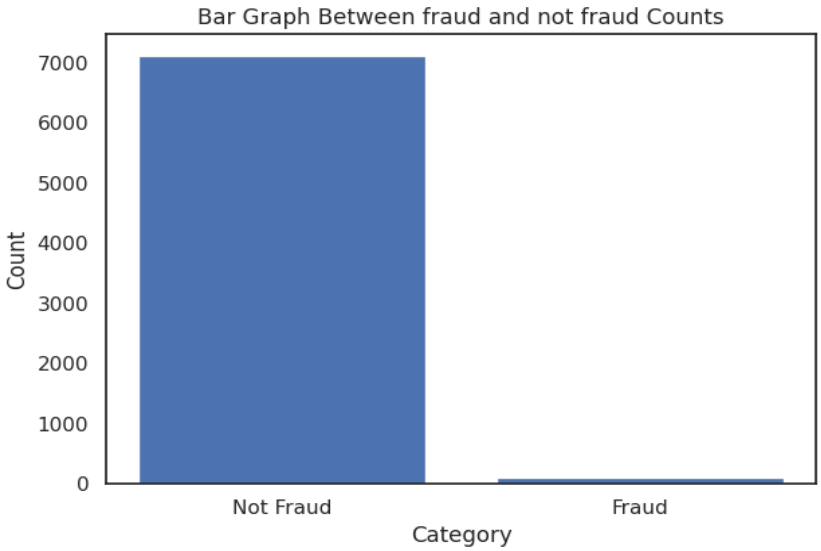

Figure 3: Bar graph of fraud and not fraud case count

The bar graph shows the count of data in the dataset facing credit card fraud and not facing credit card fraud. Figure 3 shows that more than 7000 inputs are found not facing credit card fraud issues, while the rest face credit card fraud issues from the dataset taken. This shows that the dataset is highly unbalanced, which makes the proposed model random-forest classifier a great choice for classification.

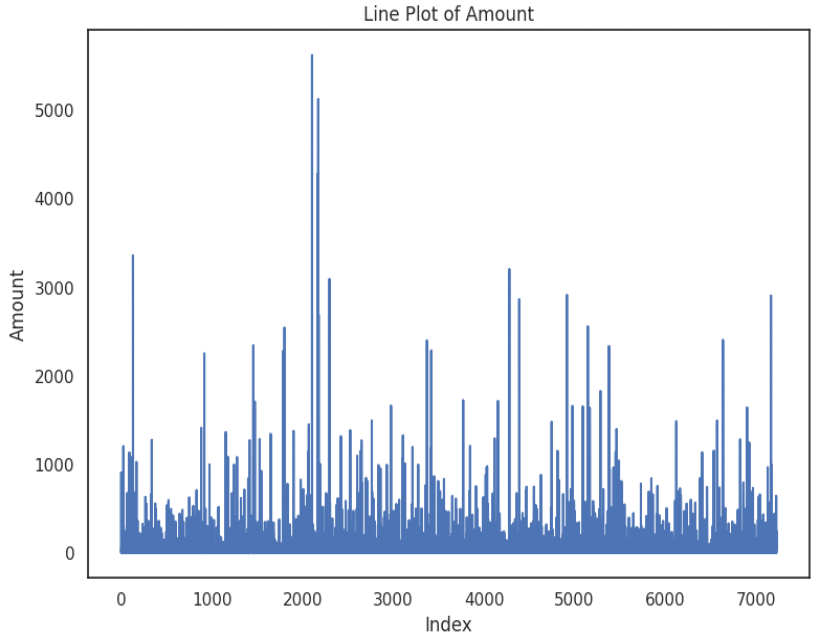

Figure 4: Line Plot of Amount

Figure 4 shows a line plot graph of the amount in the dataset. Line Plot is used to show trends and patterns in the data. This line plot shows the different numbers of amounts present in the dataset. This Line plot shows the trend and patterns of the number of transactions done by credit cards. This also helps us to visualize the highest amount, lowest amount, and average amount transacted using a credit card.

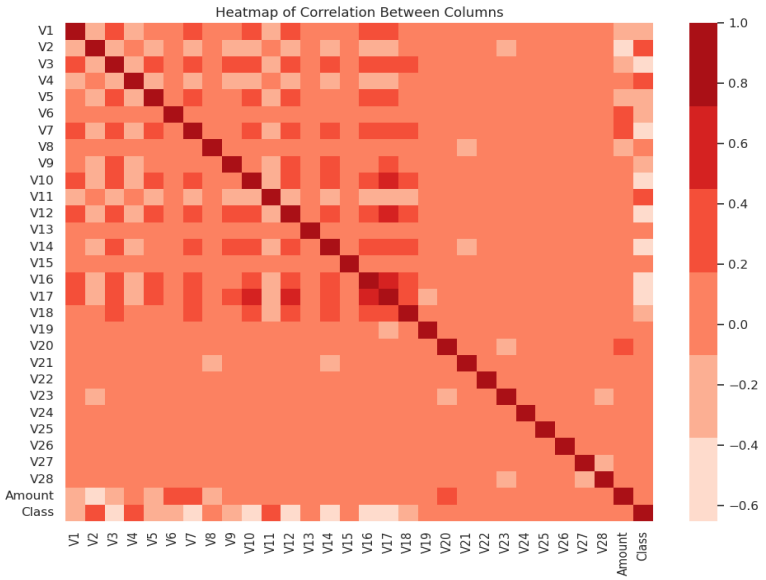

Figure 5: Heatmap between columns

Figure 5 contains the heatmap between all the columns in the dataset. A heatmap is a graphical representation of data that uses a system of color-coding to represent different values. The heatmap can visualize various data, such as the correlation between two variables or the distribution of values across a dataset. They can be used to identify trends, patterns, and correlations between variables. Here, the heatmap shows the correlation between all the variables in the columns.

5.3. Training

To avoid overfitting the data, the processed dataset is split into training and testing datasets in the ratio 80:20, respectively. The first built random forest classifier function is the random forest classifier model, which can be found in the sklearn library. The n_estimators is set to 100, max_depth is set to 10, and random_state is set to 42. The model is then trained with the training dataset. We train the model using Python modules pandas, matplotlib, numpy, and sklearn.

5.4. Evaluation

After the model has been created, and it has to be evaluated to ensure that it's working properly as it should be. Finally, the model have to be examined on unseen facts to ensure that it's miles operating effectively. This is achieved by introducing test data into the model and then analyzing the results. The model can be evaluated with the usage of numerous metrics inclusive of accuracy, precision, recall, F1 score, Specificity and Error rate. These can be found using the following formulas.

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions}$$

$$Precision = \frac{TP}{(TP + FP)}$$

Where,

TP (True Positives) is the number of positive instances predicted correctly.
FP (False Positives) is the number of cases predicted as positive but negative.

$$Recall = \frac{TP}{(TP + FN)}$$

Where,

TP (True Positives) is the number of positive instances predicted correctly.
FN (False Negatives) is the number of cases predicted as negative but positive.

$$F1\ Score = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)}$$

$$Specificity = \frac{TN}{(TN + FP)}$$

Where,

TN (True Negatives) is the number of negative instances predicted correctly.
FP (False Positives) is the number of cases predicted as positive but negative.

$$Error\ Rate = \frac{(FP + FN)}{(TP + TN + FP + FN)}$$

Where,

TP (True Positives) is the number of positive instances predicted correctly.
TN (True Negatives) is the number of negative instances predicted correctly.
FP (False Positives) is the number of cases predicted as positive but negative.
FN (False Negatives) is the number of cases predicted as negative but positive.

All these parameters are calculated and used to evaluate the model. These parameters are also visualized using matplotlib and seaborn library for easy understanding of how the model performs.

6. Results and Discussion

In this paper, we chose Python to develop our RandomForest classifier model. The proposed model was run and evaluated on Windows 11 with AMD Ryzen 7 5800x, 32 GB RAM, RTX 3060 TI using Google Collab using Python language; Google Collab offers free GPU runtime, which increases the training speed of the model, so even if you have less GPU Google Collab helps in faster training of the model. The dataset is used to train the proposed random-forest model. The model is tested and validated using a test dataset. The proposed model classifies whether or not the transaction is fraud with the data provided. The model is evaluated using Accuracy, Average Precision, Average Recall, Average F1 score, specificity, and Error rate. The model is also evaluated using the ROC Curve. The parameter values are Accuracy 99.59%, Average Precision 100%, Average Recall 88%, Average F1-score 93%, Specificity 100%, and Error rate 0.41%. A Learning Curve is used to compare training accuracy and validation accuracy.

Table 2: Classification Report

Class	**Metrics**		
	Precision	**Recall**	**F1-score**
Non-Fraud	100%	100%	100%
Fraud	100%	75%	86%

Table 2 is the classification report generated by our model. The classification report includes precision, recall, and F1-Score of both target values, as well as whether the transaction is fraudulent. The above data shows how the parameters differ for both classes due to the highly unbalanced dataset. The data can be seen by visualizing the data.

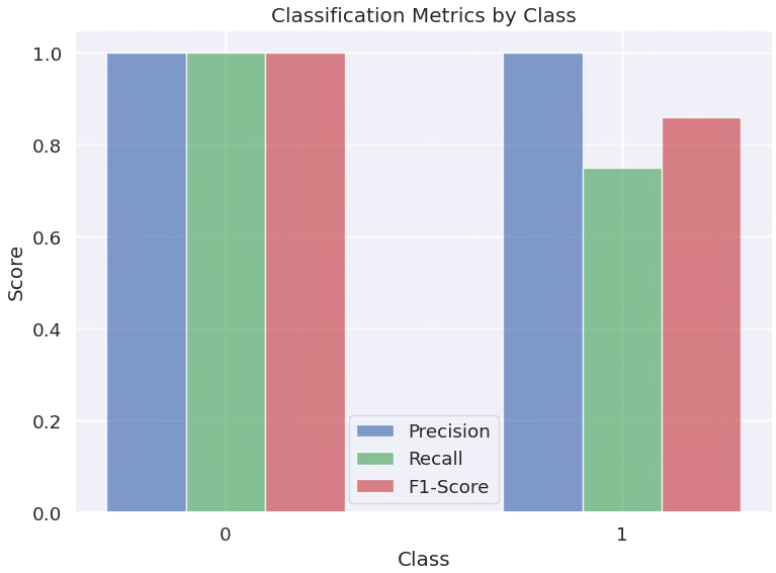

Figure 6: Bar graph between fraud class and non-fraud class result parameters

The bar graph is the graphical representation of the classification report, which can be seen in the table above. In Figure 6, 0 represents the Non-Fraud Class, and 1 represents the Fraud Class. This graphical representation clearly shows the difference between the results between the two classes. The recall and F1-score of the fraud class are low compared to the non-fraud class because the dataset contains 7120 non-fraud cases and 111 fraud cases, from which the model has huge data to learn for non-fraud cases, whereas the model learns very little data on fraud case. Due to the lack of learning data, the model cannot get high recall and F1 scores.

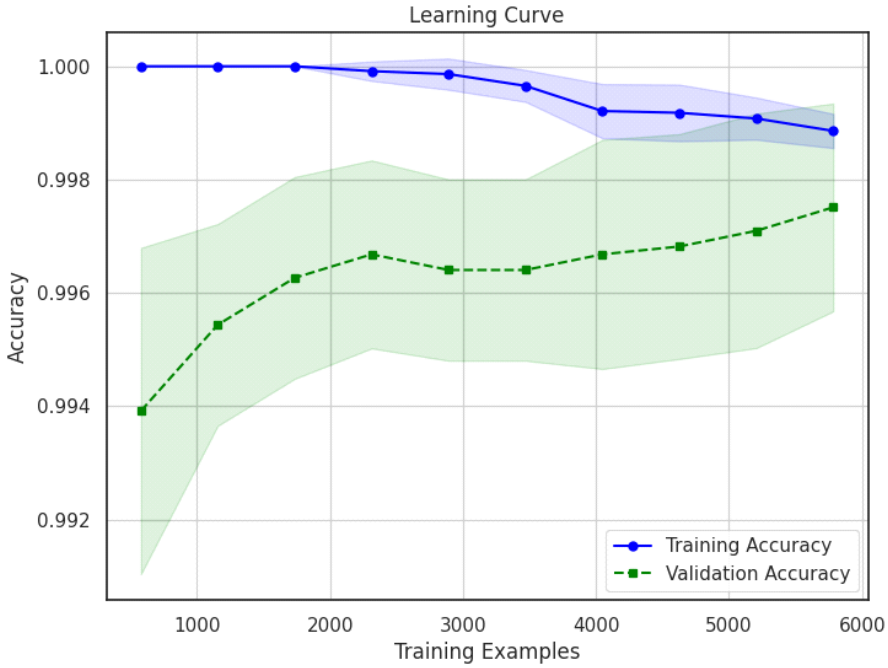

Figure 7: Learning Curve

Figure 7 is the learning curve, which indicates the accuracy of the training phase and testing phase. It clearly shows that the testing and training accuracy have less difference, which makes the model perform well. The validation accuracy starts low because the model is learning the data, and then the accuracy gradually increases, showing that the model gradually learned all the data. This low starting accuracy is due to the dataset being highly unbalanced, but the random forest classifier is best for those cases, and so can be seen in that increase in the graph.

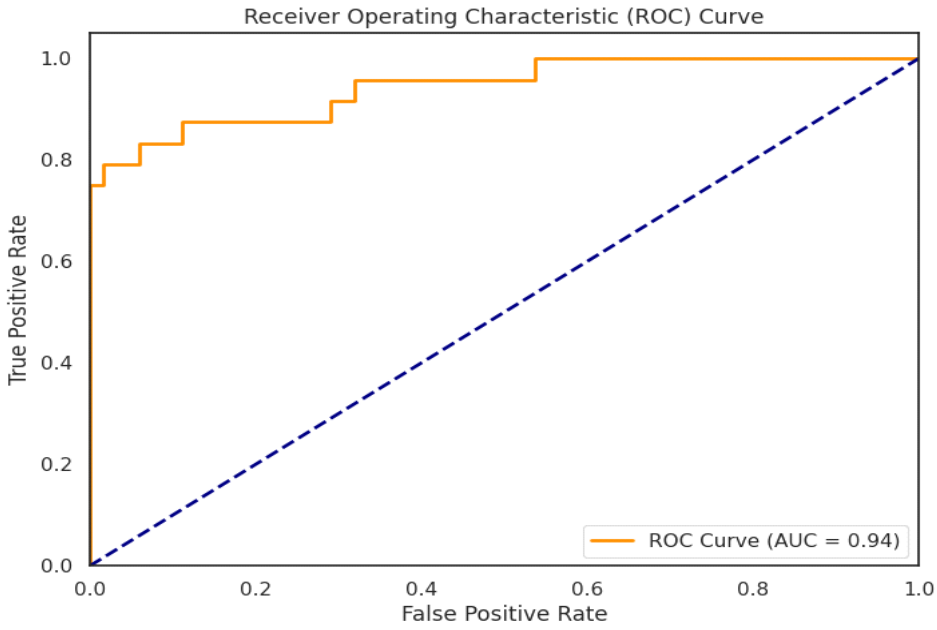

Figure 8: ROC Curve

Figure 8 shows the ROC Curve for the proposed random forest classifier model. ROC curve is a graphical representation used to check the performance of binary classification models. From the ROC Curve, we find that the AUC Score for the proposed model is 0.94, which is very high as the range of the AUC Score is from 0 to 1, and 0.94 is very close to 1, which shows the model's effective performance in distinguishing between positive and negative examples.

7. Conclusion

Credit card verification is a vital part of the search. This is because the variety of frauds in financial institutions is increasing. This problem opens the door to using artificial intelligence to create systems that may attack fraud. Creating an AI-based totally machine to hit upon fraud requires a dataset to train the system. To triumph over these troubles, a classifier based totally on RandomForest has been proposed in this publication. The model is trained with the training dataset. Finally, the proposed approach was evaluated in steps, including accuracy, average precision, average recall, average f1-score, specificity, and error rate. The proposed RandomForest classifier accomplished good results (accuracy = 99.59%, average precision = 100%, average recall = 88%, average f1-score = 93%, specificity = 100% and error rate = 0.41%). Credit card fraud is a large trouble due to constant changes and adjustments. Present devices mastering fashions for credit card fraud detection cannot manage fraud records nicely, so higher-trained models need to be expanded to capture credit card fraud more efficiently.

Acknowledgment: The support of all my co-authors is highly appreciated.

Data Availability Statement: The research contains data related to credit card transaction details with fraud. The data consists of encoded data with the amount and whether the transaction is fraudulent. The dataset is from Kaggle.

Funding Statement: No funding has been obtained to help prepare this manuscript and research work.

Conflicts of Interest Statement: No conflicts of interest have been declared by the author(s). Citations and references are mentioned in the information used.

Ethics and Consent Statement: The consent was obtained from the organization and individual participants during data collection, and ethical approval and participant consent were received.

References

1. F. K. Alarfaj, I. Malik, H. U. Khan, N. Almusallam, M. Ramzan, and M. Ahmed, "Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms," IEEE Access, vol. 10, pp. 39700–39715, 2022.
2. F. A. Ghaleb, F. Saeed, M. Al-Sarem, S. N. Qasem, and T. Al-Hadhrami, "Ensemble synthesized minority oversampling-based generative adversarial networks and random forest algorithm for credit card fraud detection," IEEE Access, vol. 11, pp. 89694–89710, 2023.
3. I. D. Mienye and Y. Sun, "A deep learning ensemble with data resampling for credit card fraud detection," IEEE Access, vol. 11, pp. 30628–30638, 2023.
4. S. N. Kalid, K.-C. Khor, K.-H. Ng, and G.-K. Tong, "Detecting frauds and payment defaults on credit card data inherited with imbalanced class distribution and overlapping class problems: A systematic review," IEEE Access, vol. 12, pp. 23636–23652, 2024.
5. E. Esenogho, I. D. Mienye, T. G. Swart, K. Aruleba, and G. Obaido, "A neural network ensemble with feature engineering for improved credit card fraud detection," IEEE Access, vol. 10, pp. 16400–16407, 2022.
6. M. Alamri and M. Ykhlef, "Hybrid undersampling and oversampling for handling imbalanced credit card data," IEEE Access, vol. 12, pp. 14050–14060, 2024.
7. N. Nguyen et al., "A proposed model for card fraud detection based on CatBoost and deep neural network," IEEE Access, vol. 10, pp. 96852–96861, 2022.
8. Y. Ding, W. Kang, J. Feng, B. Peng, and A. Yang, "Credit card fraud detection based on improved variational autoencoder generative adversarial network," IEEE Access, vol. 11, pp. 83680–83691, 2023.
9. A. A. Taha and S. J. Malebary, "An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine," IEEE Access, vol. 8, pp. 25579–25587, 2020.
10. H. Tingfei, C. Guangquan, and H. Kuihua, "Using variational auto encoding in credit card fraud detection," IEEE Access, vol. 8, pp. 149841–149853, 2020.
11. Ghosh and Reilly, "Credit card fraud detection with a neural-network," in Proceedings of the Twenty-Seventh Hawaii International Conference on System Sciences HICSS-94, Wailea, Hi, Usa, Pp. 621-630, 1994, Doi: 10.1109/Hicss.1994.323314.
12. H. Z. Alenzi and Nojood, "Fraud detection in credit cards using logistic regression," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 12, 2020.
13. A. Charmchian Langroudi, M. Charmchian Langroudi, F. Arasli, and I. Rahman, "Challenges and strategies for knowledge transfer in multinational corporations: The case of hotel 'Maria the great,'" Journal of Hospitality & Tourism Cases, 2024, Press.
14. A. Varghese, J. R. P. K. Ande, R. Mahadasa, S. S. Gutlapalli, and P. Surarapu, "Investigation of fault diagnosis and prognostics techniques for predictive maintenance in industrial machinery," Eng. Int., vol. 11, no. 1, pp. 9–26, 2023.
15. B. Priyanka, Y. Rao, B. Bhavyasree, and B. Kavyasree, "Analysis Role of ML and Big Data Play in Driving Digital Marketing's Paradigm Shift," Journal of Survey in Fisheries Sciences, vol. 10, no. 3S, pp. 996–1006, 2023.
16. B. Rashi, Y. S. Kumar Biswal, N. Rao, and D. Ramchandra, "An AI-Based Customer Relationship Management Framework for Business Applications," Intelligent Systems and Applications in Engineering, vol. 12, pp. 686–695, 2024.
17. D. D. Kumar Sharma Kuldeep, "Perception Based Comparative Analysis of Online Learning and Traditional Classroom-Based Education Experiences in Mumbai," Research Journey, Issue, vol. 330, no. 2, pp. 79–86, 2023.
18. K. M. Nayak and K. Sharma, "Measuring Innovative Banking User's Satisfaction Scale," Test Engineering and Management Journal, vol. 81, no.1, pp. 4466–4477, 2019.
19. K. Sharma and P. Sarkar, "A Study on the Impact of Environmental Awareness on the Economic and Socio-Cultural Dimensions of Sustainable Tourism," International Journal of Multidisciplinary Research & Reviews, vol. 3, no. 1, pp. 84–92, 2024.
20. K. Sharma and S. Poddar, "An Empirical Study on Service Quality at Mumbai Metro-One Corridor," Journal of Management Research and Analysis, vol. 5, no. 3, pp. 237–241, 2018.
21. K. Vora, "Factors Influencing Participation of Female Students in Higher Education w.r.t Commerce Colleges in Mumbai," International Journal of Advance and Innovative Research, vol. 5, no. 3, pp. 127–130, 2018.
22. K. Vora, Sharma Kuldeep, and P. Kakkad, "Factors Responsible for Poor Attendance of Students in Higher Education with respect to Undergraduate - Commerce Colleges in Mumbai. BVIMSR's," Journal of Management Research, vol. 12, no. 1, pp. 1–9, 2020.
23. A. Ahuja and J. Kumar, "Financial inclusion: Key determinants and its impact on financial well-being," Glob. Bus. Econ. Rev., vol. 1, no. 1, 2024.
24. J. Kumar and V. Rani, "What do we know about cryptocurrency investment? An empirical study of its adoption among Indian retail investors," The Bottom Line, vol. 37, pp. 27–44, 2024.
25. J. Kumar, "Behavioural Finance-Literature Review Summary and Relevant Issues," AAYAM: AKGIM Journal of Management, vol. 9, no. 1, pp. 42–53, 2019.

26. J. Kumar, K. Pal, S.N. Mahapatra, and S.S. Kundu, “Altman's model for predicting business failure: case study of HAFED”. Abhigyan, vol. 29, no. 3, pp. 52-62, 2011.
27. J. Kumar, S. Rana, G. Rani, and V. Rani, “How phygital customers’ experience transforms the retail banking sector? Examining customer engagement and patronage intentions,” Copmetitiveness Rev. J., vol. 34, no. 1, pp. 92–106, 2024.
28. J. Kumar, S. Rana, V. Rani, and A. Ahuja, “What affects organic farming adoption in emerging economies? A missing link in the Indian agriculture sector,” Int. J. Emerg. Mark., 2023, Press.
29. J. Kumar, M. Rani, G. Rani, and V. Rani, “Crowdfunding adoption in emerging economies: insights for entrepreneurs and policymakers,” Journal of Small Business and Enterprise Development, vol. 31, no. 1, pp. 55–73, 2024.
30. J. Kumar, M. Rani, G. Rani, and V. Rani, “What do individuals know, feel and do from a financial perspective? An empirical study on financial satisfaction,” Int. J. Soc. Econ., 2023, Press.
31. J. Kumar, V. Rani, G. Rani, and M. Rani, “Does individuals’ age matter? A comparative study of generation X and generation Y on green housing purchase intention,” Prop. Manag., 2024, Press.
32. M. Farheen, “A Study on Customer Satisfaction towards traditional Taxis in South Mumbai,” Electronic International Interdisciplinary Research Journal, vol. 12, no. 1, pp. 15–28, 2023.
33. M. Mahato and P. Kumar, “Emotional Labor - An Empirical Analysis of the Correlations of Its Variables,” European Journal of Business and Management, vol. 4, no. 7, pp. 163–168, 2012.
34. M. Mahato, “Performance Analysis of High, Medium and Low Companies in Indian Pharmaceuticals Industry,” IUP Journal of Management Research, vol. 10, no. 3, pp. 52–70, 2011.
35. M. Mandapuram, R. Mahadasa, and P. Surarapu, “Evolution of smart farming: Integrating IoT and AI in agricultural engineering,” Glob. Disclosure Econ. Bus., vol. 8, no. 2, pp. 165–178, 2019.
36. M. Modekurti-Mahato and P. Kumar, “Organizational Role Stress - Empirical Evidences from India during Economic and Political Resentment,” Purushartha - A journal of Management,” Ethics and Spirituality, vol. 7, no. 2, pp. 30–39, 2014.
37. P. G. Raju and M. M. Mahato, “Impact of longer usage of lean manufacturing system (Toyotism) on employment outcomes - a study in garment manufacturing industries in India,” Int. J. Serv. Oper. Manag., vol. 18, no. 3, p. 305, 2014.
38. P. Kakkad, K. Sharma, and A. Bhamare, “An Empirical Study on Employer Branding To Attract And Retain Future Talents,” Turkish Online Journal of Qualitative Inquiry, vol. 12, no. 6, p.7615, 2021.
39. P. Surarapu et al., “Quantum dot sensitized solar cells: A promising avenue for next-generation energy conversion,” Asia Pac. J. Energy Environ., vol. 7, no. 2, pp. 111–120, 2020.
40. P. Surarapu, R. Mahadasa, and S. Dekkati, “Examination of Nascent Technologies in E-Accounting: A Study on the Prospective Trajectory of Accounting,” Asian Accounting and Auditing Advancement, vol. 9, no. 1, pp. 89–100, 2018.
41. K. Pal and J. Kumar, “Economic value added vis-à-vis thinking of Indian corporate managers: a survey analysis,” International Journal of Financial Management, vol. 1, no. 3, p.19, 2011.
42. R. Mahadasa and P. Surarapu, “Toward Green Clouds: Sustainable practices and energy-efficient solutions in cloud computing,” Asia Pac. J. Energy Environ., vol. 3, no. 2, pp. 83–88, 2016.
43. R. Mahadasa, D. R. Goda, and P. Surarapu, “Innovations in energy harvesting technologies for wireless sensor networks: Towards self-powered systems,” Asia Pac. J. Energy Environ., vol. 6, no. 2, pp. 101–112, 2019.
44. R. Mahadasa, P. Surarapu, V. R. Vadiyala, and P. R. Baddam, “Utilization of agricultural drones in farming by harnessing the power of aerial intelligence,” Malays. J. Med. Biol. Res., vol. 7, no. 2, pp. 135–144, 2020.
45. V. Rani and J. Kumar, “Gender differences in FinTech adoption: What do we know, and what do we need to know?,” J. Model. Manag., 2023 Press.
46. S. R. Yerram et al., “The role of blockchain technology in enhancing financial security amidst digital transformation,” Asian Bus. Rev., vol. 11, no. 3, pp. 125–134, 2021.
47. T. N. Srinivasarao and N. G. Reddy, “Small and Medium Sized Enterprises Key Performance Indicators,” IOSR Journal of Economics and Finance, vol. 11, no. 4, pp. 1–06, 2020.
48. Y. Priyanka, B. Rao, B. Likhitha, and T. Malavika, “Leadership Transition In Different Eras Of Marketing From 1950 Onwards,” Korea Review Of International Studies, vol. 16, no. 13, pp. 126–135, 2023.