Bank Notes Classification: A Study of British Pound Counterfeit Notes Detection Using Binary Model Classification

Onyemelonu Nneoma Charity

Abstract—Physical cash has been a vital tool in business transactions and trades, however, there's been an impeding problem that has been affecting the finance industry and the general society. Some smart individuals have decided to go their way in creating replicas of these bank notes. Although these replicas may look similar to the original ones at a first glance, there are still some hidden features that differentiate the real from the fake, this is as a result of the different printing machines/models. In this research, a binary classification model was developed to best differentiate the real and fake features of bank notes. The dataset for building this model was made up of 1372 records of different bank notes with four inputs named Variance, Skewness, Kurtosis and Entropy respectively and an output -Authenticity. The study aims to explore the best binary clasification model that can be used to detect counterfeit banknotes in our present day This was achieved by the research hypothesis that 'There is a significant differences between the both kinds of notes. The paper also highlights some relatable past works of various researchers that have developed different methods to tackle this, in the field of Data Science and Artificial Intelligence.

Index Terms—Real and Fake Bank Notes, Binary Classification Model, SVMs, KNN, Linear Regression,

1 Introduction

Generally, in the real sense, banks go through a lot just to handle the issues faced with counterfeit notes. Moreso, in an environment like the United Kingdom, which has a population of citizens and immigrants on the rise, [1] these foreigners are not able to figure out the difference between the two due to ignorance, not enough prior experience or knowledge. There has been a need to develop and improve past models for this detection. Past detection methods have been the application of IR sensors, Discrete Wavelet Transform, Electron Microscopy, Spectrum Analysis, etc. Currently, the state of research concerning this has been on image processing, machine learning and deep learning neural network where the act of viewing and recognition is replicated into a neural network architecture and comparisons are made to already existing objects from memory or data. However, as the world becomes innovative and technology keeps taking the order of the day, so do these culprits. Hence requiring the Banking industry to top up its game. To best describe how This model will go a long way in tackling this issue as of the present, the research paper was developed in the scopes of identifying, referring, proving and concluding.

Binary Classification has been used for different purposes so far. one best definition of this kind of classification model was offered by Saurab (2017) whom he stated "binary classification is the process of classifying given document/account based on predefined classes". Normally, Classification Model types are determined by the training they undergo with data already classified into known classes. When there are two known classes it is considered binary, but when they are more than two then it can be termed Multi-Classification Model (Vedant et al 2020).

After referencing past relatable works in the literature review that best describes the different binary model classifiers, the methodological section went further to explore the analysis of the data, the steps taken to pre-process the data and how the spliting was carried out. It mentions also all the models developed and their tests for accuracy based on the two known classes of Real and Fake. Using classifiers like Random Forest and K-Nearest neighbour, the entire process will be based on answering the research question, - is there a significant difference between real and fake currency notes? This will be achieved by obtaining an accuracy score from actual and predicted positives and negatives (True Positives, True Negatives, False Positives, and False Negatives).

Followed by the results, and discussion that best portrays the end product of all the activities carried out, a deciding conclusion is generated. It is evident to say that this solution will help Banks tackle this impeding issue, it will also reduce the fall of fake currency notes into the hands of people who know little or nothing about it, for they are not to be reprimanded wrongly. When there is a rise in the detection, this will inflict a decrease in their production.

2 Literature Review

As long as Banknotes have been in use, it has been a medium of exchange for goods and services. Portraying that although they are stored in the Bank, it is even more in use in the outdoor environment. Shopping malls and large firms all make use of cash payment systems. The recognition of polymer notes is based on pattern recognition

• *N. Onyemelonu, Department of Computer Science and Electronic Engineering, Ubiversity of Essex, England. 2023.*
E-mail: no22670@essex.ac.uk
CE880 Assignment, Jan 2023

[4] where shapes, lines, inscriptions, and quality are used to identify the originality. All of these contribute to the features that make a dataset that can be trained to build a model that'll best identify real from fake notes. Currently, machine learning models have taken the rise in Big data analysis [5] and since forth generate models according to the kind of data, one of which is Binary Classifications, implemented only in cases of deducing true or false accuracies. Just as in this case study, they can be also be used for Health Diagnosis.

There are different kinds of Binary Model Classifier [6] These models are usually dependent on the features available on the set of provided data. Data must have been pre-processed via cleaning, and normalization. For better performance this dataset is usually split into a train and test data, to enable the application of the model not just on. The train data can be further split into a validation data to best predict and generate an expected accuracy score.

Some of the Binary Classification models imported in this project were: Logistic Regression: a kind of binary classification model aimed at indicating the relationship between the inputs and target data, otherwise known as independent and dependent variables. It is best for testing hypotheses concerned about the relationship of one variable to another[7].

Decision Tree Classifier: This is another classifier with the ability to reduce complex decision making-stages to simpler and easily interpreted decisions which when united generates a result that seems almost likely with the intended objective[8].

Random Forest Classifier: This considers the combination of a decision tree and a support vector machine. These trees are generated using a random vector and every tree casts a unit vote for the most popular class to be classified as the input variables[9]. K-Nearest Neighbour: This is a probabilistic distribution classifier that attributes s a decision-making rule to unclassified sample points in a featured space(cover)[10].

Support Vector Machines: This is a weighted sum of the support vectors that handles classification using parametric techniques. It develops its solution in subset of the training input [11]. All of these to best make a choice of the best kind of model that fits the dataset appropriately for the purpose of the underlying research hypothesis

Some other researchers have deemed it fit to use some other forms of programs for detection. One which has lingered, like the implementation of a neural network by Fumiaki et al [12], where Thai banknotes were scanned to images and saved as bitmap data leading to the extraction of each feature found therein for learning. In this case, front and backward propagation was carried out, and analysis of perceptrons was also considered. Some other forms of extraction like fast wavelet transforms were implemented by Eugene and Voldker[13], leading to a classification of four different classes namely, Genuine, High-Quality Forgery, Low-Quality Forgery, and Inappropriate ROI. This gave a different depth of banknotes authentication, identifying that these notes cannot only be classified into two but can be explored further into arriving at other features that can be considered in identifying the level of forgery on a set of notes. This model developed a hundred per cent accuracy in performance.

However, there were still some other researchers that have been bent on the analytical features of machine learning models and have gone out of their way to prepare better models that have carried out this detection successfully. One of whom is Yeh et al[14], who developed a multi-support vector machine to handle the recognition of fake banknotes. This SVM model was poised at reducing the rates of falsehood. An experiment of the model was carried out on the Taiwanese banknotes where the performance rate was higher than that of a single-kernel Support Vector Machine.

3 METHODOLOGY

3.1 Data Collection

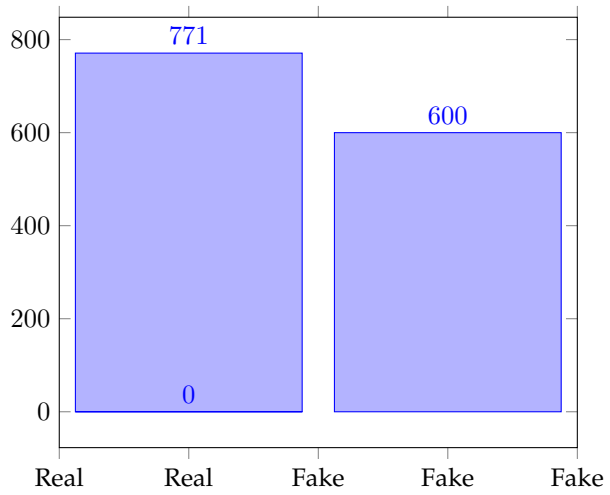
Although a text file, the data (already generated from the past and shared by Dr Raza) was imported into a data frame using pandas, having 1372 rows and 5 columns. These columns were renamed to Variance, Skewness, Kurtosis, Entropy, and Authenticity of all the columns respectively, where the first 4 columns represented the input data and the last column – authenticity was the target data having row values of 0 representing Real banknotes and 1 representing Fake banknotes.

TABLE 1
Description of the Dataset

Features Info	Functions to the Bank Notes	Data Type
Variance	Explores how one pixel variates from another pixel on a bank note[15]	Continuous
Skewness	Measures the absence of symmetry[on each bank note]	Continuous
Kurtosis	Measures the tail weights of the data relatively to the normal distribution[16]	Continuous
Entropy	Describes the quantity of bits needed to encde image data	Continuous
Authenticity	Portrays the genuineness of the banknote(real or fake) [17]	Integer

The data was then analytically explored to view relationships, correlations and

Fig 1 Number of Fake and Real Bank Notes



From the matplotlib library, a histogram was used to represent the distribution of features on each column, including the target column. Thereafter a heat map was drawn to showcase the correlation each column had on another where correlation is regarded as the [18]. A pair plot too was used to showcase the similarities and relationship among the features of the data[19]

3.2 Data Pre-processing

Using minmaxscaler from the sklearn library, The Data was pre-processed by fixing the outliers found in the kurtosis and entropy columns by scaling and normalizing the data. Thereafter, attributing the input and target columns to x and y respectively. This was then splited into a train and test sets with a test size of 0.2 The train set was further splited into two to generate a new train set (that will be used all throughout the model implementation as the training data) and a validation set with a test size of 0.1. All of these were done using the $\text{train}_t est_s plit function$.

3.3 Model Implementation

Importing the different models from sklearn, the implementation of the different binary classifiers began with the Linear Regression model that resulted in an accuracy of 0.97, 0.96, 0.95 on the Train, Test and Validation data respectively. Thereafter the Decision Tree Classifier with results 1.0, 0.97, 0.97. The Random Forest Classifier also came up with similar values of 1.0, 0.98, and 0.99. However, when the KNN model and SVMs were trained, the resulting accuracy output was 1.0, 1.0, 1.0.

4 Data Analysis

4.1 Results

After 100 percent accuracy scores were achieved by the train, validation and test data of the KNN and SVMs models, the performance was measured using a confusion matrix where a prediction was made on the test data, and the resulting metrics of f1 score, recall score and precision scores were outputted. These were deduced by computation of True Positive (TP) Rate, False Positive (FP) Rate, True Negative (TN) Rate and False Negative (FN) Rate.

A classification report on both the SVM and KNN models helped presented this in a readable format of:

Actual Class	**Predicted Class**	
	Positive	Negative
Positive	TP = 148	FP =0
Negative	FP=0	TN =127

	precision	recall	f1-score	support
Class 0	1.00	1.00	1.00	148
Class 1	1.00	1.00	1.00	127
accuracy		1.00	1.00	275
macro avg	1.00	1.00	1.00	275
weighted avg	1.00	1.00	1.00	275

4.2 Discussion

The main purpose of this research is to figure out if there is a significant difference between the fake and real British pound notes as derived from the 1372 observations given. From the results obtained, it can be deduced that this was true. Poising from the two best models, a decision was made on choosing SVM. Generally, SVMs are known for how robust they are and their ability to generalize easily and the resulting solutions they offer[20]. Moreover, SVM is now one of the most powerful classifiers in machine learning, and because it gave the required output in this case study, it'll be accepted as the best classifier.

4.3 Conclusion

In essence, to help curb the circulation of fake notes was this model developed. As much as there are improvements that need to be done in the banking sector, concerning how these culprits get hold of some essential data that facilitates their forgery of bank notes, this model can stay in the way of detecting these forged notes. Some recommendations may be on the size of the data. For better analysis, it is advised that a wide range of data be used as these culprits tend to improve daily and take all measures to avoid their latest prints go detected. This is a result of the steady rise in technology of our present day.

Appendix A
Tables

Table 1.1 Description of the Dataset

Appendix B
Graphs

Fig 1.1 Graphical representation of the fake and real notes

References

[1] Research Briefing Migration Statistics , *UK Parliament House of Common Library* LaTeX, 2022.
[2] Saurabh Kumar Srivastava "Machine Learning: A Review on Binary Classification", International Journal of Computer Applications (0975 – 8887) Volume 160 – No 7, February 2017
[3] Vedant Bahel Sofia Pillai and Manit Malhotra ''A Comparative Study on Various Binary Classification Algorithms and their Improved Variant for Optimal Performance". DOI: 10.1109/TENSYMP50017.2020.9230877 Conference: 2020 IEEE Region 10 Symposium (TENSYMP) 2020
[4] Yueqiu Ren, ''Bank Notes Recognition in Real Time using School of Engineering ANN", Computer and Mathematical Science Auckland University of Technology, 2017.

[5] Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., Muharemagic, E. ''Deep learning applications and challenges in big data analytics.'' Journal of Big Data, 2, 1. 2015.
[6] Patil, Tina R., and S. S. Sherekar. ''Performance analysis of Naive Bayes and J48 classification algorithm for data classification.'' International journal of computer science and applications 6, no. 2 256-261. 2013.
[7] Peng, Chao-Ying Joanne, Kuk Lida Lee, and Gary M. Ingersoll. ''An introduction to logistic regression analysis and reporting.'' The journal of educational research 96, no. 1 3-14. 2002.
[8] Safavian, S. Rasoul, and David Landgrebe. ''A survey of decision tree classifier methodology.'' IEEE transactions on systems, man, and cybernetics 21, no. 3 660-674. 1991.
[9] Pal, Mahesh. ''Random forest classifier for remote sensing classification.'' International Journal of Remote Sensing 26, no. 1 217-222 2005.
[10] Cover, Thomas M., and Peter E. Hart. ''Nearest neighbor pattern classification.'' IEEE transactions on information theory13, no. 1 21-27 1967.
[11] Downs, Tom, Kevin E. Gates, and Annette Masters. "Exact Simplification of Support Vector Solutions." Journal of Machine Learning Research 2 : 293–297, 2002.
[12] Fumiaki Takeda, Lalita Sakoobunthu and Hironobu Satou, "Thai Banknote Recognition Using Neural Network and Continues Learning by DSP Unit", International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, 2003.
[13] C.-Y. Yeh, W.-P. Su, and S.-J. Lee, "Employing multiplekernel support vector machines for counterfeit banknote recognition," Applied Soft Computing, vol. 11, no. 1, pp. 1439–1447, Jan. 2011.
[14] Eugen Gillich and Volker Lohweg, "Banknote Authentication", 2014.
[15] https://archive.ics.uci.edu/ml/datasets/banknote+authentication.
[16] https://www.researchgate.net/post/Where must we use variance and mean of image
[17] http://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm.
[18] http://www.astro.cornell.edu/research/projects/compression/entropy.html.
[19] https://insightsoftware.com/blog/when-and-why-to-use-heatmaps
[20] https://mylearningsinaiml.wordpress.com/2018/11/21/pairplots/
[21] Meyer, David, Friederich Leisch, and Kurt Hornik." The Support Vector Machine Under Test." Neurocomputing 55, nos. 1–2: 169–186. 2003