Black-Box Adversarial Entry in Finance through Credit Card Fraud Detection

Akshay Agarwal, Nalini Ratha

University at Buffalo, USA

Abstract

In the literature, it is well explored that machine learning algorithms trained on image classes are highly vulnerable against adversarial examples. However, very limited attention has been given to other sets of inputs such as speech, text, and tabular data. One such application where little work has been done towards adversarial examples generation is financial systems. Despite processing sensitive information such as credit fraud detection and default payment prediction, a low depiction of the robustness of the financial machine learning algorithms can be dangerous. One possible reason for such limited work is the challenge of crafting adversarial examples on the financial databases. The financial databases are heterogeneous where features might have a strong dependency on each other. Whereas image databases are homogeneous, and hence several existing works have shown it is easy to attack the classifiers trained on them. In this paper, for the first, we have analyzed the vulnerability of several traditional machine learning classifiers trained on financial tabular databases. To check the robustness of these classifiers, '*black-box and classifier agnostic*' adversarial attack is proposed through mathematical operations on the features. In brief, the proposed research for the first time presents a detailed analysis that reflects which classifier is robust against minute perturbation in the tabular features. Apart from that through the perturbation on individual features, it is shown which column feature is more or less sensitive for the incorrect classification of the classifier.

Keywords

Adversarial Attacks, Credit Card Fraud Detection, Machine Learning Classifiers, Vulnerability, Black-Box,

1. Introduction

The recent research articles claim that in the last decade from 2010 to 2020, people with the personal loan double from $11 million to $21 million [3]. At the same time, the amount of loan debt increase by three times from $55 billion to $162 billion. The processing of such a large number of loan applications and identifying any possible fraud is a tedious and time-consuming task for a human being. The possible solution to overcome the load is to utilize the power of machine learning (ML) algorithms. In the past, machine learning algorithms have shown tremendous success in solving variety of tasks ranging from object recognition [4, 5] to person identification [6, 7, 8] and solving complex medical problems [9, 10, 11]. While the machine algorithms are here to ease the human and perform the task with near perfection. However, recent research indicates that the machine learning algorithms are highly susceptible against the minute perturbation in the input data.

Imagine a scenario, where a corrupt individual came into the bank for credit card approval and the issue of a lump sum amount of money. Due to the processing of multiple applications which might be in huge numbers and the time required for handling an application, machine learning algorithms are ideal for decision making. Machine learning generally requires a significant number of feature components, which in the case of credit application can be such as age, property available, and amount paid in the previous loan if any. The corrupt personal can minutely change one or more feature components which can easily be ignored in the application by the system due to no drastic change in the feature space and hence can accept the fraud application. Apart from difficultly observing the feature space, the heterogeneous nature of the tabular databases requires expert opinion in identifying small modifications. The severity of the credit fraud or loan fraud can be seen from the recent news articles [1, 2, 12]. As per the statistics, in 2018, $24.26 Billion was lost due to payment card fraud worldwide. Among all the amount, the United Stated is one of the largest contributors with almost 38.6% reported credit card fraud cases. The sensitivity of machine learning algorithms towards minute perturbations in other domains [13] requires that ML algorithms used for tabular databases are secure to ensures the correct decision. Figure 1 shows the impact of credit card fraud in the worldwide community. Therefore, it is extremely important to extensively examine the vulnerability of machine learning algorithms before trusting their decision in the financial domain.

In this research, for the first time, we have extensively evaluated several machine learning models and their vul-

In International Workshop on Modelling Uncertainty in the Financial World (MUFin21) In conjunction with CIKM 2021,
November 1, 2021, Online
aa298@buffalo.edu (A. Agarwal); nratha@buffalo.edu (N. Ratha)
https://sites.google.com/iiitd.ac.in/agarwalakshay/home (A. Agarwal); https://nalini-ratha.github.io/ (N. Ratha)
0000-0001-7362-4752 (A. Agarwal); 0000-0001-7913-5722 (N. Ratha)

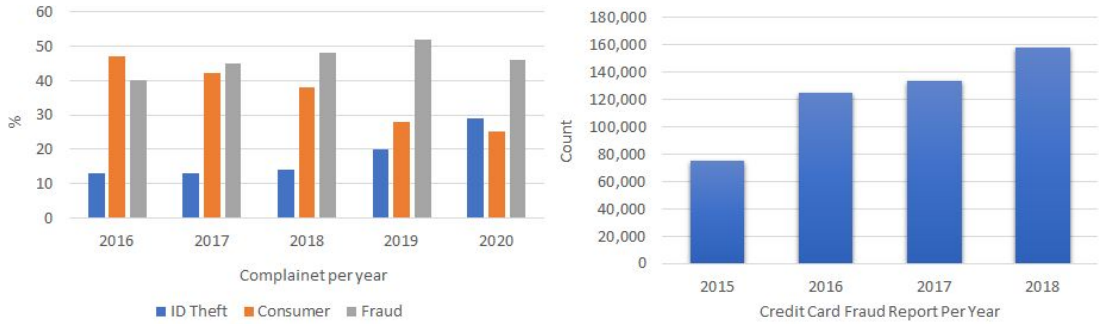

Figure 1: Reflecting the impact of credit card fraud worldwide and demand of secure deployment of automated machine learning (ML) systems. The statistics are taken from the multiple Internet sources [1, 2].

nerability against minute perturbations in the feature space (or input space). The credit card default prediction databases contain multiple features such as age, gender, payment status, and education. The individual feature can affect the classification decision due to any reason such as bias and mislabelling. For example, it might be believed that the highly educated individual might not perform fraud. Therefore, in this research, we have identified the sensitivity of various machine learning classifiers against individual features both in their raw form and under minute perturbation. While the features play an important role, the optimization function of different classifiers play an important role in learning decision boundaries. Hence, the detailed experimental evaluation of multiple machine learning classifiers has been performed to showcase which classifier is more robust or sensitive against imperceptible perturbations. In brief, the contributions of this research are:

- first-ever black-box inference time imperceptible adversarial attack on credit-card default prediction is performed;
- extensive ablation studies are conducted to find out the importance of individual feature value towards decision making;
- sensitivity analysis of multiple machine learning classifiers are presented to help in building a robust finance system utilizing robust classifier(s);
- comprehensive survey of the existing adversarial attacks developed in other data domains showcase the needs of the development of adversary to identify the vulnerabilities in the finance space as well.

In the next section, the review of the existing adversarial examples is presented followed by the description of credit card default prediction databases. In the next, exploratory database analysis has been performed to effectively examine the characteristics of the databases. Later, the machine learning classifiers chosen to perform the vulnerability analysis are described. The experimental results along with analysis are presented to showcase the impact of the proposed 'black-box and classifier agnostic' adversarial perturbation.

2. Existing Adversarial Examples Research

Since the finding of adversarial examples [14], several adversarial attack algorithms are presented in the literature. The existing adversarial attacks can be divided based on the following two criteria: (i) intention and (ii) type of learning. The type of learning can be described how much knowledge of the machine learning classifier is needed to fool it and it can be categorized into white-box and black-box. In the white-box setting, an attacker assumes the complete knowledge of the system such as its parameters and classification probabilities. On the other hand, the black-box attacks do not utilize any ML network information in creating adversarial examples. In the real world, it is extremely difficult to acquire the knowledge of the machine learning classifiers due to their security and the existence of a wide variety of machine learning algorithms. For example, Goel et al. [15, 16, 17] have utilized the concept of blockchain and cryptography to either change the structure of the networks or encrypt them to make it difficult to identify the exact parameters of the networks. Similarly, there exists a humongous number of machine learning algorithms such as supervised, unsupervised, and ensemble learning, hence, assuming the knowledge of the system an attacker wants to fool is difficult [18, 19]. Due to the above observations, a black-box attack is practical in the real world and at the same difficult to achieve. On the other hand, intention-based attacks are divided into targeted attacks and untargeted attacks. The targeted attacks aim the

input data to be misclassified by the network into one of the desired classes. For example, a credit card defaulter would like to be classified as genuine by the machine learning classifier. Whereas, the untargeted attacks aim the input data to be misclassified into 'any' class except the true class.

In the literature, several adversarial attacks are proposed. The majority of the attacks are proposed for visual object classification and limited work has been done so far for other kinds of input information such as speech and tabular data, and machine learning classifiers such as reinforcement learning. The gradient is one of the most essential information in deep network learning and utilizing this information several attacks are proposed. For example, PGD attack [20] is one of the strongest attacks for visual image classifiers. The attack is performed in multiple iterations by projecting the gradient in the direction that leads to the strong adversary. Other image-based attacks such as DeepFool [21], add the perturbation in the image iteratively so that the image can pass its corresponding class decision boundary learned by the network. The above attacks learn the manipulation for each image separately, while it is possible to learn a unique noise vector to apply on multiple images and fool the network [22, 23]. The above-described attacks are performed in the white-box setting utilizing the complete knowledge of the classifier. Another disadvantage of the white-box attack is the transferability against multiple models. As the attacks are generated utilizing the knowledge of the classifier which can be significantly different from the other unseen models, hence, leads to a poor success rate against unseen models [24, 25].

The other class of attack is the black-box attacks which are more practical in the real world and can fool multiple classifiers. The black-box attack can be further divided into query-based and generic manipulation-based. In the query-based attack, some knowledge of the system is assumed such as the decision of the classifier. By utilizing the decision of the classifier on the given input, the noise is modified leading to the desired intent of misclassification, i.e., targeted or untargeted. While the query-based attacks are more successful for unseen models whose knowledge is not available but still bounded by the number of queries that can be sent for the noise generation. Therefore, this limitation restricts the practical deployment at multiple places. Another category of attack which is general manipulation is one of the most successful attacks because not utilization of any classifier knowledge makes them agnostic to classifiers and can fool multiple classifiers. Goswami et al. [26, 27] have proposed several image manipulations for fooling face recognition networks. The manipulations are somewhat inspired by the domain knowledge of face recognition and therefore, modified the landmark features of a face image which were able to fool the recognition networks effectively. Agarwal et al. [28] have not utilized any external knowledge including perturbation vector but extract the noise inherently present in an image. The authors use an intelligent observation that due to several factors such as camera preprocessing steps, environmental factors, the noise inherently present in an image. The authors extract those noise pattern and used as an adversarial pattern. The above-mentioned attacks are performed in the image space. Limited attacks are also proposed in other categories of networks or input such as generative models [29], reinforcement learning [30], and cyberspace [31].

Table 1
Characteristics of the Credit Card databases.

Default Credit			Australian Credit	
Feature	Name	Type	Feature	Type
1	ID	Continuous	A1	Binary
2	Limit-Bal	Continuous	A2	Continuous
3	Sex	Binary	A3	Continuous
4	Education	Categorical	A4	Categorical
5	Marriage	Categorical	A5	Categorical
6	Age	Continuous	A6	Categorical
7	Pay_0	Continuous	A7	Continuous
8	Pay_2	Continuous	A8	Binary
9	Pay_3	Continuous	A9	Binary
10	Pay_4	Continuous	A10	Continuous
11	Pay_5	Continuous	A11	Binary
12	Pay_6	Continuous	A12	Categorical
13	Bill_Amt1	Continuous	A13	Continuous
14	Bill_Amt2	Continuous	A14	Continuous
15	Bill_Amt3	Continuous	A15	Binary
16	Bill_Amt4	Continuous		
17	Bill_Amt5	Continuous		
18	Bill_Amt6	Continuous		
19	Pay_Amt1	Continuous		
20	Pay_Amt2	Continuous		
21	Pay_Amt3	Continuous		
22	Pay_Amt4	Continuous		
23	Pay_Amt5	Continuous		
24	Pay_Amt6	Continuous		
25	Default Payment	Binary		

While on the one hand, adversarial attacks on machine learning classifiers especially deep learning classifiers are prevalent, the defense against them is also getting significant attention. Several defense algorithms based on the following two motives are proposed: (i) segregation of the adversarial examples from the clean examples [32] and (ii) mitigating the impact of adversarial noise [27]. The defense algorithms have shown tremendous success in countering the adversarial attacks on the image domain and show generalizability even in complex situations such as an unseen attack, unseen database, and unseen model [33, 34]. The survey of the existing research on adversarial examples can be further referred from the survey papers [35, 36, 37].

It is interesting to observe from the above discussion that adversarial machine learning is one of the fasted growing communities; however, only a few works exist towards the robustness in the financial domain. The prime reason for such low existence can bethink from

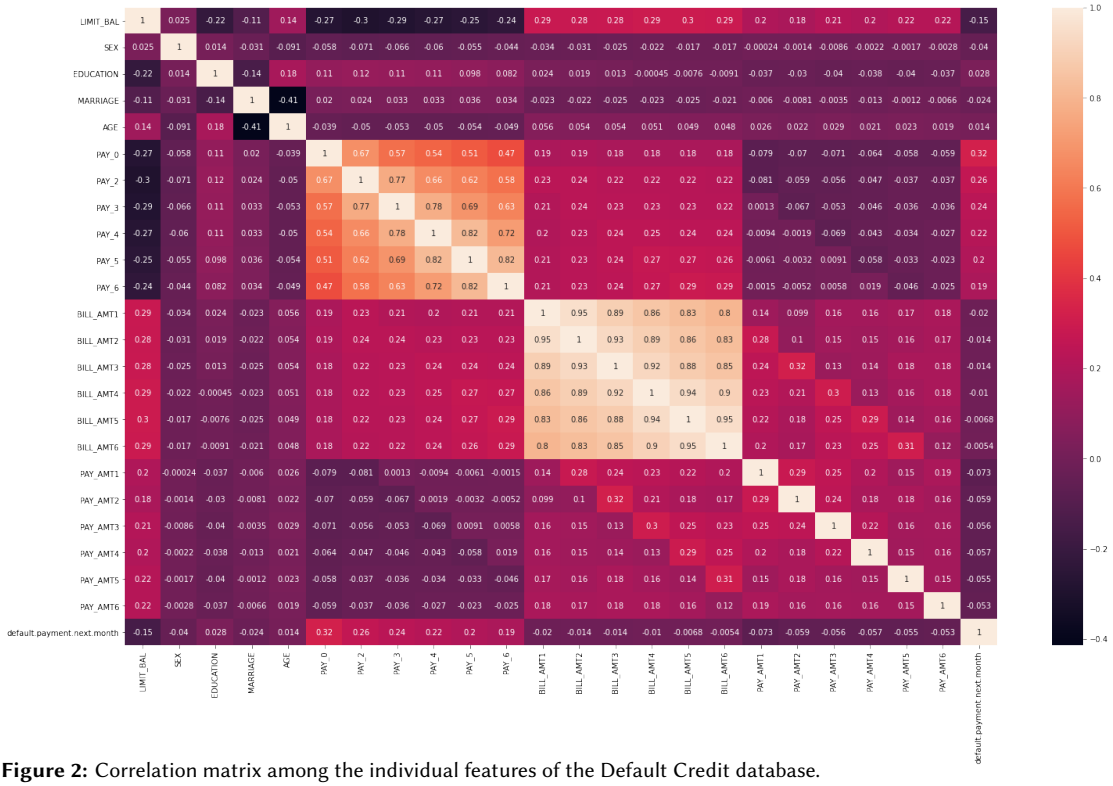

Figure 2: Correlation matrix among the individual features of the Default Credit database.

the point of the type of input. The financial data especially tabular databases are heterogeneous as compared to homogeneous image databases. Tabular features are not interchangeable in contrast to the pixels of an image. Apart from that, the images are rich in visual information and hence humans can predict the information by looking at them and easy to identify whether any manipulation has been performed. Whereas tabular data are less interpretative and it is complex to identify the minor modification in individual value. In the literature, few research works are proposed for crafting adversarial noise on tabular databases. Ballet et al. [38] and Levy et al. [39] have proposed an imperceptible adversarial attack by minimizing the norm of the perturbation. The critical drawback of the attacks is that the attacker assumes the complete knowledge of the classifier for learning the perturbation and hence less practical for real-world deployment. Another drawback is that the norm-based perturbation on the tabular features can yield unrealistic transformations [40]. Apart from that, the above attacks on tabular data are evaluated on a single classifier, i.e., a shallow neural network or decision forest.

To overcome the limitations of the existing adversarial study on the tabular databases, we have proposed an adversarial manipulation method based on mathematical operators. The proposed attacks work in the black-box setting and do not utilize any information of a classifier. Therefore, the proposed attack is classifier agnostic and can be applied against '**any**' classifier. In contrast to the existing research reported on the limited classifier, the proposed research study the adversarial strength against multiple classifiers and shows that the proposed attack can fool each of them. Apart from this, the proposed manipulation also aims to reveal the role of individual tabular features in the classification.

3. Finance Databases

In this research, we have used two popular credit card default prediction databases namely Default Credit Database [41, 42] ad Australian Credit Database [12]. The default credit database is one of the largest databases for the binary prediction of the default payment category. The database contains 30, 000 data points belonging to two categories of default payment, i.e., yes or no. In total, the database consists of 24 features belonging to multiple types such as binary (0 or 1), categorical (1 to n), and continuous. The ID is a feature to represent an individual in the database and hence no role in the classification of the data point. Therefore, the ID feature is

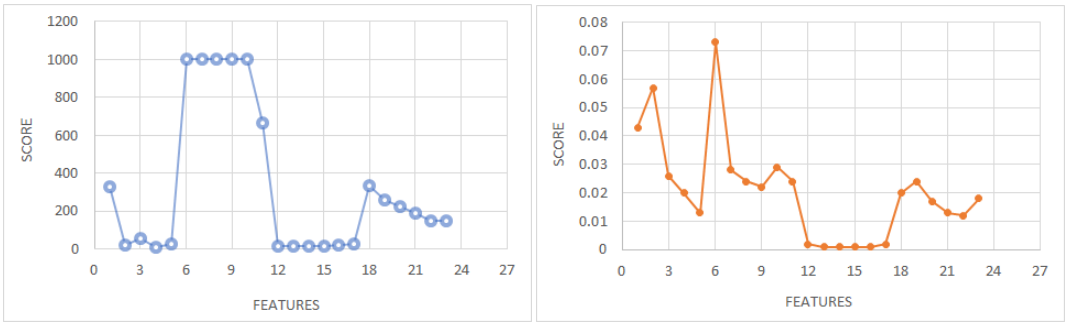

Figure 3: Score of an individual feature computed using UFR (left) and MRMR (right) algorithm on the Default Credit Card database.

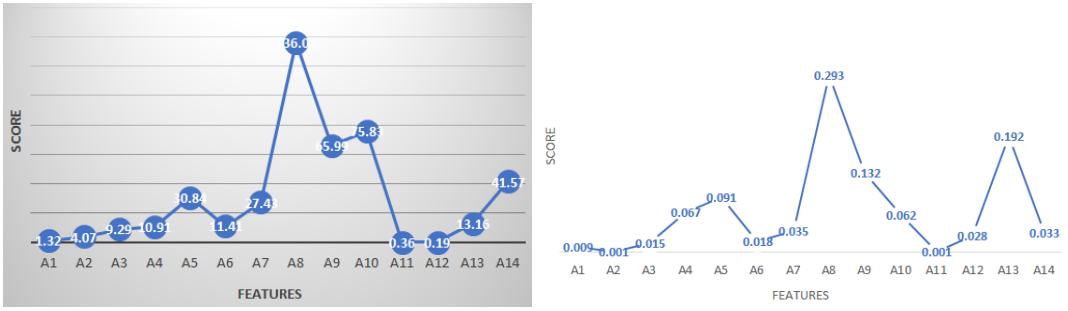

Figure 4: Score of an individual feature computed using UFR (left) and MRMR (right) algorithm on the Australian Credit Card database.

discarded from the Default Credit database. It is clear from the description that each feature has a different scale and hence, it is important to bring each feature into the same range, such as between 0 to 1. We have performed the min-max normalization to bring the scale of each feature to the same range. The Australian Credit database contains 14 features aiming to classify the data into binary categories of default payment. Similar to the Default Credit database, the Australian database consists of the features of different scales and hence normalized using min-max scaling. The characteristics of both the databases are given in Table 1. Contrary to few available pieces of research [38] which drops few features for adversarial learning on credit database, we have utilized each feature in the database and analyze their impact on adversary generation.

4. Exploratory Data Analysis

Before performing the adversarial attack on the input features of the Credit Card databases, we have performed the exploration studies on the features such as correlation among the features and relevance of the features.

4.1. Correlation Analysis

Figure 2 shows the correlation heatmap among each feature in the Default Credit Card database. It is clear from the heatmap that, no feature exhibits a strong correlation with the class variable (default payment). Whereas, the features that belong to the same category such as 'Pay_' and 'BILL_AMT' show a strong correlation among themselves. For example, 'Pay_0' have the positive correlation value of 0.67 with variable 'Pay_1'. 'Pay_0' feature represents the repayment status in September 2005 and the value of the feature ranges between -1 to 9. Other pay features represent the repayment status between April to August 2005. The correlation among them shows the repayment status of the current month and in turn, the credit default payment is somewhat dependent on the status of the last month. However, as compared to repayment status, 'BILL_AMT' features have very strong correlation values among themselves. The correlation value of at least 0.8 is observed between different features. 'BILL_AMT' represents the amount of bill statement between April 2005 to September 2005.

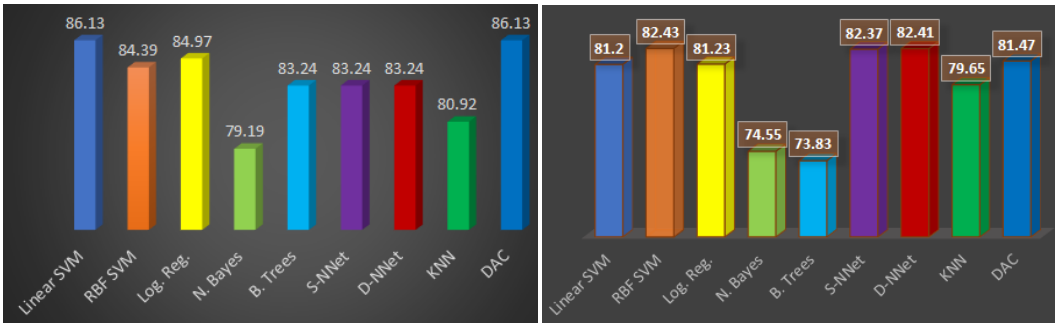

Figure 5: Credit Card default prediction accuracy on clean test set of Australian Credit Card (left) and Default Credit Card (right) database. Log. Reg. represents the logistic regression and B. Trees represents the binary trees. S-NNet and D-NNet represent the shallow and deep neural networks, respectively.

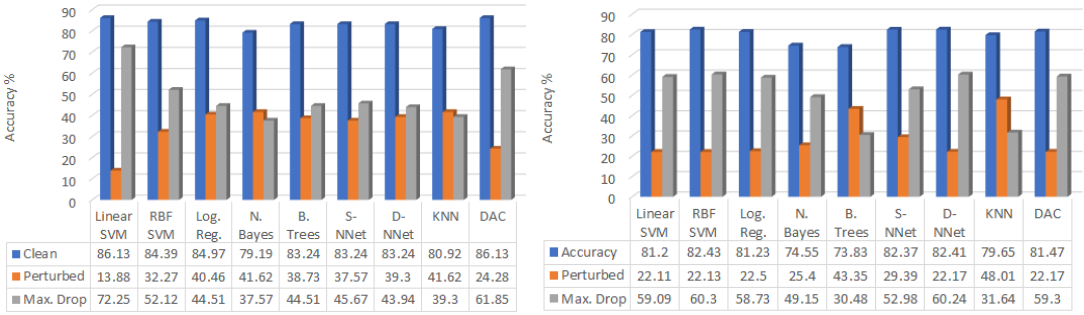

Figure 6: Credit Card default prediction accuracy on the clean test set, perturbed set of reflecting maximum drop, and the difference in accuracy using Australian Credit Card (left) and Default Credit Card (right) database. Log. Reg. represents the logistic regression and B. Trees represents the binary trees. S-NNet and D-NNet represent the shallow and deep neural networks, respectively. Clean shows the accuracy of the classifier on the original clean test set, whereas, perturbed shows the accuracy when any of the features is perturbed which leads to a maximum drop in the accuracy on the test set. Max. Drop is the difference between clean and perturbed set accuracy.

4.2. Feature Importance

Another data exploratory analysis has been performed by examining the importance of individual features concerning the class label. For that two feature selection or feature weight assignment algorithms namely Univariate Feature Ranking (UFR) and Minimum Redundancy Maximum Relevance (MRMR) [43], are utilized. The advantage of both the algorithm is that they accept categorical and continuous features for the classification problem. The UFR algorithm measures the independence of each feature concerning the class variable using the chi-square test between them. The smaller the p-value on a particular feature represents the higher the dependence between the feature and class label and the importance of the feature for classification. MRMR algorithm iteratively examines the features to find the features which are mutually and maximally dissimilar to each other but effective for decision making. The algorithm achieves its goal of selecting the important features by reducing the redundancy among the features and weighting the relevant features. For that, the MRMR algorithm computes the mutual information among the features and between feature and class label. The MRMR algorithm selects the best feature set (S) for classification by maximizing the relevance score $|V_s|$ between feature x and class label y. At the same time, the algorithm aims to minimize the redundancy score $|W_S|$ between two feature values x and z. The $|V_S|$ and $|W_S|$ can be defined using the following equations:

$$V_S = \frac{1}{|S|} \sum_{x \in S} I(x, y)$$

$$W_S = \frac{1}{|S|^2} \sum_{x \in S} I(x, z)$$

where, $|S|$ represents the number of features in the optimal subset S. Finally, mutual information quotient (MIQ)

Table 2
Adversarial vulnerability of multiple machine learning classifiers against the proposed perturbation defined in Equation 1 on Australian Credit Card Database. Colored box represents the sensitive features and drop in accuracy of classifier on the corresponding feature.

Perturb Feature	SVM		Logistic Regression	Naive Bayes	Binary Trees	Neural Network		KNN	DAC
	Linear	RBF				Shallow	Deep		
1	86.13	85.55	84.39	79.77	84.97	83.81	87.28	82.08	86.13
2	86.13	82.66	84.39	77.46	83.81	77.46	81.50	78.61	86.13
3	86.13	84.39	83.81	71.67	79.77	80.35	84.97	80.35	84.97
4	86.13	85.55	64.16	74.57	41.04	43.93	63.00	82.10	80.92
5	86.13	86.13	70.52	83.81	79.77	72.25	71.10	78.61	82.08
6	86.13	86.13	84.97	83.81	82.10	78.61	86.13	72.83	85.55
7	86.13	84.97	84.39	41.62	84.39	75.14	83.24	50.87	85.55
8	13.88	32.27	40.46	67.63	38.73	37.57	39.30	44.51	24.28
9	86.13	84.39	85.55	79.19	76.88	83.24	83.24	80.35	86.70
10	86.13	85.55	41.62	41.62	83.23	63.58	87.28	41.62	50.29
11	86.13	87.28	86.70	78.61	83.23	85.55	82.66	82.66	85.29
12	86.13	85.55	84.39	63.00	83.23	70.52	83.81	58.96	86.13
13	86.13	83.24	80.92	79.19	79.19	78.61	82.66	77.46	84.97
14	86.13	84.97	41.62	41.62	84.39	50.29	86.13	41.62	41.62

is calculated to select the subset of features using the following equation:

$$MIQ_x = \frac{V_x}{W_x}$$

where, V_x and W_x are the relevance and redundancy value of a feature x, respectively.

The earlier adversarial studies discarded few features and hence do not provide the complete picture on the credit card domain. We want to highlight that the proposed research is the first work explaining detailed analysis helpful both crafting the attack and mitigating it by protecting the important features. Figure 3 and Figure 4 show the score plot of the features from the Default Credit and Australian Credit Card database, respectively. On the Default Credit database, feature 6 (i.e., Age as shown in Table 1) shows the highest importance irrespective of the feature selection algorithm. Feature 8 is found most relevant in the Australian database using both UFR and MRMR feature selection algorithms.

5. Vulnerable Machine Learning Algorithms

In this research, we have used several machine learning algorithms to carefully investigate the impact of adversarial manipulation on the feature space of Credit Card databases. To present a first-ever detailed study, in total, nine different classifiers are used for extensively investigating the adversarial fraud in the finance domain. Further, we describe each of the algorithms used for binary classification on clean and manipulated features:

1. Support vector machine (SVM) [44]: It is one of the most popular machine learning classifier because of strong mathematical foundation. SVM learns the decision hyperplanes by maximizing the distance between the nearest points of each class. SVM classifier have significant success in the binary classification tasks such as presentation attack detection [45, 46] and adversarial examples detection [34, 33]. Based on its success on binary classification tasks, SVM is an ideal choice to be used for credit card default prediction as well. SVM works on the following optimization function:

$$minimize \|w\| \;\; such \;\; that \;\; y_i(W^T x_i - b) \geq 1 \quad for\, i \in 1, ..., n$$

where, w is the vector on the separating hyperplane and b is the bias term. x_i and y_i are the i^{th} data point and label, respectively. n represents the total training data points. Upon solving the above equation, the classifier obtained is: $X \rightarrow sign(w^T x + b)$. In this research, we have used two variants of SVM referring to basically two kernels namely 'linear' and 'radial basis function' (RBF) used for learning the separating hyperplane.
2. Logistic Regression [47]: It is another simple and popular classifier for the task of binary classification problem. It uses the logistic function to model the probabilities of the binary classes. The class of the input is predicted by taking the maximum of the probabilities of the classes output by the model. The probability of each class can be

computed using the following logistic formula:

$$\pi(X) = \frac{\exp \beta_0 + \beta_1 x_1 + ... + \beta_k x_k}{1 + \exp \beta_0 + \beta_1 x_1 + ... + \beta_k x_k}$$

where, βs are the parameters of the classifier and xs are the feature values.

3. Naive Bayes Classifier: It is based on the popular Bayes theorem which can be written as follows:

$$P(B/A) = \frac{P(A/B)P(B)}{P(A)}$$

where A and B are two independent events and $P(A) \neq 0$. It assumes the features are independent from each other and observation follows multivariate distribution.

4. Binary Trees: It works on segregating the classes based on the features at each level of the tree. At each level, data is partitioned into classes and the features for level are selected based on the impurity function such as Gini impurity. The classifier works iteratively and stops when either all the data points are exhausted or leaf nodes arrive. As the name suggested, at each level of the tree, only two nodes are allowed at most.

5. Neural Network [48]: It is another most successful machine-learning architecture that works on mapping the input features to the output classes through multiple layers in between. The intermediate layers are popularly referred to as hidden layers and they can vary from 1 to any number. In this research, we use shallow NN (S-NNet) with one hidden layer and deep NN (D-NNet) with 2 hidden layers each having the neurons equal to half of the size of the neurons in the previous layer. Both NNs are trained using a stochastic gradient descent algorithm.

6. K Nearest Neighbor (KNN) [49]: It is the simplest and training-free classifier that works on the measurement of the distance between test points and training data points. The test point is classified into the class from which it has the lowest distance. In this research, we have used the $K = 5$ nearest data points to find the closes class using the 'Euclidean' distance.

7. Discriminant Analysis Classifier (DAC): It is based on the assumption that the data points of different classes contain different parameters of the Gaussian distribution. For classification, the Gaussian parameters of each class are found out using the training set. To identify the class, the posterior probability of point belonging to each class is calculated as follows:

$$P(x|k) = \frac{1}{((2\pi)^d |\Sigma_k|)^{1/2}} exp(\frac{-1}{2}(x - \mu_k)|\Sigma_k|^{-1} (x - \mu_k)^T)$$

where, $P(x, k)$ is the probability of the point x belonging to class k. μ_k and Σ_k are the Gaussian parameters of class k.

6. Proposed Adversarial Attack and Experimental Results

In this research, we have studies the impact of manipulation on individual features for credit default prediction. We have applied several mathematical operations to obtain the manipulated features. Broadly, the proposed classifier agnostic and black-box attack can be defined using the following equation:

$$x_{pert} = \{XOR(x_{clean}, 1), \quad if \ x_{clean} \ is \ binary \\ x_{clean} + \eta \quad if \ x_{clean} \ is \ continuous\} \tag{1}$$

where, x_{pert} is the perturbed variant of the clean feature x_{clean}. η is defined using several function such as: $\eta = C$, where C is a constant value between 0 to 1. Other mathematical function which are explored for η are $exp(max(x_{clean}) - min(x_{clean})$ and $log(max(x_{clean}) - min(x_{clean})$. Apart from that, another simple mathematical attack on the feature value termed as 'feature dropout' here can be defined as:

$$x_{pert} = x_{clean} * 0$$

In this section, we describe the adversarial manipulation results and analysis using the constant value modification as the attack. The databases are divided into training and testing, where the training set contains randomly selected 75% of the total data point. The remaining 25% data points are used for the evaluation of each of the classifiers trained on the training set.

The analysis of the results can be divided into the following parts: ❶ accuracy on the clean images, ❷ robustness of a classifier, and ❸ sensitive features for adversarial goal. The credit card default payment results of each classifier on a clean test of both the databases are reported in Figure 5. On the Australian database, as compared to the non-linear classifiers such as RBF SVM and Neural Network, the linear classifier such as linear SVM performs better. Whereas, on the default credit card database, the RBF SVM performs best as compared to other linear and non-linear classifiers. It is interesting to note from going shallow to a deep neural network, no significant improvement in accuracy notices on both the databases. In another observation, the Naive Bayes classifier performs the worst on the Australian credit card database and second-worst on the default credit card database.

Analysis concerning classifiers: In terms of the sensitivity of the classifier, it is found that the SVM classifier is the least robust in terms of the magnitude of the accuracy

Table 3
Adversarial vulnerability of multiple ML classifiers against the proposed perturbation defined in Equation 1 on Default Credit Card Database. Colored box represents the sensitive features and drop in accuracy of classifier on the corresponding feature.

Perturb Feature	SVM		Logistic Regression	Naive Bayes	Binary Trees	Neural Network		KNN	DAC
	Linear	RBF				Shallow	Deep		
1	81.19	78.45	78.55	80.10	69.25	78.60	80.30	78.69	78.95
2	81.19	81.16	81.81	64.17	73.13	81.58	82.47	79.29	81.92
3	81.19	78.76	78.61	81.10	69.27	79.48	81.68	78.43	79.04
4	81.19	79.88	79.13	75.76	71.61	81.44	81.89	80.33	79.47
5	81.19	80.21	82.32	34.41	66.40	77.79	81.35	77.57	81.81
6	22.11	22.13	22.50	25.40	43.35	77.15	22.17	48.01	22.17
7	41.76	77.91	81.36	25.96	61.28	81.63	32.99	57.56	79.05
8	81.19	80.10	82.40	26.10	72.10	54.41	31.95	74.22	82.07
9	81.19	80.00	81.89	25.77	71.60	29.39	75.21	74.36	81.99
10	81.19	82.69	82.17	25.37	64.24	79.79	80.81	70.03	82.28
11	81.19	81.61	81.35	25.75	69.63	74.36	80.91	72.25	81.48
12	81.19	81.65	77.89	38.68	60.20	78.67	81.31	77.91	77.88
13	81.19	82.15	30.83	33.40	68.25	42.83	40.40	61.17	71.67
14	81.19	81.61	25.37	81.10	67.71	42.97	27.89	77.88	53.25
15	81.19	82.12	79.28	42.78	67.93	80.61	78.15	78.17	79.19
16	81.19	82.10	80.84	30.78	71.11	78.17	79.09	78.91	80.05
17	81.19	81.93	81.41	63.55	66.99	78.85	30.00	79.80	81.41
18	81.19	78.35	77.89	77.89	71.35	77.89	77.91	77.89	77.89
19	81.19	79.10	77.89	77.89	71.30	76.23	80.36	77.89	77.90
20	81.19	80.55	77.91	77.89	73.13	71.79	56.09	77.89	80.31
21	81.19	78.61	78.11	77.89	69.38	82.33	78.11	77.89	78.83
22	81.19	78.65	77.89	77.89	69.23	77.95	78.35	77.91	78.59
23	81.19	78.48	77.96	77.89	68.47	77.91	79.85	77.89	79.33

drop on both Australian credit card and default credit card databases. On the Australian database, the accuracy of the linear SVM drops from 86.13% to 13.88%. The relative drop in the accuracy is 72.25% which is the highest among all the classifiers used for credit default prediction. On the other hand, the Naive Bayes classifier which performs the worst on the Australian database shows the least drop in accuracy when the features are perturbed using the proposed black-box and model agnostic attack. In other words, the Naive Bayes classifier is found most effective in handling the perturbation. The accuracy of each classifier on the clean images, least accuracy obtained under perturbation, and difference reflecting a maximum drop in the accuracy is reported in Figure 6 (left) on the Australian database. On the default credit card database, the non-linear RBF classifier found the highest vulnerable and the relative drop in the performance is went to 60.3%. KNN classifier is found most robust in terms of the relative drop in the performance when the features are compared as compared to the accuracy on the clean features. It is interesting to note that, even on the Australian database, KNN shows the second-best robustness on the perturbed features. Figure 6 (right) shows the maximum sensitivity of each classifier on the default credit card database.

Analysis concerning features: The default payment database contains 23 features by removing the ID feature which is simply a sequence reflecting the observation number and class variable, i.e., default payment. Whereas, the Australian database contains 14 features for classification. We want to mention that in this research, we have shown the adversarial strength by perturbing a single feature only. On the Australian database, feature 8 is found most sensitive feature, and perturbing that feature affects the performance of each classifier significantly. Apart from affecting the performance of each classifier, feature 8 shows the highest reduction in the accuracy of each classifier. Feature 8 contains the binary values and we have modified the binary values through the 'XOR' operator as shown in the proposed attack equation 1. The second worst feature is the feature 14 which contains the continuous values. However, interesting both linear and non-linear SVM, binary trees, and deep neural networks are found robust against the slight modification on it.

On the default credit card database, feature 6 found the weakest point of each classifier except for shallow neural network (S-NNet). The perturbation of the feature 6 significantly dropped the accuracy of the affected classifiers. The RBF SVM classifier is found sensitive

against feature 6 only. We want to highlight that both the feature selection algorithms give the highest score to the features 6 as shown in Figure 3 on the default credit card database. Similarly, on the Australian database, each classifier has been found highly sensitive to the highest relevance features reported by the feature selection algorithms as shown in Figure 4. The detailed analysis on the sensitivity of individual features is given in Tables 2 and 3.

Other Manipulations: We want to highlight that the other mathematical operations such as exp and log mentioned in Section 6 yield similar adversarial phenomena are observed on each classifier.

6.1. Unwanted Phenomena for Attacker

It is interesting to observe that the adversarial perturbation does not always reduce the performance of a classifier. Apart from that, another interesting point is that the features which are least important for classification, perturbing them can inversely affect the goal of an attacker. The importance of the features can be calculated using the feature selection algorithm. For example, on the Australian database, the feature 11 was found least relevant by both UFR and MRMR feature selection algorithm. Interestingly, perturbing this feature significantly improves the performance of multiple classifiers. For example, the performance of the RBF SVM, logistic regression, and shallow neural network (S-NNet) improves by 2.89%, 1.73%, and 2.31%, respectively. Similarly, the features which are found less relevant by the feature selection algorithms on the default database, perturbing them shows the performance improvement. For example, features 1 and 14 are among the least important feature in the default payment database. However, perturbing them drastically increased the performance of the Naive Bayes classifier. The performance of Naive Bayes shows at least 5.55% jump in the classification performance when perturbing these features. From the above analysis, we suggest careful attention is required while perturbing a feature, a random perturbation of any feature set might not be fruitful for an attacker. Although further analysis can reveal future directions to improve the performance of a classifier by securing only the relevant features.

7. Conclusion

Adversarial vulnerability of the visual classifiers is extensively explored and paves the way for improving their robustness for secure real-world deployment. However, limited work has been done on financial databases especially tabular databases. The probable reason might be the heterogeneous nature of the databases and the low degree of freedom for perturbation. The degree of perturbation can be defined in the terms of the number of values available for manipulation. For example, an image contains a significantly large number of values (pixels) available for manipulation and is easily interchangeable. Whereas, the tabular finance databases contain a low number of features and can not be easily interchanged with each other. Few works exist to identify the vulnerability of ML algorithms on tabular databases. However, limitations of the existing attacks are that they require white-box access of the classifiers and result in unwanted transformations of the features. In this research, we have proposed a first-ever black-box attack on the tabular credit card default prediction databases. We have evaluated a broad number of machine learning classifiers as compared to a few classifier vulnerability assessments in the existing works. The proposed attack proves its classifier agnostic strength by fooling each classifier. Apart from the evaluation of multiple classifiers, we have also studied the sensitivity concerning individual features of the databases. Interestingly, it is observed that perturbation of every feature might hurt the aim of an attack, and therefore, intelligent consideration is required. We hope the proposed research opens multiple research threads both towards finding the vulnerabilities of tabular classifiers and improving their robustness.

References

[1] https://shiftprocessing.com/credit-card-fraud-statistics/, Credit card fraud statistics in the united stated, https://mk0shiftprocessor1gw.kinstacdn.com/wp-content/uploads/2019/10/CC-Fraud-reports-in-US-2-e1571769539315.jpg, 2020.

[2] Facts + statistics: Identity theft and cybercrime, https://www.iii.org/fact-statistic/facts-statistics-identity-theft-and-cybercrime, 2020.

[3] Personal loan statistics for 2020, https://www.fool.com/the-ascent/research/personal-loan-statistics/, 2020.

[4] S. Girish, S. R. Maiya, K. Gupta, H. Chen, L. S. Davis, A. Shrivastava, The lottery ticket hypothesis for object recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 762–771.

[5] M. Mandal, L. K. Kumar, M. S. Saran, et al., Motionrec: A unified deep framework for moving object recognition, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 2734–2743.

[6] M. Singh, S. Nagpal, M. Vatsa, R. Singh, Enhancing fine-grained classification for low resolution images, arXiv preprint arXiv:2105.00241 (2021).

[7] M. Singh, S. Nagpal, R. Singh, M. Vatsa, Derivenet for (very) low resolution image classification, IEEE Transactions on Pattern Analysis and Machine Intelligence (2021).

[8] S. Ghosh, R. Singh, M. Vatsa, Subclass heterogeneity aware loss for cross-spectral cross-resolution face recognition, IEEE Transactions on Biometrics, Behavior, and Identity Science 2 (2020) 245–256.

[9] I. Nigam, R. Keshari, M. Vatsa, R. Singh, K. Bowyer, Phacoemulsification cataract surgery affects the discriminative capacity of iris pattern recognition, Scientific reports 9 (2019) 1–9.

[10] Alphafold: a solution to a 50-year-old grand challenge in biology, https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-\challenge-in-biology, 2020.

[11] F. O. Geraldes, Pushing the boundaries of computer-aided diagnosis of melanoma, The Lancet Oncology 22 (2021) 433.

[12] Statlog (australian credit approval) data set, https://archive.ics.uci.edu/ml/datasets/Statlog+%28Australian+Credit+Approval%29, 2020.

[13] F. Pierazzi, F. Pendlebury, J. Cortellazzi, L. Cavallaro, Intriguing properties of adversarial ml attacks in the problem space, in: 2020 IEEE Symposium on Security and Privacy (SP), IEEE, 2020, pp. 1332–1349.

[14] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, arXiv preprint arXiv:1312.6199 (2013).

[15] A. Goel, A. Agarwal, M. Vatsa, R. Singh, N. Ratha, Deepring: Protecting deep neural network with blockchain, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 0–0.

[16] A. Goel, A. Agarwal, M. Vatsa, R. Singh, N. Ratha, Securing cnn model and biometric template using blockchain, in: 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS), IEEE, 2019, pp. 1–7.

[17] A. Goel, A. Agarwal, M. Vatsa, R. Singh, N. K. Ratha, Dndnet: Reconfiguring cnn for adversarial robustness, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 22–23.

[18] K. Das, R. N. Behera, A survey on machine learning: concept, algorithms and applications, International Journal of Innovative Research in Computer and Communication Engineering 5 (2017) 1301–1309.

[19] S. Pouyanfar, S. Sadiq, Y. Yan, H. Tian, Y. Tao, M. P. Reyes, M.-L. Shyu, S.-C. Chen, S. Iyengar, A survey on deep learning: Algorithms, techniques, and applications, ACM Computing Surveys (CSUR) 51 (2018) 1–36.

[20] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, arXiv preprint arXiv:1706.06083 (2017).

[21] S.-M. Moosavi-Dezfooli, A. Fawzi, P. Frossard, Deepfool: a simple and accurate method to fool deep neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2574–2582.

[22] K. R. Mopuri, A. Ganeshan, R. V. Babu, Generalizable data-free objective for crafting universal adversarial perturbations, IEEE transactions on pattern analysis and machine intelligence 41 (2018) 2452–2465.

[23] J. Hayes, G. Danezis, Learning universal adversarial perturbations with generative models, in: 2018 IEEE Security and Privacy Workshops (SPW), IEEE, 2018, pp. 43–49.

[24] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, A. L. Yuille, Improving transferability of adversarial examples with input diversity, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2730–2739.

[25] X. Wang, K. He, Enhancing the transferability of adversarial attacks through variance tuning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1924–1933.

[26] G. Goswami, N. Ratha, A. Agarwal, R. Singh, M. Vatsa, Unravelling robustness of deep learning based face recognition against adversarial attacks, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018.

[27] G. Goswami, A. Agarwal, N. Ratha, R. Singh, M. Vatsa, Detecting and mitigating adversarial perturbations for robust face recognition, International Journal of Computer Vision 127 (2019) 719–742.

[28] A. Agarwal, M. Vatsa, R. Singh, N. K. Ratha, Noise is inside me! generating adversarial perturbations with noise derived from natural filters, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 774–775.

[29] J. Kos, I. Fischer, D. Song, Adversarial examples for generative models, in: IEEE Security and Privacy Workshops, 2018, pp. 36–42.

[30] Y.-C. Lin, Z.-W. Hong, Y.-H. Liao, M.-L. Shih, M.-Y. Liu, M. Sun, Tactics of adversarial attack on deep reinforcement learning agents, in: International Joint Conference on Artificial Intelligence, 2017, pp. 3756–3762.

[31] I. Rosenberg, A. Shabtai, L. Rokach, Y. Elovici, Generic black-box end-to-end attack against state of the art api call based malware classifiers, in: International Symposium on Research in Attacks, In-

trusions, and Defenses, Springer, 2018, pp. 490–510.
[32] A. Agarwal, R. Singh, M. Vatsa, N. Ratha, Are image-agnostic universal adversarial perturbations for face recognition difficult to detect?, in: 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS), IEEE, 2018, pp. 1–7.
[33] A. Agarwal, R. Singh, M. Vatsa, N. K. Ratha, Image transformation based defense against adversarial perturbation on deep learning models, IEEE Transactions on Dependable and Secure Computing (2020).
[34] A. Agarwal, G. Goswami, M. Vatsa, R. Singh, N. K. Ratha, Damad: Database, attack, and model agnostic adversarial perturbation detector, IEEE Transactions on Neural Networks and Learning Systems (2021).
[35] J. Zhang, C. Li, Adversarial examples: Opportunities and challenges, IEEE transactions on neural networks and learning systems 31 (2019) 2578–2593.
[36] R. Singh, A. Agarwal, M. Singh, S. Nagpal, M. Vatsa, On the robustness of face recognition algorithms against attacks and bias, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 13583–13589.
[37] A. Serban, E. Poll, J. Visser, Adversarial examples on object recognition: A comprehensive survey, ACM Computing Surveys (CSUR) 53 (2020) 1–38.
[38] V. Ballet, J. Aigrain, T. Laugel, P. Frossard, M. Detyniecki, et al., Imperceptible adversarial attacks on tabular data, in: NeurIPS 2019 Workshop on Robust AI in Financial Services: Data, Fairness, Explainability, Trustworthiness and Privacy (Robust AI in FS 2019), 2019.
[39] E. Levy, Y. Mathov, Z. Katzir, A. Shabtai, Y. Elovici, Not all datasets are born equal: On heterogeneous data and adversarial examples, arXiv preprint arXiv:2010.03180 (2020).
[40] E. Erdemir, J. Bickford, L. Melis, S. Aydore, Adversarial robustness with non-uniform perturbations, arXiv preprint arXiv:2102.12002 (2021).
[41] I.-C. Yeh, C.-h. Lien, The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients, Expert Systems with Applications 36 (2009) 2473–2480.
[42] M. Lichman, Uci machine learning repository, https://archive.ics.uci.edu/ml, 2013.
[43] C. Ding, H. Peng, Minimum redundancy feature selection from microarray gene expression data, Journal of bioinformatics and computational biology 3 (2005) 185–205.
[44] C. Cortes, V. Vapnik, Support-vector networks, Machine learning 20 (1995) 273–297.
[45] A. Agarwal, M. Vatsa, R. Singh, Chif: Convoluted histogram image features for detecting silicone mask based face presentation attack, in: 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS), IEEE, 2019, pp. 1–5.
[46] A. Agarwal, D. Yadav, N. Kohli, R. Singh, M. Vatsa, A. Noore, Face presentation attack with latex masks in multispectral videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 81–89.
[47] R. E. Wright, Logistic regression. (1995).
[48] J. J. Hopfield, Artificial neural networks, IEEE Circuits and Devices Magazine 4 (1988) 3–10.
[49] K. S. Fu, T. M. Cover, Digital pattern recognition, volume 3, Springer, 1976.