HIGH SCHOOL EDITION
Journal of Student Research
Volume 11 Issue 3 (2022)

Banknote Authentication Using Logistic Regression and Artificial Neural Networks

Alexander Wang[1], Dr. Guillermo Goldsztein[2#] and Dr. Zhaonan Sun[#]

[1]West Windsor – Plainsboro High School North
[2]Georgia Institute of Technology
[#]Advisor

ABSTRACT

Banknotes are special notes authorized by the government and carry monetary value. As a result, there are incentives for criminals to create counterfeit. The goal of this study is to create models using machine learning techniques that can accurately classify a banknote as authentic or fake. The methods used were logistic regression and artificial neural networks. An open-source data set was obtained and split into 7 sub-datasets, and multiple models were created to model the data. There was a total of 7 logistic regression models, each corresponding to one of the 7 sub-datasets. Additionally, an artificial neural networks model was used on the 7th sub-dataset. Both the neural networks model and the logistic regression model achieved accuracies greater than 99%.

Introduction

Even as more and more people are transferring to digital payments, banknotes are still an important part of our lives (Marino, 2021). Banknotes are a representation of the legal currency of a country/region, and they typically have security features that allow them to be identified as authentic. In the United States (US), only the Federal Reserve is allowed to print banknotes, helping ensure that all authentic banknotes are consistent and difficult to replicate (Kagan, 2021).

Banknotes provide users with a convenient way of payment. Because of their widespread use (they are a very common form of currency, with 50.3 billion US dollar notes in circulation as of 2020 (Federal Reserve Board, 2021)), they are a crucial part of our financial system.

Introduction to Banknote Authentication and the Goals of this project

Banknote authentication is the process of analyzing a banknote to determine whether it is authentic or a forgery. Traditionally, checking if a banknote is real or not involves carefully examining the texture, serial numbers, its security features, and comparing it with other banknotes (Lewis, 2022). Not only is this process limited in efficiency, but there can also be inevitable human error where well-forged banknotes go undetected.
In comparison, machine learning-based methods bypass both obstacles. Not only are they faster, but a well-trained model is also more accurate when it comes to detecting forgery.

The goal of this research project is to create a model that can accurately differentiate between authentic banknotes and fake banknotes.

Some methods that other people have used include artificial neural networks (ANN) and back propagation neural networks (BPNN). Shahani et al. (2018) found that back propagation artificial neural networks is the best method. Ravi Kumar et al. (2018) found that back-propagation neural networks with Latent Dirichlet Allocation

(LDA) yields the best results. Another study by Yasar et al. (2016) found that artificial neural networks achieve results higher than 99%.

An Introduction to Artificial Neural Networks and Logistic Regression

Artificial Neural Networks

Artificial Neural Networks and Logistic Regression are both popular machine learning techniques. In artificial neural networks, an input layer, hidden layers, and an output layer are used (figure 1).

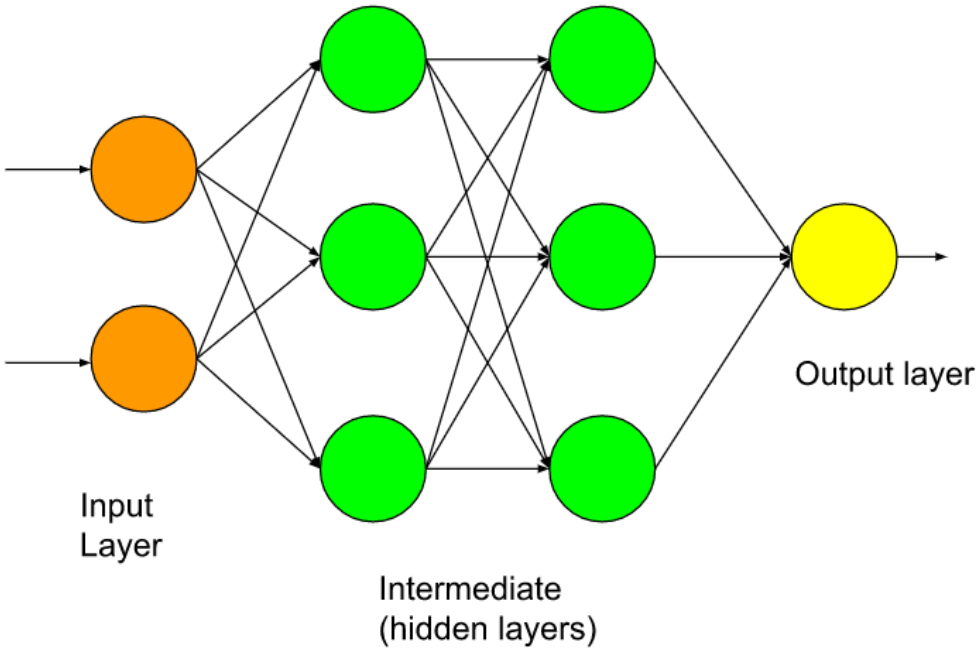

Figure 1. An example of an artificial neural network with 2 intermediate layers. As seen in figure 1, the artificial neural network has an input layer, intermediate layers (hidden layers), and an output layer.

The input propagates through the different intermediate hidden layers of the model, and the goal of neural networks is to find a relationship between the input and the output. If an artificial neural network model has more nodes, then it is better able to model the data.

Logistic Regression

Logistic regression is a machine learning method that takes data and categorizes data into 2 categories. Logistic regression takes in the input and parameters and outputs a probability that is between 0 and 1 (Figure 2). The formula for logistic regression is

$$P = \frac{e^{a+bX}}{1 + e^{a+bX}}$$

where P is the probability, e is the base of the natural logarithm, X is the input, and a and b are parameters (Brannick, n.d.). If the output probability is greater than 0.5, it is categorized as the category for 1; on the other hand, if the output probability is less than 0.5, it is categorized as the category for 0.

Logistic regression is a type of sigmoid, which is a type of function that takes numbers and maps them to a probability between 0 and 1 (Naeem, n.d.). The formula for the sigmoid function is

$$P = \frac{1}{1 + e^{-x}}$$

where P is the probability and e is the base of the natural log (Naeem, n.d.).

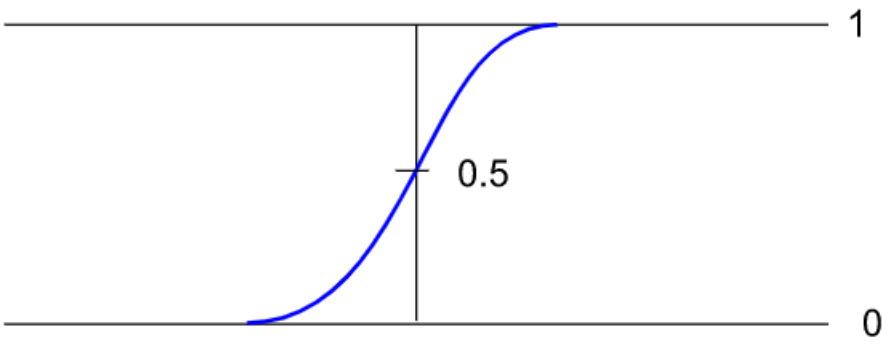

Figure 2. The sigmoid function curve. As seen in figure 2, logistic regression returns a probability that's between 0% and 100% (0 and 1). The halfway point is marked by 0.5, signifying 50%. If the probability is above 0.5, then it will be classified as 1. Otherwise, if it's below 0.5, then it will be classified as 0.

Method

Information about the data

A dataset on banknote authentication was obtained from UCI ML Repository dataset (Saluja, 2018). The data is in the form of a CSV file. The dataset's features are variance, skewness, kurtosis, entropy, and class. The predictors are secondary variables that were extracted from the images of banknotes. More specifically, variance refers to how the pixels vary compared to their neighboring pixels (Rooks, 2022). Skewness refers to the symmetry of the image (Caban, 2010). Kurtosis works with noise reduction of an image (Image Engineering, 2011), and entropy refers to randomness in an image (Rooks, 2022). Class categorizes whether the image is an image of an authentic banknote or a fake banknote. The data for variance, skewness, kurtosis and entropy are continuous while class is either a 1 or a 0, which categorizes the banknote as authentic or fake.

The data was extracted from 1372 images taken from genuine and forged banknote-like specimens. An industrial camera usually used for print inspection was used for digitalization. The gray-scale pictures have a resolution of about 660 dpi and are 400 by 400 pixels. Wavelet Transform tool was used to extract features from images. (Saluja, 2018).

Sub-dataset usage

The data was split into multiple sub-datasets. The data was split into separate sub-datasets to understand better the size and shape of the data, the clustering of the data, and possible unexpected, interesting aspects of the data. Each sub-dataset contained only a few selected features and all the data points of the images that are contained in the selected features. Another sub-dataset containing all the features acts as a control. This is to determine if any of the features are unnecessary and whether feature selection was needed. If the model that used the sub-dataset which contained all the features had a higher accuracy than each of the other sub-datasets' models that used only select features, then all the features are important and feature selection is not needed.

The dataset was split into 7 sub-datasets. The first 6 sub-datasets each only included 2 out of the 4 features, while the 7th sub-dataset contained all the features. For the first 6 sub-datasets, all distinct pairs of features were chosen, creating all possible combinations by choosing two random features from four features. This created a total of 6 sub-datasets with only 2 features, which correspond to 6 logistic regression models. Finally, the 7th dataset that contained all the features was used twice; it was first used to create another logistic regression model before it was used again to create a model using artificial neural networks.

Models

For each logistic regression model, the two categories are authentic and fake, and the activation function is sigmoid. In the artificial neural network model, there are a total of 4 layers: 1 input layer, 2 intermediate layers, and 1 output layer. The 1st intermediate layer has 15 nodes and has an activation function of relu. The 2nd intermediate layer has 3 nodes and also has an activation function of relu. The output function of the neural networks model has an activation function of sigmoid.

Matplotlib and seaborn were used to plot the data so it can be separable with a line. Sklearn was used to split the data into a training set and a validation set, and TensorFlow.Keras was used to create the model.

For each sub-data set, the training and validation set was split into a ratio of 3:1. The seed used in the random generator before splitting the dataset is 4.

SkLearn.metrics was used to evaluate the model. The most important factor of our metrics is our accuracy, which is the number of correct predictions over the number of predictions. The formula for accuracy is

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

where TP is true positives, TN is true negatives, FP is false positives, and FN is false negatives (MyDataModels, 2021).

True positives are positives that are predicted to be positive, true negatives are negatives that are predicted to be negative, false positives are negatives that are predicted positives, and false negatives are positives that are predicted negatives. TP, TN, FP, and FN can be summarized in the confusion matrix (Table 1).

Table 1. Confusion Matrix showing TP, TN, FP, and FN.

	Predicted Value		
Actual Value		**Positive**	**Negative**
	Positive	TP	FN
	Negative	FP	TN

Results

After conducting a series of tests, the following results were obtained. The first sub-dataset only contained the features of variance and skewness (figure 3). The model used logistic regression. The model returned an accuracy of 86% on the training set and 88% on the validation set (Table 2).

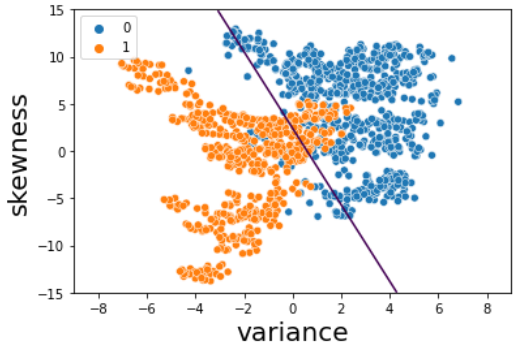

Figure 3. Logistic regression on the first sub-dataset where only skewness and variances were used to predict banknote authenticity. Each data point's coordinates correspond to a certain variance and skewness, and its color corresponds to its authenticity (results are shown in the figure). Orange signifies that the banknote with a specific variance and skewness is authentic and blue means that the banknote is fake. The line graphs the 50% probability mark from logistic regression where it categorizes future predictions of whether the banknote is authentic or fake.

Table 2. The accuracies of the logistic regression model on the training set and the validation set of the 1st sub-dataset.

Training set:	**Accuracy**
	0.86
Validation set	0.88

The second sub-dataset contained the features of variance and kurtosis (figure 4). The model used logistic regression.

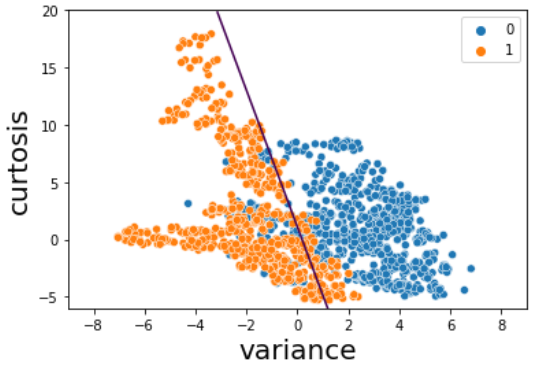

Figure 4. Logistic regression on the second sub-dataset where only kurtosis and variances were used to predict banknote authenticity (results are shown in the figure). Each data point's coordinates correspond to a certain variance and kurtosis, and its color corresponds to its authenticity. The method used to obtain the line that predicts whether a banknote is authentic is the same as before.

Table 3. The accuracies of the logistic regression model on the training set and the validation set of the 2nd sub-dataset.

Training set:	**Accuracy**
	0.87
Validation set	0.90

The third sub-dataset contained the features of variance and entropy (figure 5). The model used logistic regression.

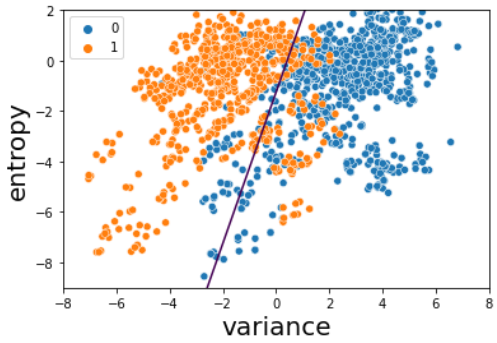

Figure 5. Logistic regression on the third sub-dataset where only entropy and variances were used to predict banknote authenticity (results are shown in the figure). Each data point's coordinates correspond to a certain variance and entropy, and its color corresponds to its authenticity. The method used to obtain the line that predicts whether a banknote is authentic is the same as before.

Table 4. The accuracies of the logistic regression model on the training set and the validation set of the 3rd sub-dataset.

Training set:	**Accuracy**
	0.88
Validation set	0.86

The fourth sub-dataset contained the features of skewness and kurtosis (figure 6). The model used logistic regression.

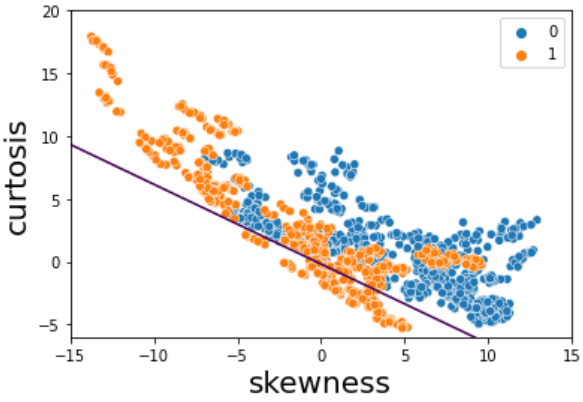

Figure 6. Logistic regression on the fourth sub-dataset where only kurtosis and skewness were used to predict banknote authenticity (results are shown in the figure). Each data point's coordinates correspond to a certain kurtosis and skewness, and its color corresponds to its authenticity. The method used to obtain the line that predicts whether a banknote is authentic is the same as before.

Table 5. The accuracies of the logistic regression model on the training set and the validation set of the 4th sub-dataset.

Training set:	**Accuracy**
	0.76
Validation set	0.76

The fifth sub-dataset contained the features of skewness and entropy (figure 7). The model used logistic regression.

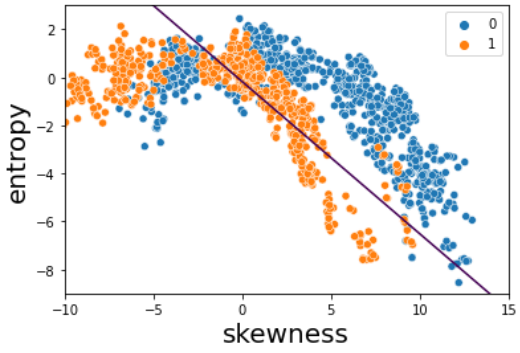

Figure 7. Logistic regression on the fifth sub-dataset where only kurtosis and skewness were used to predict banknote authenticity (results are shown in the figure). Each data point's coordinates correspond to a certain entropy and skewness, and its color corresponds to its authenticity. The method used to obtain the line that predicts whether a banknote is authentic is the same as before.

Table 6. The accuracies of the logistic regression model on the training set and the validation set of the 5th sub-dataset.

Training set:	**Accuracy**
	0.73
Validation set	0.70

The sixth sub-dataset contained the features of kurtosis and entropy (figure 8). The model used logistic regression.

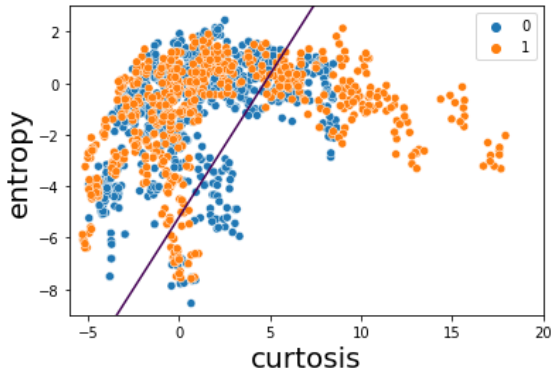

Figure 8. Logistic regression on the sixth sub-dataset where only kurtosis and skewness were used to predict banknote authenticity (results are shown in the figure). Each data point's coordinates correspond to a certain kurtosis and entropy, and its color corresponds to its authenticity. The method used to obtain the line that predicts whether a banknote is authentic is the same as before.

Table 7. The accuracies of the logistic regression model on the training set and the validation set of the 6th sub-dataset.

Training set:	**Accuracy**
	0.59
Validation set	0.60

The seventh and last sub-dataset contained all 4 features. The model used logistic regression.

Table 8. The accuracies of the logistic regression model on the training set and the validation set of the 7th sub-dataset.

Training set:	**Accuracy**
	0.98
Validation set	0.99

Additionally, another model was created using the 7th sub-dataset and it used artificial neural networks instead of logistic regression.

Table 9. The accuracies of the artificial neural networks model on the training set and the validation set of the 7th sub-dataset.

Training set:	**Accuracy**
	1.00
Validation set	1.00

Discussion

Analysis of Data

Results suggest that the artificial neural network model is the most accurate out of all the models. It is a perfect model with an accuracy of 100%, compared to the most accurate logistic regression model which has an accuracy of 99%. The neural network model has the highest accuracy because it has a significant number of nodes, so it can model complicated data more accurately.

The logistic regression model with all 4 features is more accurate than each of the logistic regression models with only 2 features because the logistic model has more features to work with, thus allowing it to consider all the factors when classifying each banknote. Additionally, the accuracy of sub-dataset 7 (the sub-dataset with all the features) is greater than each of the accuracies of subsets 1 through 6 (sub-datasets with only 2 out of the 4 features), so no feature selection is necessary.

The lowest accuracies are from sub-dataset 6 (table 7), with an accuracy of 59% on the training set and a 60% accuracy on the validation set. The scatterplots of sub-datasets 1, 2, and 3 are relatively clustered into an authentic and a fake group, while the scatterplots of sub-datasets 4, 5, and 6 are relatively mixed. This means that the logistic regression function returns a higher accuracy for sub-datasets 1, 2, and 3 compared to 4, 5, and 6.

Conclusions and Future Study

The unexceptional accuracies of sub-datasets 1 through 6 compared with the near-perfect accuracies of 7 and 8 shows that 2 features are not enough to properly distinguish between fake and authentic banknotes, but a model that uses all 4 features is enough to determine a banknote's authenticity. This means theoretically with any given picture of a banknote, the only features necessary to determine its authenticity are variance, skewness, kurtosis, and entropy.

It is suggested that future researchers focus on improving the accuracy of the logistic regression model. This is because the logistic regression model is more lightweight than the neural networks model, and the model can be more easily distributed to banks and ATMs across the country. Another suggestion is to improve the size of the dataset to include banknotes from other countries to find general trends of counterfeit money. The dataset should be updated regularly to keep ahead of counterfeiters who will try to create ever more realistic banknotes.

Acknowledgements

I would like to thank Dr. Guillermo Goldsztein, a professor at Georgia Tech, for guiding me through the research of this project. I am also extremely grateful for Dr. Zhaonan Sun instructing and supporting me throughout the construction of this paper. This project and paper would not have been possible without them, and I truly appreciate their help.

References

Caban, J. J. (2010, September 29). *Introduction to image statistics*. Image Statistics. Retrieved August 19, 2022, from https://www.csee.umbc.edu/~caban1/Fall2010/CMSC691/Schedule_files/Docs/08-ImageStatistics.pdf

Federal Reserve Board., U. S. (2021, April 21). Currency and Coin Services. Federal Reserve Board - Currency in Circulation: Volume. Retrieved August 4, 2022, from https://www.federalreserve.gov/paymentsystems/coin_currcircvolume.htm

Image Engineering. (2011, October 28). What is kurtosis? Retrieved August 4, 2022, from https://www.image-engineering.de/library/technotes/740-what-is-kurtosis

Kagan, J. (2021, September 8). *What is a banknote?* Banknote Definition. Retrieved August 4, 2022, from https://www.investopedia.com/terms/b/banknote.asp

[Kumar, G. R., & Nagamani, K. (2018). Banknote authentication system utilizing deep neural network with PCA and LDA machine learning techniques. International Journal of Recent Scientific Research, 9(12), 30036-30038.

Brannick, M. T. (n.d.). *Logistic Regression*. Logistic regression. Retrieved August 19, 2022, from http://faculty.cas.usf.edu/mbrannick/regression/Logistic.html

Marino, K. (2021, July 16). *The pandemic fueled the decline of cash*. Axios. Retrieved August 19, 2022, from https://www.axios.com/2021/07/16/legal-cash-economy-decline-pandemic

MyDataModels. (2021, April 6). *How good is your machine learning algorithm?* MyDataModels. Retrieved August 19, 2022, from https://www.mydatamodels.com/learn/how-good-is-your-machine-learning-algorithm/

Naeem, A. (n.d.). *What is sigmoid and its role in logistic regression?t*. Educative. Retrieved August 19, 2022, from https://www.educative.io/answers/what-is-sigmoid-and-its-role-in-logistic-regression

Lewis, M. R. (2022, June 27). *4 ways to detect counterfeit US money*. wikiHow. Retrieved August 4, 2022, from https://www.wikihow.com/Detect-Counterfeit-US-Money

Rooks, M. (2022, February 10). What Is Entropy In Image Processing? Retrieved August 19, 2022, from https://www.icsid.org/uncategorized/what-is-entropy-in-image-processing/

Rooks, M. (2022, February 22). What Is Variance Image Processing? Retrieved August 19, 2022, from https://www.icsid.org/uncategorized/what-is-variance-image-processing/

Saluja, R. (2018, November 30). *Bank note authentication UCI Data*. Bank Note Authentication UCI data | Kaggle. Retrieved August 4, 2022, from https://www.kaggle.com/datasets/ritesaluja/bank-note-authentication-uci-data

Shahani, S., Jagiasi, A., & Priya, R. L. (2018). Analysis of Banknote Authentication System using Machine Learning Techniques. International Journal of Computer Applications, 975, 8887.

Yasar, A., Kaya, E., & Saritas, I. (2016). Banknote classification using artificial neural network approach. International Journal of Intelligent Systems and Applications in Engineering, 4(1), 16-19.