© Universiti Tun Hussein Onn Malaysia Publisher's Office

IJIE

Journal homepage: http://penerbit.uthm.edu.my/ojs/index.php/ijie
ISSN : 2229-838X e-ISSN : 2600-7916

The International Journal of Integrated Engineering

Comparative Study of Machine Learning Algorithms and Correlation Between Input Parameters

Muhammad Syafiq Alza Alias[1], Norazlin Ibrahim[1], Zalhan Mohd Zin[1*]

[1]Industrial Automation Section,
Section 14, Sect. 14, Jalan Teras Jernang 43650 Bandar Baru Bangi, Selangor, MALAYSIA

* Corresponding Author

DOI: https://doi.org/10.30880/ijie.2019.11.04.009
Received 23 April 2019; Accepted 8 July 2019; Available online 5 September 2019

Abstract: The availability of big data and computing power have triggered a big success in Artificial Intelligence (AI) field. Machine Learning (ML) becomes major highlights in AI due to the ability of self-improved as it is fed with more data. Therefore, Machine Learning is suitable to be applied in financial industry especially in detecting financial fraud which is one of the main challenges in financial system. In this paper, 15 different types of supervised machine learning algorithms are studied in order to find the highest accuracy that should be able to detect credit card fraudulent transactions. The best algorithm among these algorithms is then further used and studied to find the correlation between the input variables and the accuracy of the results produced. The results have shown that Multilayer Perceptron (MLP) produced the highest accuracy among the 15 other algorithms with 98% accuracy of detection. Besides that, the input parameters also play an important role in determining the accuracy of the results. Based on the result, when input parameter known as 'V4' decreased, the recorded accuracy has increased to 99.17%.

Keywords: Artificial Intelligence, Machine Learning, Financial Fraud, Input Parameter.

1. Introduction

Financial fraud is a crucial crime that has become one of the main concerns to governments throughout the world. In 2017, the Consumer Sentinel Network took in nearly 2.7 million re-ports with people losing nearly $905 million just for fraud cases with an increase of $63 million over what consumers reported losing in 2016 [1]. What is more worrying than the fact of mil-lion-dollar lost is that a 2016 study from the Better Business Bureau stated that more than half of those who suffered a fraud-related financial loss had a college or university degree [2]. The Association of Certified Fraud Examiners (ACFE) shares the definition of fraud as "the use of one's occupation for personal enrichment through the deliberate misuse or application of the employing organization's resources or assets" [3].

One of the most studied kind of financial fraud is Credit Card Fraud [4] and it can be described as the activity of achieving financial advantage or causing economic loss to the credit card holder without their acknowledgement [5]. Fraudsters can commit a credit card fraud in multiple ways. Generally, according to [6], credit card fraud can be categorized into three categories which are traditional card related frauds, merchant related frauds and Internet frauds.

Today, Artificial Intelligence (AI) has come to play an integral role and proven to be beneficial in financial industry [7]. It helps to save time, reduce costs and add value [8]. Contrary to what people might think, AI is hardly a new topic. It was pro-posed by J. McCarthy in 1956 when the seminal summer workshop was organized at Dartmouth College, New Hampshire, US [8]. For years, AI remained a subject of scholarly study or an inspiration for science-fiction writers. However, there has been a significant acceleration in recent years such as availability of data, computing power and efficiency of algorithm [9]. One of the highlights in this AI revolution are machine learning

*Corresponding author: zalhan@unikl.edu.my algorithms, software that self-improves as it is fed more and more data, a trend that the financial industry can benefit from immensely.

Although machine learning algorithms is proven good in classification of data, it is still a challenging task to detect fraud. This is due to a few reasons such as the fraudsters will try to make every fraudulent transaction legitimate [10], highly imbalanced dataset because the data have more legitimate transactions compared to the fraudulent [11] and others. Therefore, this research will focus on detecting credit card fraudulent transactions by applying different supervised machine learning algorithms to get the suitable machine learning algorithm with high accuracy of detection. Besides that, this research will also be conducted to study the correlation between the input variables and the accuracy of the result produced.

2. Literature Review

Machine Learning can be classified into 3 main categories which are supervised learning, unsupervised learning and reinforcement learning [12]. Supervised learning the most common form of machine learning [13] and can be defined as the use of a marked feature set to retain some classification function or can be said as labelled trained data set [14]. Overview of the machine learning algorithms used in this research is as follow:

2.1 Multi-Layer Perceptron (MLP)

The most utilized algorithm in machine learning is MLP due to its ability to approximate any non-linear function with high accuracy of result [15]. MLS also known as multi-layer feed forward neural network with basically consist of 3 types of layers (Fig. 1). The first one is called input layer and it is used for the input variables. The second is the hidden layer which might consist of more than 1 layer which function is to determine non-linear relationships among the input variables. The result of the network or the predicted output value is given in the last layer known as output layer. Each layer is important in determination of the result.

In [25], the result shows that MLP performs better than Chebyshev Functional Link Artificial Neural Network (CFLANN) and Decision Tree in both of their experiment with Australian Cred-it Card dataset and German Credit Card dataset with the accuracy of 85.4% and 84.12% respectively

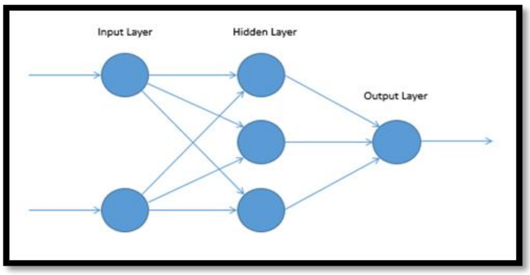

Fig. 1 - The architecture of MLP [16].

2.2 Logistic Regression (LR)

Logistic regression is used to determine the correlation between binary outputs and independent variables by giving out a probability value as the predicted value of the dependent variable [17]. It is based from logistic function or can be called as sigmoid function which take any real value and map it into its S-shape curve between value 0 and 1. Mathematically, it can be represent as the following:

$$y = e^\wedge(b0 + b1*x) / (1 + e^\wedge(b0 + b1*x)) \qquad (1)$$

Definition 2.2: where y is the predicted value, x is the input value, b0 is the bias and b1 is the coefficient of the input value.

2.3 K-Nearest Neighbour (KNN)

Almost like Euclidean, Mahanttan and Minkowski distance functions, the K-nearest neighbour is an instance-based learning which carries out its classification based on a similarity measure [18]. It is among the simplest machine learning algorithm. In this research the Euclidean distance will be used to find distance between two input vectors. The equation of Euclidean distance can be written as:

$$D = \sqrt{\sum_{k=1}^{n}(Xa - Xb)^2} \quad k = 1,2,\ldots,n \tag{2}$$

Definition 2.3: where D is the Euclidean distance, Xa and Xb are input vectors and n is the number of closest training example.

2.4 Support Vector Machine (SVM)

Another popular machine learning algorithm is SVM which can be described as algorithm which finds an optimized hyper-plane that separates training samples of different class as much as possible [19]. Fig. 2 shows the dataset is separated between two class with a linear line.

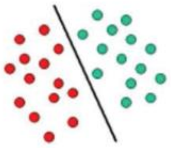

Fig. 2 - Linear Hyperplane.

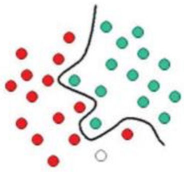

Fig. 3 – Non-Linear Hyperplane.

However, nowadays, the data is more complex and cannot be separated with a linear line. Therefore, Kernel function is introduced and used to map the nonlinear separable classes of data into a higher dimension (Fig. 3) [20]. The Kernel function can be expressed as [21]:

$$K(x, y) = [p(x), p(y)] \tag{3}$$

Definition 2.4: where K is the kernel function, x and y are n dimensional inputs, f is a map from n-dimension to m-dimension space. [x,y] denotes the dot product.

According to work done in [26], by using Microarray Breast Cancer data, SVM has recorded the highest accuracy among 6 other machine learning algorithms which are KNN, Decision Tree, Random Forest, Logistic Regression, Adaboost and Gradient Boosting Machines (GBM). SVM logged 99.23% of accuracy after feature selection methods followed by Logistic Regression, KNN, Random Forest, Adaboost, GBM and Decision Tree.

2.5 Decision Tree (DT)

DT can draw well-defined rules from data set having a root node and number leaf nodes. Each class is correlate with the leaf node [15]. In DT, each internal node splits the attributes into different branches based on the number of discrete values from that attribute. This technique’s simplicity, transparency and self-exploratory makes it very popular in classification and regression [22].

2.6 Random Forest (RF)

Basically, RF is developed based on DT due to single-tree model is sensitive to specific training data and easy to overfit [23]. It works by having multiple DT in the training phase and the out-put class is based on the majority vote from each tree [24]. Fig. 4 illustrated the architecture of RF algorithm.

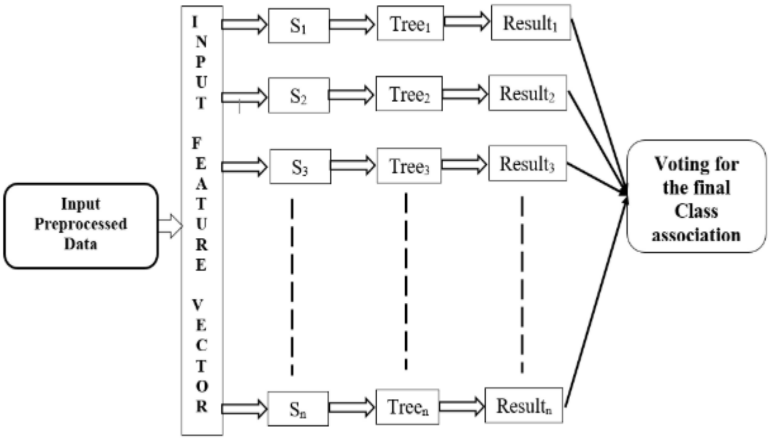

Fig. 4 – Architecture of RF algorithm.

3. Methodology

In this research, 15 machine learning algorithms will be used to detect credit card fraudulence activity. Overview of this research methodology is shown in Fig. 5.

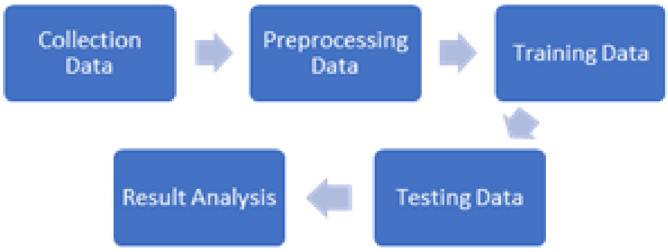

Fig. 5 – Overview of research methodology.

The simulations of this research basically begin with the collection of data. The data is collected from Kaggle website which owned by Google Inc which allows user to find and publish dataset. The data which is used in this research contain credit cards transactions made by European cardholders in September 2013. These two days transaction dataset is highly imbalanced with only 492 frauds out of total of 284,807 transactions. All the input variables contain only numerical value from the result of a PCA transformation. However, the original features are not provided except for Time and Amount due to confidentiality of data. All the features are renamed to Features V1, V2, V3 until V28 (Table 1). The output of the data whether the transaction is fraudulence or not is tabulated in Feature 'Class. Value 1 is assigned when the transaction is a fraudulence transaction and 0 is for genuine transaction.

Table 1 - Sample data from Credit Card Transaction dataset.

No	Time	V1	V2	…	V26	V27	V28	Amount	Class
0	0.0	-1.359807	-0.072781	...	-0.189115	0.133558	-0.021053	149.62	0
1	0.0	1.191857	0.266151	...	0.125895	-0.008983	0.014724	2.69	0
2	1.0	-1.358354	-1.340163	...	-0.139097	-0.055353	-0.059752	378.66	0
3	1.0	-0.966272	-0.185226	...	-0.221929	0.062723	0.061458	123.50	0
4	2.0	-1.158233	0.877737	...	0.502292	0.219422	0.215153	69.99	0

In the pre-processing phase, the dataset will be checked for any missing values. Then, the features in the data is transformed by scaling each feature to a given range. After that, the data is split into training and testing data. In this research, 3,000 transactions data have been used for the simulations which consist of 2,508 legitimate transactions and all the 492 frauds transactions. The training data had 2,400 transactions with 1,989 legitimate transactions and 411 fraudulence transactions. The rest of 600 transactions data was used for testing with 519 legitimate trans-actions and 81 fraudulence transactions data.

The training data is used to train the machine learning algorithm for classification. All the machine learning algorithms is trained, and then, the result is validated by using the testing data. The accuracy of the algorithm is checked by comparing the results of testing data with actual label result of the data.

Lastly, this research will study further on the relationship be-tween input parameter and the result produced. In each simulation, one of the input parameters will be drop and the accuracy of fraud detection will be observed.

4. Results and Discussion

The dataset consists of imbalanced data (Fig. 6). Based on the data, 99.83% are from legitimate transaction class while 0.17% are from the fraudulence class. It is good sample data for this research because the real data trend will look similar as this data.

The results of this study are presented in Table 2. Based on the table, it can be concluded that MLP algorithm has the most accurate result compared to others followed by Logistic Regression and eXtreme Gradient Boosting (XGBoost). The MLP works well due to the structure of the algorithm which provides numerous of nonlinearity function such as rectified linear units (ReLU), Sigmoid function, Adaptive Moment Estimation (Ad-am), Dropout and others.

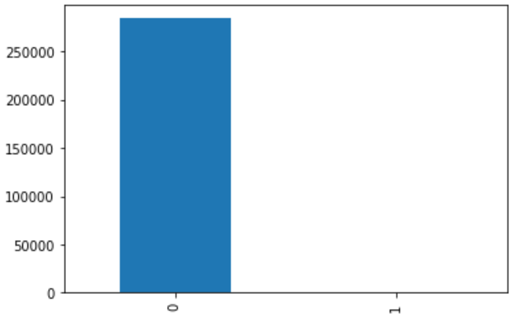

Fig. 5 – Imbalanced dataset.

Since MLP algorithm works well with this dataset, this research continues to study the correlation of input variables with the accuracy produced. Table 3 showed the result of fraud transac-tion detection when certain input parameter is drop out from the dataset. The first input parameter which is Time can be described as the most important parameter for this dataset. This is because when dropping the parameter from the dataset table, the accuracy of fraud detection is the lowest at 97.83%. Mean-while, the most irrelevant feature in the dataset can be is V4 due the highest accuracy of result produced after this column is drop from the dataset table. Therefore, it can be concluded that input parameter plays significant role in determining the accu-racy of result. Some of the features are not important to classify the data need to be drop.

Table 2 - Accuracy comparison between machine learning algorithms.

Algorithms	Accuracy
MLP	0.9800
Logistic Regression	0.9783
XGBoost	0.9700
K-Fold Cross Validation	0.9680
Random Forest	0.9630
Bagging Classifier	0.9630
Gradient Boosting Classifier	0.9620
VotingClassifier	0.9620
AdaBoostClassifier	0.9600
ExtraTreesClassifier	0.9600
KNN	0.9583
Decision Tree	0.9570
SVM	0.9500
GaussianProcessClassifier	0.9480
Gaussian Naïve Bayesian	0.9380

Table 3 - The accuracy of fraud detection when certain input parameter is drop from dataset.

Parameter Drop	Accuracy
Time	0.9783
V1	0.9867
V2	0.9867
V3	0.9867
V4	0.9917
V5	0.9867
V6	0.9850
V7	0.9867
V8	0.9867
V9	0.9867
V10	0.9867
V11	0.9900
V12	0.9867
V13	0.9867
V14	0.9867
V15	0.9867
V16	0.9867
V17	0.9867
V18	0.9867
V19	0.9883
V20	0.9850
V21	0.9867
V22	0.9867
V23	0.9867
V24	0.9867
V25	0.9867
V26	0.9867
V27	0.9867
V28	0.9883
Amount	0.9867

Fig. 7 reveals the training history of MLP algorithm when V4 parameter is drop from the dataset and the confusion matrix is shown in Fig. 8.

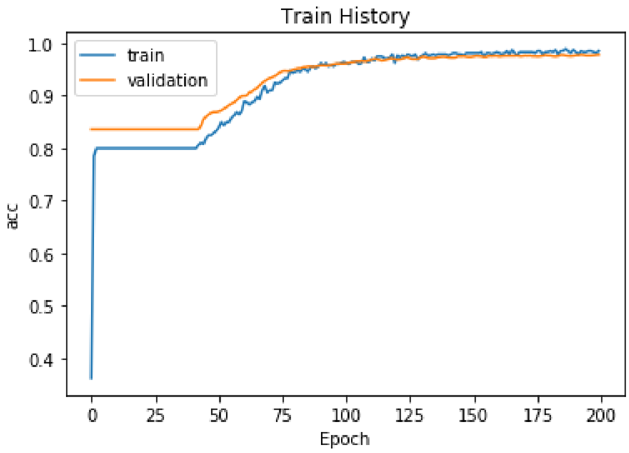

Fig. 7 – The training history of MLP algorithm when V4 is drop.

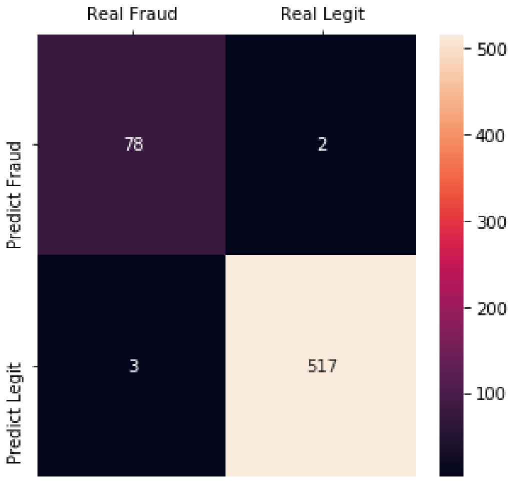

Fig. 8 – The confusion matrix of 600 testing data when V4 parameter is drop.

The train history for the MLP algorithm when Time parameter is drop is displayed in Fig. 9 with its confusion matrix in Fig. 10.

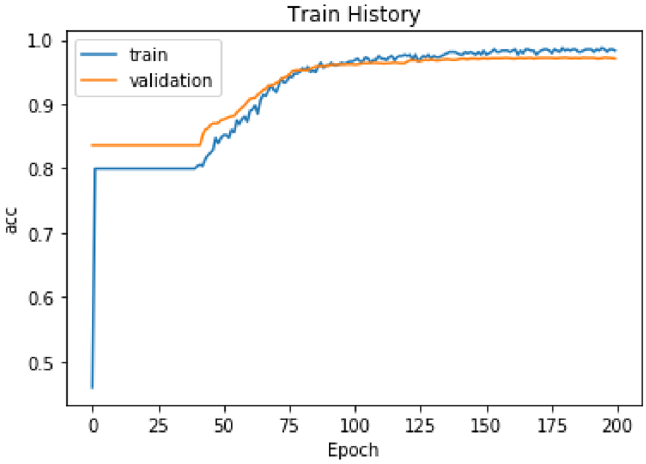

Fig. 9 – The training history of MLP algorithm when Time is drop.

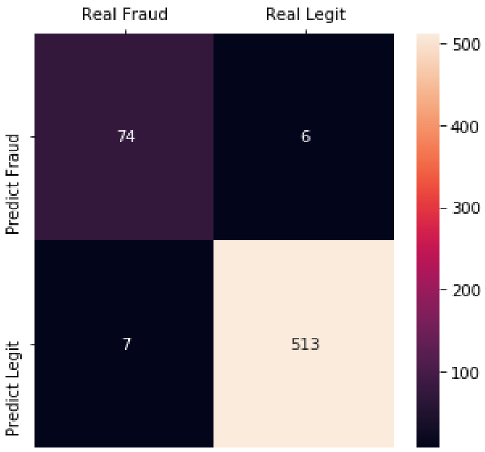

Fig. 10 – The confusion matrix of 600 testing data when Time parameter is drop.

Based on Fig, 7 and 9, the trend of the graph for training history of MLP algorithm is nearly the same. However, when input parameter 'V4' is dropped, the validation accuracy is slightly higher and produced result which almost the same as the training accuracy compared to the testing on 'Time' parameter dropped. Therefore, it can be said than 'V4' parameter is not an important feature in determining the trend of fraudulent trans-action data.

Although the performance on MLP is good with accuracy of more than 97%, the false negative value from the confusion matrix in Fig. 8 and Fig, 10 is still a worrying issue. This is because the banks or financial institutes cannot afford to slip any of the fraudulent transactions. In this research, by dropping the ‘V4’ parameter, 5 data are not classified correctly. There are 2 legitimate transactions are classified as fraud and 3 fraudulent data are classified as legitimate transactions. By referring to Fig. 10, 7 out of 519 legitimate transactions are wrongly categorized as fraudulence transactions and 7 from 81 fraudulence transactions is considered as legitimate transaction by the MLP algorithm.

The histogram of feature for parameter V4 is illustrated in Figure 11 and for Time parameter is in Figure 12. These 2 histograms help to in visualizing the data for human interpretation. Although in Fig. 11 shows more clear difference between fraud and legitimate transaction by human eye compared to histogram in Fig 12, however, the result gives a different conclusion. This is because MLP algorithm able to generate a complex function to detect the real trend of the data.

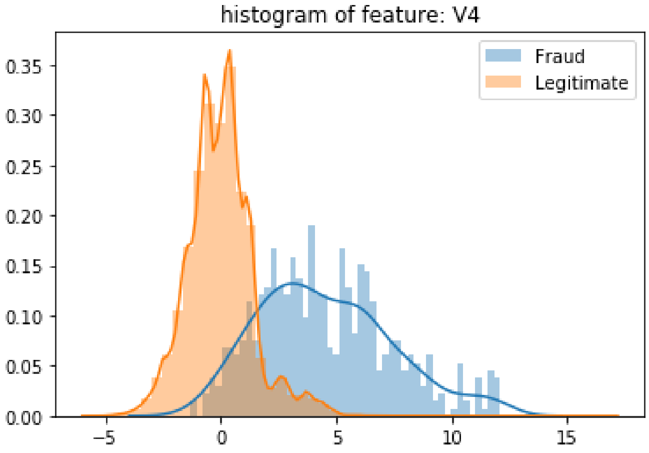

Fig. 11 – The histogram of feature for parameter V4.

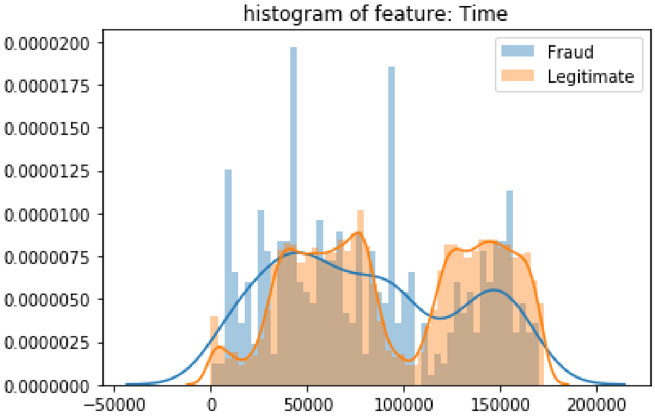

Fig. 12 – The histogram of feature for parameter Time.

5. Conclusion

In this paper, the accuracy of 15 machine learning algorithms on detecting fraudulent credit card transaction is presented. Based on the result MLP algorithm has the highest accuracy compared to others followed by LR, XGBoost, K-Fold Validation, Random Forest, Bagging classifier, Gradient Boosting classifier, Voting classifier, AdaBoostClassifier and ExtraTreesClassifier. Most of the results have good result with accuracy of detection more than 95%. However, the GaussionProcessClassifier and Gaussian Naïve Bayesian is doing less than that and the results recorded are 94.8% and 93.8% respectively. Therefore, it can be said that, the MLP algorithm works well to detect the Credit Card Fraudulence transaction. Besides that, input parameter is also essentials part in during the training of MLP algorithm. Selecting suitable input will generate a better performance in term of accuracy of fraudulence transaction detection.

In future works, the MLP algorithm will be studied to further increase the performance of detection. Each factor in MLP algorithm such as number of neurons, hidden layers, percentage of drop out and others will be tested to find out the most suitable design of MLP algorithm to detect fraudulence transaction. Besides that, the research will be conducted with real data based on the local area to ensure MLP algorithm trained can perform well with respective trend of data.

Acknowledgement

The authors would like to thank Ministry of Higher Education Malaysia and Universiti Kuala Lumpur (UniKL) for funding this research through research Grants (FRGS/1/2016/ICT02/UNIKL/02/2) and (UniKL STR15142).

References

[1] F. T. Commission, “Consumer Sentinel Network Data Book 2017,” 2018.
[2] E. Fletcher and R. Pessanha, *Cracking the Invulnerability Illusion*. 2016, p. 15.
[3] “Reporting/Investigating Fraudulent Acts,” 2017.
[4] P. K. Chan, W. Fan, A. L. Prodomidis, and S. J. Stolfo, “Distributed Data Mining in Credit Card Fraud Detection,” *IEEE Intell. Syst.*, vol. 14, p. 67, 1999.
[5] G. E. Melo-Acosta, F. Duitama-Munoz, and J. D. Arias-Londono, “Fraud Detection in Big Data using Supervised and Semi-supervised Learning Techniques,” 2017.
[6] R. Patidar and L. Sharma, “Credit Card Fraud Detection Using Neural Network,” *Int. J. Soft Comput. Eng.*, vol. 1, no. June 2011, pp. 35–38, 2011.
[7] Solon Angel, “How AI Could Protect Your Business From Financial Fraud,” 2019. [Online]. Available: https://www.forbes.com/sites/forbesfinancecouncil/2019/02/14/how-ai-could-protect-your-business-from-financial-fraud/#22d87b776116. [Accessed: 22-Feb-2019].
[8] H. I. Bülbül and Ö. Ünsal, “Comparison of Classification Techniques Used in Machine Learning as Applied on Vocational Guidance Data,” *Proc. - 10th Int. Conf. Mach. Learn. Appl. ICMLA 2011*, vol. 2, pp. 298–301, 2011.
[9] D. Neil, “Deep Neural Networks and Hardware Systems for Event-driven Data,” 2017.
[10] K. Randhawa, C. H. U. K. Loo, M. Seera, C. P. Lim, and A. K. Nandi, “Credit Card Fraud Detection Using AdaBoost and Majority Voting,” *Digit. Object Identifier*, pp. 14277–14284, 2018.
[11] S. Dhankhad, E. A. Mohammed, and B. Far, “Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection : A Comparative Study,” *2018 IEEE Int. Conf. Inf. Reuse Integr.*, pp. 122–125, 2018.
[12] S. B. S, V. G. Biju, and C. M. Prashanth., “Kappa and Accuracy Evaluations of Machine Learning Classifiers,” *2017 2nd IEEE Int. Conf. Recent Trends Electron. Inf. Commun. Technol.*, pp. 20–23, 2017.
[13] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[14] B. K. Bhavitha, A. P. Rodrigues, and N. N. Chiplunkar, “Comparative study of machine learning techniques in sentimental analysis,” *Proc. Int. Conf. Inven. Commun. Comput. Technol. ICICCT 2017*, no. Icicct, pp. 216–221, 2017.
[15] M. K. Mishra and R. Dash, “A comparative study of chebyshev functional link artificial neural network, multi-layer perceptron and decision tree for credit card fraud detection,” *Proc. - 2014 13th Int. Conf. Inf. Technol. ICIT 2014*, vol. 228, no. August 2013, pp. 228–233, 2014.
[16] Nonki Takahashi, “Neural Network Extension – Small Basic Featured Thread – Small Basic,” 2014. [Online]. Available: https://blogs.msdn.microsoft.com/smallbasic/2014/11/06/neural-network-extension-small-basic-featured-thread/. [Accessed: 22-Feb-2019].
[17] G. Rushin, C. Stancil, M. Sun, S. Adams, and P. Beling, “Horse race analysis in credit card fraud - Deep learning, logistic regression, and Gradient Boosted Tree,” *2017 Syst. Inf. Eng. Des. Symp. SIEDS 2017*, pp. 117–121, 2017.
[18] J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare, “Credit card fraud detection using machine learning techniques: A comparative analysis,” *Proc. IEEE Int. Conf. Comput. Netw. Informatics, ICCNI 2017*, vol. 2017–Janua, pp. 1–9, 2017.
[19] W. C. Chung, C. Y. Chang, and C. C. Ko, “A SVM-Based committee machine for prediction of Hong Kong horse racing,” *Ubi-Media 2017 - Proc. 10th Int. Conf. Ubi-Media Comput. Work. with 4th Int. Work. Adv. E-Learning 1st Int. Work. Multimed. IoT Networks, Syst. Appl.*, no. 1, pp. 3–6, 2017.
[20] S. Ekiz and P. Erdogmus, “Comparative study of heart disease classification,” *2017 Electr. Electron. Comput. Sci. Biomed. Eng. Meet. EBBT 2017*, pp. 1–4, 2017.
[21] Lili Jiang, “What are kernels in machine learning and SVM and why do we need them?,” 2016. [Online]. Available: https://www.quora.com/What-are-kernels-in-machine-learning-and-SVM-and-why-do-we-need-them/answer/Lili-Jiang?srid=oOgT. [Accessed: 24-Feb-2019].
[22] E. Drabiková and E. F. Škrabul’Aková, “Decision trees - A powerful tool in mathematical and economic modeling,” *2017 18th Int. Carpathian Control Conf. ICCC 2017*, pp. 34–39, 2017.

[23] S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang, and C. Jiang, "Random forest for credit card fraud detection," *ICNSC 2018 - 15th IEEE Int. Conf. Networking, Sens. Control*, pp. 1–6, 2018.

[24] I. Ahmad, M. Basheri, M. J. Iqbal, and A. Rahim, "Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection," *IEEE Access*, vol. 6, pp. 33789–33795, 2018.

[25] M. K. Mishra and R. Dash, "A comparative study of chebyshev functional link artificial neural network, multi-layer perceptron and decision tree for credit card fraud detection," *Proc. - 2014 13th Int. Conf. Inf. Technol. ICIT 2014*, vol. 228, no. August 2013, pp. 228–233, 2014.

[26] S. Turgut, M. Dagtekin, and T. Ensari, "Microarray breast cancer data classification using machine learning methods," *2018 Electr. Electron. Comput. Sci. Biomed. Eng. Meet. EBBT 2018*, pp. 1–3, 2018.