IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Comparative Analysis Of Algorithms And Hybrid Approaches: Credit Card Fraud Detection

[1]Tippabhotla Sowmya Sri, Student in Dept. Of Master of Computer Applications, at Miracle Educational Society Group of Institutions

[2]Mr. B. Mahendra Roy, Assistant Professor at Miracle Educational Society Group of Institutions

[3]Sattaru Suresh Babu, Associate Professor at Miracle Educational Society Group of Institutions

ABSTRACT

Today almost every person uses a credit card, but fraudulent activities are a great concern to both the customers and the financial institutions. This project compares the performance of several machine learning algorithms such as Decision Tree, KNN, Logistic Regression, SVM, Random Forest, and also XGBOOST, as well as a combined approach, which introduces the use of deep learning CNN. The main issue tackled is the problem on dataset which is fairly skewed, and hence, the normal transactions are a majority, while the fraudulent transactions are few. PCA for feature selection and SMOTE for data balancing techniques are applied for this purpose. A combination of the two, wherein CNN is combined with Decision Tree, increases all detection accuracy to 100%. This project offers valuable contributions as it highlights sample solutions to the problem of credit card fraud by using the Canadian Credit Card Dataset in a fast and accurate way.

Keywords: CNN, XGBOOST, SVM

INTRODUCTION:

The issue of Card Not Present (CNP) fraud has become prevalent as economic activity shifts from cash to a more digital approach. It entails the use of stolen, lost or fake credit cards to carry out transactions without physically presenting the card. As the world before CCF indicates a cashless society economic growth this would make sense as credit cards and other electronic means become more the order of business for the world. However, CCF poses alarming risks for any financial system serving the consumers, the businesses and even the society as a whole. Identity theft crimes increased by 113%, with 44.6% of this figure being predominantly credit card scams. The global theft of payment through cards stood at $24.26 billion and the United States can be said to be at the forefront of this neck breaking trend. Solving this problem requires adoption of AI based algorithms for the task of fraud detection. Using historical data helps detect fraudulent transactions in a certain ecosystem while improving statistical indicators. This also helps prevent similar cases from occurring subsequently and facilitates better financial systems.

GAP IDENTIFIED BASED ON LITERATURE SURVEY:

The reviewed journals have dealt fraud detection with aspects of machine learning that are helpful; however, these fail to provide sufficient emphasis on majority and minority classes thus still leading to class bias. Although promising, deep learning has its setbacks since most publications do not act on feature extraction, which in effect limits performance of the algorithms.

Key Gaps:

• Uneven Data Scale: The previous work fails in developing standard approaches to biased datasets without changing the data characteristics.

• Integration of ML and DL: There has been little work on the machine learning aspect and the machine learning and deep learning aspect in future research.

• Feature Reduction: Not much consideration is placed on the use of PCA for feature reduction in order to cut down on computing power.

• Scalability and Real Time Use Case: Models mostly do get implemented well enough for real time detection across the large databases.

PROBLEM STATEMENT:

Fraudsters using stolen credit card credentials can substantially undermine an organization's finances and reputation. The crime of fraud tends to be rare and therefore has a noticeable impact on the dataset and even more the imbalance that exists.

Key Challenges:

• High Class Imbalance: There are around a few 492 fraudulent transactions against more than two hundred thousand normal ones, which makes the models to consistently underperform.

• Feature Identification: Finding and selecting important features so the model can be enhanced.

• Achieving Maximum Accuracy: Finding a balancing act where accurate detection can be attained while still being able to get as many true positive rates as possible.

• Adjustability: It has to be able to use large sets of real information concerning transaction data.

• Overfitting Risk: During training, and especially while applying synthetic data generated by SMOTE, it is important that algorithms do not get overfitted.

PROPOSED METHOD:

• High Class Imbalance: There are around a few 492 fraudulent transactions against more than two hundred thousand normal ones, which makes the models to consistently underperform.

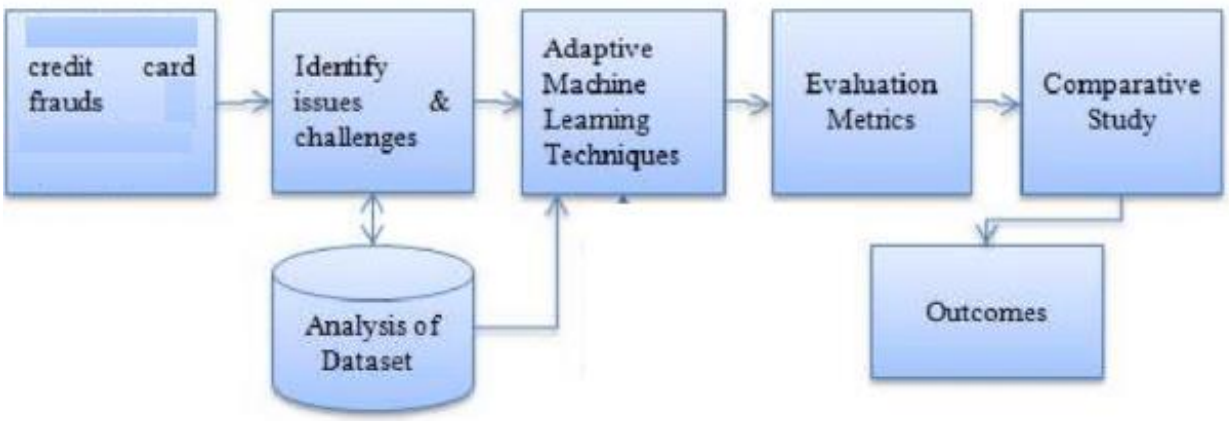
• Feature Identification: Finding and selecting important features so the model can be enhanced.

• Achieving Maximum Accuracy: Finding a balancing act where accurate detection can be attained while still being able to get as many true positive rates as possible.

• Adjustability: It has to be able to use large sets of real information concerning transaction data.

• Overfitting Risk: During training, and especially while applying synthetic data generated by SMOTE, it is important that algorithms do not get overfitted.

ARCHITECTURE:



CANADIAN CREDIT CARD DATASET:

The used dataset is a Canadian Credit Card Dataset which has two broad categories, a normal transaction which is more than 200,000 records, and a fraudulent transaction with 492 records. Given the extreme imbalance, it creates significant problems for effective fraud detection. Amount features and anonymized numerical PCA derived attributes are included in the dataset. The Imbalance problem is addressed through SMOTE technique which helps create fraudulent samples with ease, thus a balanced dataset is acquired. This processing guarantees that the machine learning and deep learning models aimed at classifying fraudulent transactions are not biased to any particular category thereby enabling true predictions.

METHODOLOGY:

1. **Dataset Analysis**

   The work commences in analyzing the Canadian Credit Card Dataset with more than 200,000 normal transactions and only 492 fraudulent, thus the analysis emphasizes the class imbalance. In order to visualize class imbalance, trends and the distribution of the features there is a need to conduct an EDA. This step guarantees the understanding of the dataset and what the challenges of fraud detection are likely to be.

2. **Data Preprocessing**

   To ready the data for modeling purposes, a number of preprocessing methods are put in place:

   Normalization : All features are transformed into a common range so that algorithms which are greatly affected by feature magnitudes especially KNN can operate optimally.

   Dimensionality Reduction: Effective reduction in mathematics dimensionality is possible in PCA technique since it retains the important features while discarding the less important ones.

3. SMOTE Technique to Address Class Imbalance in Datasets

To counter the imbalance in datasets, SMOTE is used. This technique interpolates between the existing samples of the class which is less represented (the fraudulent transactions in this case) in order to create synthetic samples so that equal representation of both classes is achieved. Once SMOTE technique is applied, the proportionate dataset is verified to ensure that all the classes are well represented.

4. Dataset Partitioning

The data is divided into components for training, which forms 80% while 20% is for testing so that the models are tested on data which is new for them. To eliminate bias, data is shuffled during the process of dividing the data. This is additionally supported by cross-validation to check the reliability of the model over varying data subsets.

5. The Initial Models of Machine Learning

Decision Tree: Classifications are done using a set of rules that follow a hierarchical structure.

K-Nearest Neighbors (KNN): A machine learning method that classifies the transactions according to their distance in the feature space.

Logistic Regression: An algorithm that computes the odds of the transactions being a fraudulent transaction.

Support Vector Machine (SVM): Algorithm that searches for a hyperplane in an n-dimensional space that distinctly classifies the classes.

Random Forest: A group of decision trees to achieve better accuracy in classification.

XGBOOST: An implementation of gradient boosting which is geared for efficiency.

6. Deep Learning of imaging data using CNN

In identifying fraudulent transactions, some authors place emphasis on the ability of CNNs to extract features:

In order to analyze the effect of balancing, CNNs are trained on both imbalanced and balanced data sets. This consists of convolutional layers for feature learners, pooling layers for dimensionality reduction and fully connected layers for classification. Binary cross-entropy is the formulated loss function whereas Adam optimizer is used for fine-tuning the models.

7. Hybrid Model Building

A hybrid model consisting of CNN and Decision Tree is built up as follows:

The features from the balanced CNN model are extracted. These features are consequently used in the training of a Decision Tree classifier where the pattern recognition capability of the CNNs and the interpretability of Decision Tree is harnessed.

8. Hybrid Model Assessment

All the models are assessed using the hybrid model approach and accuracy, precision, recall, F1 score, and confusion matrices are used as the key metrics. The hybrid model which combined CNN and Decision Tree was able to record 100 percent accuracy which is better than each of the models on their own.

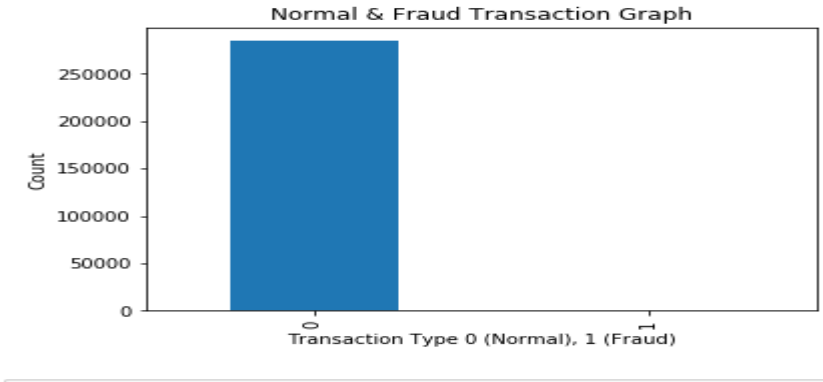
9. Visualization and Results

Bar graphs are used to represent the results attained for the accuracy performance while the heat map is used on the confusion matrices. These diagrams make it easy to understand the effectiveness of the models and point at the best one which was the hybrid model.

10. Automation and Integration

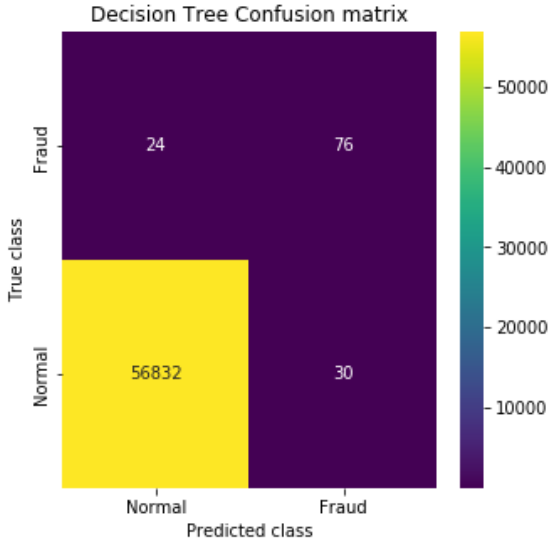
This hybrid model has been embedded within an automated platform for effective and instant fraud identification making it easier to use in banks or financial institutions.

RESULTS:



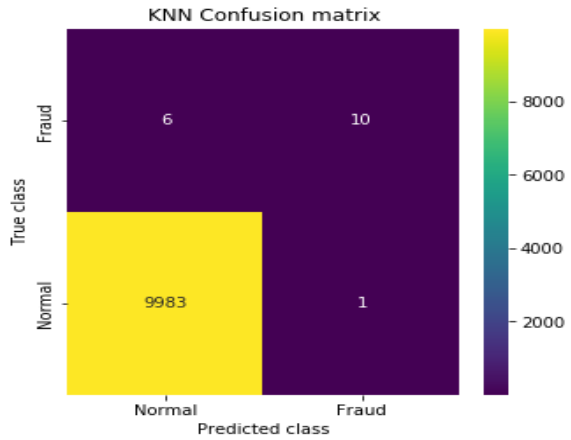
We are finding and plotting graph of normal and fraud transaction count

```
Decision Tree Accuracy  : 99.90519995786666
Decision Tree Precision   : 85.82795065189515
Decision Tree Recall      : 87.97362034399072
Decision Tree FSCORE      : 86.86946094042544
```



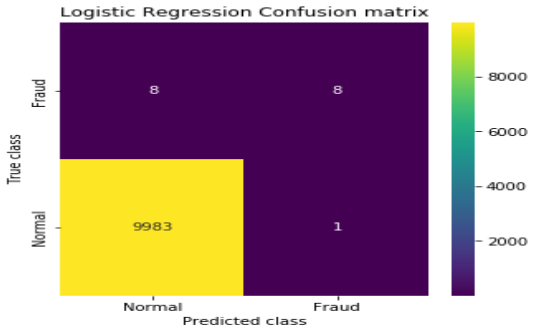
Training decision tree algorithm and we got its accuracy as 99.90%

```
KNN Accuracy  : 99.92999999999999
KNN Precision   : 95.42451241820548
KNN Recall      : 81.24499198717949
KNN FSCORE      : 87.01951338010016
```



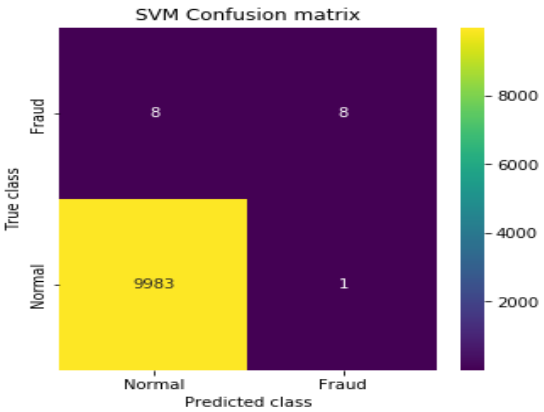
In above screen with KNN we got 99.92% accuracy and you can see other metrics also

```
Logistic Regression Accuracy  : 99.91
Logistic Regression Precision   : 94.40440841201524
Logistic Regression Recall      : 74.99499198717949
Logistic Regression FSCORE      : 81.97747183979975
```



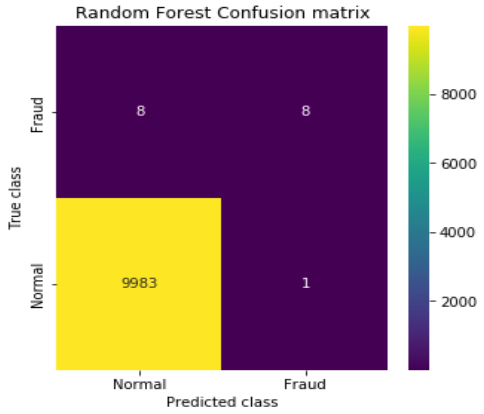
Training logistic regression got 99.91% accuracy

```
SVM Accuracy  : 99.91
SVM Precision   : 94.40440841201524
SVM Recall      : 74.99499198717949
SVM FSCORE      : 81.97747183979975
```



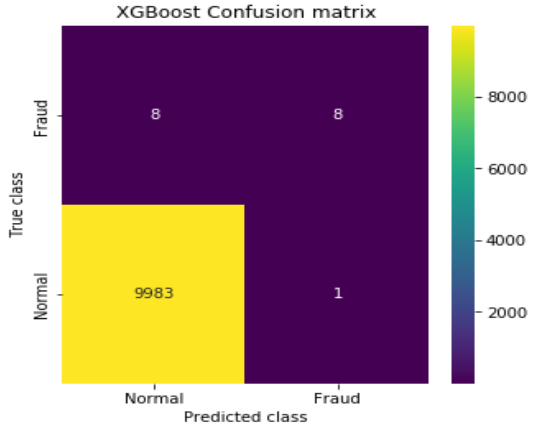
In above screen SVM got 99.91% accuracy

```
Random Forest Accuracy  : 99.91
Random Forest Precision   : 94.40440841201524
Random Forest Recall      : 74.99499198717949
Random Forest FSCORE      : 81.97747183979975
```



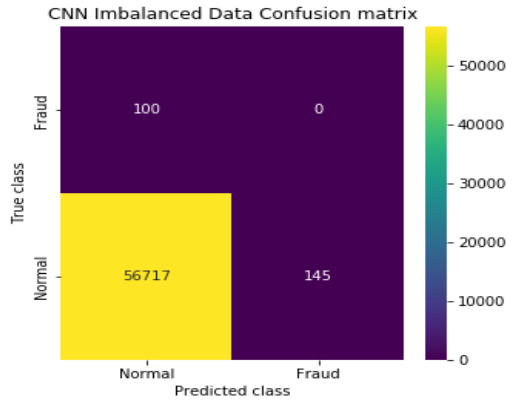
Random forest also got 99.91% accuracy

```
XGBoost Accuracy  : 99.91
XGBoost Precision   : 94.40440841201524
XGBoost Recall      : 74.99499198717949
XGBoost FSCORE      : 81.97747183979975
```



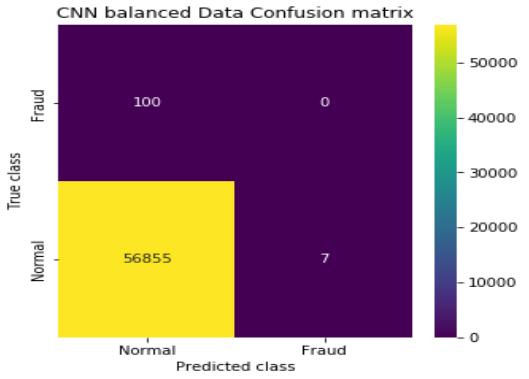
XGBOOST got 99.91% accuracy

```
CNN Imbalanced Data Accuracy  : 99.56988869772832
CNN Imbalanced Data Precision   : 49.911998169561926
CNN Imbalanced Data Recall      : 49.87249832928845
CNN Imbalanced Data FSCORE      : 49.89224043139014
```



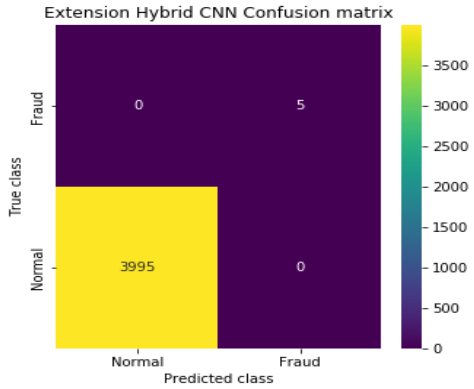
CNN imbalanced data we got 99.56% accuracy

```
CNN balanced Data Accuracy  : 99.8121554720691
CNN balanced Data Precision   : 49.912211394960934
CNN balanced Data Recall      : 49.99384474693117
CNN balanced Data FSCORE      : 49.95299471959373
```

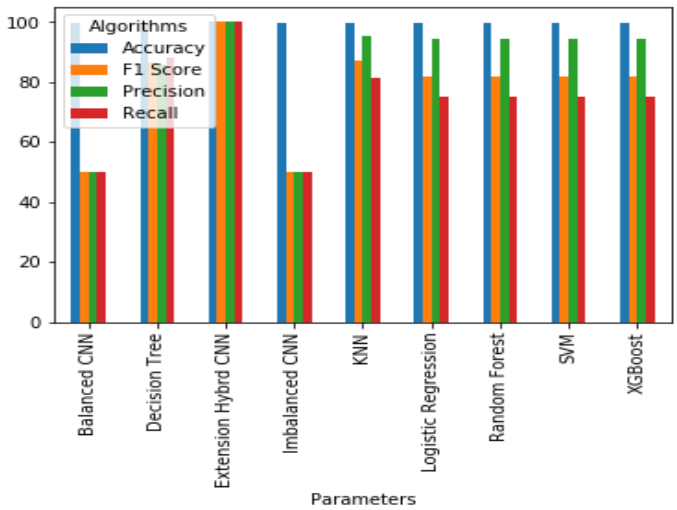


Balanced CNN we got 99.81% accuracy

```
Extension Hybrid CNN Accuracy  : 100.0
Extension Hybrid CNN Precision   : 100.0
Extension Hybrid CNN Recall      : 100.0
Extension Hybrid CNN FSCORE      : 100.0
```



We are extracting features from balanced CNN as hybrid CNN and then retraining with decision tree and we got accuracy as 100%



In above graph x-axis represents algorithm names and y-axis represents accuracy and precision with different colour bar represents different metrics

Prediction:

```
Test Data : [ 7.54300000e+03  3.29594333e-01  3.71288930e+00 -5.77593511e+00
  6.07826551e+00  1.66735901e+00 -2.42016841e+00 -8.12891249e-01
  1.33080118e-01 -2.21431131e+00 -5.13445447e+00  4.56072011e+00
 -8.87374836e+00 -7.97483600e-01 -9.17716637e+00 -2.57024775e-01
 -8.71688490e-01  1.31301363e+00  7.73913873e-01 -2.37059945e+00
  2.69772776e-01  1.56617169e-01 -6.52450441e-01 -5.51572219e-01
 -7.16521635e-01  1.41571662e+00  5.55264740e-01  5.30507389e-01
  4.04474055e-01  1.00000000e+00] =====> Contains Fraud Transaction Signature
Test Data : [ 0.00000000e+00 -1.35980713e+00 -7.27811733e-02  2.53634674e+00
  1.37815522e+00 -3.38320770e-01  4.62387778e-01  2.39598554e-01
  9.86979013e-02  3.63786970e-01  9.07941720e-02 -5.51599533e-01
 -6.17800856e-01 -9.91389847e-01 -3.11169354e-01  1.46817697e+00
 -4.70400525e-01  2.07971242e-01  2.57905802e-02  4.03992960e-01
  2.51412098e-01 -1.83067779e-02  2.77837576e-01 -1.10473910e-01
  6.69280749e-02  1.28539358e-01 -1.89114844e-01  1.33558377e-01
 -2.10530535e-02  1.49620000e+02] =====> Contains Cleaned Signatures
```

Test data and after prediction output

CONCLUSION

This project presents a novel robust technique for credit card fraud detection using hybrid machine learning and deep learning approaches. By applying SMOTE to the imbalance present in the data and PCA for the reduction of features, the models achieve a decent level of accuracy and great level of reliability. The hybrid model of CNN and Decision tree outperforms the stand alone algorithms with a score of 100% on accuracy. This system is an efficient one that can apply for real life fraud detection and help maintain security of the finances. More work can be done on the optimization of feature engineering and further application of real life datasets to widen the uses.

REFERENCES:

[1] Y. Abakarim, M. Lahby, and A. Attioui, ‘‘An efficient real time model for credit card fraud detection based on deep learning,’’ in Proc. 12th Int. Conf. Intell. Systems: Theories Appl., Oct. 2018, pp. 1–7, doi: 10.1145/3289402.3289530.

[2] H. Abdi and L. J. Williams, ‘‘Principal component analysis,’’ Wiley Interdiscipl. Rev., Comput. Statist., vol. 2, no. 4, pp. 433–459, Jul. 2010, doi: 10.1002/wics.101.

[3] V. Arora, R. S. Leekha, K. Lee, and A. Kataria, ‘‘Facilitating user authorization from imbalanced data logs of credit cards using artificial intelligence,’’ Mobile Inf. Syst., vol. 2020, pp. 1–13, Oct. 2020, doi: 10.1155/2020/8885269.

[4] A. O. Balogun, S. Basri, S. J. Abdulkadir, and A. S. Hashim, ‘‘Performance analysis of feature selection methods in software defect prediction: A search method approach,’’ Appl. Sci., vol. 9, no. 13, p. 2764, Jul. 2019, doi: 10.3390/app9132764.

[5] B. Bandaranayake, ‘‘Fraud and corruption control at education system level: A case study of the Victorian department of education and early childhood development in Australia,’’ J. Cases Educ. Leadership, vol. 17, no. 4, pp. 34–53, Dec. 2014, doi: 10.1177/1555458914549669.

[6] J. Błaszczyński, A. T. de Almeida Filho, A. Matuszyk, M. Szelg¸, and R. Słowiński, ‘‘Auto loan fraud detection using dominance-based rough set approach versus machine learning methods,’’ Expert Syst. Appl., vol. 163, Jan. 2021, Art. no. 113740, doi: 10.1016/j.eswa.2020.113740.

[7] B. Branco, P. Abreu, A. S. Gomes, M. S. C. Almeida, J. T. Ascensão, and P. Bizarro, ‘‘Interleaved sequence RNNs for fraud detection,’’ in Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2020, pp. 3101–3109, doi: 10.1145/3394486.3403361.

[8] F. Cartella, O. Anunciacao, Y. Funabiki, D. Yamaguchi, T. Akishita, and O. Elshocht, ‘‘Adversarial attacks for tabular data: Application to fraud detection and imbalanced data,’’ 2021, arXiv:2101.08030. [9] S. S. Lad, I. Dept. of CSERajarambapu Institute of TechnologyRajaramnagarSangliMaharashtra, and A. C. Adamuthe, ‘‘Malware classification with improved convolutional neural network model,’’ Int. J. Comput. Netw. Inf. Secur., vol. 12, no. 6, pp. 30–43, Dec. 2021, doi:

10.5815/ijcnis.2020.06.03. [10] V. N. Dornadula and S. Geetha, ‘‘Credit card fraud detection using machine learning algorithms,’’ Proc. Comput. Sci., vol. 165, pp. 631–641, Jan. 2019, doi: 10.1016/j.procs.2020.01.057.

[11] I. Benchaji, S. Douzi, and B. E. Ouahidi, ‘‘Credit card fraud detection model based on LSTM recurrent neural networks,’’ J. Adv. Inf. Technol., vol. 12, no. 2, pp. 113–118, 2021, doi: 10.12720/jait.12.2.113-118.

[12] Y. Fang, Y. Zhang, and C. Huang, ‘‘Credit card fraud detection based on machine learning,’’ Comput., Mater. Continua, vol. 61, no. 1, pp. 185–195, 2019, doi: 10.32604/cmc.2019.06144.

[13] J. Forough and S. Momtazi, ‘‘Ensemble of deep sequential models for credit card fraud detection,’’ Appl. Soft Comput., vol. 99, Feb. 2021, Art. no. 106883, doi: 10.1016/j.asoc.2020.106883. [14] K. He, X. Zhang, S. Ren, and J. Sun, ‘‘Deep residual learning for image recognition,’’ 2015, arXiv:1512.03385.

[15] X. Hu, H. Chen, and R. Zhang, ‘‘Short paper: Credit card fraud detection using LightGBM with asymmetric error control,’’ in Proc. 2nd Int. Conf. Artif. Intell. for Industries (AII), Sep. 2019, pp. 91–94, doi: 10.1109/AI4I46381.2019.00030.

[16] J. Kim, H.-J. Kim, and H. Kim, ‘‘Fraud detection for job placement using hierarchical clusters-based deep neural networks,’’ Int. J. Speech Technol., vol. 49, no. 8, pp. 2842–2861, Aug. 2019, doi: 10.1007/s10489-019-01419-2.

[17] M.-J. Kim and T.-S. Kim, ‘‘A neural classifier with fraud density map for effective credit card fraud detection,’’ in Intelligent Data Engineering and Automated Learning, vol. 2412, H. Yin, N. Allinson, R. Freeman, J. Keane, and S. Hubbard, Eds. Berlin, Germany: Springer, 2002, pp. 378–383, doi: 10.1007/3-540-45675-9_56.

[18] N. Kousika, G. Vishali, S. Sunandhana, and M. A. Vijay, ‘‘Machine learning based fraud analysis and detection system,’’ J. Phys., Conf., vol. 1916, no. 1, May 2021, Art. no. 012115, doi: 10.1088/1742-6596/1916/1/012115.

[19] R. F. Lima and A. Pereira, ‘‘Feature selection approaches to fraud detection in e-payment systems,’’ in E-Commerce and Web Technologies, vol. 278, D. Bridge and H. Stuckenschmidt, Eds. Springer, 2017, pp. 111–126, doi: 10.1007/978-3-319-53676-7_9.

[20] Y. Lucas and J. Jurgovsky, ‘‘Credit card fraud detection using machine learning: A survey,’’ 2020, arXiv:2010.06479.