Issues in Information Systems
Volume 25, Issue 3, pp. 47-61, 2024

DOI: https://doi.org/10.48009/3_iis_2024_105

Examining ensemble models to detect credit card fraudulent transactions

Queen E. Booker, *Metropolitan State University Minnesota, queen.booker@metrostate.edu*
Carl M. Rebman, Jr., *University of San Diego, carlr@sandiego.edu*

Abstract

Fraudulent credit card transactions impact both consumers and card issuers. The ability to detect fraudulent credit card transactions can reduce the cost of credit card use. Prior research has shown that machine learning and ensemble models can identify fraudulent transactions with good accuracy. However, no study has been found that compares heterogeneous and homogeneous models. This research study examines and compares machine learning algorithms with multiple ensemble models for detecting fraudulent credit card transactions using data available from a U.S.-based credit card issuer. The results show that heterogeneous ensemble models can better detect fraudulent and non-fraudulent transactions than individual and homogeneous models. The results suggest that underlying individual algorithms are used in the ensemble matter. Specifically, heterogeneous models that use both random forest modeling and neural network modeling tend to outperform individual models and ensemble models that do not utilize both.

Keywords*:* credit card fraud detection, ensemble models, machine learning, suitability

Introduction

Credit card transactions are a common method of business transactions. The four major credit card networks - Visa, Mastercard, American Express, and Discover – utilize AI-powered user authentication analysis, Europay, Mastercard, and Visa (EMV) chips, and card identification numbers (CIDs or CVVs) to help protect against credit card fraud. However, despite these network protections, consumers still are victims of credit card fraud. Fraud is defined by Webster's Dictionary as "wrongful or criminal deception intended to result in financial or personal gain." According to the Cornell School of Law Legal Information Institute (2022) credit card fraud is "a form of identity theft that involves an unauthorized taking of another's credit card information to charge purchases to the account or remove funds from it." According to the Federal Trade Commission, credit card fraud is the most common form of identity theft in the United States. (Maxwell, 2023)

Nikolina Cveticanin at dataprof.net (2022) states:

- By 2023, retailers will lose about $130 billion each year on card-not-present transactions.
- Card-not-present fraud is now 81% more common than point-of-sale fraud.
- 54% of businesses are only "somewhat confident" they would be able to detect fraudulent activity on time.

With the increasing use of credit cards, algorithms to improve fraud detection would benefit the financial services industry. For this reason, machine learning algorithms, and specifically, ensemble models, may be a part of the solution. Prior research has been performed using machine learning algorithms. However, the models developed used extensive data that is unlikely to be available at a point-of-sale transaction. Thus,

additional research is needed to identify effective algorithms to mitigate fraudulent transactions with minimum data. This paper investigates and compares the use of ensemble models against the individual machine learning models to determine if an ensemble can outperform individual models in detecting fraudulent transactions. This research builds on prior research that has shown that ensembles perform better than individual models in predicting fraudulent credit card transactions.

The contribution of our study is the evaluation of both homogeneous and heterogeneous combinations of ensemble models. The rest of this paper is structured as follows. In the literature review, previous research is discussed. Next, our research methodology is described, followed by the results. Last are the conclusions, limitations, and next steps.

Literature Review

Credit Card Fraud

The credit card industry maintains its profits in three ways: by collecting interest on balances carried month after month, by charging transaction fees for using the card, and by keeping fraudulent transactions at a minimum. Although credit card networks have a zero-liability policy for fraudulent transactions, the card issuer takes financial responsibility for fraudulent transactions. The number of fraudulent transactions is small when compared to the total number of transactions. However, any fraudulent transaction increases the cost of using credit cards for both the issuer and the consumer. Thus, it is important to minimize the number of times a fraudulent transaction is approved. According to Kultur and Calayan (2017), credit card fraud detection tools are rule-based, utilizing heuristics and, in some cases, data mining algorithms. They state "fraud experts define the rules applied in these systems based on historical fraud cases and their investigation results. When a new transaction matches one or more of the previously defined rules, the system raises an alarm to indicate that the new transaction is potentially fraudulent.

The rule-based approach is successful in identifying fraudulent transactions that follow previously observed fraud patterns, but it lacks agility (Booker & Rebman, 2024). Before a new rule is added to the existing rule set, a considerable number of fraudulent transactions matching that rule have typically already occurred. Krivko (2010) adds that long delays required before a rule can be added can quickly cause a rule to become obsolete. To detect fraud, issuers upload essential data such as card details, IP addresses, and e-mail addresses to a centralized database. Merchants and banks cannot see the information of other banks' cardholders but can get a certain risk score from the centralized system (Kultur & Calayan, 2017). There are many reported approaches in the literature for credit card fraud detection. The most recent approaches are based on machine learning algorithms.

Machine Learning Algorithms

Machine learning is a subfield of artificial intelligence. It is concerned with the development and use of algorithms that can learn from data and generalize to unseen data, and thus perform tasks without explicit instructions. Malone et al (2020) state "The function of a machine learning system can be descriptive, meaning that the system uses the data to explain what happened; predictive, meaning the system uses the data to predict what will happen; or prescriptive**,** meaning the system will use the data to make suggestions about what action to take." Machine learning starts with almost any type of data that is gathered and prepared to be used as training data. When using machine learning, analysts begin with a model to work with, supply the data, and let the computer model train itself to find patterns or make predictions. The analyst may tweak the model to improve its performance (Kultur & Calayan, 2017).

Machine learning algorithms are divided into three categories: supervised, semi-supervised, and unsupervised (Mahesh, 2020). In supervised learning, the algorithm is labeled or maps the input data to generate the required output. This algorithm examines the training data and creates an inferred function to map new examples (Cunningham et al., 2008; Liu et al., 2021). Supervised learning models fall into three categories: classification, regression, and forecasting. Semi-supervised learning is similar to supervised learning except it uses labeled data to ground predictions, and unlabeled data to learn the shape of the larger data distribution (Bewtra, 2022). In unsupervised learning, the algorithm does not label or map the input data to generate the required output. There is no training data; the algorithm self-discovers the natural patterns (James et al., 2021). Unsupervised learning falls into two categories: clustering and dimension reduction.

Ensemble Models

Polikar (2012) and Petrakova et al. (2015) describe ensemble models as a machine learning technique that combines base models to create a single optimal predictive model to improve a model's performance or reduce the likelihood of an incorrect selection. Brown (2010) explains that the underlying principle of ensemble learning is all models have limitations and will make mistakes in real-world situations. By combining individual models, ensemble learning can minimize each model's weaknesses while maximizing each model's strengths, resulting in a more accurate decision-making process.

Ensemble learning models can be either homogeneous or heterogeneous. Homogeneous models have individual algorithms from the same class of models. Heterogeneous ensembles combine models from different classes of machine learning algorithms. For example, a homogeneous ensemble would only have models from the neural network family whereas the heterogeneous ensemble would have models from multiple families of machine learning. These are just examples. Most ensemble models have an odd number of individual models with a minimum of three. These ensembles have been shown to exhibit more predictive performance than single-model predictions (Hsu et al. 2009).

Credit Card Fraud Detection and Machine Learning Algorithms

Credit card fraud detection with machine learning algorithms has been extensively studied by researchers. Most of the research used supervised learning techniques which is not surprising given the classification of the dependent variable. Supervised learning techniques include but are not limited to regression (LR), naïve Bayes (NB), kth nearest neighbor (KNN), neural networks (NN), decision trees (DT), random forest (RF), boosting (BT), bootstrapping (BR), and support vector machines (SVM). Self-organizing maps (SOM) was one unsupervised method found. A summary of the results is shown in Table 1.

Table 1: Summary of Research Studies on Credit Card Fraudulent Transactions and Machine Learning

Study Authors	Year	Models Evaluated	Results
Chen et al.	2005	SVM, SOMNN, and BPNN	Back Propagation Neural Network (BPNN) had the best accuracy in detecting fraudulent transactions with an accuracy of 78% for the BPNN compared to SVM at 67%
Quah and Sriganes	2008	SOM	SOM had an accuracy rate of 93%.
Bhattacharyya et al.	2011	SVM	96% accuracy rate
Zareapoor et al.	2012	KNN and NN	KNN had a higher accuracy rate
Zareapoor et al.	2012	SVM	98% accuracy rate
Seeja and Zareapoor	2014	KNN and LR	KNN had a higher accuracy rate
Olszewski	2014	SOM	SOM had an accuracy of 90%.

Study Authors	Year	Models Evaluated	Results
Seeja and Zareapoor	2014	SVM	96% accuracy rate
Rushin et al	2017	NN, LR, and Gradient Boosted Tree (GBT)	NN performed better than both GBT and LR
Tran et al	2018	Multiple	NN performed better than other models as data size increased reaching 97% accuracy
Sohony et al.	2018	RF, FNN	FNN had better results than RF
Armel and Zaidouni	2019	DT, RF, and NB	RF models outperformed DT and NB
Makki et al	2019	LR, DT, SVM, and NN	NN had the highest accuracy rate
Singh and Jain	2020	NB, KNN, RF	NB had a minimum accuracy of 95.15, KNN had a minimum accuracy of 93.75, and a hybrid model of KNN and RF had a minimum accuracy of 81.97%
Carrasco et al	2020	NN	92% accuracy rate

There have been several studies applying ensemble models to the credit card fraud detection problem. These studies have been summarized in Table 2**.**

Table 2: Summary of Research Studies on Credit Card Fraudulent Transactions and Ensemble Models

Study Authors	Year	Models Used	Results
Sohony	2018	RF and FNN	Combined model outperformed individual models reaching an accuracy rate of 92%
Kim et al.	2019	champion-challenger framework; NN	Combined model outperformed individual models
Kokkinaki	1997	RF and NN	Combined model outperformed individual models
Gadi et al.	2008	NN, Bayesian Network (BN), NB, and DT	Combined model outperformed individual models
Sadgali et al	2019	NN, NB, DT	Combined model outperformed individual models with 96.3% accuracy
Prusti and Rath	2010	DT, NN, SVM	Ensemble outperformed individual models; 96% accuracy
Xie et al	2021	(LR), SVM, NB, DT, RF, AdaBoost, and XGBoost	Ensemble had accuracy rate of 98%
Kultur et al	2017	DT, RF, Bayesian Network, NB, SVM, KNN	Ensemble outperformed individual models
Booker and Rebman	2024	NB, SVM, RF, NN	Ensemble outperformed individual models

Other significant findings from prior research include Kultur et al's (2017) finding that alternative voting strategies remove the barrier of an odd number of individual models. Booker and Rebman (2024) found that the combination of individual models impacts the performance of the ensemble.

Limitations of the literature review

None of the studies found in the literature used a common data set or the same variables even when using the publicly available Kaggle dataset. This limits the ability to discuss the impact of different independent variables on the model outcomes.

Research Methodology

The study follows the research methodology used by Carrasco et al (2020). It begins with the research purpose followed by a description of the data, the data analysis, and the results.

Purpose Statement and Hypotheses

Prior research (Kultur et al, 2017, Xie et al, 2021, Booker & Rebman, 2024) has already demonstrated that ensemble models perform better than individual models in detecting fraudulent transactions. The purpose of this research study is to compare and contrast the performance of individual models, a homogenous ensemble, and heterogeneous ensembles comprised of highly regarded models from the literature. Specifically, the study examines if a homogenous ensemble can outperform a heterogeneous ensemble. Based on the literature review, we selected neural networks, random forest (RF), and naïve Bayes (NB) for the individual models. Because most studies did not specify which neural networks were used in the study, three common networks were chosen for this study: recurrent neural network (RNN), feedforward neural network (FNN), and convolutional neural network (CNN). Each ensemble included three of the individual models. The three neural networks were used to develop the homogeneous model and all combinations of neural networks, RF, and NB were used for the heterogeneous models. The resulting ensemble were:

- **HOE**: RNN, FNN, CNN
- **HEE1**: RF, RNN, NB
- **HEE2**: RF, FNN, NB
- **HEE3**: RF, CNN, NB
- **HEE4**: RF, FNN, CNN
- **HEE5**: RF, FNN, RNN
- **HEE6**: RF, CNN, RNN
- **HEE7**: NB, FNN, CNN
- **HEE8**: NB, FNN, RNN
- **HEE9**: NB, CNN, RNN

The goal was to identify a model that would simultaneously minimize false positives and false negatives. Using the results from Xie et al (2021) and Kulfur (2017) in reporting that ensemble models significantly outperform their individual models, our first research hypothesis is:

H0: *Ensemble models will significantly outperform their individual models.*

In addition, based on the results reported by previous studies, neural networks tend to outperform other methods. Using those results our second hypothesis is:

H1: *The homogeneous ensemble model will significantly outperform the heterogeneous ensemble models. Specifically, we expect HOE to outperform HEE1, HEE2, HEE3, HEE4, HEE5, HEE6, HEE7, HEE8, and HEE9.*

Individual and ensemble models were first evaluated using standard suitability measures. Suitability measures are important for assessing whether a particular data mining technique or algorithm is appropriate and effective for a given dataset and analytical task. It involves evaluating various factors to determine if the chosen method will produce meaningful and useful results. Standard suitability measures include accuracy, specificity, Type I error score, Type II error score, recall, precision, F-score, and area under the

curve (AUC). Many of the measures are associated with the confusion matrix which is a table used in data mining and machine learning to describe the performance of a classification model. It presents a summary of the predictions made by a model compared to the actual known outcomes in the dataset. The basic confusion matrix is structured as shown in Figure 1.

		Actual	
		Positive (0)	**Negative (1)**
Prediction	**Positive (0)**	True positive	False positive
	Negative (1)	False negative	True negative

Figure 1: Basic Confusion Matrix for a Binary Decision Model

- **True Positive (TP)**: Instances where the model correctly predicts the positive class.
- **False Negative (FN)**: Instances where the model incorrectly predicts the negative class.
- **False Positive (FP)**: Instances where the model incorrectly predicts the positive class.
- **True Negative (TN)**: Instances where the model correctly predicts the negative class.

Accuracy, specificity, recall, Type I Error, Type II Error, precision and F-measure are calculated based on the results from the confusion matrix. The description and formula for each measure is shown in Table 3.

Table 3: Suitability Measures Descriptions and Formula

Suitability Measure	Description	Formula
Accuracy	Measures the proportion of correct predictions made by the model out of all predictions made	(TP + TN)/(TP + TN + FP + FN)
Specificity	Measures the proportion of actual negative instances that are correctly identified as negative by the model out of all true negative instances.	TN/(TN + FP)
Recall	Measures the ability of a model to correctly identify positive instances out of all actual positive instances in the dataset.	TP/(TP + FN)
Type I Error	The number of false positives	FP
Type II Error	The number of false negatives	FN
Precision	Measures the accuracy of positive predictions made by a model, indicating the proportion of correctly predicted positive instances out of all instances predicted as positive.	TP/(TP + FP)
F-measure	Combines precision and recall into a single value. It is particularly useful when you want to balance the trade-off between precision and recall and have a single metric that summarizes both.	2 x (Precision x Recall)/(Precision+Recall)
Area under the curve (AUC)	Quantifies the overall performance of a binary classification model based on its ROC curve	No relationship to the confusion matrix.

According to the literature, there is general agreement that suitability measures are defined by an Accuracy score over 90%, Specificity score over 85%, Type I Error score under 10%, Type II Error score under 10%, Recall score over 85%, Precision score over 85%, F measure score over 85%, and an AUC near to 1 (Chen et al., 2021; Demraoui et al., 2022; Karthiban et al., 2019; Lusinga et al., 2021; Li et al., 2017; Ndayisenga, 2021; Orji et al.; Peiris, 2022; Pimcharee & Surinta, 2022, Booker & Rebman, 2024).

Suitability measures are helpful in determining if a model is suitable for predicting future values. However, all the measures can be influenced by imbalanced class distributions. A model can receive an acceptable score by having a high TP and virtually no TN. It is for this limitation multiple measures are examined to determine a model's suitability. Because of the limitations of suitability measures, it is important to test the significant difference between predictive power.

To test the significant difference between each individual model and the ensemble models, a t-test was performed comparing error rates. The t-test is a fundamental statistical tool used to assess whether differences between two groups are statistically significant. It provides a rigorous method for researchers and analysts to compare means and make informed decisions based on empirical data by quantifying the difference between the groups relative to the variation within the groups.

Description of the Data

The data used in the study is from the credit card transactions for a card issuer based in Mississippi. The card issuer has over seventy million transactions. A random selection of 250,000 non-fraudulent transactions and 250,000 fraudulent transactions were selected from January 2015 to December 2023 transactions to be used for the training and testing of the models. 80% of the data was used for training and 20% used for testing. Although the selection was purposeful to ensure a balanced dataset, it must be noted that the overall data from the issuer is imbalanced as there are significantly more non-fraudulent transactions than fraudulent ones. Another random selection of 100,000 transactions were selected to validate the models. The features of the data set are shown in Table 4. All the features are numerical with used_security number, mobile purchase, online_order, repeat_retailer, and fraud as binary values and the others real numbers.

Table 4: Data set Features

Feature	Explanation
distance_from_home	Distance from the credit card holder zip code
mobile_purchase	1 if on a mobile device, 0 if not
online_purchase	1 if online purchase or not
distance_from_last_transaction	Distance from the last transaction
ratio_to_median_purchase_price	Ratio of the purchase to consumer purchases
repeat_retailer	Whether or not the consumer previously shopped with the retailer
time_between_transaction	Amount of time between transactions
used_security_number	Whether or not the consumer used the CVV or pin
Fraud	Whether or not the transaction was deemed fraudulent

A descriptive analysis for the variables with real numbers showed high kurtosis and skewness for each, indicating some outliers. Rather than eliminate the outliers, the variables were recoded using log e to normalize the data. The results of the recoding are shown in Table 5. The recoded data significantly reduced the skewness, normalizing the data.

Table 5. Descriptive Statistics for Variables with Real Number values recoded using log e.

Feature	*distance_from_ home_ln*	*distance_from_last_ transaction_ln*	*ratio_to_median_ purchase_price_ln*
Mean	2.33	0.03	0.32
Median	2.30	0.00	0.00
Standard Deviation	1.41	1.78	1.13
Sample Variance	1.96	3.24	1.21
Kurtosis	0.01	0.01	0.01
Skewness	0.01	0.01	-0.01
Range	12.28	12.19	10.18

Results

The study involved analyzing five individual machine learning models and ten ensemble models. Each model was developed using Python and analyzed using Excel using a Dell Precision 3660 with a 13th generation Intel core processor, 16 GB of memory, and 512 GB of storage, running the Windows 11 operating system. After the development of each model, confusion matrices and suitability measures were calculated to determine if the model met previously reported suitability standards. Suitability standards were calculated based on the training and test data sets. Next, each model was validated by comparing the original outcome with the predicted outcome of the entire dataset to compare how well the models performed against each other.

Recall the models studied: RF, FNN, RNN, CNN, NB, HOE, HEE1, HEE2, HEE3, HEE4, HEE5, HEE6, HEE7, HEE8, and HEE9. Fifty percent of the transactions in both the training dataset and test dataset are fraudulent. Also, recall the configuration of the ensemble models:

- **HOE**: RNN, FNN, CNN
- **HEE1**: RF, RNN, NB
- **HEE2**: RF, FNN, NB
- **HEE3**: RF, CNN, NB
- **HEE4**: RF, FNN, CNN
- **HEE5**: RF, FNN, RNN
- **HEE6**: RF, CNN, RNN
- **HEE7**: NB, FNN, CNN
- **HEE8**: NB, FNN, RNN
- **HEE9**: NB, CNN, RNN

The variable Fraud is the dependent variable. If the variable is coded as 0 it is non-fraudulent. If it is coded as 1 the transaction was fraudulent. Each model provides a predicted value between 0 and 1 for the variable Fraud. The predicted value determines whether a transaction is determined to be fraudulent or not based on a particular threshold. For example, if a predicted value was .68 and the threshold was .6 or .5, the prediction was coded as fraudulent whereas if the threshold was .7, the prediction was coded as not fraudulent. During the testing of the models, we evaluated many thresholds. A threshold of 0.5 had the best results for all models. Table 6 shows the percentage of each model predicted as fraudulent and non-fraudulent for the training data. Table 7 shows the percentage of each model in predicting the test data at the .5 threshold.

Table 6: Predicted Results for Training Data

Model	Threshold	
	0.5	
	Predicted Fraudulent	Predicted Not Fraudulent
RF	52%	48%
CNN	53%	47%
FNN	52%	48%
RNN	54%	46%
NB	56%	44%
HOE	51%	49%
HEE1	50%	50%
HEE2	52%	48%
HEE3	51%	49%
HEE4	51%	49%
HEE5	51%	49%
HEE6	50%	50%
HEE7	52%	48%
HEE8	52%	48%
HEE9	52%	48%

Table 7: Predicted Results for Testing Data

Model	Threshold	
	0.5	
	Predicted Fraudulent	Predicted Not Fraudulent
RF	52%	48%
CNN	52%	48%
FNN	52%	48%
RNN	55%	45%
NB	56%	44%
HOE	51%	49%
HEE1	50%	50%
HEE2	52%	48%
HEE3	51%	49%
HEE4	51%	49%
HEE5	51%	49%
HEE6	50%	50%
HEE7	52%	48%
HEE8	52%	48%
HEE9	52%	48%

Table 8 shows the performance of each model with the validation dataset. The validation dataset also had 50% fraudulent and 50% non-fraudulent transactions. Given the best performance for all models was with a .5 threshold, a .5 threshold was used to validate the models.

Table 8: Actual vs Predicted Results for Validation Data Set

Model	Threshold	
	0.5	
	Predicted Fraudulent	Predicted Not Fraudulent
RF	54%	46%
CNN	60%	40%
FNN	59%	41%
RNN	55%	45%
NB	71%	29%
HOE	51%	49%
HEE1	50%	50%
HEE2	51%	49%
HEE3	52%	48%

Model	Threshold	
	0.5	
	Predicted Fraudulent	Predicted Not Fraudulent
HEE4	51%	49%
HEE5	51%	49%
HEE6	50%	50%
HEE7	52%	48%
HEE8	52%	48%
HEE9	52%	48%

As discussed earlier, there are many measures to determine if a model is suitable to use as a predictive model, specifically, accuracy, specificity, Type I error, Type II error, Precision, F-measure, and area under the curve (AUC) and suitability measures are defined by an Accuracy score over 90%, Specificity score over 85%, Type I Error score under 10%, Type II Error score under 10%, Recall score over 85%, Precision score over 85%, F measure score over 85%, and an AUC near to 1. The suitability analysis is shown in Table 9.

Table 9: Suitability Analysis

Model	Suitability Measures (rounded to the nearest two decimal places)							
	Acc (>.90)	Pre (>.85)	Recall (>.85)	Spec (>.85)	F meas (>.85)	AUC (close to 1)	T I E (<.10)	T II E (<.10)
RF	0.93	0.86	0.90	0.94	0.86	0.68	0.10	0.10
CNN	0.84	0.79	0.78	0.84	0.49	0.67	0.02	0.08
FNN	0.86	0.74	0.76	0.74	0.28	0.63	0.04	0.03
RNN	0.82	0.72	0.81	0.63	0.61	0.66	0.02	0.09
NB	0.60	0.69	0.84	0.83	0.48	0.66	0.31	0.09
HOE	0.90	0.86	0.85	0.97	0.86	0.63	0.01	0.01
HEE1	0.95	0.64	0.89	0.88	0.92	0.94	0.10	0.10
HEE2	0.93	0.66	0.90	0.81	0.87	0.85	0.12	0.11
HEE3	0.92	0.68	0.87	0.80	0.86	0.84	0.18	0.16
HEE4	0.95	1.00	0.98	0.99	0.95	0.99	0.02	0.03
HEE5	0.92	1.00	0.95	0.99	0.91	0.95	0.02	0.02
HEE6	0.90	1.00	0.95	0.98	0.89	0.95	0.03	0.02
HEE7	0.89	0.81	0.89	0.93	0.98	0.89	0.11	0.09
HEE8	0.87	0.83	0.89	0.88	0.97	0.87	0.09	0.10
HEE9	0.88	0.82	0.91	0.81	0.97	0.86	0.11	0.11

Using the suitability measures, none of the individual models meet all the standards. RF meets most of the measures followed by the neural networks. The HOE meets all the measures except AUC. HEE4, HEE5, and HEE6 meet all the measures. HOE was comprised of all the neural networks. HEE4, HEE5, and HEE6 were comprised of at least two neural networks and RF.

Our final analysis was to compare the performance of the models using t-tests, comparing samples assuming equal variances. Table 10 shows the t-statistics for the test. Table 11 shows the results between the different models. High t-statistics indicate a substantial difference between the two models: the higher the value, the stronger the difference. All the t-statistics with a p-value of less than .01 are shown with an asterisk after the t-statistic. The neural network models showed strong differences between the individual model and the ensembles, indicating that there are different strengths in the individual neural networks that contribute to the improvements in the ensembles. Ensembles with RF had some of the lowest t-statistics between the individual model and the ensembles. This suggests that RF contributes strongly to the ensemble model with strong true positives or true negatives than its counterparts.

Table 10: T-test Results Between the Individual Models and the Ensemble Models

	RNN	FNN	CNN	RF	NB
HOE: RNN, FNN, CNN	57.346 *	57.518 *	56.544 *		
HEE1: RF, RNN, NB	29.123 *			22.647 *	35.838 *
HEE2: RF, FNN, NB		31.357 *		21.737 *	28.612 *
HEE3: RF, CNN, NB			16.214 *	22.390 *	28.737 *
HEE4: RF, FNN, CNN		31.618 *	63.911 *	22.113 *	
HEE5: RF, FNN, RNN	58.218 *	58.835 *		21.218 *	
HEE6: RF, CNN, RNN	62.691 *		51.357 *	21.814 *	
HEE7: NB, FNN, CNN		21.838 *	27.346 *		39.647 *
HEE8: NB, FNN, RNN	29.191 *	22.737 *			32.357 *
HEE9: NB, CNN, RNN	28.212 *		21.134 *		32.168 *

Table 11: T-test Results Between Ensemble Models

	HEE1	HEE2	HEE3	HEE4	HEE5	HEE6	HEE7	HEE8	HEE9
HOE	37.390 *	39.391 *	29.390 *	25.838 *	28.134 *	27.178 *	1.006	1.032	0.905
HEE1		11.191 *	12.141 *	38.121 *	39.135 *	37.910 *	21.191 *	23.146 *	26.193 *
HEE2			1.532	31.104 *	38.011 *	41.231 *	19.043 *	23.701 *	28.501 *
HEE3				29.131 *	15.697 *	16.687 *	27.903 *	26.757 *	19.773 *
HEE4					1.146	1.006	30.945 *	29.530 *	19.835 *
HEE5						1.032	17.218 *	20.153 *	24.787 *
HEE6								28.667 *	28.314 *
HEE7								0.569	0.323
HEE8									0.569

Within the ensembles, there was no significant difference between HEE4, HEE5, and HEE6, and between HEE7, HEE8, and HEE9. The results suggest that the ensembles that use both RF and NN outperform other combinations of ensemble and individual models. There was also a significant difference between the HOE and all the HEEs except HEE7, HEE8, and HEE9.

Recall our hypotheses:

H0: *Ensemble models will significantly outperform their individual models.*

H1: *The homogeneous ensemble model will significantly outperform the heterogeneous ensemble models.*

For H0, we can accept the hypothesis that all the ensemble models outperformed their individual models. However, for H1, we reject that the HOE outperformed the other models as there is no significant difference between the performance of the HOE and HEE7, HEE8, and HEE9. What this implies is that the neural network combinations are stronger together than they are alone in ensemble models, and that NB is not a strong contributor to this particular problem.
.

Conclusions, Limitations, and Next Steps

The purpose of this study was to examine if different combinations of ensemble models were better predictors of credit card fraud than their individual models, and if a homogenous model would outperform heterogenous models. The study applied the models to actual credit card transactions. As expected, the ensemble models outperformed their individual counterparts. The models with both the neural network and random forest models outperformed other ensembles except the homogeneous ensemble, validating that the underlying individual models are important when companies are planning to use ensembles to reduce approved fraudulent transactions.

This study has several limitations. While the models performed well, there is room for improvement. For example, there is a need to study if there are seasonal characteristics to the data as well as how well ensembles trained on seasonable data can capture those differences. Further, there are many supervised machine learning algorithms as well as semi-supervised and unsupervised algorithms. There is a need to truly delve into the different models to determine which situations are best suited to which types of ensemble models.

The next step for this study is to continue to examine the different types of individual models and build more ensembles to determine if (1) more individual models improve outcomes, (2) a more universal model can be developed, and (3) explore various ensemble models on a combination of credit card transactions instead of a single data set.

References

Armel, A. & Zaidouni, D. (2019). Fraud Detection Using Apache Spark. *Proceedings of the 5th International Conference on Optimization and Applications (ICOA), pp. 1–6,* Kenitra, Morocco.

Bewtra, A. (2022). The Ultimate Guide to Semi-Supervised Learning. https://www.v7labs.com/blog/semi-supervised-learning-guide.

Booker, Q. & Rebman, C. (2024). Applying Heterogeneous Ensemble Models to Detect Credit Card Fraudulent Transactions. *Proceedings of the 2024 Southwest Decision Sciences Conference (SWDSI),* Galveston, TX.

Brown, G. (2010). Ensemble Learning. *Encyclopedia of Machine Learning and Data Mining.* https://doi.org/10.1007/978-0-387-30164-8_252

Brown, S. (2021). Machine Learning, Explained. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained.

Carrasco, S. M. & Sicilia-Urban, M. A. (2020). Evaluation of Deep Neural Networks for Reduction of Credit Card Fraud Alerts. *IEEE Access*, (8*)*, pp. 186421–186432.

Chen, R.C., Luo, S.T., Liang, X., and Lee, V. (2005). Personalized Approach Based on SVM and ANN for Detecting Credit Card Fraud. *International Conference on Neural Networks and Brain, IEEE,* 810-815.

Chen, Z., Lin, G., & Jin, X. (2021). Credit Approval Prediction Based on Boruta-GBM Model. *2021 7th International Conference on Systems and Informatics (ICSAI)*. https://doi.org/10.1109/icsai53574.2021.9664026

Demraoui, L., Eddamiri, S., & Hachad, L. (2022). Digital Transformation and Costumers Services in Emerging Countries: Loan Prediction Modeling in Modern Banking Transactions. *Lecture Notes on Data Engineering and Communications Technologies*, 627–642. https://doi.org/10.1007/978-3-030-90618-4_32

Cunningham, P., Cord, M., & Delany, S. J. (2008). Supervised Learning. *Springer eBooks*, 21–49. https://doi.org/10.1007/978-3-540-75171-7_2

Cveticanin, N. 2022. "*Credit Card Fraud Statistics: What Are the Odds?*", https://dataprot.net/statistics/credit-card-fraud-statistics/

Gadi, M. F. A.; Wang, X.; do Lago, A. P. (2008). Credit Card Fraud Detection with Artificial Immune System. *International Conference on Artificial Immune Systems*, Springer, 119-131, 2008

Hsu, K.-L., Moradkhani, H., and Sorooshian, S. (2009), A Sequential Bayesian Approach for Hydrologic Model Selection and Prediction, Water Resource. Res., 45, W00B12

Jaafari, A., Panahi, M., Pham, B. T., Shahabi, H., Bui, D. T., Rezaie, F., & Lee, S., (2019). Meta Optimization of an Adaptive Neuro-fuzzy Inference System with Grey Wolf Optimizer and Biogeography-based Optimization Algorithms for Spatial Prediction of Landslide Susceptibility. *Catena*, (175), pp. 430-445.

James, G. M., Witten, D., Hastie, T., & Tibshirani, R. (2021). Unsupervised Learning. *Springer Texts in Statistics*, 497–552. https://doi.org/10.1007/978-1-0716-1418-1_12

Jiang, C., Song, J., Liu, G., Zheng, L., & Luan, L. (2018). Credit Card Fraud Detection: A Novel Approach Using Aggregation Strategy and Feedback Mechanism. *IEEE Internet Things J.*, (5:5).

Karthiban, R., Ambika, M., & Kannammal, K. E. (2019). A Review on Machine Learning Classification Technique for Bank Loan Approval. *International Conference on Computer Communication and Informatics*. https://doi.org/10.1109/iccci.2019.8822014

Kim, E., Lee, J. & Shin, H. (2019). Champion-challenger analysis for credit card fraud detection: Hybrid ensemble and deep learning. *Expert Systems with Applications*. 128. pp. 214–224.

Kokkinaki, A. I. (1997). On Atypical Database transactions: Identification of Probable Frauds Using Machine Learning for User Profiling. *Knowledge and Data Engineering Exchange Workshop, IEEE proceedings*, 107-113.

Kultur,Y. & Calayan, M. U. (2017). Hybrid Approaches for Detecting Credit Card Fraud. *Expert Systems*, 34(2), np–n/a. https://doi.org/10.1111/exsy.12191

Li, Z., Tian, Y., Li, K., Zhou, F., & Yang, W. (2017). Reject Inference in Credit Scoring Using Semi-Supervised Support Vector Machines. *Expert Systems with Applications*, *74*, 105–114. https://doi.org/10.1016/j.eswa.2017.01.011

Lusinga, M., Mokoena, T., Modupe, A., & Mariate, V. (2021). Investigating Statistical and Machine Learning Techniques to Improve the Credit Approval Process in Developing Countries. *AFRICON*. https://doi.org/10.1109/africon51333.2021.9570906

Mahesh, B. (2020). Machine Learning Algorithms- A Review. *International Journal of Science and Research (IJSR).[Internet]*, *9*, 381-386. https://t.ly/Scmc

Makki, S., Assaghir, Z., Taher, Y., Haque, R., Hacid, Y., & Zeineddine, H. (2019). An Experimental Study with Imbalanced Classification Approaches for Credit Card Fraud Detection. *IEEE Access*. (7), pp. 93010-93022

Malone, T., Rus, D., & Laubacher, R. (2020). Artificial Intelligence and the Future of Work. https://workofthefuture.mit.edu/wp-content/uploads/2020/12/2020-Research-Brief-Malone-Rus-Laubacher2.pdf

Maxwell, T. (2023) How major credit card networks protect customers against fraud. Bankrate.com

Minegishi, T. & Niimi, A. (2011). Detection of Fraud Use of Credit Card by Extended VFDT. *Internet Security (WorldCIS), World Congress on IEEE*, pp 152-159.

Ndayisenga, T. (2021). Bank Loan Approval Prediction Using Machine Learning Techniques. *[Doctoral dissertation, University of Rwanda]*. http://www.dr.ur.ac.rw/handle/123456789/1437

Olszewski, D. (2014). Fraud Detection Using Self-Organizing Map Visualizing the User Profiles. *Knowledge-Based Systems*, 70, 324-334.

Orji, U. E., Ugwuishiwu, C. H., Nguemaleu, J. C. N., & Ugwuanyi, P. O. (2022). Machine Learning Models for Predicting Bank Loan Eligibility. *2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON)*. https://doi.org/10.1109/nigercon54645.2022.9803172

Peiris, M. P. C. (2022). *Credit Card Approval Prediction by Using Machine Learning Techniques* [Doctoral dissertation, University of Colombo School of Computing]. https://dl.ucsc.cmb.ac.lk/jspui/handle/123456789/4593

Petrakova, A., et al. (2015). Heterogeneous versus Homogeneous Machine Learning Ensembles. Information Technology and Management Science, 18(1), 135–140.

Pimcharee, K., & Surinta, O. (2022). Data Mining Approaches in Personal Loan Approval. *Engineering Access*, *8*(1), pp. 15-21. doi: 10.14456/mijet.2022.2. https://ph02.tci-thaijo.org/index.php/mijet/article/view/244392

Polikar, R. (2012). Ensemble learning. Ensemble Machine Learning, 1-34. https://doi.org/10.1007/978-1-4419-9326-7_1

Prusti, D. & Rath, S.K. (2019). Web Service-Based Credit Card Fraud Detection by Applying Machine Learning Techniques. *Proceedings of the TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, Kochi, India, pp. 492-497.

Rushin, C., Stancil, M., Sun, S., Adams, & P. Beling, (2017). Horse Race Analysis in Credit Card Fraud—Deep Learning, Logistic Regression, and Gradient Boosted Tree. *Systems and Information Engineering Design Symposium (SIEDS)*, pp. 117–121.

Sadgali, I., Sael, N., & Benabbou, F. (2019). Fraud Detection in Credit Card Transactions Using Neural Networks. *Proceedings of the 4th International Conference on Smart City Applications (SCA '19). Association for Computing Machinery, New York, NY, USA*, pp. 1-4.

Seeja, K. & Zareapoor, M.(2014). Fraudminer: A Novel Credit Card Fraud Detection Model Based on Frequent Item Set Mining. *The Scientific World Journal*, ID 252797, 2014.

Singh, A. & Jain, A. (2020). An Empirical Study of AML Approach for Credit Card Fraud Detection-Financial Transactions. *International Journal of Computers Communications & Control*, vol. 14, no. 6, pp. 670–690, 2020.

Sohony, I., Pratap, R., & Nambiar, U. (2018). Ensemble Learning for Credit Card Fraud Detection. *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data Association for Computing Machinery*, pp. 289-294.

Stehman, S. V. (1997). Selecting and Interpreting Measures of Thematic Classification Accuracy.

Tran, P. H., Tran, K.P., Huong, T.T., Heuchenne, C., HienTran, P., & Le, T.M.H.. (2018). Real Time Data-Driven Approaches for Credit Card Fraud Detection. *Proceedings of the 2018 International Conference on E-Business and Application. Association for Computing Machinery*, New York, NY, USA, pp. 6-9.

Yalong Xie, Aiping Li, Liqun Gao, & Ziniu Liu. (2021). A Heterogeneous Ensemble Learning Model Based on Data Distribution for Credit Card Fraud Detection. *Wireless Communications and Mobile Computing*, *2021*. https://doi.org/10.1155/2021/2531210

Zareapoor, M.; Seeja, K.; Alam, M. A.(2012). Analysis on Credit Card Fraud Detection Techniques: Based on Certain Design Criteria. *International Journal of Computer Applications,* 52 (3), 35-42, 2012