المجلة الإلكترونية الشاملة متعددة التخصصات
العدد التاسع والعشرون شهر(10) 2020

Machine learning to detect identity fraud

Name: Munayfah Mamdouh Nawi Al-shammari

A master's student in cyber security at Al-Jouf University

E-mail: 401205813@ju.edu.sa

Abstract:

Identity fraud is one of the most dangerous frauds and problems that threaten individuals and threaten their stability, as it is considered a kind of malicious harm that individuals face, for several reasons including, theft, extortion and many bad acts.

Many government and private transactions require the presentation of an ID card, starting from completing the largest government transaction to the least government transaction, so a method must be found that seeks to detect identity fraud using machine learning.

Through reference to previous studies, practical research, scientific papers, and master's and doctoral letters, it was found that machine learning has a high ability to detect impersonation.

Key word: machine learning, fraud, detect identity, and detect fraud.

Introduction:

The peoples of the world have been keen, from the beginning of humanity until today, to preserve their social, national, and cultural distinction and uniqueness. Therefore, they have been keen on having an identity that helps elevate individuals in societies. From one another, identity is an integral part of the formation of individuals from their birth until their departure from life (Van Knippenberg, A., 2020).

The existence of the idea of identity contributed to the expression of a set of characteristics of individuals' personalities; Because identity adds to the individual individuality and personality, as it is the image that reflects his culture, language, belief, civilization, and history, and also contributes to building bridges of communication between all individuals, whether within their societies, or with societies that differ from them partially, or depending on the language. Culture, or thought, or a complete difference in all fields without exception (Hermans, H. J, 2018).

Identity is based on historical and social implications, which in turn helps people to know their identity as well as the identity of others through communication with family, peers, institutions, organizations, the media and other means of communication in our daily life (T., Lam, et al, 2019).

There is no doubt that many governmental and private transactions require the presentation of the ID card, so starting with the completion of the largest government transaction to the least government transaction, all of them require the presentation of the original ID card to be completed (Burnes, D., DeLiema, M., & Langton, L., 2020).

However, it seems that some parties are lenient, first of all, in preserving their image of this important national document, and secondly using them for the benefit of other people.

Any person, at any time, can be exposed to identity fraud. Adolescents, in particular, are at risk of falling victim to theft which leads to the spread of information without knowing the risks it may face, there was no interest in the crime of identity fraud, but in recent years' things have to change as it has become clear that the crime of fraud and identity theft has become the fastest growing crime compared to other crimes in the world.

However, there are many methods that help detect identity fraud, and one of these methods is by indicating whether an image received from a source (eg, a security camera placed at an entrance) (Lebel, H., et al,2019), or using a plurality of historical identity records (Coggeshall, S., et al, 2010), ...etc. In our study, however, we will use machine learning to detect identity fraud.

Research objectives:

In general, this study aims to identify the methods of machine learning to discover identity fraud, through the main objective, the I sought to make comparisons between studies to demonstrate the best methods that help in detecting identity fraud, to show the most important challenges to discover identity fraud, and the future direction to discover Identity fraud.

Machine learning:

Machine learning is a type of artificial intelligence, which allows software applications to become more accurate in predicting results without explicitly programming them. It is the automatic improvement of the computer learning process based on previous computer experiences, but without programming it, that is, without human assistance. This process begins with entering high-quality data. The choice of algorithms depends on the type of data and the type of work that we plan to automate (Hutter, F., et al, 2019).

The primary focus of machine learning is building algorithms that can import input data, and use statistical analysis to predict outcomes within an acceptable range.

The interest in machine learning is due to its ability to increase the amount of data available, cheaper and more powerful computing, and store data more cost-effectively (Alpaydin, E., 2020).

All this means that it is possible to quickly and automatically create models that can analyze larger and more complex data and provide faster and more accurate results, on a very large scale. By creating highly accurate models, a higher percentage can be obtained to identify profitable opportunities and avoid any unknown risks (Mohri, M., Rostamizadeh, A., & Talwalkar, A., 2018).

Machine learning algorithms are categorized into: supervised learning ,and unsupervised learning. In supervised learning, humans provide the required input and output, in addition to providing the accuracy of predictions during training the algorithm. Once the algorithm finishes learning,

It will apply what learned to new data. In unsupervised learning, there is no need to train the algorithm with the required outputs, and instead, it uses an iterative approach called: (deep learning); unsupervised learning algorithms are used for more complex processing tasks than supervised learning systems.

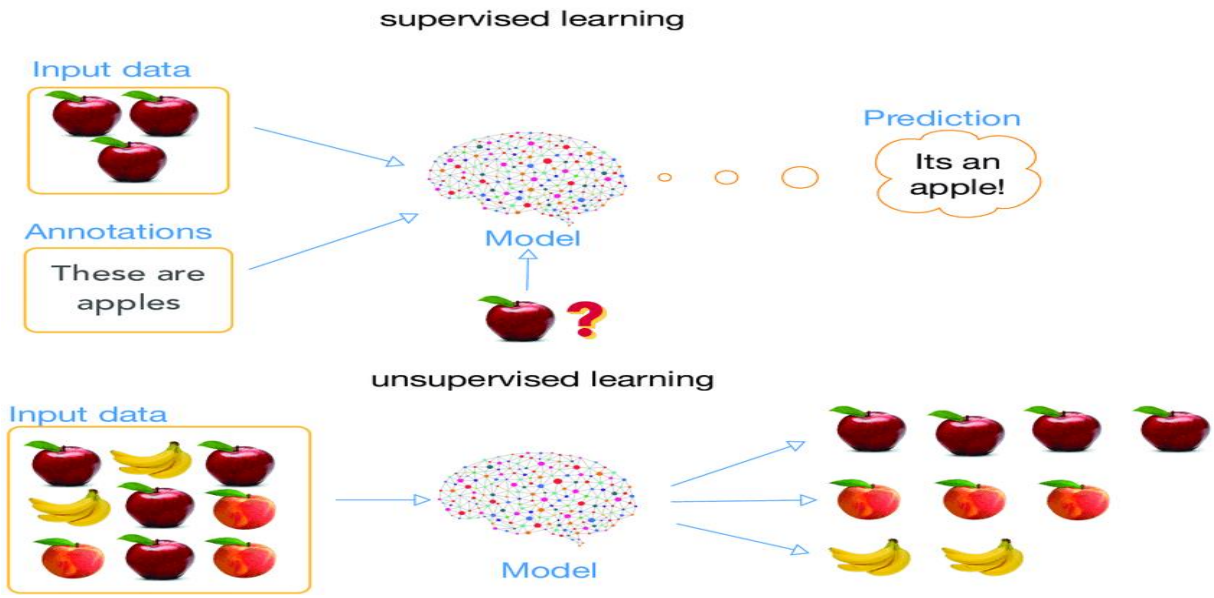

Figure 1: supervised and unsupervised machine learning (Ma, Y., et al , 2018)

Machine Learning Algorithm:

Machine learning algorithms from the perspective describe the commonalities of the algorithms (such as function, mode of operation). Below we categorize them according to common factors of algorithms, and based on the algorithms that aid fraud detection:

Regression algorithm:

Is a type of algorithm that obtains the best combination of input features by reducing the gap between the expected value and the actual result value? For continuous value

prediction, there is linear regression, etc., and for discrete value / category prediction, we can also consider logistic regression as a kind of regression algorithm. Common regression algorithms are as follows: Ordinary Least Squares Regression (OLSR), Linear Regression, Logistic Regression, Stepwise Regression, Locally Estimated Scatterplot Smoothing (LOESS), and Multivariate Adaptive Regression Splines (MARS) (Verrelst, J., et al, 2012).

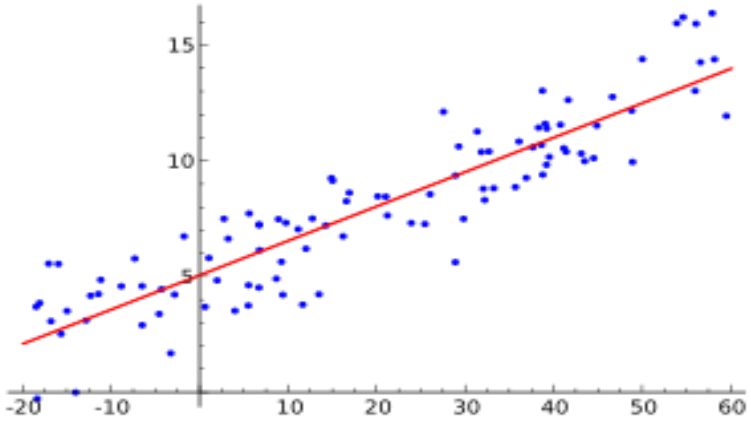

Figure 2: Regression algorithm (Wikipedia)

K Nearest Neighbor (KNN)

KNN is a very simple and very effective algorithm. The KNN model is represented by a complete set of training data. Very simple, isn't it? The prediction of a new point is done by finding the nearest K neighbors in the data set and collecting the output variable for these K instances. The only question is how to determine the similarity between the data instances. If all signs have the same scale (for example, centimeters), the easiest way is to use Euclidean distance - a number that can be calculated based on differences with each input variable (Zhang, S., et al, 2017).

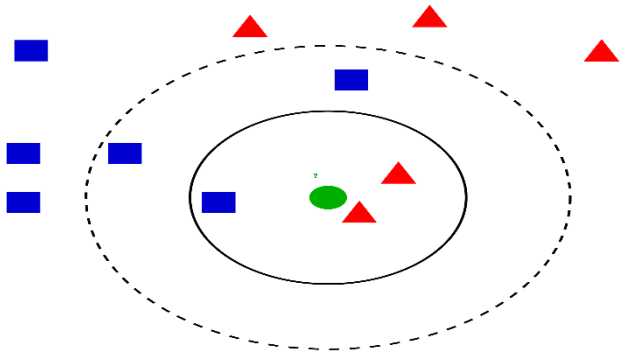

Figure 3: KNN (Srivastava, T., 2018)

Support vector method (SVM):

The support vector method is perhaps one of the most popular and discussed machine learning algorithms. Support Vector is the space dividing line for the input variables. In the support vector method, the trailing plane is chosen so that the points in the plane of the input variables are better separated by their class: 0 or 1.

In the two-dimensional plane, this can be represented as a line that completely separates the points of all classes, during training, the algorithm looks for parameters that help separate layers better with an excessive plane (Zhang, Y. et al, 2016).

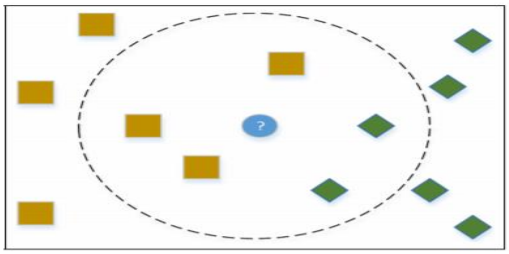

Figure 4: KNN Illustration of k-nearest neighbors (the K value is yellow) (Zhang, Y., et al, 2016)

Decision Tree:

A decision tree is a special tree structure in which each node represents a decision, and is linked to all decision options through a top-down flow structure.

A decision tree are useful structures for regression and classification problems. As you can see, each node is divided into two paths. In this case, each node in our binary decision tree is a feature of our data set (Brijain, M., et al, 2014).

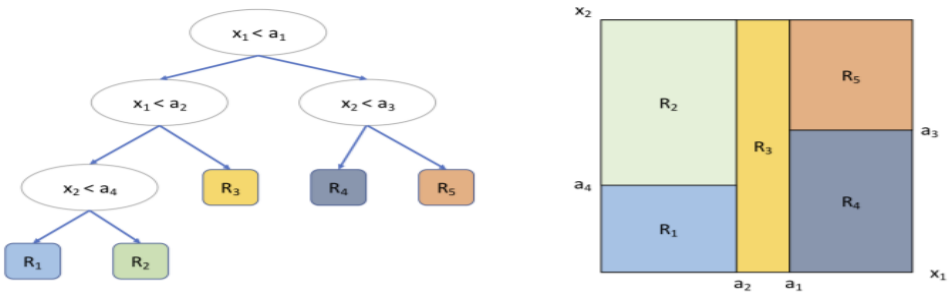

Figure 5: Decision Tree (Lamb, K. D., 2019)

Random forest:

The random forest is based on the aforementioned concept, by introducing randomness into the parameters of each node in the split binary decision tree. Decision trees are greedy, which means they use a greedy algorithm to determine the optimum parameter value that will be segmented to reduce errors. For our random forest to function, all trees must be as unrelated as possible. Therefore, random forest modified the CART model and divided it into a large number of random values in different subsamples of data, so the beauty of the random forest is that when we randomly run our model on a subsample of the data, the error we get for each prediction is random. These errors can be modeled by randomization. The mean value of the random distribution is 0. Therefore, when we run our random branching and take the average of all predictions, we will obtain a model that minimizes errors under ideal conditions (Liaw & Wiener, 2002).

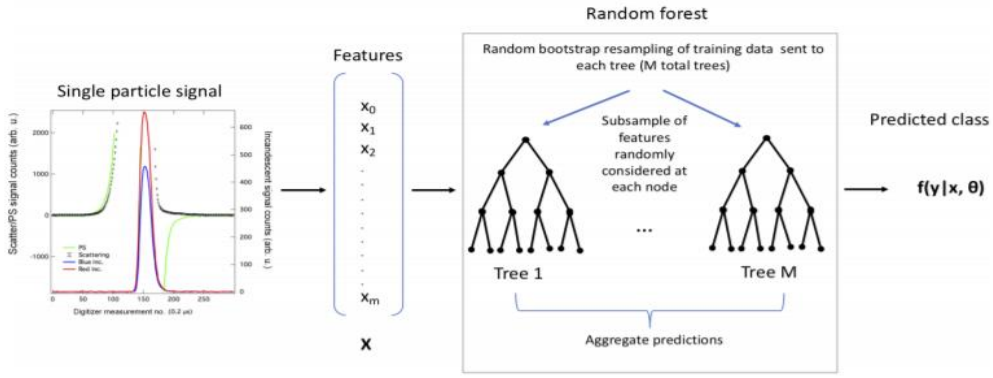

Figure 6: Random forest (Lamb, K. D., 2019)

Machine Learning to detect fraud: How it works?

Identity fraud began to appear and spread recently, it was associated with advanced technology, which is computer technology, the Internet, machine learning and artificial intelligence, which resulted in distinguishing it with a set of facts that made it different from other crimes.

In most organizations, fraud is often discovered after it has occurred. Ideally, measures should be taken before they occur, or at least before major damage occurs. Our fraud detection platform allows organizations to detect fraud before it occurs, using machine learning techniques; The platform detects fraud in all types of sectors, including but not limited to the banking sector, the insurance sector, the government sector, and the healthcare sector (Davis, B., & Conwell, W., 2007).

Although uncovering crimes usually requires knowing the nature of the crime you are looking for, machine learning is now able to discover it better than humans. Machine learning detects patterns of behavior that humans cannot observe, and it can intervene to stop suspicious activity before it gets out of control.

The following figure summarizes the mechanism by which machine learning works to detect fraud:

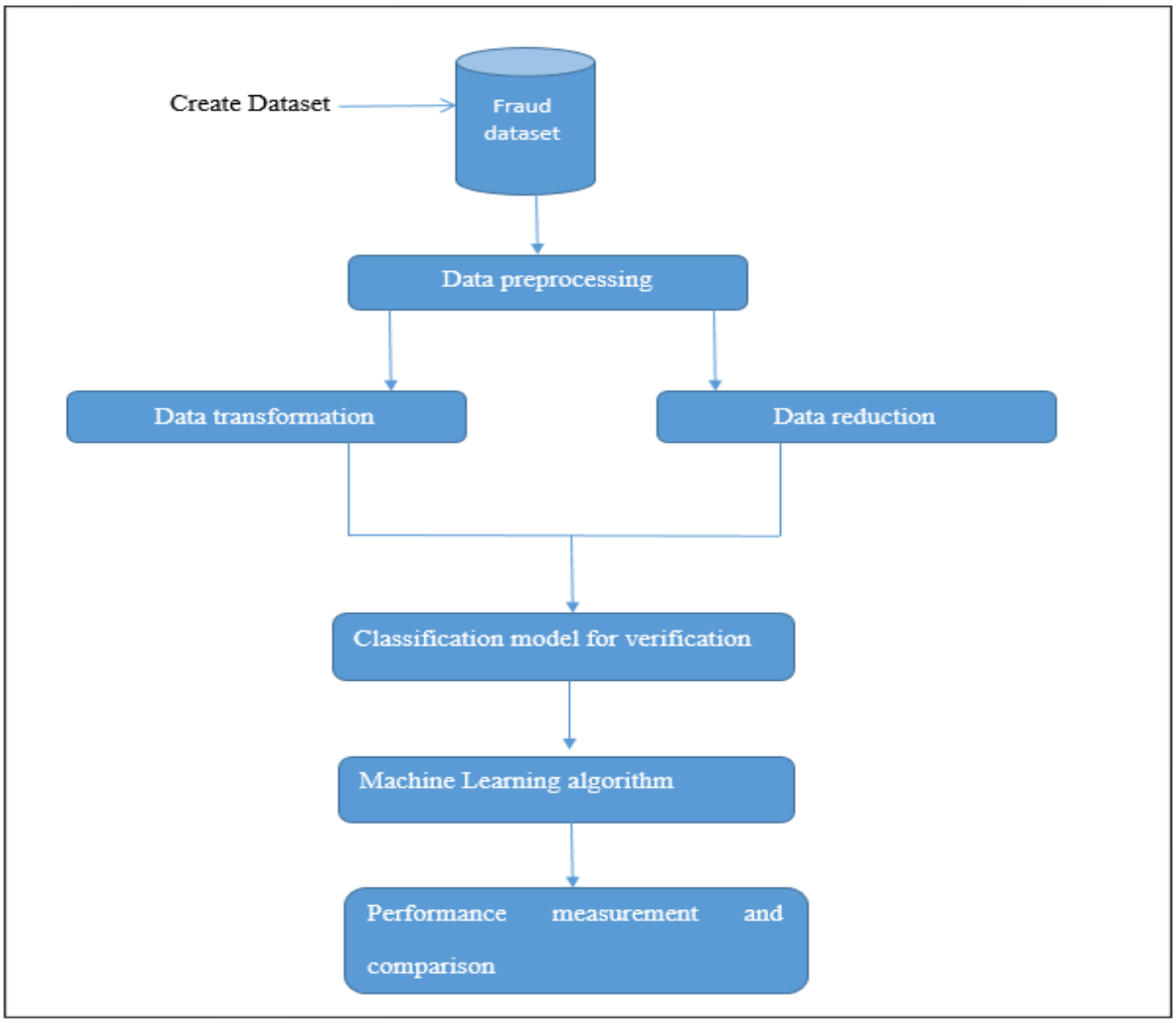

Figure 7: detect Fuard methodology (Yee, O. S., et al, 2018)

In the beginning, there must be a comprehensive dataset containing fraud and non-fraud identities in order to conduct studies and distinguish between them.

In the process of data processing, through which operations are performed that lead to making the data in a way that can be dealt with and draw conclusions,

As the unordered and unclassified data (dirty data) and containing duplicates cannot lead to clear and good-quality results. After that we clean the data, in this step it is possible to fill in the gaps in the resulting data as in the use of the most likely data (setting averages or zero in some cases) for the purpose of ridding the final results from fragmentation, also includes dealing with outliers, as well as including them Remove duplicate data. As for data transfer, in this step, the data is converted into appropriate information forms, meaning that the normalization process occurs (García, S., Luengo, J., & Herrera, F., 2015).

Subsequently, the practical application of the dataset is performed by applying machine learning algorithms to it, such as KNN, Naïve Bayes, Random forest, decision tree, regression algorithm, and SVM.

After reaching the required results, it is evaluated by relying on several rules and equations, including accuracy, precision, recall, and error rate.

Comparison of machine learning algorithms to detect identity fraud:

In the following table, studies are summarized that aim to detect identity fraud using machine learning:

Table 1: Machine learning Algorithms

Num	references	objectives	Algorithm	result
1	Raghavan, P., & El Gayar, N.	- Fraud Detection to detecting unusual activities using data mining	Using three algorithm of machine learning algorithm which are k-	Two methods for measurement were used to evaluate the

	(2019)	techniques. - The datasets which will be used are the European (EU) Australian and German dataset.	nearest neighbor (KNN), random forest, and support vector machines (SVM).	results which are Matthews Correlation Coefficient, and Accuracy. 1- Where got MCC for KNN is 0.2487 and AUC is 0.6047. 2- Where got MCC for Random forest is 0.2912 and AUC is 0.6437. 3- Where got MCC for SVM is 0.4038 and AUC is 0.6857. 4- In general the best algorithm based on the accuracy is SVM.
2	Sadgali, I., Sael, N., & Benabbou, F. (2019)	The aim of this study is to identify the techniques and methods that give the best results that have been perfected to detect the financial fraud.	Using machine learning algorithm which are: Bayesian Belief Networks, Genetic algorithm, Support Vector	1- we found the PNN was the best performing that get 98.09% ,

		The datasets which will be used NSL-KDD dataset.	Machine, Classification and Regression Tree (CART), Multilayer Feed Forward Neural Network (MLFF), Genetic Programming (GP), and Naives bays	another one is Genetic algorithm (95%) who gave marginally lower accuracies in most cases. 2- after that the Naives bays and SVM gives good results which get 99,02%, 98,8%.
3	Wu, S. H., et al (2015).	This study aims to detecting the Identity Fraud on Social Network. The set D contains 278 instances of which 178 are positive, and 100 are negative. Each instance is represented by an 139-dimension feature vector.	Use the support vector machine (SVM).	Our experiment we can achieve higher than 80% detection accuracy within 2 min, and over 90% after 7 min of observation time.
4	Yee, O. S., Sagadevan, S., & Malim, N. H. A. H. (2018).	This study aims to detect the Credit Card Fraud Detection Using Machine Learning. Using two data set the first one dummy dataset and the	using five machine learning algorithm which are: Naïve Bayes, K2, TAN, Logistics, and J48	We found the accuracy of the result based on the algorithm which are Naïve Bayes= 96.7%, K2 = 95.8%, TAN= 99.7%,

		second is a newly transformed.		Logistics= 100.0%, and J48= 100.0% Overall, all the Bayesian classifiers achieved significantly better results after being fed with filtered data.

Future direction:

Identity fraud is a community problem that must be solved, so in the future we have directed to spread awareness among people to show the most important risks that occur as a result of identity fraud. In future studies, researchers could seek to develop machine learning techniques and apply them to more than one database to show the highest rate Accuracy can be accessed, in addition to the application of artificial intelligence techniques and their combination with machine learning to achieve a high ability to detect identity fraud.

Conclusion:

One of the main problems that threaten individuals' safety and stability is recognizing and identifying identity as it is realized an image of predictable individual's malicious harm.

There are several methods that help in detecting identity fraud, as they have relied on machine learning techniques. After going back to previous studies and analyzing them, it was found that machine learning has a high ability to detect identity fraud.

References:

[1] Alpaydin, E. (2020). Introduction to machine learning. MIT press. https://books.google.com/books?hl=en&lr=&id=tZnSDwAAQBAJ&oi=fnd&pg=PR7&dq=Machine+learning+&ots=F3RT714rAg&sig=-SZzUfCj-6hblJVvJsaBh-IRdCg

[2] Brijain, M., Patel, R., Kushik, M., & Rana, K. (2014). A survey on decision tree algorithm for classification.

[3] Burnes, D., DeLiema, M., & Langton, L. (2020). Risk and protective factors of identity theft victimization in the United States. Preventive Medicine Reports, 17, 101058.

[4] Coggeshall, S., Jost, A., Blue, J., DiChiara, C. J., Cook, M., & Shao, X. (2010). U.S. Patent No. 7,793,835. Washington, DC: U.S. Patent and Trademark Office.

[5] Davis, B., & Conwell, W. (2007). U.S. Patent Application No. 11/613,891. https://patents.google.com/patent/US20120123959A1/en

[6] García, S., Luengo, J., & Herrera, F. (2015). Data preprocessing in data mining (pp. 195-243). Cham, Switzerland: Springer International Publishing.

[7] Haslam, C., Steffens, N. K., Branscombe, N. R., Haslam, S. A., Cruwys, T., Lam, B. C., ... & Yang, J. (2019). The importance of social groups for retirement adjustment: evidence, application, and policy implications of the social identity model of identity change. Social issues and policy review, 13(1), 93-124.

[8] Hermans, H. J. (2018). Society in the self: A theory of identity in democracy. Oxford University Press.

[9] Hutter, F., Kotthoff, L., & Vanschoren, J. (2019). Automated machine learning: methods, systems, challenges (p. 219). Springer Nature.

[10] Lamb, K. D. (2019). Classification of iron oxide aerosols by a single particle soot photometer using supervised machine learning. Atmospheric Measurement Techniques, 12(7), 3885-3906.

[11] Lebel, H., Friedman, O., Danino, Y. U., Littman, R., & Yehuda, Z. B. (2019). U.S. Patent Application No. 15/886,017.

[12] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

[13] Ma, Y., Liu, K., Guan, Z., Xu, X., Qian, X., & Bao, H. (2018). Background augmentation generative adversarial networks (BAGANs): Effective data generation based on GAN-augmented 3D synthesizing. Symmetry, 10(12), 734.

[14] Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). Foundations of machine learning. MIT press.

[15] Raghavan, P., & El Gayar, N. (2019, December). Fraud Detection using Machine Learning and Deep Learning. In 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE) (pp. 334-339). IEEE.

[16] Sadgali, I., Sael, N., & Benabbou, F. (2019). Performance of machine learning techniques in the detection of financial frauds. Procedia computer science, 148, 45-54.

[17] Srivastava, T. (2018). Introduction to k-Nearest Neighbors: A powerful Machine Learning Algorithm (with implementation in Python & R). Analyticsvidhya. com.

[18] Van Knippenberg, A. (2020). Strategies of identity management. In Ethnic minorities (pp. 59-76). Garland Science.

[19] Verrelst, J., Muñoz, J., Alonso, L., Delegido, J., Rivera, J. P., Camps-Valls, G., & Moreno, J. (2012). Machine learning regression algorithms for biophysical

parameter retrieval: Opportunities for Sentinel-2 and-3. Remote Sensing of Environment, 118, 127-139.

[20] Wu, S. H., Chou, M. J., Tseng, C. H., Lee, Y. J., & Chen, K. T. (2015). Detecting in situ identity fraud on social network services: A case study with facebook. IEEE Systems Journal, 11(4), 2432-2443.

[21] Yee, O. S., Sagadevan, S., & Malim, N. H. A. H. (2018). Credit card fraud detection using machine learning as data mining technique. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 10(1-4), 23-27.

[22] Yee, O. S., Sagadevan, S., & Malim, N. H. A. H. (2018). Credit card fraud detection using machine learning as data mining technique. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 10(1-4), 23-27.

[23] Zhang, S., Li, X., Zong, M., Zhu, X., & Cheng, D. (2017). Learning k for knn classification. ACM Transactions on Intelligent Systems and Technology (TIST), 8(3), 1-19.

[24] Zhang, Y., Lu, S., Zhou, X., Yang, M., Wu, L., Liu, B., ... & Wang, S. (2016). Comparison of machine learning methods for stationary wavelet entropy-based multiple sclerosis detection: decision tree, k-nearest neighbors, and support vector machine. Simulation, 92(9), 861-871.