Counterfeit Currency Detection

Team members: Dhruv Gupta - Taahaa Dawe - Ghizlane Rehioui

Abstract

The global menace of counterfeit currency constitutes a substantial risk to worldwide financial systems and security. This research tackles the pressing requirement for sophisticated detection mechanisms through the utilization of machine learning techniques. Precisely, robust models are constructed employing logistic regression, decision tree, and Support Vector Machine to adeptly classify banknotes as genuine or counterfeit. Utilizing an open-source dataset, multiple logistic regression, decision tree, and Support Vector Machine models are developed. Notably, all models demonstrate accuracies surpassing 98%. This study emphasizes the efficacy of these techniques in the battle against counterfeit currency and underscores the pivotal role of advanced detection methods in upholding financial integrity.

Counterfeit Currency Detection

Team members: Dhruv Gupta - Taahaa Dawe - Ghizlane Rehioui

1. Introduction

The proliferation of forged banknotes has become an increasingly widespread phenomenon, presenting a substantial challenge to governments, businesses, and individuals, as incidents are reported worldwide. Counterfeit currency creates a complex problem with far-reaching consequences for various stakeholders. Financial institutions, businesses, and individuals face the risk of unknowingly accepting counterfeit money, resulting in financial losses. The dissemination of fake banknotes also undermines public trust in the currency, erodes the credibility of financial systems, and impedes the effectiveness of monetary policies. Furthermore, the proceeds from counterfeiting often contribute to illegal activities, raising serious concerns for law enforcement agencies.

Counterfeit currency involves the creation of imitation money designed to closely resemble genuine banknotes to deceive. Counterfeiters employ advanced printing techniques and materials to replicate the appearance of authentic currency, making it difficult for both individuals and professionals to differentiate between real and fake notes. These fraudulent reproductions vary from low-quality imitations to highly sophisticated copies, further complicating efforts to combat this illicit activity.

This research paper aims to delve into the issue of counterfeit currency by making use of a readily available dataset from the UCI Machine Learning Repository. This dataset will be run through a series of machine learning algorithms whose accuracies in detecting fake notes will be compared. An additional proposal in this research paper will be the development of a simple application for the extraction of features from a user-uploaded image using the Wavelet Transform to check its authenticity using the most accurate machine learning algorithm. This report is organized into six sections starting with an introduction of the topic, the research questions, a brief overview of the literature, the main methods used in this research, the results

main findings in addition to an application proposal, and finally the limitations, conclusion, and future work.

2. Research Questions

This study seeks to explore the intricate realm of counterfeit currency detection through a series of in-depth inquiries. These encompass ethical considerations in the deployment of models, examining the synergies between computer vision and conventional methods, comprehending challenges in real-time implementation, evaluating the impact of supplementary data on model performance, utilizing Python for image processing techniques, and quantifying crucial visual features that differentiate genuine from counterfeit currency. The objective of this research is to offer thorough insights into the multifaceted dimensions of identifying counterfeit banknotes.

The main research questions to be addressed in this study are as follows:

- **RQ1:** How can image processing techniques in Python be utilized to extract relevant features from currency notes for counterfeit detection?
- **RQ2:** What are the key visual features that distinguish genuine currency from counterfeit copies, and how can they be quantified?
- **RQ3:** Which Machine Learning model/models best detect counterfeit currency?
- **RQ4:** What ethical considerations should be taken into account when deploying counterfeit currency detection models, and how can these concerns be addressed with Python?

2.1 Objective

To address the initial research question, it is essential to acquire a grasp of image processing techniques in Python, aimed at extracting crucial features from currency notes. This contributes to the formulation of effective models for detecting counterfeit currency. Subsequently, the quantification of key visual features from images facilitates their seamless transfer to a set of Machine Learning models, enabling an assessment of how well each model distinguishes genuine currency from counterfeit copies. Following this, the algorithm can be supplemented with additional data or uploaded as images.

In terms of ethical considerations in model deployment, the goal is to identify and mitigate ethical concerns such as privacy and biases linked to the establishment of counterfeit currency detection models. This involves leveraging Python for a responsible and transparent implementation. Another ethical dimension involves determining the accessibility of this modeling or detection software and ensuring its usage aligns with moral principles.

This paper also aims to comprehend the technical challenges associated with achieving real-time counterfeit currency detection. This involves considering factors like image pre-processing, processing speed, hardware limitations, and variations in the environment.

2.2 Rationale

The motivation for these research questions stems from the imperative to tackle the intricacies of counterfeit currency detection. Through an examination of ethical considerations, an exploration of synergies among diverse detection methodologies, an understanding of challenges in real-time implementation, and the enhancement of model performance with additional data and features, the research endeavors to provide valuable insights to the field. The integration of Python and Machine Learning algorithms results in the development of transparent, efficient, and ethically sound systems for counterfeit currency detection.

3. Literature Review

The body of literature focused on counterfeit currency detection demonstrates an increasing acknowledgment of the paramount significance of protecting financial systems from the widespread menace posed by counterfeit banknotes. Scholars have delved deeply into diverse aspects of this domain, offering valuable insights and methodologies through their extensive exploration.

In the field of banknote authentication or counterfeit detection, the research community is constantly building datasets and testing out various algorithms. The Wavelet Transform is a commonly used algorithm to extract four main image features from note images to build those datasets (Ashok et al., 2010). The digitization process, which precedes the feature extraction

using the Wavelet Transform, as well as the feature extraction is extensively explained by Gillich and Lohweg (2010).

The main features to be extracted are variance, skewness, entropy, and kurtosis. Some datasets rely on less than those four features (variance and skewness) and clustering algorithms such as the K-Means (E, 2020). Additionally, the notes are into two categories while other datasets categorize them into three or four categories namely Genuine, High-Quality Forgery, and Low-Quality Forgery (Gillich & Lohweg, 2010). The study of Gillich and Lohweg (2010) leads to the classification of notes into four different clusters: ”Genuine”, ”High-Quality Forgery”, ”Low-Quality Forgery”, and ”Inappropriate ROI” using the Support Vector Machine (SVM in short). Other machine learning algorithms used in this topic include the Probabilistic neural network (PNN), Multi-layer Perceptron (MLP), Radial Basis Function (RBF), Decision Tree (DT), and Naïve Base with accuracy rates exceeding 98% except for the RBF (Kumar et. al., 2015).

For the same dataset used in the current paper, Shahani et. al. (2018) found that the accuracy of the Back Propagation Neural Network (BPN) is 100%, which was greater than that of the SVM. Logistic Regression and Artificial Neural Networks were also used on the same dataset by varying the validation and training sets, reaching the conclusion that for better counterfeit detection, including all four features is preferred (Wang et. al., 2023).

Surprisingly, not much is published in the literature around the ethical aspect of counterfeit detection. This calls for urgency in dealing with this issue because the counterfeit currency in circulation can only be expected to grow as data science tools are readily available for everyone to use and develop.

4. Methods

4.1 Data Description and Pre-Processing

The dataset utilized in this study originates from the UCI Machine Learning Repository and consists of 1372 rows and 5 columns. It encompasses four continuous features, namely variance,

skewness, kurtosis, and entropy, which play a pivotal role in determining the authenticity or class of banknotes. The classification is indicated in the "class" column, represented by integer values 0 for genuine banknotes and 1 for counterfeit ones. Notably, these four features, detailed in Table 4.1.1, were extracted from images through the application of the Wavelet Transform.

Variance	How each pixel varies from the nearby pixels and classifies them into different regions.
Skewness	Measures the lack of symmetry.
Kurtosis	Measures to what extent the data are, relative to a normal distribution, light-tailed or heavy-tailed.
Entropy	The amount of information that a compression algorithm must code for.
Class	0 representing a genuine note. 1 representing a fake note.

Table 4.1.1 Definitions of the Dataset Features.

The Wavelet Transform is a mathematical algorithm specifically crafted to convert a waveform, derived from an image in the spatial domain, into a series of coefficients. These coefficients are determined based on an orthogonal basis of small finite waves, called wavelets (Ashok et al., 2010). In the case of this dataset, the information was derived from digitized grayscale images of size 400x400 pixels and a resolution of 660 dpi. These images were captured using an industrial camera (*UCI Machine Learning Repository*, n.d.). The data is then standardized before running the machine learning models and the data is split into two parts: training and test data.

Upon conducting data exploration, it was observed that the dataset initially comprised 55.5% genuine notes (class 0). Following a check for duplicates, it was identified that there were 24 duplicate rows. Consequently, the revised dataset, after cleaning, now reflects a slightly adjusted percentage of 54.7% for authentic notes, as illustrated in the figure below.

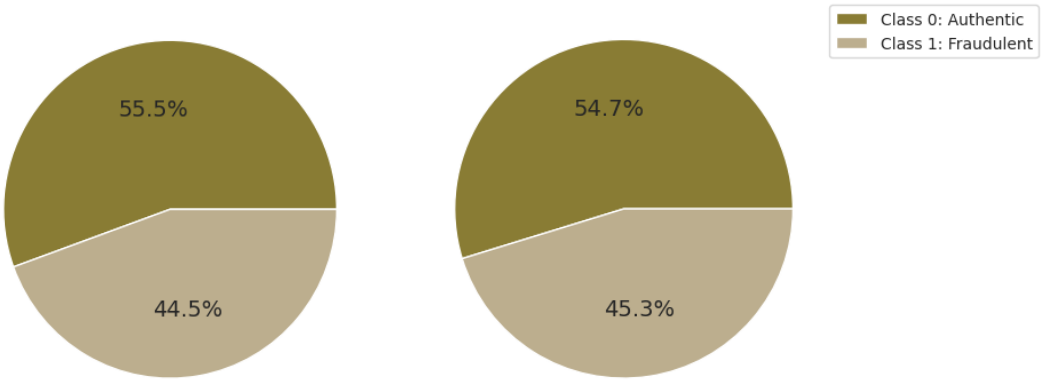

Figure 4.1.1. Percentage of class 0 and 1 before (on left) and after (on right) cleaning

In the data exploration process, another crucial step involves checking for correlations. The correlation coefficient serves as a metric for gauging the linear relationship between two variables, ranging from -1 to 1. A value of -1 signifies a perfect negative correlation, 1 indicates a perfect positive correlation, and 0 implies no correlation. The figure below illustrates moderate to strong correlations among all pairs of variables. The most substantial negative correlation is observed between skewness and kurtosis (-0.79), succeeded by the correlations between variance and class (-0.72), as well as skewness and entropy (-0.53).

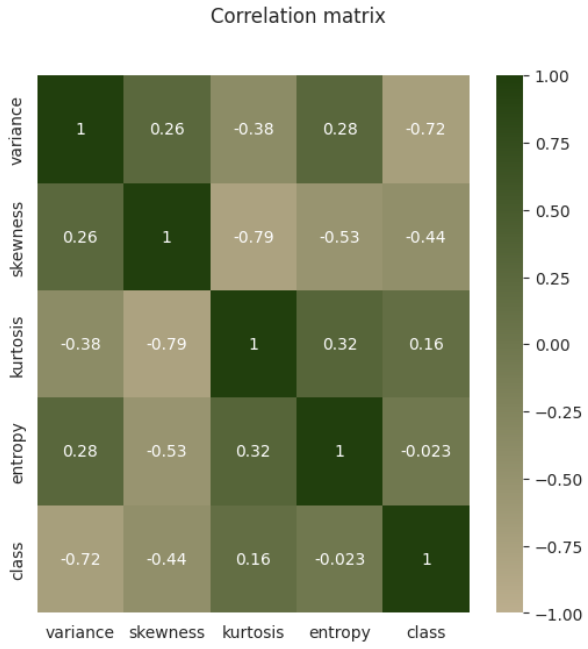

Figure 4.1.2. Correlation Matrix.

The pair plot presented below offers more insightful observations, considering the distinctions between class 0 (authentic) and class 1 (inauthentic) for each of the four variables. Notably, skewness and kurtosis appear to exhibit a negative correlation for both classes, with class 0 (authentic) displaying higher skewness values but lower kurtosis values compared to class 1. The relationship between variance and entropy seems somewhat positive (correlation coefficient of 0.27), and there is a noticeable separation between the two classes.

Examining the distribution of each variable per class adds another layer of insight. For entropy, it appears that the distributions of both classes overlap. In terms of kurtosis, class 0 has a narrower and taller distribution, while class 1 exhibits a longer right tail. For skewness, class 0 is positioned to the right, indicating higher values compared to the other class, and it appears taller. The same is observed for variance.

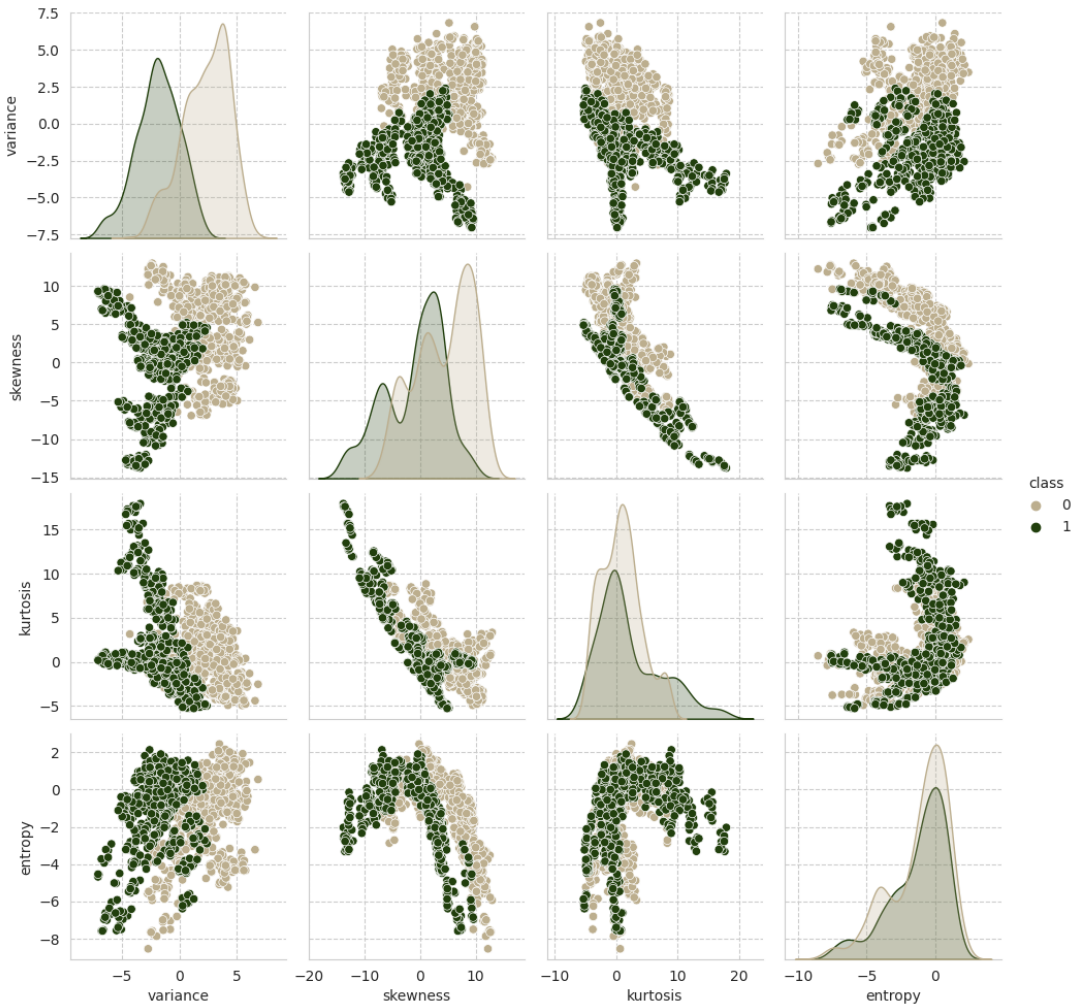

Figure 4.1.3. Pair Plot of Features.

4.2 Statistical Methods and Software

4.2.1. Logistic regression

Logistic Regression is a machine learning approach that processes data and assigns it to two distinct categories. The analysis of independent variables is undertaken to ascertain the binary outcome, with the final results falling into one of two categories (Wolff, 2020). This is made by calculating probabilities as follows:

$$\mathbf{P(Y=1|X) \text{ or } P(Y=0|X)}$$

In the formula above, the likelihood of the dependent variable Y assuming one of two values (0 or 1) based on the independent variable X. While the independent variable X can be either categorical or numeric, the dependent variable Y is consistently categorical (Wolff, 2020). In the context of the current study, the two categories align with the *class* column, where 0 signifies genuine and 1 denotes fake.

4.2.2. Random Forest

The Random Forest is a supervised machine-learning algorithm that aggregates the outputs of multiple decision trees to produce a unified result. The algorithm involves several steps, beginning with the selection of a subset of data points (n random records from a total of k records) and a subset of features (m features) for building each decision tree. The second step entails constructing individual decision trees for each sample. Subsequently, each decision tree generates an output. In the final step, the ultimate output is determined through Majority Voting for classification tasks or Averaging for regression tasks (R, 2023).

4.2.3. Support Vector Machines (SVM)

SVM, which stands for Support Vector Machines, is a machine learning algorithm that aims to distinctly classify the data points by finding a hyperplane in an N-dimensional space, based on the number of features (*Support vector machines*, n.d.).

Selecting SVM for detecting counterfeit money makes sense because it performs well when dealing with various features, especially in scenarios involving pictures or images. Secondly, it excels at handling complex situations where there is a lot of information but limited examples to

learn from. Another noteworthy feature is SVM's ability to navigate tricky situations where the relationship between features is not straightforward. Additionally, it employs a technique known as the 'kernel trick' to better comprehend intricate relationships in the data (*Support vector machines*, n.d.).

4.2.4. Software and Application Development

Python is the programming language used to execute the aforementioned models. Libraries that are relevant to this study include *numpy*, *pandas*, *matplotlib*, *seaborn*, *pickle*, and *sklearn*.

An additional open-source Python library called Streamlit will be useful to run the Note Authentication Application, which is a complementary part of the project. The Application will allow any user to insert the banknote whose authenticity they would like to check in image format (.jpeg or .jpg). Screenshots of the application interface are shown below and the code for the whole project is available on GitHub.

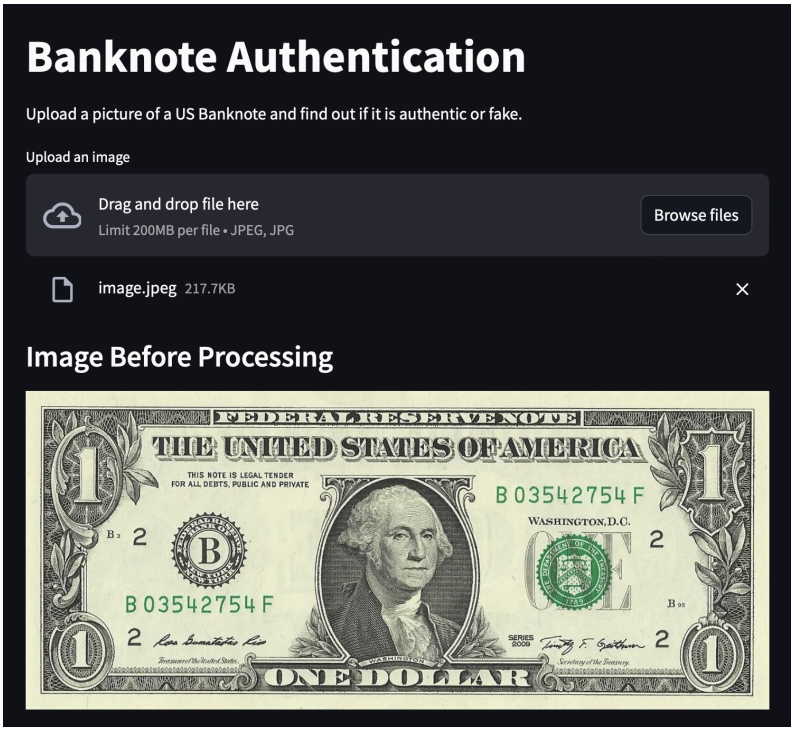

Figure 4.2.1 Genuine image as input to the counterfeit detection application.

After uploading the image to check for genuineness, the application processes the image in the same way the dataset of this paper was, extracts the features, displays the processed image as

well as the features, as seen in figure 4.2.2, along with a message saying the note is genuine or fake.

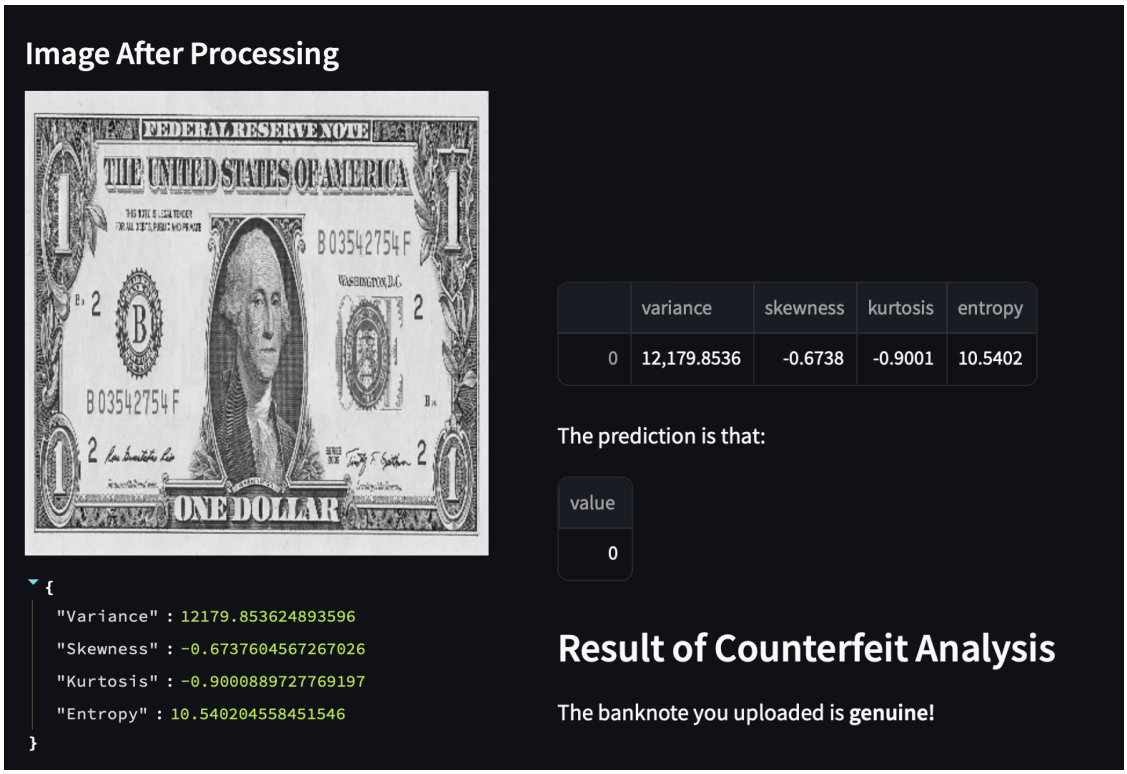

Figure 4.2.2 Output of uploading genuine images to the counterfeit detection application.

5. Results

The following measures have been used to measure the performance of the models implemented.

Accuracy – Accuracy refers to the test's capability to distinguish between authentic and counterfeit note cases correctly.

$$\textbf{Accuracy} = (TP + TN)/(TP + TN + FP + FN)$$

Precision – The precision of a test lies in its capacity to identify the count of notes labeled as genuine by the classifier, accurately classifying them as genuine.

$$\textbf{Precision} = TP/(TP + FP)$$

Where:

- True Positive (TP) = the number of cases correctly identified as genuine notes.
- True negative (TN) = the number of cases correctly identified as fake notes.
- False positive (FP) = the number of cases incorrectly identified as genuine notes.
- False negative (FN) = the number of cases incorrectly identified as fake notes

Techniques	**Accuracy (%)**	**Precision (%)**
Logistic Regression	98.3	100
Random Forest	98.54	100
SVM	100	100

Table 5.1 Accuracy and Precision Score Comparison between algorithms.

5.1 Classification Report Comparison between Algorithms

Within sci-kit-learn, the `classification_report` function is integrated into the metrics module, serving to produce a textual summary encompassing diverse classification metrics pertinent to a classification model. It proves especially valuable in appraising a classifier's performance on a given dataset.

When invoked, the '*classification_report*' function requires the true class labels and predicted class labels as input. It subsequently calculates various metrics such as precision, recall, f1-score, and support for each class within the classification task. These metrics offer insights into the classifier's effectiveness across different classes.

	precision	recall	f1-score	support
0.0	1.00	0.96	0.98	158
1.0	0.95	1.00	0.97	117
accuracy			0.98	275
macro avg	0.98	0.98	0.98	275
weighted avg	0.98	0.98	0.98	275

Figure 5.1.1 Logistic Regression

	precision	recall	f1-score	support
0.0	1.00	0.98	0.99	240
1.0	0.97	0.99	0.98	172
accuracy			0.99	412
macro avg	0.98	0.99	0.99	412
weighted avg	0.99	0.99	0.99	412

Figure 5.1.2 Random Forest

	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	233
1.0	1.00	1.00	1.00	179
accuracy			1.00	412
macro avg	1.00	1.00	1.00	412
weighted avg	1.00	1.00	1.00	412

Figure 5.1.3 SVM

6. Conclusion

6.1 Findings

The SVM algorithm demonstrated superior effectiveness when applied to this dataset, which relies heavily on image quality for its generation. Ongoing research in this sector is actively expanding, employing a variety of image-processing techniques to consistently improve result accuracy. Importantly, the proposed techniques have broader applicability, extending their utility to the extraction of features from other currencies as well.

6.2 Limitations

Ensuring high image quality is crucial for the detection and classification of input images. Camera-captured images need to focus on the currency, covering approximately 80% of the image area, with the note positioned front-facing and devoid of any damage or marks. The continuous evolution of technology introduces challenges in distinguishing between genuine and counterfeit notes, exemplified by instances such as Supernotes. Additionally, the limited size of the dataset poses a potential compromise to accuracy, especially given the diverse patterns that

may influence the model's effectiveness. Addressing these considerations is paramount for optimizing the performance of image detection and classification systems.

6.3 Future work and studies

The exploration of wearable devices and portable detectors involves the development of a concept centered on a wearable device incorporating a camera for image capture, with subsequent classification facilitated through an associated application. This envisioned device is designed to capture images, which are then transmitted to the application for classification. To augment the datasets, the inclusion of data from supernotes is proposed. Combining this extended dataset with the implementation of Neural Networks holds the potential to significantly enhance accuracy, especially in dealing with complex datasets that may exhibit intricate patterns and characteristics. Additionally, more attention should be given to the ethical implications of having access to these counterfeit detection algorithms.

7. References

Ashok, V., Balakumaran, T., Gowrishankar, C., Vennila, I., & Kumar, A. (2010). The Fast Haar Wavelet transforms for signal & image processing. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1002.2184

E, R. (2020). *Banknote authentication analysis using Python K-means clustering - IJISRT*. Banknote Authentication Analysis Using Python K-Means Clustering. https://ijisrt.com/assets/upload/files/IJISRT20OCT060.pdf.pdf

Gillich, E., & Lohweg, V. (2010). *Banknote authentication*. Banknote Authentication. https://www.researchgate.net/profile/Eugen-Gillich-2/publication/266673146_Banknote_Authentication/links/5436b8140cf2643ab9887bca/Banknote-Authentication.pdf

Kumar, C., & Dudyala, A. K. (2015). *Banknote authentication using decision tree rules and machine learning techniques*. Banknote authentication using decision tree rules and machine learning techniques. https://ieeexplore.ieee.org/document/7164721

UCI Machine Learning Repository. (n.d.). https://archive.ics.uci.edu/dataset/267/banknote+authentication

R, S. E. (2023, October 26). Understand random forest algorithms with examples (Updated 2023). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

Shahani, S., R L, P., & Jagiasi, A. (2018). Analysis of banknote authentication system using machine learning. https://www.researchgate.net/publication/323223299_Analysis_of_Banknote_Authentication_System_using_Machine_Learning_Techniques

Support vector machines. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/svm.html#:~:text=Support%20vector%20machines%20(SVMs)%20are,Effective%20in%20high%20dimensional%20spaces.

Wang, A., Goldsztein, G., & Sun, Z. (2023, December 4). *Banknote authentication using logistic regression and Artificial Neural Networks*. Journal of Student Research. https://www.jsr.org/hs/index.php/path/article/view/3777

Wolff, R. (2020, August 26). *5 types of classification algorithms in machine learning*. MonkeyLearn Blog. https://monkeylearn.com/blog/classification-algorithms/