Enhance Credit Card Fraud Detection Models using Rough Set Theory

Noor Khalid Hussein
[1]*Department of Computer Science*
University Of Technology
Baghdad, Iraq
[2]College of Biotechnology Al-Nahrain University Baghdad, Iraq
cs.19.82@grad.uotechnology.edu.iq

Bashar Saadoon Mahdi
Department of Computer Science
University Of Technology
Baghdad, Iraq
110043@uotechnology.edu.iq

Ayad Rodhan Abbas
Department of Computer Science
University Of Technology
Baghdad, Iraq
ayad.r.abbas@uotechnology.edu.iq

***Abstract*— in recent years, fraud technologies became more improved and easier for fraud. This is why a wide variety of machine learning methods were implemented and advanced for the recognition of the transactions of fraudulent credit cards. The fundamental issue with failing any detection technique on any fraud operation represents the results' accuracy. The present study included a discussion of the ways for improving the efficiency of fraud detection with the use of the approaches of machine learning and feature reduction with the use of the Rough Set Theory, followed by the use of the Voting approach for choosing the most suitable approach for inclusion in the systems fraud detection. It has also provided comprehensive research on the European customer database and how the classifiers interact with it by applying three classification algorithms. In addition to that, they were utilizing Python language tool to apply machine learning algorithms with the approach of Voting for choosing the optimal model of a classifier. Experimental results have shown that the DT algorithm's use is the optimal one as it has been capable of achieving 99.82% accuracy and 0.8 sec. Processing time.**

Keywords*— Classification; Credit Card; Fraud Detection; Machine Learning; Voting; Decision Tree; Naive Bayes; Support Vector Machine; RST.

I. Introduction

Fraud is a label for any malicious activity that wants to harm people by stealing their money, identity or other things. In contrast, digital money has become common use these days; fraud operations also become more powerful and more efficient to keep pace with this development[1]. For any conducting electronic financial transactions (digital money) in the last few years, credit card has become one of the widely used to that purpose, To protect these credit cards the systems for any institution like banks, companies or other else must have high-security techniques to detect any fraud operations[2]. The protect operations of credit card have a big challenge to prevent fraud from occurring on them, and machine learning one of the wildest techniques used to detect fraud transactions on credit cards.

Machine learning (ML) is an abroad scientific field based on computer science, mathematics, statistics, engineering, and many other areas of mathematics and science. The main idea of using ML approaches recognizes if the transactions are fraudulent or not, there is a four type of machine learning supervised learning, Unsupervised Learning, Semi-supervised Learning, and Reinforcement learning, in this research focus on supervised learning[3], supervised learning depends on the historical behaviour of transaction data for all users in the system to predict the rules base, this rule base is used to check any new transaction and defined it either fraudulent or safety transaction[4]. To apply machine learning algorithms, there are many tools like ORANGE, O3, WEKA, etc. However, this research uses Python as a tool to apply different types of machine learning algorithms.

Historically, a wide variety of environments and programming languages were utilized for enabling the research of application development and machine learning. None-the-less, due to the fact that the general-purpose Python language underwent a massive popularity growth in the scientific computing community throughout the past ten years, the latest libraries of deep learning and machine learning are presently Python-based[5].

With its primary emphasis on readability, Python can be described as a high-level interpreted programming language that has been widely recognized due to the fact that it is easy to learn. However, it still has the ability for harnessing the systems-level programming language power whenever required. Besides the advantages of the actual language, the community around available libraries and tools has made Python especially suitable for the work-loads in machine learning, data science, and scientific computing. Based on one of the latest KDnuggets polls, which has surveyed over 1,800 participants for data science, analytics, and machine learning preferences, Python has been capable of maintaining its status as the most commonly utilized language of the year 2019 [6]. Three different classification algorithms on our credit card dataset are used, and the comparing results dependent on accuracy and time to find the best algorithm that achieves high accuracy and low time—in this study, using three algorithms categorization like naive Bayes, decision trees, and support vector machine. Then select the best classifier.

In the rest of this paper, section II presents the related works, where section III offers the dataset description, our proposed method presented in section IV, result and dissection present in section V, and finally, conclusion present in section VI.

II. RELATED WORKS

Frauds are increasing significantly because of this. Every day the losses increase more and more; this led to increased fraud-related studies to reduce them as much as possible. like Awoyemi et al. [7], some reviews have studied the impacts of the feature selection on two sets of features with four classifier algorithms DT, SVM, NB, and NN. Feature selection used the correlation analysis technique. These algorithms were implemented on the raw dataset and feature selection dataset. Two datasets have been used, which were Taiwan and European bank. These datasets were obtained from the ULB and UCI repository that contains 284807 and 30000 transactions, respectively. The comparative analysis results showed that DT variants classifiers had outperformed the NB, SVM, and NN algorithms, respectively. The characteristic ranked, and raw datasets of European credit card fraud data recorded the decision trees' maximum performance metrics.

Awoyemi et al. [8] have used three classifier NB, KNN, and LR, on highly credit card fraud. A dataset of the credit card transactions has been obtained from the European card-holders, which contained 284807 transactions. The three methods were implemented on the pre-processed as well as raw data. Results have shown that NB, K-NN, and LR classifiers' optimal accuracy were 97.92%, 97.69% and 54.86%. The comparative results have demonstrated that KNN has performed much better than both NB and LR classifications.

.Kavipriya and Geetha [9]have proposed a new method for the analysis, detection and recognition of fraudulent transactions. This approach has been based upon sufficient classification and clustering approaches like the SVMs and apriori, respectively. The credit card transaction dataset used has 600 instances and 23 attributes downloaded from the UCI repository. The result showed that the suggested approach provided more sufficient results, which has helped obtain the high fraud coverage in combination with a low rate of false alarms compared to the existing Hidden Markov Model (HMM).Dataset Description.

Lakshmi and Kavila [10] have applied three machine learning methods: LR, DT, and RF, for credit card fraud detection. This study has utilized the dataset of the credit card transactions from the European bank. This dataset has been obtained from Kaggle, and it contained a totally of 284808 credit card transactions. The result showed that the accuracy for DT, LR, and RF classifier were 94.30%, 90%, and 95.50%, respectively. The comparative results showed that RF had performed better compared to LR and DT classifier.

M. Ummul Safa and R. M. Ganga in 2019 [11] built a model to compare three algorithms techniques to find the best one among them. The algorithms in this model were Naïve Bayes, Logistic Regression and K-nearest neighbor .the result shown that Logistic Regression gives the highest accuracy with 97.69%. Unfortunately, this algorithm has a limitation in time; this algorithm's processing time reaches 38.1 seconds. Still, the Naïve Bayes algorithm spent just 10 seconds to find the result but with less accuracy reached 83%. The researchers' suggestion was to use Logistic Regression as the best algorithm for frauds detection

Appen Pumsirirat, Liu Yan in 2018 [12] create a model depended on deep learning by using a deep neural network with Auto-encoder (AE) and restricted Boltzmann machine (RBM) that model can reconstruct legal transactions for finding suspicious ones from normal patterns. To evaluate this model, the accuracy, True Positive Rate and False Positive Rate, was used.

The database used in this research was taken from the website https://archive.ics.uci.edu. It has also been released in Kaggle, a community of machine learners and data scientists. It includes the record related to credit card transactions made via European card-holders and occurred in 2 days in Sept. 2013. Also, the dataset contains 284807 transactions, 492 of which were fraudulent. Furthermore, the dataset was highly unbalanced as the positive class responsible for just 0.172% of total transactions. Table 1 shows the unbalanced class distribution. The dataset includes the numerical. Yet, because of the issue regarding confidentiality, original features weren't disclosed. Also, there were 30 features; 28 were created via PCA, as it was one of the dimensionality reduction approaches, where a lot of original variables were condensed into a small sub-set of feature variables. Furthermore, the only features which weren't transformed into principal components have been 'Time' and' Amount'.

TABLE 1: DATASET CLASS DISTRIBUTION

No.	Class	Number
1	0	284807
2	1	429

III. PROPOSED METHODOLOGY

The proposed methodology, which may be represented in Figure 1, will be introduced in the present section. Firstly, the idea is to pre-process and feeds the split data, then split our data into training and testing. Then build our model used three methods such as naïve Bayes, decision tree, and function classifier. All these methods in the Python programing language. The proposed method is evaluated using an evaluation metric to check each method's prediction, then select the best method for the Fraud detection problem.

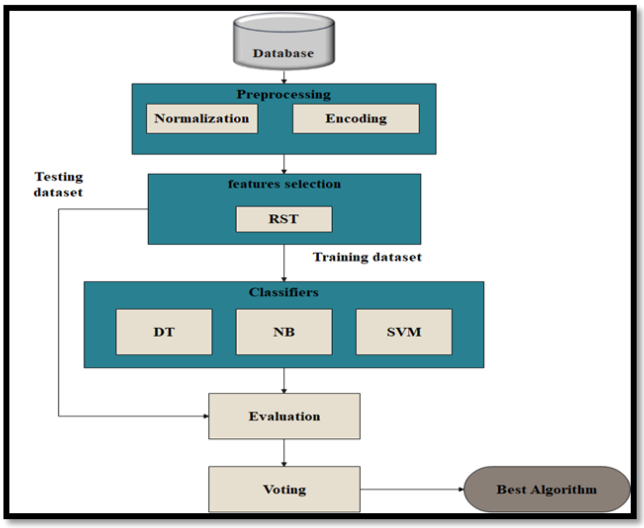

Figure 1: Proposed method

A. *Pre-processing*

The wide range of this dataset can badly affect any classifier algorithms, which has led to an inefficient system. To solve these problems, encoding and normalization methods should be used.

1. Encoding

This thesis used the encoding technique for converting data into an integer domain. The integer domain was needed in this work because it deals with the rough set theory that is why integer value is required. The range of value started from 0 to *n*. the algorithm of encoding shows below.

Algorithm 1: encoding dataset

Input:
D = read dataset
L = length of dataset (D)
C = number of columns (D)
Initialization:
For $c_{i} \in C$ *do*
For $l_{j} \in L$ *do*
Group to the same value;
Each group was given an integer value;
End for
End for
Output:
Encoding dataset

2. Normalization

This thesis has used the max-min normalization method, as indicated in chapter two. This approach was also applied in each column, and the value was ranged between [0-1]. This method was used only with the original dataset because if it was used with rough set theory, it did not work because the rough set worked with an integer value. Here, it was wanted to see the effect of normalization on the original dataset[13].

There were various normalization approaches, such as Z-score normalization, Decimal scaling normalization and Min-Max normalization.

Min-Mix Normalization is the approach providing linear transformations on the original range of data. This approach is keeping the relations between original data. Also, it has the ability for fitting the data into the predefined boundary with predefined boundary

$$A' = \left(\frac{A - \min value\ of\ A}{\max value\ of\ A - \min value\ of\ A}\right) * (D - C) + C$$

In which,
A' contains Min-Max Normalized data one
In the case where the predefined boundary is [C, D]
In the case where the A is the range of original data
With this technique, the unstructured data can be normalized [14]

B. *Building Model*

We built three algorithms for every categorize type in Python. Every model has been tested, results have been derived, and the optimal model was selected. After that, they evaluated every model and compared each algorithm's results and determined the best model that could be used.

Classification techniques on the most popular method used in machine learning and deals with different types of the dataset, the main goal for use Classification Techniques to classified data accordion to some conditions to be more useful in detection operations. The data will be more understand and more valuable to predict events, rules, and so on. Classification techniques contain two main approaches supervised learning technique that can detect fraud operation dependent on fraudulent patterns of the last transactions and unsupervised technique that can detect fraud operation dependent on comparative and testing computed data to find unexpected transactions the details of it describe by Fabrizio Carcilloa, YannA¨el Le Borgne, Olivier Caelenb, Yacine Kessacib, Fr' ed' eric Obl'eb, Gianluca Bontempi[15]. Each supervised and unsupervised learning technique has many algorithms that deal with the dataset differently; this part will describe the main techniques used in this research.

1. *Naive Bayes*

Naive Bayes is a type of algorithms in machine learning; naive Bayes is a supervisor algorithm and very widely used for classification, also it a simple to understood and use. The basis for this algorithm depends on the probabilistic theory[16]. For more details of Naive Bayes as described[17], shown in equation 1.

$$p(c / f) = \frac{p(f / c)}{p(f)} \quad , \mathrm{P}\ (\mathrm{f})>0$$

$$p(c / f) = \frac{p(f / c)}{p(f)}$$

Where in f refers to the features, and c denotes the class.

2. *Decision Tree*

Decision trees it just like our brains when trying to decide, so it's easy to understand, simple to belting. The decision tree techniques work with a continuous or independent data set and give all possible solutions. Decision trees have the same contents as a real tree that contains is (root node, branches, and left nodes)[18]. During the process ML, this procedure might be seen as the assembly regarding computational approaches and mathematical equations for help in depiction, generalization and classification regarding the selected dataset. Usually, order trees include different non-leaf nodes straightforwardly associated with the principle leaf nodes with arcs [19]. Regarding such a situation, each of the non-leaf nodes is seen as an input feature, whereas the arcs were ordered as feature values. Each one of the leaf nodes belongs to the target class value or probability distribution.

.

3. Support Vector Machine (SVM)

These are considered as one of the majorly used classifiers [20]. SVM's main concept is using hyperplanes to separate different classes [21]. Also, SVMs are presenting the advantage of high precision levels when handling linearly-separable data, yet they don't equally perform in terms of non-linearly separable data. This issue's solution is in using kernel functions to shift the data to broader dimensional space. Thus, the data will be separated (in a linear manner). SVM's primary idea is selecting an adequate kernel function and adjusting the kernel parameters' parameters. Concerning calculations, one of the optimization issues is to seek the most suitable decision plan. One of the adequate decision plans might be facilitating the linear decisions' generation via a kernel function through a non-linear transformation, as can be seen in the equation below.

$$f(x) = \mathcal{W}^T x_i + b$$

$$f(x) = \sum_{i=0}^{N} \lambda_i l_i (\mathcal{W}_i^T x + b)$$

$\mathcal{W}^T$ Denotes the vector's weight, $f(x)$ denotes the feature sets regarding the two classes,, λ_i represents the dual function returned following training, *x* indicates the training dataset, *l* represents the classes (output), and *b* bias indicates omega θ.

$\mathcal{W}^T$ The vector weight represents the feature sets regarding the two classes representing the dual function returned following training, *x* represents the training dataset, *l* represents the classes (output), and *b* bias represents omega θ. Linearly-separable situations were excellent for SVM's optimal performance.

C. Feature selection

The Features Selection method is used to minimize the number of features within the database without adversely affecting the model's performance. FS is considered one of the most important data mining methods, as it helps to improve learning and increase output accuracy [22].

a. Rough Set Theory (RST)

Pawlak, cited in other works [23, 24], has used RST as a useful mathematical tool for handling incomplete, inexact and uncertain information. Many studies are focusing on RST development and applications [25-29]. Concerning data analysis, one of the significant benefits of RST is the non-requirement for primer or extra data in terms of information such as grade of membership associating with the fuzzy set theory, essential probability task relating to the Dempster–Shafer theory and the probability distribution that is related to statistics In this thesis. The best features might be identified via specifying the dependency between the decision feature and conditional feature. Furthermore, features with high dependency values were taken in the final sub-set related to the best features for that proposed reduction rule was used.

1- **Reduction rule of RST**: This method is based on RST and dealing with dataset inconsistency. Also, inconsistency indicates a minimum of 2 objects with the same description (attribute values' vector) yet with different classification. Often, this happens naturally-collected (not created in an artificial way) datasets, which typically specifies that the description was probably inadequate for distinguishing objects. Also, inconsistency was specified in RST via the quality of approximation related to classification or, shortly, the classification quality [30].

.

D. Evaluation Metric

Researchers have used the same approach of validation method for the determination of the percentages of the classifier. They have utilized the dataset that had an approximately equal size and class distribution. For each one of the folds, the classifier has been trained with the help of the 10. Here, the researchers have explained the performance measurements used for the machine learning classification issue[31, 32]; based on the confusion matrix, numerous measurements may be utilized to examine the model efficiency regarding the accuracy, which has been determined (2-17) using the below mentioned in Table 2. The recall has been used to determine the accuracy of every class known. Precision has also been inaccurately classified with the use of the following equation, which has been helpful to calculate F1 scores.

TABLE 2: METRIC EQUATIONS

Metric name	Equation
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Recall	$\frac{TP}{TP+FN}$
Precision	$\frac{TP}{TP+FP}$
F1score	$2*\frac{\mathrm{Re}\,call*\mathrm{Pr}\,ecesion}{\mathrm{Re}\,call+\mathrm{Pr}\,ecesion}$

Accuracy is typically determined using the confusion matrix, which has been dependent upon the choice of datasets. Researchers have utilized the table of contingency to improve accuracy and performance.

V. RESULTS AND DISCUSSION

The researcher applied the PyCharm Community Edition 2020.2.2 x64 (Python language) tool, which offered the processes like libraries and the visualization and validation results. This study applied machine learning algorithms. The researcher used default parameters for machine learning as it is. Then used the confusion matrix for determining accuracy. The researchers implemented three tests on 31 features and almost 300000 instances. The result of the reduction role of RST shown in Table 3. The work tested for each main category.

Table 3 Reduction Rule of RST

Data-base Features	No of features	Reduction Rule Features	No of Features Reduction
Time ,V1,V2,V3,….,V28, amount, payment-month	**31**	**Time, V10,V22, payment-month**	**4**

Table 4 has shown the accuracy of Naïve Bayes. The table has delivered the first categorization of our method.

TABLE 4: *NAIVE BAYES*

State	class	Precision %	Recall %	Fscore %	Average accuracy %	Time (S)
Without normalization	0	1.00	0.68	0.81	67.55	0.18
	1	0.00	0.40	0.00		
With normalization	0	1.00	0.74	0.85	74.16	0.15
	1	0.00	0.32	0.00		
With normalization And feature selection (R-NB)	0	1.00	0.78	0.87	77.5710	0.0937
	1	0.00	0.36	0.01		

Table 4 showed the naïve Bayes method, where normalization and RST with naïve Bayes achieve 77.57 % accuracy higher than the general method. R-NB updateable achieve a higher time of 0.09 milliseconds. Table 5 utilization decision tree method.

Table 5: Decision tree

State	class	Precision %	Recall %	Fscore %	Average accuracy %	Time (S)
Without normalization	0	1.00	1.00	1.00	99.7085	49.3057
	1	0.19	0.20	0.19		
With normalization	0	1.00	1.00	1.00	99.8103	4.2498
	1	0.24	0.04	0.07		
With normalization And feature selection (R-DT)	0	1.00	1.00	1.00	99.8261	0.8749
	1	0.00	0.00	0.00		

Table 5 showed the Decision Tree method, where normalization and RST with DT achieve 99.82 % accuracy higher than the general method. Rough-Decision Tree updateable achieve a higher time of 0.8 milliseconds. The Table 6 shows the Support Vector Machine classification methods.

Table 6: Support Vector Machine

State	class	Precision %	Recall %	Fscore %	Average accuracy %	Time (S)
Without normalization	0	1.00	1.00	1.00	99.8261	3055.1195
	1	0.00	0.00	0.00		
With normalization	0	1.00	1.00	1.00	99.8261	179.2896
	1	0.00	0.00	0.00		
With normalization and feature selection (R-SVM)	0	1.00	1.00	1.00	99.8261	953.9286
	1	0.00	0.00	0.00		

Table 6 showed the SVM method, where normalization and RDT with SVM achieve the same accuracy of 99.81 % as the general method. R-SVM updateable achieve time 953.92 seconds while the general method was more than three thousands second. The following figure shows the comparing diagram among those three methods accuracy then chosen the best performance classification methods.

Figure 2: Model accuracy

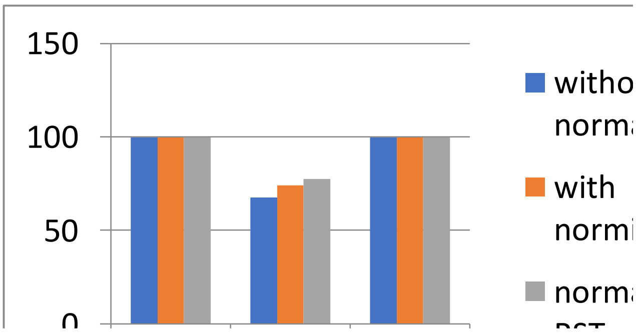

Figure 2 shown comparing accuracy for DT, NB and SVM classification methods. The voting method chooses DT and SVM as the best accuracy because it had the highest rate that achieved 99.82%, where DT and SVM achieved the same accuracy while NB achieved 77.57% a. The following figure shows the comparing diagram among those three methods accuracy then chosen the best performance classification methods.

Figure 3: Model time

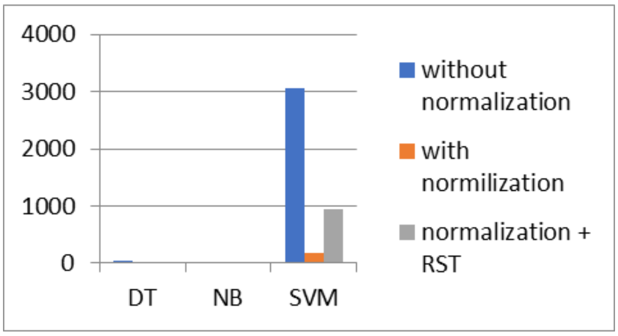

Figure 3 shown comparing the time for DT, NB and SVM classification methods. The voting method chooses NB as the best time because it had the highest rate that achieved 0.09 milliseconds, where DT and SVM achieved 0.8 and 953.92 seconds.

As a Voting, the DT will be the best method with Europe Database transactions for both accuracy and time.

vi. COMPARING OF PROPOSAL MODEL AND OTHER STUDIES

The table below shows five studies for fraud detection systems on credit cards, and these studies chased from different years. Each one of them applies various techniques to enhance the detection system and to achieve the highest accuracy with fast processing time. The proposed model of this thesis was the best of them where the accuracy reached 99.82%, and the time was just 0.8 milliseconds

Table I Comparing result's studies

Year	Name of study	Metric	Accuracy %	Time Durations /s
2021	Thesis Proposal model	Rough set theory with DT, NB and SVM	99.8261	0.8749
2018	Credit Card Fraud Detection using DL based on Auto-Encoder and Restricted Boltzmann Machine	deep Auto-encoder and restricted Boltzmann machine (RBM) with neural network	96.03	Study without timer
2019	Credit Card Fraud Detection Using Machine Learning	Spiting with naive Bayes, k-nearest neighbor and logistic regression	97.69	38.1

vi. CONCLUSION AND FUTURE WORK

This research focuses on enhancing classification methods to detect frauds transactions by using RST and choosing the right classification algorithm that fits the dataset can give an efficient system; choosing the right classification model leads to improving the system's efficiency. As a future work for this research, and after studying and knowing the results of these classification algorithms and comparing their performance to find the most suitable and best for the Europe database, it will be appropriate. Rough set theory was used as feature selection to enhance the accuracy where RS-DT has achieved higher than other techniques in the future to conduct a study to find a suitable improvement method applied to the algorithms with higher voting in this research, which helps in improving the performance of the algorithms and thus improving the performance of detection systems Fraud.

REFERENCES

[1] Y. Jain, S. NamrataTiwari, and S. Jain, "A comparative analysis of various credit card fraud detection techniques," *International Journal of Recent Technology and Engineering,* vol. 7, pp. 402-407, 2019.

[2] D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic, and A. Anderla, "Credit card fraud detection-machine learning methods," *2019 18th International Symposium INFOTECH-JAHORINA (INFOTECH),* 2019, pp. 1-5.

[3] A. A. Soofi and A. Awan, "Classification techniques in machine learning: applications and issues," *Journal of Basic and Applied Sciences,* vol. 13, pp. 459-465, 2017.

[4] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC medical informatics and decision making,* vol. 19, pp. 1-16, 2019.

[5] S. Raschka, J. Patterson, and C. Nolet, "Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence," *information,* vol. 11, p. 193, 2020.

[6] G. Piatetsky, "Python leads the 11 top Data Science," *Machine Learning Platforms: Trends and Analysis,* 2019.

[7] J. Awoyemi, A. Adetunmbi, and S. Oluwadare, "Effect of Feature Ranking on Credit Card Fraud Detection," 2018.

[8] J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis," in *2017 International Conference on Computing Networking and Informatics (ICCNI),* 2017, pp. 1-9.

[9] T. Kavipriya and N. Geetha, "An identification and detection of fraudulence in credit card fraud transaction system using data mining techniques," *International Research Journal of Engineering and Technology,* vol. 5, 2018.

[10] S. W. Ling, N. K. Batcha, and R. Logeswaran, "Machine Learning Model for Predicting Potential Donors Using Logistic Regression," *Journal of Applied Technology and Innovation (e-ISSN: 2600-7304),* vol. 4, p. 34, 2020.

[11] M. U. Safa and R. Ganga, "Credit Card Fraud Detection Using Machine Learning."," 2019.

[12] A. Pumsirirat and L. Yan, "Credit card fraud detection using deep learning based on auto-encoder and restricted Boltzmann machine," *International Journal of advanced computer science and applications,* vol. 9, pp. 18-25, 2018.

[13] C.-N. Jiao, Y.-L. Gao, N. Yu, J.-X. Liu and L.-Y. Qi, "Hyper-graph regularized constrained NMF for selecting differentially expressed genes and tumor classification," *IEEE journal of biomedical and health informatics,* vol. 24, pp. 3002-3011, 2020.

[14] S. Gopal Krishna Patro and K. K. Sahu, "Normalization: A Pre-processing Stage," *arXiv,* p. arXiv: 1503.06462, 2015.

[15] F. Carcillo, Y.-A. Le Borgne, O. Caelen, Y. Kessaci, F. Oblé, and G. Bontempi, "Combining unsupervised and supervised learning in credit card fraud detection," *Information Sciences,* 2019.

[16] E. A. Baruah, S. Baruah, and J. Goswami, "A Comparative Analysis of Different Classification Algorithms based on Students' Academic Performance Using WEKA."

[17] D. H. Abd, A. T. Sadiq, and A. R. Abbas, "Political articles categorization based on different naïve Bayes models," in *International Conference on Applied Computing to Support Industry: Innovation and Technology,* 2019, pp. 286-301.

[18] H. H. Patel and P. Prajapati, "Study and analysis of decision tree-based classification algorithms," *International Journal of Computer Sciences and Engineering,* vol. 6, pp. 74-78, 2018.

[19] N. A. Nnamoko, F. N. Arshad, D. England, and J. Vora, "Meta-classification model for diabetes onset forecast: A proof of concept," in *Bioinformatics Biomedicine (BIBM), 2014 IEEE International Conference on,* 2014, pp. 50-56.

[20] B. Liu, *Web data mining: exploring hyperlinks, contents, and usage data*: Springer Science & Business Media, 2007.

[21] R. Moraes, J. F. Valiati, and W. P. G. Neto, "Document-level sentiment classification: An empirical comparison between SVM and ANN,"

Expert Systems with Applications, vol. 40, pp. 621-633, 2013.

[22] M. Dua, "Attribute Selection and Ensemble Classifier based Novel Approach to Intrusion Detection System," *Procedia Computer Science*, vol. 167, pp. 2191-2199, 2020.

[23] S. Broumi, F. Smarandache, and M. Dhar, *Rough neutrosophic sets*: Infinite Study, 2014.

[24] S. Rizvi, H. J. Naqvi, and D. Nadeem, "Rough Intuitionistic Fuzzy Sets," in *JCIS*, 2002, pp. 101-104.

[25] S. Lee and G. Vachtsevanos, "An application of rough set theory to defect detection of automotive glass," *Mathematics and computers in simulation*, vol. 60, pp. 225-231, 2002.

[26] F. Questier, I. Arnaut-Rollier, B. Walczak, and D. Massart, "Application of rough set theory to feature selection for unsupervised clustering," *Chemometrics and Intelligent Laboratory Systems*, vol. 63, pp. 155-167, 2002.

[27] J. Dai and Q. Xu, "Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification," *Applied Soft Computing*, vol. 13, pp. 211-221, 2013.

[28] H. Yu, G. Yang, M. Lin, F. Meng, and Q. Wu, "Application of rough set theory for NVNA phase reference uncertainty analysis in a hybrid information system," *Computers & Electrical Engineering*, vol. 69, pp. 893-906, 2018.

[29] B. Tripathy, R. Mohanty, and T. Sooraj, "Application of uncertainty models in bioinformatics," in *Biotechnology: Concepts, Methodologies, Tools, and Applications*, ed: IGI Global, 2019, pp. 141-155.

[30] J. Jelonek, K. Krawiec, and R. Slowiński, "Rough set reduction of attributes and their domains for neural networks," *Computational intelligence*, vol. 11, pp. 339-347, 1995.

[31] D. H. Abd, A. T. Sadiq, and A. R. Abbas, "Classifying political Arabic articles using support vector machine with different feature extraction," in *International Conference on Applied Computing to Support Industry: Innovation and Technology*, 2019, pp. 79-94.

[32] I. S. Al-Mejibli, D. H. Abd, J. K. Alwan, and A. J. Rabash, "Performance evaluation of kernels in support vector machine," in *2018 1st Annual International Conference on Information and Sciences (AiCIS)*, 2018, pp. 96-101.