INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Optimized Stacking Ensemble (OSE) for Credit Card Fraud Detection using Synthetic Minority Oversampling Model

Karen Charly Veigas[1], Durga Srilekha Regulagadda[1], Sujatha Arun Kokatnoor[2]*

1 Student, Department of Computer Science and Engineering, School of Engineering and Technology, CHRIST (Deemed to be University), Bengaluru, 560074, India
2 Assistant Professor, Department of Computer Science and Engineering, School of Engineering and Technology, CHRIST (Deemed to be University), Bengaluru, 560074, India

OPEN ACCESS

Received: 10.05.2021
Accepted: 03.09.2021
Published: 01.10.2021

Citation: Veigas KC, Regulagadda DS, Kokatnoor SA (2021) Optimized Stacking Ensemble (OSE) for Credit Card Fraud Detection using Synthetic Minority Oversampling Model. Indian Journal of Science and Technology 14(32): 2607-2615. https://doi.org/10.17485/IJST/v14i32.807

* **Corresponding author**.

sujatha.ak@christuniversity.in

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: Credit fraud is a global threat to financial institutions due to specific challenges like imbalanced datasets and hidden patterns in real-life scenarios. The objective of this study is to propose a model that effectively identifies fraudulent transactions. **Methods:** Methods such as Synthetic Minority Oversampling Technique (SMOTE) and Generative Adversarial Networks (GAN) that artificially generate synthetic data are used in this paper to approximate the distribution of data among the two classes in the original dataset. After balancing the dataset, the individual models Multi-Layer Perceptron (MLP), k-Nearest Neighbors algorithm (kNN) and Support Vector Machine (SVM) are trained on the augmented dataset to establish an initial improvement at the data level. These base-classifiers are further incorporated into the Optimized Stacked Ensemble (OSE) learning process to fit the meta-classifier which creates an effective predictive model for fraud detection. All base-classifiers and the final Optimized Stacked Ensemble (OSE) have been implemented to critically assess and evaluate their performances. **Findings:** Empirical results obtained in this paper show that the quality of the final dataset is considerably improved when Synthetic Minority Oversampling Technique (SMOTE) and Generative Adversarial Networks (GAN) are used as oversampling algorithms. The Multi-Layer Perceptron model showed an increase of 10% in the F1 Score while kNN and SVM showed an increase of 3% each. The optimized model is built using a Stacking Classifier that combines the GAN-improved Multi-Perceptron Model with the other standard classification models such as KNN and SVM. This ensemble outperforms the existing enhanced Multi-Layer Perceptron with near-perfect accuracy (99.86%) and an increase of 16% in F1 Score, resulting in an effective fraud detection mechanism. **Novelty:** For the current dataset, the Optimized Stacked Ensemble model shows an increase of 16% in F1 Score as compared to the existing Multi-Perceptron model.

Keywords: Ensemble; Credit Card; Fraud Detection; GAN; SMOTE; MLP

1 Introduction

The usage of counterfeit or stolen credit cards is referred to as Credit card fraud and is closely related to the crime of identity theft. Institutions such as banks are responsible for detecting and blocking such kinds of transactions. With the increasing use of online transactions and online banking, these frauds have increased in number as well. The continued diversity in the behavior and pattern of these fraudulent transactions makes it more and more difficult to detect these transactions; this leads to huge losses for both banks and customers alike. Therefore, it is very important to have an efficient and robust method to detect fraudulent transactions in real-time.

Due to the recent advancements in data science, various models have been proposed to help solve this problem. There are a few stand-alone methods and algorithms, such as anomaly detectors, which show decent accuracy in classifying the non-fraudulent transactions but tend to fail with classifying the fraudulent ones due to the lack of insufficient data [(1)]. This is tested further in the paper. Secondly, simple classifiers like Random Forest Classifier and Support Vector Machine tend to show higher accuracy and F1 score due to their successful classification of the majority of the non-fraudulent transactions [(2)]. The purpose of this paper is to create an ensemble that uses the decisions made by classifiers and Neural Networks for better accuracy and F1 score. The objective of the OSE is to rectify the imbalance in the credit card datasets and accurately classify unseen transactions. While measures like ROC show the accurate classification of fraudulent data, it is important that a balance between the true positives and true negatives are found that are identified by the OSE. The F1 score combines the precision and the recall to a single metric in accordance with its harmonic mean. The main purpose of this study is to compare the classifiers' performance. The classifier's F1 scores are used to assess which classifier generates better results. Since the F1 score takes into account both the recall and precision, it provides the trade-off that is being looked for in this study, and is considered best suited for real-life transactional scenarios.

Researchers are committed to employing machine learning and data mining approaches to discover an efficient solution in this domain, especially due to recent accelerated breakthroughs in these technologies. Pattern recognition methods like Decision Trees [(3)] and Neural Networks provide a steady scientific basis for anti-fraud. These methods work decently for rule-based detection systems which catch fraud transactions but they require excessive manual work to enumerate all possible detection rules. Unsupervised learning [(4)] has an edge over conventional supervised learning methods as they are unrivalled at finding implicit correlations between data and hidden fraud patterns. Unfortunately, unsupervised learning lags in accuracy when compared to supervised learning. Supervised learning [(1,5)] learns from preceding examples that are generated while training on labelled data. Labelled data have tags that help the model differentiate between patterns that are related to fraudulent transactions and patterns which represent normal behavior. Hence, the individual strengths of MLP (unsupervised), K-nearest Neighbor's algorithm (kNN) and Support Vector Machine (SVM) are combined into a collective ensemble to create an efficient fraud detection system.

In light of the difficulties that traditional classification approaches face concerning the imbalanced datasets, the availability of minority class data in the dataset is increased. These minority classes typically contribute less in minimizing the objective function in standard classification methods. This is done by generating synthetic fraud data and appending these new values to the original dataset thus doubling the number of fraud transactions [(6)]. The augmented dataset is used to test the effectiveness of an Autoencoder model and Multi-Layer Perceptron model. Autoencoder and MLP are both unsupervised algorithms. In [(7)], three unsupervised models are used for anomaly detection, more specifically Credit Card Fraud - and it is shown that Autoencoder outperforms the other two models (One-Class SVM and robust Mahala Nobis). Autoencoder is the most commonly used unsupervised model for anomaly detection but it still falls short in performance when compared to the Multi-Layer Perceptron model.

Based on the experiments, it is observed that Multilayer Perceptron is a better fit for the final ensemble. The final ensemble model proposed makes use of the stacking classifier to take the outputs of base-learners as input and attempts to learn the best possible way to combine these input predictions to obtain better output prediction.

The main contributions of this paper are characterized as below:

- Implemented a model based on Multi-Layer Perceptron (MLP) and GAN to distinguish fraudulent transactions from normal transactions and observed a 10% increase in F1 score when the augmented dataset is tested during experimental study.
- A comparative analysis is performed between an Auto-Encoder model with SMOTE oversampling and the GAN-improved MLP.
- Proposed a model based on stacking classifier which integrates the MLP model along with kNN and SVM. This ensemble is then tested on the augmented dataset and found to have a 90% F1 score.

The structure of the paper is as follows. Section II presents the Related Work; Section III introduces the Proposed Model with SMOTE and GAN methods along with the final ensemble. Afterwards, Section IV provides a detailed explanation of our model. The evaluation and analysis of all models are shown in Section V. Finally, Section VI concludes the paper.

2 Related Work

The field of study referred to in this paper is fascinating because datasets that are generated in real-time scenarios are highly imbalanced and hinder models from training accurately as they are largely biased towards genuine transactions and overlook fraudulent transactions. A cost-sensitive decision tree algorithm is modeled that considers misclassification costs as a method of minimizing bias in well-known traditional classification models [(3)]. To solve these imbalanced dataset issues, the initial focus is to generate artificial fraudulent data that balances the overall data presented to the models for training. Although Random oversampling technique provides a decent starting point, its main drawback is the creation of smaller decision regions which may further contribute to the overfitting problem [(7)]. This observation leads to the use of oversampling methods that specifically generate synthetic data that aid this domain.

Generative Adversarial Networks (GANs) has a generator that feeds on noise as input, and produces realistic (synthetic) copies of the data as training progresses. The underlying patterns or data structures found within datasets are not directly observable and are challenging to extract using other strategies [8]. GANs are able to detect these patterns in the trained data and generate remarkably similar synthetic data values. They have also been proven to generate convincing images and are able to correctly perform discrimination when new instances are introduced [9]. They can be used as an oversampling method to generate synthetic tabular data based on only the fraud transactions to help balance out the ratios in the dataset.

As the imbalance in the current dataset stands, [10] also implements a similar approach where a GAN is trained to output artificial minority class data which are combined with the original training dataset. This forms the augmented training set that is used to improve classifier performance. However, by injecting these examples in a training set there is an increase in observed false positives. This problem is remedied in the proposed model by further implementing the Optimized Stacked Ensemble.

Oversampling techniques such as SMOTE (Synthetic Minority Oversampling Technique) are based on the concept of the minority class(es) in the dataset. This strategy avoids the problem of classifier overfitting, and the decision boundaries allocated to the minority class are spread deeper into the majority class space [11]. SMOTE works under the concept of interpolating between several minority class features of the dataset that lie together. Rather than just duplicating the minority class data points, it focuses on the features. Operations such as rotation and skewing are ways that can help perturb the training data. Synthetic examples are then generated and are introduced onto the dataset along the "lines" joining the minority class neighbors [12]. Hence, the working of SMOTE can be said to be operating in the "feature space" and not the "data space". In [12], the effects of SMOTE are discussed extensively and is proven to have improved the algorithm when trained with a balanced dataset created using SMOTE. Other variants of SMOTE include MWMOTE, SMOTE-ENN and Safe-level smote, which are discussed in [13,14]. In this paper, a hybrid version of oversampling and under-sampling is used called Smote-Tomek. Tomek Links, developed by Tomek in the year 1976, is an under-sampling technique that uses Euclidean distance to choose data points of majority class that are closer to the minority class and keep them; while discarding ones that are farther away in the data space [15].

The method of inverse random under-sampling method and stacking also called the SIRUS is proposed in [16]. Inverse random under-sampling is used to generate multiple datasets which contain all minority class samples and randomly chosen majority class samples. For each of these datasets, classifiers are trained and tested and the best combination of first-level learners (i.e., the classifiers for the ensemble) is chosen. Although this showed positive results, the ensemble only uses classifiers for the ensemble. The addition of a Neural Network, like what is being discussed in this paper, makes the model more robust than with just classifiers.

Hybrid techniques that are not based solely on the working of oversampling are tested side by side in [17], while proposing a new model called Constructive Covering Algorithm (CCA) which takes a new methodology to delete samples that are overfitted in the data space. This model is proven to be better than almost all hybrid SMOTE models with binary classification. Although this model is very efficient with omitting data points that can cause the algorithm to overfit or the class space to overlap, the complexity of the model is larger than other existing models such as Tomek Links.

Another model discussed in [18] is called the DBSM model - a hybrid balancing method between Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and SMOTE. While the former algorithm is trained to only produce negative samples, SMOTE produces positive samples for the dataset. Outputs from both the algorithms are combined to create a new balanced dataset. The model lacks in explaining how it would work with n-class datasets, as it is solely focused only on binary classification.

Many comparative studies over well-established feature models have been done at an application level. AdaBoost, Random Forest, Naive Bayes, and PART machine learning techniques are few techniques that have been used in the domain of financial fraud detections and these techniques are compared using the five parameters to determine their performance [1]. Although these models perform extremely well, Neural Networks are preferred over regular feature models. One main reason to choose Neural Networks over regular feature models is that they tend to learn complex relationships between data and even nonlinear structures. This is vital for real-life scenarios as most relationships between inputs and outputs are non-linear as well as complex. Hence, the Multi-Layer Perceptron model is introduced as one of the key members of the stacked ensemble. Various standard ensemble techniques already exist as presented in [19,20] such as Bagging, AdaBoost, Random Forest and Gradient boosting classifier ensembles. However, these have considerable drawbacks and aren't efficient in providing an effective fraud detection system.

In [21], a comparative analysis between various types of algorithms is done using a real-time dataset. Thirty per cent of the real-time dataset is used as a test case while the other seventy percent is under-sampled to help train the models better. It is evident that the models need a balanced dataset to work efficiently. The conclusion of [21] states that using Logistic Regression as the meta classifier for the stacking classifier showed the most promising results for predicting fraudulent transactions with an accuracy of over 95%.

It is important to optimize the detection accuracy to avoid any catastrophic losses that can be faced due to misclassification of credit card transactions. In stacked classifiers, various classification models are combined to help reduce the generalization error. In [2], it is mentioned that stacking various classification models combines machine learning classifiers which are conceptually different and uses either a Hard Vote (majority voting) or a Soft Vote (average predicted probabilities) to predict the class labels. Although the author has provided a precision-recall score for a stacked classifier, they have only incorporated regular supervised methods of classification [2,4,5] and the final class label is predicted using the weighted majority voting technique in [4]. Similarly, [5] also sees a minimal difference in accuracy between the standalone models, but a decent increase in the predictive accuracy percentage is noticed. The aim is to further improve this hybridized model by implementing oversampling methods on the ensemble of learners to increase the scalability of classification after the proper model training.

3 Proposed Model

This section contains a detailed description of the model proposed in this paper. The framework of OSE is seen in Figure 1 and shows how the model works. Using Python's extensive packages, the dataset is first preprocessed and then fed into the few selected machine learning algorithms for predictive learning. The preprocessed data is sent through oversampling methods for balancing the dataset. SMOTE-Tomek improved dataset is then used to train the SVM classifier and GAN improved dataset to train the kNN classifier and the MLP model. Next, test data is input to the classifiers for predictions and these predictions are sent to the OSE. The OSE uses Logistic Regression as its meta classifier and the stacked prediction from the OSE is considered as the final output and is cross-validated with the original data to reveal the F1 score and accuracy of the model.

3.1 Generative Adversarial Networks (GANs)

GANs are deep learning technologies that learn hierarchies of concepts by layering abstractions on top of one another. GANs have been widely successful in synthesizing real-looking images and convincing tabular data. GANs are composed of two models, one being a generator and the other discriminator, that compete against each other in a zero-sum minimax game. The discriminator estimates the probabilities that a sample comes from the original training data or generated synthetic data, while the generator learns the distribution of samples in the dataset. Usually, both models are multilayer Neural Networks that are trained until the discriminator is unable to distinguish between real and generated data, that is, global optimality is achieved. The minimax loss function is shown in equation (1).

$$L_G = E_x[\log\log(D(x))] + E_z[loglog(1 - D(G(z)))] \quad (1)$$

Here, L_G is the minimax loss function value, E_x is the expected value over all real data instances and $D(x)$ is the discriminator's estimate of the probability that data instance x is real. Similarly, E_z is the expected value over all random inputs to the generator, while $G(z)$ is the generator's output over the given noise z, which in this case is the fraud value in the original dataset.

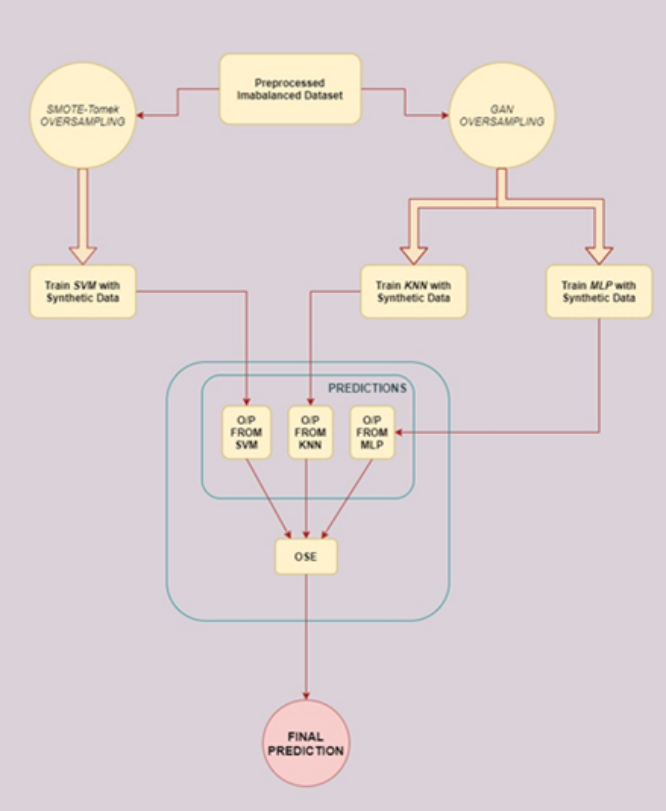

Fig 1. Architecture of OSE Model

3.2 Synthetic Minority Oversampling Technique (SMOTE) and Tomek Links

SMOTE, an oversampling model, and Tomek, an under-sampling model, together form a hybrid model that is more efficient than when compared to an under-sampling method or a Random Over Sampler (ROS). This model removes all the overlapping data points for both of

the classes distributed in data space. The first step is where SMOTE algorithm oversamples the minority class data points. SMOTE selects the minority class instance at random, which in this case is the fraudulent transactions, and finds its k-Nearest Neighbors in the minority class. After choosing one instance, it finds the nearest neighbor to this instance at random from one of the neighbors and forms the synthetic instances by connecting the two original instances forming a line segment in the feature space. Hence, it can be said that the synthetic instances are generated as a curvilinear combination of the two instances that are chosen from the original data points.

This process continues until the synthetic minority class samples along with the actual minority samples reach the threshold of samples given to the algorithm. In this case, the threshold is the same as the majority class samples count. The approach is considered effective because of synthetic examples that are relatively close to the original instances in the feature space. After oversampling is done by SMOTE, the class clusters may invade each other's space. This implies that the classifier model may be overfitting.

To avoid overfitting and overlapping of data points, Tomek is applied to the dataset. Tomek links are formed between the opposite class paired samples that are overlapping or are negligibly near to each other. The pair is then removed from the data space to increase the class separation between both the classes. This in turn makes it easier for the algorithm to classify data points. Hence, Tomek links are applied to oversampled minority class samples done by SMOTE.

3.3 Optimized Stacking Ensemble (OSE) Model

To begin with the ensemble model, the F1 scores and accuracies of the classifiers and the Neural Network models are compared separately. The optimized classifiers' predictions at the base level of the framework are compiled into an ensemble using the technique of stacked generalization, and the augmented dataset is trained using a different learning algorithm (One Vs Rest Classifier). Stacked generalization enables the larger model, in this case, the OSE, to embed the initial Neural Networks as sub-models and use them for training and prediction while simultaneously using Logistic Regression as the meta learner. The stacking process functions on two separate levels: Level 0 contains the three base-level learners which are individually used to predict the classes of unseen transactions from the validation set. The base learners are combined along with their predicted target classes and are then passed on to the meta-learning phase as input. At Level 1, the individual classification outputs and the expected class for each transaction instance are now considered as new features for every transaction and are presented to the meta-level classifier as the training set. This method allows the OSE to use the new features in the training set while maintaining the original target attribute.

The stacking classifier takes inputs from two classifiers and one Neural Network; the meta classifier used is Logistic Regression. The predictive performance of the ensemble's resultant stacked classifier is then evaluated on the test set. During the experimental study, SVM showed better performance in both parameters. SVM outperforms most models due to its ability to work well with high-dimensional datasets which is a key factor in Credit Card Fraud Detection. kNN is robust to noisy training data, which in this case is the presence of a large number of genuine transaction examples. RFC, on the other hand, is better at dealing with categorical features and not entirely suitable for datasets like the one that is being dealt with for this problem. Because of these reasons, they performed less. As for choosing which Neural Network to incorporate in the ensemble, in Figure 2 it can be seen that although the autoencoder model does well concerning the majority class, it fails when compared to the MLP for the minority class. Hence, MLP is chosen for the ensemble, along with SVM and kNN as the base classifiers and Logistic Regression as the meta classifier.

4 Results and Discussion

4.1 Dataset

The dataset used in this paper contains transactions made by European credit cardholders on two consecutive days in September 2013 [22]. The set has 284,807 transactions and 31 columns, 28 of these columns are redacted to maintain confidentiality. Out of these transactions, only 492 of them have been classified as fraudulent cases (only 0.17% of the dataset) while the rest are classified under non-fraudulent or genuine transactions, making it very imbalanced. The available dataset contains purely numeric columns. This is mentioned in the source webpage as a result of Principal Component Analysis (PCA) transformation of the original data. This redaction process is done to respect the confidentiality issues and the sensitive nature of the data; the original features and more background information about the data are not provided.

4.2 Data Preprocessing

The dataset used in this study contains transactions in which only 492 of them have been classified as fraudulent cases (only 0.17% of the dataset) while the rest are classified under non-fraudulent or genuine transactions, making it very imbalanced. In [20], it is speculated that the variables might fall under different categories, each category containing the statistics of a transaction characteristic such as regional, transactional, merchant type, time/amount based and time/frequency of transactions based. The remaining labeled columns are Time, Amount, and Class. Time column points to the time elapsed between the first transaction and the present transaction (in seconds). The amount column contains the transaction amount; it comes in handy while training a cost-sensitive detection model. Class is a reference feature that points to whether the transaction is fraudulent or not [1: Fraud, 0: Genuine]. The amount column is heavily skewed and is hence normalized.

SMOTE and Vanilla GAN are experimented to oversample the minority data in the binary classification. Although the fundamental concept of how SMOTE and GAN work to oversample the minority dataset differ, the effects they have on the datasets are very similar. The effects of both the aforementioned oversampling techniques can be seen when both datasets are fitted and tested using a Logistic Regression classifier. Logistic Regression classifier is an optimal choice for this comparison because it forms no biases about the distributions of classes in a specific feature space.

4.3 Performance Analysis

A preliminary comparison between the two types of Neural Networks is conducted in this study. Multi-Layer Perceptron (MLP) is trained on a GAN-improved dataset and Autoencoder is run on a SMOTE improved dataset. These models are then tested on 20% of the dataset that are set aside. Figure 2 indicates that MLP has a great margin of F1 score at 76% and hence, is chosen as a key member in the final stacking ensemble. Similarly, one round of testing is done on supervised models namely kNN, SVM, and Random Forest Classifier (RFC) based on the literature review. From Figure 3 , it is observed that after testing on the imbalanced dataset

, kNN and SVM perform much better as compared to RFC. Although SVM outperformed the other classifiers, it is not suitable as a standalone classifier for two main reasons: 1. In real life scenarios, there is a growing number of samples at any given point in time due to the exponential growth of transactions and SVM does not bode well with large datasets; 2. SVM is good with classifying when there is a clear separation in classes, which is not always present in the scenario of credit card fraud data points. For this very reason, the OSE model is essential. It uses the classifying power of three algorithms to make a final prediction.

The first step to building the OSE framework is to test the effectiveness of GAN as an oversampling method. To do this, the initial experiment is set up such that the MLP model is trained on the original dataset and its performance is observed as 66% in terms of F1 Score. After generation and injection of the synthetic fraud data into the original dataset, the same model is trained and observed a 10% increase in F1 Score at 76%. Similarly, the performance of each model is evaluated before and after applying the oversampling methods and it is observed that there is a clear increase in performance across the board in terms of F1 score as shown in Figure 4 . Evaluation metrics such as Accuracy and F1 Score are used to evaluate the performance of the proposed model. The classification performances (based on F1 score and Accuracy) of the OSE along with the base level classifiers, i.e., SVM, MLP, and kNN classifiers, are graphically shown in Figure 3.

In the experiment, the train to test data is split in an 80:20 ratio. This test data is used to check if the proposed framework's meta-classifier (Logistic Regression) classifies well on unseen data. Since the test data is not used in the classifier's training process, it is used to provide an unbiased estimate of the classification performance of the OSE. As seen in Figure 5 , the F1 Score performance of the OSE is at 0.905 and its Accuracy is at 99.8% showing enhanced fraudulent/genuine classification accuracies. The MLP classifier, with 94 percent accuracy, and the SVM classifier, with 93 percent accuracy, trail the OSE. kNN shows the highest F1 score of 0.96 and an accuracy of 95%. Despite a slightly lower F1 score than kNN, the OSE is preferred due to its ability to harness the abilities of unsupervised MLP which works best with finding hidden patterns of fraudulent transactions in real-life scenarios.

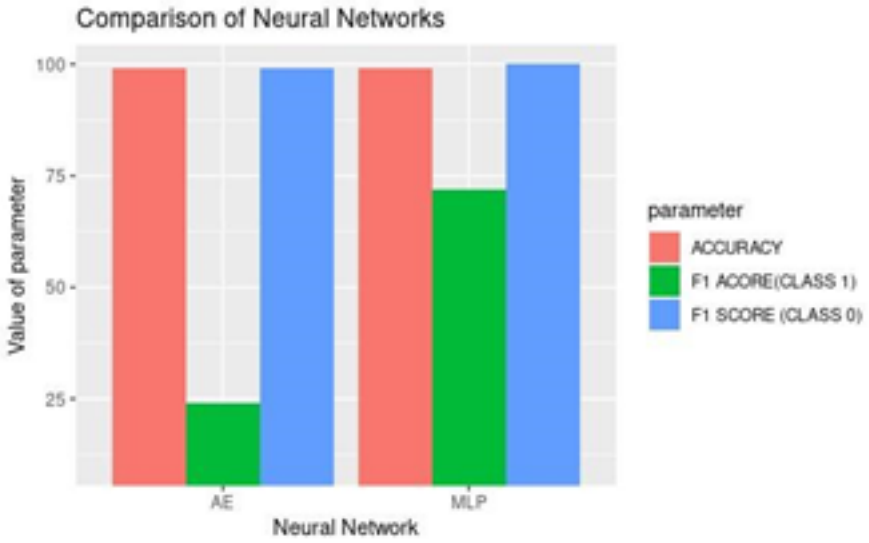

Fig 2. Comparison of Neural Networks

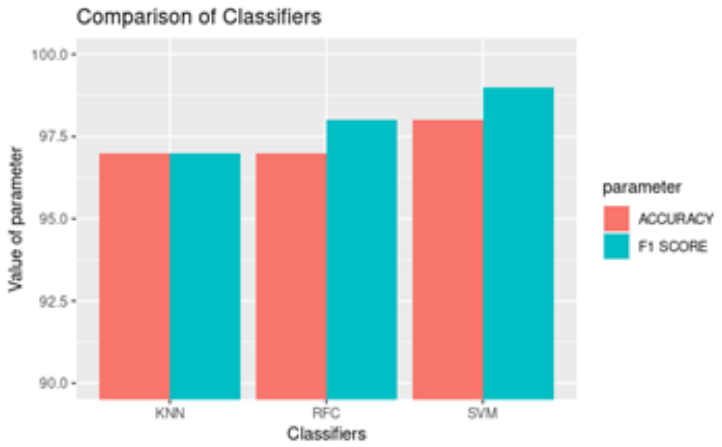

Fig 3. Comparison of Supervised Classifiers

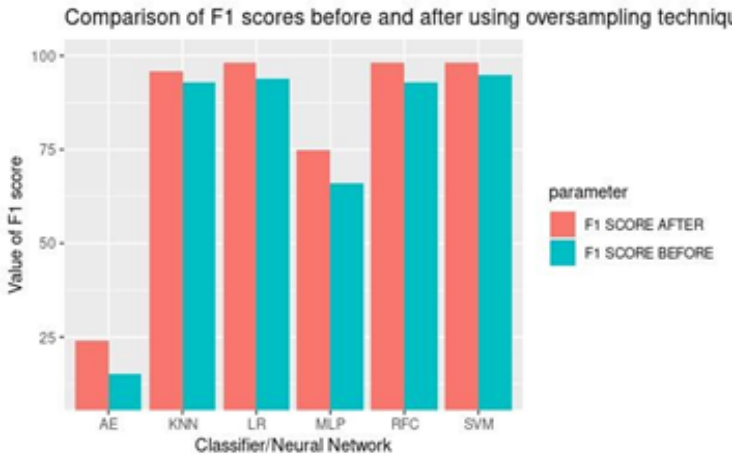

Fig 4. F1 scores after Oversampling

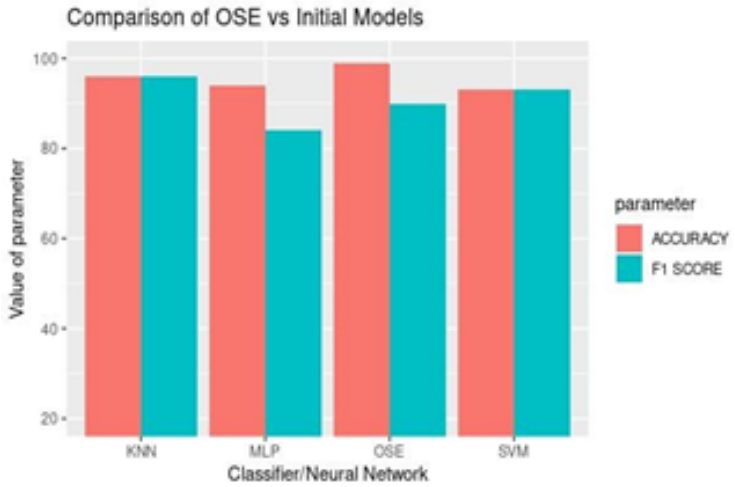

Fig 5. F1 scores of Stacking Ensemble vs Conventional Models

5 Conclusion and Limitations

5.1 Conclusion

The study of imbalanced datasets and ensemble learning paradigms is crucial in the field of fraudulent deductions and other similar studies. The motivation behind this paper is due to a lack of ensembles learning paradigms that are trained with not just imbalanced datasets, but with synthetic datasets that can help provide more information on the minority class features. In this paper, the negative effects that a heavily skewed, highly imbalanced dataset can have on classification algorithms and Neural Networks models are reduced. An ensemble model (OSE) is built which has both a Neural Network and supervised classification algorithms to help detect fraudulent transactions in varying dataset fragments. This approach harnessed the strength of stacked classifiers in handling highly negatively skewed data. The F1 score and the accuracy of the ensemble model are impressive and the model is proven to be more robust than the stand-alone classifiers because of its high accuracy in classifying fraudulent transactions. The proposed model also worked well with high class-imbalanced datasets, which is very important for classifying real-time datasets.

5.2 Limitations

One major limitation regarding the OSE would be the time taken for it to train with the given data. With the credit card data being a non-stationary one, it is harder to run a model with mostly pre fitted and pre trained parameters. The OSE, like all other simple classifiers, needs to constantly learn. The involvement of two classifiers and one neural network somewhat resolves this issue and can be further combated by making the learning process can have much larger intervals as compared to a simple classifier. However, this is still not recommended over large periods of time.

Future Scope

To improve the model further, the concept of weighted voting can be applied to the predictions from the first layer of classifiers in the OSE. Another method would be to use boosting algorithms which can be trained on the synthetic data that is being generated using the oversampling techniques discussed previously. Aside from this, the concept of non-stationary data can be handled in the future. Non-stationary data refers to the continuous data that is being produced every single moment with a new feature behavior. Non-stationary data compounded with the imbalanced data set problem does not provide ideal situations for classifiers to perform well.

References

1) Singh A, Jain A. Adaptive Credit Card Fraud Detection Techniques Based on Feature Selection Method. In: Bhatia S, Tiwari S, Mishra K, Trivedi M, editors. Advances in Computer Communication and Computational Sciences. Advances in Intelligent Systems and Computing;vol. 924. Springer. 2019. Available from: https://doi.org/10.1007/978-981-13-6861-5_15.
2) Mishra A, Ghorpade C. Credit Card Fraud Detection on the Skewed Data Using Various Classification and Ensemble Techniques. In: IEEE International Students' Conference on Electrical, Electronics and Computer Science. 2018;p. 1–5. Available from: https://ieeexplore.ieee.org/document/8546939.
3) Sahin Y, Bulkan S, Duman E. A Cost-Sensitive Decision Tree Approach for Fraud Detection. *Expert Systems with Applications*. 2013;40(15):5916–5923. doi:10.1016/j.eswa.2013.05.021.
4) Ali I, Aurangzeb K, Awais M, Khan RJUH, Aslam S. An Efficient Credit Card Fraud Detection System using Deep-Learning based Approaches. In: 2020 IEEE 23rd International Multitopic Conference. 2020;p. 1–6. Available from: https://ieeexplore.ieee.org/document/9318202.
5) Pun JK, Lawryshyn Y. Improving Credit Card Fraud Detection using a Meta-Classification Strategy. *International Journal of Computer Applications*. 2012;56(10):41–46. doi:10.5120/8930-3007.
6) Douzas G, Bacao F. Effective Data Generation for Imbalanced Learning Using Conditional Generative Adversarial Networks. *Expert Systems with applications*. 2018;91:464–71. Available from: https://doi.org/10.1016/j.eswa.2017.09.030.
7) Rezapour M. Anomaly Detection using Unsupervised Methods: Credit Card Fraud Case Study. *International Journal of Advanced Computer Science and Applications*. 2019;10(11):1–8. Available from: https://thesai.org/Downloads/Volume10No11/Paper_1-Anomaly_Detection_using_Unsupervised_Methods.pdf.
8) Sasank JVVS, Sahith GR, Abhinav K, Belwal M. Credit Card Fraud Detection Using Various Classification and Sampling Techniques: A Comparative Study. In: 2019 International Conference on Communication and Electronics Systems. 2019;p. 1713–1718. Available from: https://ieeexplore.ieee.org/document/9002289.
9) Zhou K, Wang W, Hu T, Deng K. Application of Improved Asynchronous Advantage Actor Critic Reinforcement Learning Model on Anomaly Detection. *Entropy*. 2021;23(3):274–274. Available from: https://doi.org/10.3390/e23030274.
10) Marra F, Saltori C, Boato G, Verdoliva L. Incremental Learning for the Detection and Classification of Gan-Generated Images. In: 2019 IEEE International Workshop on Information Forensics and Security. 2019;p. 1–6. Available from: https://arxiv.org/abs/1910.01568.
11) Fiore U, Santis AD, Perla F, Zanetti P, Palmieri F. Using Generative Adversarial Networks for Improving Classification Effectiveness in Credit Card Fraud Detection. *Information Sciences*. 2019;479:448–455. Available from: https://doi.org/10.1016/j.ins.2017.12.030.
12) Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of artificial intelligence research*. 2002;16:321–57. doi:10.5555/1622407.1622416.
13) Shamsudin H, Yusof UK, Jayalakshmi A, Khalid MN. Combining Oversampling and Undersampling Techniques for Imbalanced Classification: A Comparative Study using Credit Card Fraudulent Transaction Dataset. In: 2020 IEEE 16th International Conference on Control & Automation. 2020;p. 803–808. Available from: https://ieeexplore.ieee.org/document/9264517.
14) Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *The Journal of Machine Learning Research*. 2017;18(1):559–63. Available from: https://www.jmlr.org/papers/volume18/16-365/16-365.pdf.
15) Sisodia DS, Reddy NK, Bhandari S. Performance Evaluation of Class Balancing Techniques for Credit Card Fraud Detection. In: 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering. 2017;p. 2747–2752. Available from: https://ieeexplore.ieee.org/document/8392219?denied=.
16) Yan Y, Liu R, Ding Z, Du X, Chen J, Zhang Y. A Parameter-Free Cleaning Method for SMOTE in Imbalanced Classification. *IEEE Access*. 2019;7:23537–23585. doi:10.1109/ACCESS.2019.2899467.
17) Zhang Y, Liu G, Luan W, Yan C, Jiang C. Application of SIRUS in Credit Card Fraud Detection. In: Computational Data and Social Networks. Springer. 2018;p. 66–78. Available from: https://link.springer.com/chapter/10.1007/978-3-030-04648-4_6.
18) Jonathan B, Putra PH, Ruldeviyani Y. Observation Imbalanced Data Text to Predict Users Selling Products on Female Daily with SMOTE, Tomek, and SMOTE-Tomek. In: Proceedings - 2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology, IAICT 2020. 2020;p. 81–85. Available from:

https://ieeexplore.ieee.org/document/9172033.
19) Sanguanmak Y, Hanskunatai A. DBSM: The combination of DBSCAN and SMOTE for Imbalanced Data Classification. In: 2016 13th International Joint Conference on Computer Science and Software Engineering. 2016;p. 1–5. Available from: https://ieeexplore.ieee.org/document/7748928.
20) Dhankhad S, Mohammed E, Far B. Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study. In: 2018 IEEE International Conference on Information Reuse and Integration. 2018;p. 122–125. Available from: https://ieeexplore.ieee.org/document/8424696.
21) Novakovic J, Markovic S. Classifier Ensembles for Credit Card Fraud Detection. In: and others, editor. 2020 24th International Conference on Information Technology. 2020;p. 1–4. Available from: https://ieeexplore.ieee.org/document/9070534.
22) Lenka SR, Pant M, Barik RK, Patra SS, Dubey H. Investigation into the Efficacy of Various Machine Learning Techniques for Mitigation in Credit Card Fraud Detection. In: Bhateja V, Peng SL, Satapathy SC, Zhang YD, editors. Evolution in Computational Intelligence;vol. 1176. Springer. 2021;p. 255–254. Available from: https://link.springer.com/chapter/10.1007/978-981-15-5788-0_24.