CLASSIFYING OBSERVE AND REVERSE SIDES OF ANCIENT ROMAN COINS

WITH MACHINE LEARNING AND DEEP LEARNING ALGORITHMS

ARDA CANSER ADALI

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
BACHELOR OF SCIENCE IN COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE

DEPARTMENT OF
COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

STUDENT NUMBER

2038999

COMMITTEE

Dr. Sharon Ong
Drs. Stijn Rotman

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

June 22, 2023

WORD COUNT

7190

ACKNOWLEDGMENTS

My thanks to dr. Sharon Ong for her guidance during the whole process of writing this thesis.

CLASSIFYING OBSERVE AND REVERSE SIDES OF ANCIENT ROMAN COINS

WITH MACHINE LEARNING AND DEEP LEARNING ALGORITHMS

ARDA CANSER ADALI

Abstract

Classification of ancient coins is a challenging and time-consuming process for both numismatists and machine learning algorithms. The proposed method of classifying the observe and the reverse sides of the ancient coins as accurately as possible can be helpful for both numismatists and coin classifier algorithms. It can reduce the workload of numismatists by pre-classifying the coins and detecting errors in the datasets that cause accuracy drops in machine learning algorithms. The research is conducted on a highly representative dataset with random forest, support vector machine, and CNN. Additionally, several methods for increasing the accuracy of these classifiers, such as; feature extraction methods, contrast enhancements, and data augmentations were also tested. The experiments showed that the pre-trained CNN was very successful at the given task and achieved very promising results that demonstrate the potential benefits of the goal of this research. Machine learning algorithms also achieved promising results when in combination with the feature extraction methods.

0.1 *Source/Code/Ethics/Technology Statement Example*

The dataset used for the research was the publicly avaliable RRC-60 dataset by Aslan, Vascon, and Pelillo (2020). Some of the figures are directly from the dataset RRC-60 while the others were created via this dataset. For the implementation of the coding components python language was used with the Google Colab environment. Several libraries, packages and APIs were used for this research, including; Joblib, Numpy (Harris et al. (2020)), Pandas (McKinney et al. (2010)), Matplotlib Hunter (2007), Scikit-learn (Pedregosa et al. (2011)), Scikit-image (Van Der Maaten and Postma (2006)), TensorFlow (Abadi et al. (2015)) and Keras (Chollet et al. (2015)). The reused/adapted code fragments are clearly indicated in the notebook.

Figure 1: The observe and the reverse sides of an ancient coin from Aslan et al. (2020).

1 INTRODUCTION

1.1 *Project Definition*

In this study, several machine learning and deep learning methods will be used and compared for classifying the observe[1] and reverse[2] sides of ancient Roman coins. Ancient coins possess invaluable information about the time that they were minted. Therefore, classifying and segmenting ancient coins is a crucial part of numismatics. This is a long and tedious process that requires expert knowledge. Determining which side of the coin is which, is the first step in the classification of ancient coins. This task distinguishes the observe (front) side of the coins which usually depicts a bust from the reverse (back) side of the coins which usually depicts a motif, see Figure 1. The purpose of this study is to implement a computer vision method that does this classification. There is potential for using this classifier with the application of further models. For instance, classifiers can be trained on the reverse side's bust image to determine the corresponding emperor, aiding in estimating the coin's age. The recent developments in computer vision can be of great use in this task. Convolutional Neural Networks are known for their great efficiency in the classification of images. CNNs will be used with several approaches to optimize the performance of the CNNs. Also, machine learning algorithms with different feature extraction methods will be used to be compared with the CNN models. The dataset that will be used is the Roman Republican Coins 60 (RRC-60) dataset.

[1] The observe side of a coin refers to the front of the coin, usually with the depiction of a bust figure.

[2] The reverse side of a coin refers to the back of the coin, usually the side that is opposite to the observe side.

1.2 *Motivation*

Coins have been the primary tool for trade for thousands of years. Civilizations all over the world minted coins throughout history to standardize the value of their currency. The minting of coins is also an ideal tool for rulers to impose their influence on the people. That's why there have been thousands of minted coins with unique features throughout history. These features give invaluable information about the period that the coins were minted such as the influence of rulers, military and political events, beliefs of the society, economic welfare of the civilization, etc. Arandjelović (2010). The research on ancient coins requires efficient and accurate classification of the coins. This is a crucial and laborious aspect of numismatic research that requires great theoretical knowledge and expertise. Ancient coins can be in really bad condition due to centuries of wear. Thus even with great experience, the process is extremely time-consuming and prone to errors Anwar, Anwar, Zambanini, and Porikli (1908).

The observe side of almost every coin class depicts a bust of a ruler or a god while the reverse side almost always depicts motifs of people, animals (horse, eagle, elephant, etc.), or tools (axe, vase, scepter, etc.). The first step in classifying ancient coins is to determine the observe and reverse sides of the coin. Even though the automatic classification of observe and reverse sides of coins can have several useful applications for researchers and numismatists, there aren't any related works in this field. Therefore the goal of this research is to automatize the first step of classifying ancient coins, determining the sides of the coins, as accurately as possible. This pre-classification has several practical applications. It can be used by museums while dealing with large hoards of ancient coins. Additionally, this can help researchers identify patterns and trends that may have been missed through traditional manual methods, leading to discoveries and a deeper understanding of history. Finally, It can be used to increase the accuracy of a more specific classification of ancient coins by ensembling algorithms.

With the recent developments in the field of computer vision, classification algorithms are becoming more and more effective in overcoming the previously mentioned challenges of classifying ancient coins. Machine learning algorithms are commonly used for classification problems which will also be used in this study. This brings us to our first research question:

> [RQ-1:] *To what extent can the machine learning algorithms classify the sides of ancient Roman coins?*

To reduce the dimensionality of the dataset, two different feature extraction methods will be used. Principal Component Analysis is very commonly used with machine learning classifiers and Histogram of Oriented Gradients returns information about the orientations in coins which is commonly used while classifying coins. So our first sub-question is:

How does the feature extraction algorithms Principal Component Analysis (PCA) and Histogram of Oriented Gradients (HOG) affect the performance of machine learning classifiers?

Another approach to image classification is using deep learning models. CNNs can achieve great accuracy on image classifications and after training with a large dataset, it can be extremely useful in classifying ancient Roman coins. The inclusion of CNN models in this research leads us to the second research question:

[RQ-2:] *To what extent can a Convolutional Neural Network classify the observe and the reverse sides of ancient Roman coins in comparison to machine learning classifiers?*

Pre-trained CNN models are usually created by using the best available architectures and training techniques with high amounts of intensive training processes. This results in excellent image classification accuracy and for this reason; pre-trained CNNs were also implemented in this research:

How does the performance of traditional CNN models compare to pre-trained CNN models on ancient coin classification?

Even though CNNs are great at image classification tasks, several preprocessing methods can be applied to the dataset to improve the accuracy of the classification. Contrast enhancement is one of the methods that could help classify ancient coins because of the many advantages that it provides. This method can be useful for classification with machine learning algorithms as well. Additionally, increasing the image resolution can be of help, since it would increase the information that is fed to the CNN. Similarly, increasing the size of the training dataset by applying augmentations on the dataset is an option for feeding the CNN more information. The effect that these preprocessing methods have on the accuracy of the ancient coin classifiers will be observed. The final sub-questions of this research are:

To what extent does the contrast enhancement method (CLAHE) affect the performance of machine learning and deep learning classification of ancient Roman coins?

How does the image resolution and image augmentation affect the performance of CNNs while classifying the sides of ancient Roman coins?

1.3 *Summary of Findings*

The pre-trained CNN model outperformed the traditional CNN in terms of classification accuracies. Achieving an impressive accuracy of &97, the proposed real-life applications for classifying the observed and reverse sides of ancient coins appear highly feasible. In contrast, machine learning algorithms such as random forest and SVM yielded similar but significantly poorer results compared to CNNs. This outcome aligns with the recognized effectiveness of CNNs in computer vision implementations. Contrast enhancement on images had no significant impact on CNNs, while its effects varied when used with different feature extraction methods in machine learning algorithms. Specifically, HOG feature extraction performed better with contrast-enhanced images and also outperformed PCA in general. Increasing image resolution enhanced CNN accuracy, while image augmentation introduced excessive variation, leading to reduced CNN accuracy

2 RELATED WORK

2.1 *Challenges of Classifying Ancient Coins*

Ancient coin classification with learning-based classifiers requires a high amount of discriminatory data and intensive training of the classifiers. Ancient coins have less distinctive features that differentiate them, compared to modern coins for classification purposes. There are several reasons for this; most ancient coins wear out and change colors due to being exposed to harsh environmental conditions for centuries. In addition to this, ancient coins are hand-made and thus, usually are disproportionate and unaligned Schlag and Arandjelovic (2017). They are found and processed in different places around the world that's images taken for research usually vary in illumination conditions and qualities. All of these factors cause ancient coins to have high inter-class variations and also low intra-class variations. This makes it very challenging to train classifier algorithms with ancient coin datasets Zambanini, Kavelar, and Kampel (2014). These challenges are one of the reasons that classifying the observe sides from the reverse sides of the ancient coins would be very helpful as the first step of broader research.

2.2 *Machine Learning For Classifying Coins*

There are many ways of classifying modern coins with machine learning algorithms with varying levels of success. In the eigenspace approach by Huber, Ramoser, Mayer, Penz, and Rubik (2005), multiple eigenspaces from the discriminative features of each coin are created. Afterward, Bayesian Fusion is used to generate a probability that encompasses both the observe and the reverse sides of coins. The results showed an %92 accuracy with a dataset of over 11 thousand coins. Another method proposed by Reisert, Ronneberger, and Burkhardt (2006), utilized the discretized gradient directions technique for extracting orientation-related information from the coins and performed classification with the nearest neighbor algorithm which achieved %97 percent accuracy with a dataset of 10 thousand images. One of the first attempts at classifying ancient coins was done by Zaharieva, Kampel, and Zambanini (2007), with existing algorithms for modern coin classification. In this research, multiple features were compared for ancient coin classification and SIFT features with nearest neighbor algorithms were used for classification. The results showed %93, %80, and %79 classification accuracy of three different coin types.

2.3 *Deep Learning For Classifying Ancient Coins*

Deep learning methods have been the common method for classifying coins in the recent years due to their capabilities of efficiently processing high amounts of data. In the research by Schlag and Arandjelovic (2017), a CNN was used to identify the specific emperors that are depicted on the observe sides of ancient coins. They achieved a recognition rate of %71 on the RIC-Cond dataset which is much higher than the previously used machine learning classifiers. A different approach for classifying ancient coins was done by Kim and Pavlovic (2017), where CNN classifiers were trained with the images of observe sides, reverse sides and both sides. The training was done for over 40000 iterations and with the RIC dataset. The results showed that using both sides for training resulted in much higher accuracy (%76) than training with only the observe sides (%70) and training with only the reverse sides (%62). Another research by Cooper and Arandjelović (2020), used CNNs for classifying the motifs on the reverse sides of ancient coins. Ancient coins that depicted different animals and objects were used and the model achieved accuracy ranging from %84 to %72 for different classes. There were also instances of using pre-trained CNNs for ancient coin classification, Kiourt and Evangelidis (2021). This research used multiple pre-trained CNNs that were trained with the ImageNet dataset. The models were fine-tuned by transfer learning with a limited-sized AnCoins-12 dataset that was proposed. The results of this research showed that even with small datasets, pre-trained CNNs can reach high accuracy while classifying ancient coins.

2.4 *Contrast Enhancement Methods For Classification*

There are also several augmentation techniques that can be performed on ancient coin images to improve the performance of the classifier algorithms. Contrast enhancement is a method that is widely used in medical image classifications. It can be helpful in classifying the motifs on ancient coins because it increases the visibility of the details on images and aids the extraction of important features from images Pizer et al. (1987). The edges of coins possess a lot of discriminative features and enhancement of the edges and contrast in general can help improve the performance of classification Van Der Maaten and Postma (2006) Contrast Limited Adaptive Histogram Equalization (CLAHE) is a contrast enhancement method that excels at images with localized contrast variations and images with uneven illumination Dalara, Sindhu, and Vasanth (2022). This makes CLAHE especially useful to apply to ancient coin datasets because these datasets are usually gathered from multiple smaller datasets that were collected in different conditions which results in high inter-class variations Aslan et al. (2020).

2.5 *Feature Extraction Methods For Classifying Ancient Coins*

There are several preprocessing methods that can be performed to reduce the dimensionality of the dataset. Early machine learning methods for classifying ancient coins made use of local feature extractions such as scale-invariant feature transform (SIFT) and achieved promising results with a limited number of coin classes. SIFT method has the ability to match the local features from images and is invariant to different scales and rotations which makes it a great option for extracting useful information from ancient coins Zaharieva et al. (2007). Histogram of Oriented Gradient (HOG) is a feature extraction method that is very popular in computer vision applications in areas of human detection and object detection because of its effectiveness in capturing local edges. In a study by Dadi and Pillutla (2016), HOG was proposed as an alternative to Principal Component Analysis (PCA) when performing face detection. The results show that HOG-based SVM has achieved around %9 better accuracy than PCA-based SVM. PCA is a commonly used feature reduction technique in computer vision, it extracts the components of an image that explain the most variance in the data. PCA was also used in multiple research on coin classification, for example, the aforementioned eigenspace-based approach by Huber et al. (2005).

2.6 *Dataset Augmentation For Image Classification*

It is commonly the case that as a machine learning algorithm has access to more data, it gets more accurate at classification. Data augmentations are a way of increasing the size of a dataset by manipulating the images. The traditional augmentation techniques are flipping, rotation, shifting, zooming, and distorting. The research by Perez and Wang (2017) showed that traditional augmentation techniques are highly effective at increasing the classification accuracy of CNNs for basic object detection tasks. In the research, newer augmentation methods that made use of GANs were also proposed and showed promising results.

2.7 *Literature Gap*

The literature motivates the use of CNNs and pre-trained CNNs which were implemented in this research. Contrast enhancement and data augmentation methods have also been shown to be effective at increasing the accuracy of image classifiers. Although there is related work in machine learning with different feature extraction methods for ancient coin classification, there is no work for classifying the observe and the reverse sides of ancient coins that also explores the performance of HOG or PCA feature extraction with random forest or SVM classifiers.

3 METHOD

3.1 *Dataset Description*

The dataset that was used for this research is the "Roman Republican Coins - 60" dataset or shortly the RRC-60. This dataset was created by Aslan et al. (2020), with images that were collected from acsearch.info and Coinage of the Roman Republic Online (CRRO). The RRC-60 dataset consists of images of 60 different coin types which were based on a publicly available dataset from Zambanini and Kampel (2013). The RRC-60 dataset consists of 60 unique coin types and contains 200 images (100 for the observe side and 100 for the reverse side) of each unique coin type. For the purposes of classifying the observe and the reverse sides of ancient coins the different coin classes can be united under two folders: "The observe coins" and "The reverse coins". The two class folders contain 6000 images each which totals 12000 unique images. The coins in this dataset are from 144BC to 7BC and are mainly from the late Roman Republic era and some are from the early Roman Empire era.

Figure 2: The observe and reverse sides of two exceptional coins from Aslan et al. (2020)

There are samples of coins in the dataset where both sides of an ancient coin depict a bust or where both sides of the coin depict a motif. These exceptional coins can be tricky for deep learning and machine learning models to classify since they don't follow the common norms of ancient coins. Since there is precedent for these types of coins in the world, it was decided to keep the exceptional coins as a part of the data set. The data set contains 7 exceptional coins (4 of which have busts on both sides and 3 of which have motifs on both sides) to accurately represent the ancient Roman coins. An example of the exceptional ancient coins can be seen in Figure 2.

The data set has high variations in images of some coins that are caused by a variety of factors such as; the wear of the coins, the materials of the coins, the illumination differences, disoriented/shifted minting, color changes due to oxidation, etc. These variations are expected to be seen in ancient coins because of environmental conditions and long preservation times, see Figure 3. These conditions can pose additional challenges for the classifiers; however, they are a natural part of coin classification and must be addressed.

3.2 *Dataset Preparation*

A detailed visualization of methods can be seen in Figure 4. The dataset was rearranged to fit the purposes of this research; the original dataset consisted of 2 main classes: observe and reverse, and 60 sub-classes, for each coin type. The sub-classes were merged, and during this step, multiple images with 4 color channels (RGBa) were converted into RGB images. The images were resized to 128 by 128 arrays and stored in a pickle file for easy access during the training of machine learning algorithms. Additionally, during the training of the CNNs, the images were resized to both 128 by 128 and 299 by 299 arrays to evaluate the impact of resolution on the accuracy of the CNNs.

Figure 3: The example of high inter-class variations due to wear (Observe sides on top and Reverse sides on bottom) from Aslan et al. (2020).

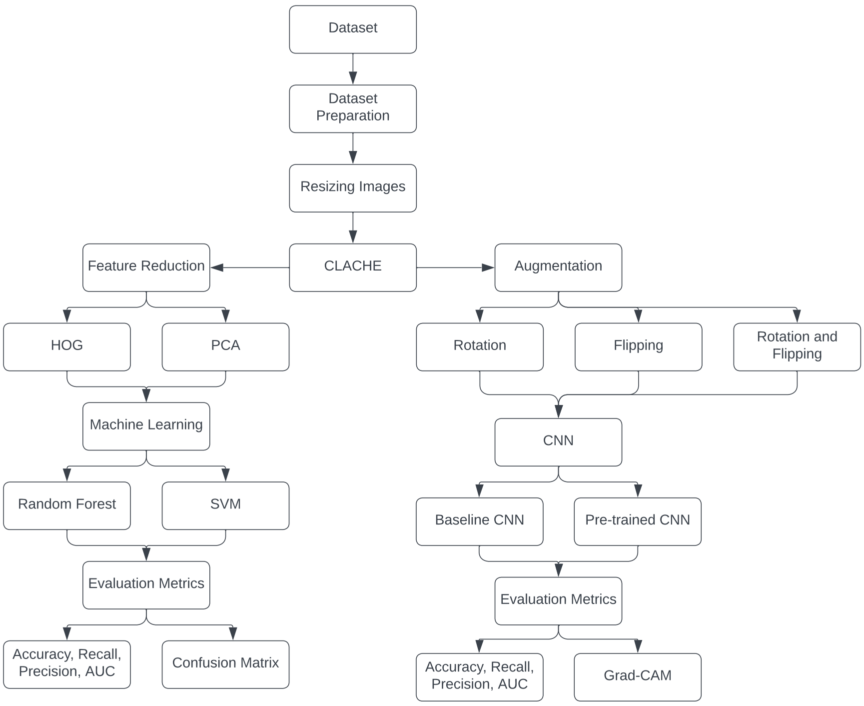

Figure 4: Process Overview, showing the steps taken for data preparation, feature reduction, training algorithms and evaluation.

Figure 5: The images of coins before and after CLAHE

One of the sub-goals of this study is to analyze the impact of augmentations on the effectiveness of CNN classification. The images in the dataset were augmented in three different ways: they were flipped horizontally and vertically, they were rotated in 6 different angles and they were both flipped and rotated.

Besides basic manipulations and rearrangements, two main types of data preprocessing were performed in this research. The first type involves sharpening and enhancing the edge details of the images using Contrast Limited Adaptive Histogram Enhancement (CLAHE). The second type focuses on dimensionality reduction techniques, primarily aimed at improving the efficiency of machine learning classifiers. Specifically, Histogram of Oriented Gradients features (HOG features) and Principal Component Analysis (PCA) were used to reduce the dimensionality of the images.

3.3 *Contrast Limited Adaptive Histogram Enhancement*

Histogram equalization is the most commonly used technique for enhancing the contrast levels in images. Enchanced contrast levels is expected to increase the accuracy of classification tasks. The CLAHE method was performed on the whole dataset, for increasing the performance of both the CNN models and the machine learning models. An example of coins before CLAHE and after CLAHE can be seen in Figure 5.

3.4 *Feature Reduction*

The purpose of feature reduction is to systematically reduce the dimensionality of images in big datasets while still maintaining as much information about variations as possible. This helps in increasing the efficiency of the models and decreasing the computational costs. In this research, two different types of feature reduction techniques were performed on the dataset to observe the effects on the performance of the models. All of the images were normalized and flattened before the feature reduction

Figure 6: The visualisation of the images after HOG transformation, the features reduced to 20 percent and 50 percent respectively. (Higer resolution versions of the images can be seen in Appendix A (page 30).

procedures. The package used for PCA and HOG transformation was Scikit-Learn Pedregosa et al. (2011).

3.4.1 *Principal Component Analysis*

PCA was performed on both the original dataset and the contrast-enhanced dataset. The images used for machine learning are sized 128 by 128 pixels which totals 16384 components in total. After PCA, the number of components was reduced to cover %75 explained variation and %95 explained variation. The number of components required in both of the datasets to reach %75 and %95 percent variation is 870 and 3227 components for the CLAHE dataset and 470 and 2558 components for the original dataset.

3.4.2 *Histogram of Oriented Gradients*

HOG features were extracted for each image in the dataset with the help of transformers. HOG features were calculated by dividing the images into blocks of 16 by 16 pixels and 24 by 24 pixels and calculating the orientation inside those blocks in 9 different directions for the gradients. This resulted in 8440 and 3664 components respectively. Examples of the image of a coin before and after HOG transformation with different numbers of components can be seen Figure 6. As can be seen in the figures, HOG transformation outlines the edges of the images in a way that still keeps as much information about the images as possible.

3.5 *Machine Learning*

The machine learning models used in this research were Random Forest Breiman (2001) and Support Vector Machine (SVM) Lowe (2004). The classifiers were trained with features that were extracted from PCA and HOG methods. The dataset was split into train and test sets with a %80 to %25 ratios. The hyperparameters of the classifiers were fine-tuned by using GridSearchCV Pedregosa et al. (2011).

3.5.1 *Random Forest*

Random Forest is an ensemble classifier which runs multiple decision trees in parallel with different subsets of features. It has the advantage of referring to multiple decision trees. It is also great for dealing with high dimensional data. It was used in this research because it is easy to implement and it has been shown to perform well in ancient coin classification tasks Ma and Arandjelović (2020). The tested hyper-parameters for Random forest are "max_depth" of 2, 6, 10, and "n_estimators" of 100, 400 and 1000. The optimal results were max_depth of 10 and n_estimators of 1000.

3.5.2 *Support Vector Machine*

SVMs are often used for ancient coin classification especially in combination with SIFT feature extraction Kim and Pavlovic (2017). A linear SVM can form linear decision boundaries that can discriminate between features. The tested hyper-parameters for SVM are "C" of 1, 8, and 20. The optimal results were "C" of 1. The kernel used was linear kernel.

3.6 *Convolutional Neural Network*

The CNNs for this research were created and trained using TensorFlow Abadi et al. (2015) and Keras Chollet et al. (2015) API. CNN classifiers are known for their great performance in computer vision problems. The purpose of implementing a CNN approach subsequent to traditional machine learning classifiers was to assess and compare the effectiveness of both of the methods. Two different CNN classifiers were used: a basic CNN and a pre-trained CNN. The basic CNN was made up of several common convolutional blocks. The traditional CNN was created as a baseline to be compared with the pre-trained CNN. A pre-trained CNN is a saved network trained with a large dataset that can be fine-tuned to specific tasks of classifications. In this case, the pre-trained CNN was trained with a larger dataset and optimized for the classification of ancient coins. The optimizer used for both of the CNNs was the Adam algorithm. The CNNs were trained up to 10 epochs. The datasets used for training the models were the original unaltered RRC-60 and the CLAHE-transformed version of the RRC-60. Several augmentations were also tested while training the CNNs, these included flipping, rotating, and both flipping and rotating.

The dataset was split into training, test, and validation sets respectively %80, %10 and %10 ratios.

3.6.1 *Baseline CNN*

CNNs are composed of layers of neurons that contain associated parameters and weights. The architecture of these layers and the selected hyperparameters is a crucial factor that effects the performance of the CNN. The CNN model that was built for this research was inspired by the findings of Schlag and Arandjelovic (2017) which used a few small kernels (3x3) that were stacked instead of one big kernel which helped in increasing the describability of the CNN and reduced computational costs. The architecture of the baseline CNN is as follows: 3 convolutional layers, 3 maximum pooling layers, 1 flattening layer, 1 dropout layer, and 2 dense layers. All of the functions that were applicable used the ReLu activation function which improves convergence speed and generalization, except for the last dense layer where the Sigmoid function was preferred.

3.6.2 *Pre-trained CNN*

The pre-trained CNN classifier used in this research was the InceptionV3 model by Szegedy, Vanhoucke, Ioffe, Shlens, and Wojna (2016). It was trained with the ImageNet database of over a million images. The pre-trained CNN can classify images into 400 categories. Transfer learning on the pre-trained CNN is performed to make it suitable for classifying ancient coins. The step-by-step process is as follows: Downloading the Inception-v3 model. Freezing its layers to prevent unwanted distortions during training. Adding new trainable layers to the CNN that are suitable for the task. Then, the newly added top layers of the CNN are trained with the RRC-60 dataset. After this the base model is unfrozen and the whole CNN is trained again.

3.7 *Evaluation Metrics*

3.7.1 *Accuracy, Precision, Recall and Area Under the Curve*

Accuracy is the measure of correctly classified classes over all of the classifications done. It is the most commonly used evaluation method for categorical classifications. The AUC metric gives valuable information when it is used in binary classifications. This metric measures the area under the receiver operating characteristic (ROC) curve which is the curve of true positives against false positives. AUC score shows the success rate of the classifier, an AUC value of 1 indicates perfect classification while a value of 0.5 indicates random classification. The precision score shows the correctly classified positive classes over all of the positively predicted classes. The recall score shows the correctly classified positive classes over all of the positive classes. Accuracy, precision, recall, and AUC score of both the machine learning and deep learning classifiers was measured. Additionally, Confusion matrices for the machine learning classifiers were created by using information from these metrics. The goal of these confusion matrices is to visualize the false positives and false negatives during classification. Confusion matrices for CNNs aren't created because the predicitions of these algorithms are evaluated in depth with the help of the Grad-CAM method.

3.7.2 *Gradient-weighted Class Activation Mapping*

Grad-CAM is a method that was created by Selvaraju et al. (2017). It is a tool for interpreting the results of CNN classification in more detail. It generates a heatmap that highlights the distinguishing features and regions of an image, providing insights into the model's decision-making process. Thus, we are able to conclude that our model is actually activating around correct patterns and regions which shows that the and the model actually learned from the dataset it was trained on. The grad-CAM method was used with the CNN classifiers to evaluate the correctly classified and misclassified coins in more detail.

4 RESULTS

In this section, the performance of the SVM, the random forest, the baseline CNN and the pre-trained CNN will be evaluated with respect to the different variables that were used. A short summary of the best results from the classifiers in Table 1 shows that both of the CNNs performed better than the machine learning classifiers. This was an expected result because of the high capabilities of CNNs to process images. The pre-trained CNN achieved significantly better results than the baseline CNN with a near-perfect accuracy of 0.973 compared to 0.933. The machine learning classifiers achieved great accuracies as well, especially the random forest classifier which outperformed the SVM classifier and achieved accuracies that were close to the baseline CNN.

The contrast enhancement method CLAHE was the common factor in the highest score by all four of the classification methods. Even though this result indicates that the CLAHE method is ultimately beneficial in the classification of observe and reverse sides of the coins, this wasn't the case

Table 1: Short summary of the best results for the machine learning and the CNN classifiers.

Model	Dataset (Resolution)	Specifics of Training	Accuracy	AUC
Baseline CNN	CLAHE (299)	Not Augmented	0.933	0.973
Pre-trained CNN	CLAHE (299)	Not Augmented	0.973	0.980
Random Forest	CLAHE (128)	HOG (%50)	0.902	0.957
SVM	CLAHE (128)	HOG (%50)	0.872	0.947

Table 2: The accuracy, precision, recall, and AUC scores of baseline CNN and pre-trained CNN with different resolutions and contrast enhancement.

Model	Dataset (Resolution)	Accuracy	Precision	Recall	AUC
Baseline CNN	Original (128)	0.925	0.921	0.929	0.978
	Original (299)	0.929	0.925	0.939	0.973
	CLAHE (128)	0.923	0.922	0.920	0.976
	CLAHE (299)	**0.933**	0.940	0.926	0.973
Pre-trained CNN	Original (128)	0.966	0.965	0.965	0.971
	Original (299)	0.971	0.977	0.967	0.978
	CLAHE (128)	0.969	0.957	0.981	0.972
	CLAHE (299)	**0.973**	0.990	0.954	0.980

with several different parameters used in the research. These discrepancies can be seen in Table 2, Table 4, and Table 5 which will be evaluated further.

The results for the CNN classifiers without performing augmentations on the data can be seen in Table 2. The accuracy scores were very high overall which is promising for the purpose of this research. All of the findings had an AUC score of a minimum of 0.97 which shows that the CNNs have a great ability to distinguishing between the observe and the reverse sides of ancient coins. The negligible impact of various CNN implementations on the AUC score can be attributed to the scores already being exceptionally high, thus making further improvements increasingly challenging.

The pre-trained CNN achieved better results overall than the baseline CNN with every different implementation and with every evaluation metric. The different implementations of the pre-trained CNN had a minimal impact on the classifier's accuracy. This observation is understandable considering that the classifier already achieved an extremely high accuracy with lower-resolution images and without contrast enhancement. As the classifier becomes more proficient at classifying, further enhancements to its performance become increasingly challenging.

Figure 7: The effect of contrast enhancement on worn ancient coins.

The baseline CNN showed more visible performance variations with different implementations. Firstly, the accuracy increased significantly as the resolution of images increased from 128 by 128 to 299 by 299. However, the effect of contrast enhancement on the accuracy of the CNN was inconsistent as it both improved and decreased the accuracy of the classifier in different implementations. This inconsistent effect of contrast enhancement can be explained by a few factors. It may be the case that the images on the original dataset contained sufficient contrast and the CLAHE method didn't provide significant changes. It can also be caused by the CLAHE method increased the contrast of the images excessively and thus the discriminatory details of the images were lost in the process. Taking into account the example images before and after contrast enhancement in Figure 5, the former explanation better explains the results. The example images of worn ancient coins in Figure 7 that are presumably more susceptible to loss of valuable information due to an increase in noise after contrast enhancement also support this explanation. Further research can be conducted on contrast enhancement using datasets of ancient coins with a variety of conditions to gain a deeper understanding of the impact of CLAHE on ancient coin classification.

Additionally, three different augmentation techniques were performed on the training data for the CNNs, see Table 3. The results show a decrease in the performance of the CNNs that were trained with the augmented data. The results were especially lower when both of the augmentation techniques, flipping and rotation were applied at the same time. These unexpected results can be caused by a few factors: The most probable reason is that the excessive amount of augmentation on the dataset can lead the learning process of the CNN to be more challenging and thus resulting in a decrease in overall accuracy. An alternative option is that the CNN may be over-fitting, it may be over-specialized due to the high amount of variations introduced by the augmented data. However, the training and validation loss curves of the CNNs, seen in Appendix B (page 30), don't support this idea. Another probable explanation that is also supported by the training and validation accuracy curves, also seen in

Table 3: The accuracy, precision, recall, and AUC scores of baseline CNN and Pre-trained CNN models with augmented dataset.

Model	Augmentation	Accuracy	Precision	Recall	AUC
Baseline CNN	Rotation	0.896	0.869	0.925	0.962
	Flip	0.889	0.897	0.886	0.952
	Both	0.817	0.772	0.885	0.908
	None	0.923	0.922	0.920	0.976
Pre-trained CNN	Rotation	0.953	0.943	0.967	0.962
	Flip	0.938	0.969	0.908	0.951
	Both	0.927	0.938	0.913	0.944
	None	0.966	0.965	0.965	0.971

Appendix B (page 30), is that the CNN classifiers were not trained long enough with enough epochs to achieve their maximum accuracy results. The high amounts of additional variations by data augmentations and the effect of this on the learning speed of the CNNs may be underestimated by this research. Decreasing and simplifying the augmentations performed on the dataset can be a solution to these problems.

In addition to assessing general statistics, analyzing individual examples of images specific to the dataset can provide more profound insights into the results of the CNNs. The observe and reverse sides of ancient coins possess inherent differences and it is more logical to evaluate them separately in this matter. The observe sides of coins are usually more plain and follow a similar pattern when compared with the reverse sides of coins. Most observe sides of ancient coins depict a bust that contains features that are similar to each other: lips, nose, and hair. These features logically should make it easier to classify the observe sides of the ancient coins. The class activation maps of the observe sides of the coins can be seen in Figure 8, showing higher levels of activation in these areas. This supports the claim that the observe sides of ancient coins have more predictable features. The results in Table 2 also support that it is easier to classify observe sides of ancient coins correctly. The precision scores are generally higher than the accuracy scores, showing that the observe sides of the coins were consistently classified correctly. One example reaching the highest precision score of 0.99. The recall scores being lower than the accuracy scores also points to the misclassification of reverse coins being high.

Both the baseline CNN and the pre-trained CNN had difficulties with classifying the reverse sides of the coins. The reverse sides of the coins are usually more complicated and more unique than the observe sides of the coins. It should also be noted there were a notable number of worn coins on both the observe and the reverse sides. However, when a complicated

Figure 8: The Grad-CAM activation maps of correctly classified observe-sided coins (baseline CNN).

motif on the reverse side is worn out, it becomes even more intelligible and thus, difficult to accurately classify. Examples of commonly misclassified worn reverse-sided coins by the CNNs can be seen in Figure 9, along with their respective activation maps. The activation maps show that some of the specimens[3] were so worn out that the CNNs activated randomly all over the images.

Examining the images and activation maps of the misclassified observe coins also showed that the wear on coins was the major reason for misclassifications. It was previously mentioned that the classification of the observe-sided coins was mainly done by the identification of three components: lips, nose, and hair of the busts. Examples of coins where at least one of these components was disfigured which resulted in the misclassification of the sides are presented in Figure 10. Some of the activation maps show activation around one or two of the three mentioned components and are still misclassified. This shows the significance of all three being present in the image together. There are also activation maps of some extremely worn coins that were similar to the activation maps of the extremely worn reverse-sided coins where the CNN would activate randomly.

Another common pattern of misclassified ancient coins was discovered upon further examination. The exceptional coins that were explained in the dataset description, with motifs or busts on both sides of the coins were often misclassified. The classification of these exceptional coins was a task in which there were observable differences between the performance of the baseline CNN and the pre-trained CNN. The baseline CNN misclassified some of the exceptional coin types more often than the pre-trained CNN.

[3] A specimen of coin refers to a specific individual coin.

Figure 9: Reverse-sided coins that were misclassified due to wear (baseline CNN).

Figure 10: Observe-sided coins that were misclassified due to wear (baseline CNN).

Figure 11: Observe-sided coins that were misclassified due to exceptional depictions (baseline CNN).

These were the observe-sided coins with depictions of motifs seen in Figure 11 with respective activation maps. Especially the coin types with the plain depictions of tools were hard to classify accurately for the baseline CNN where the classifier would fail to activate on specific points of the images.

The reverse-sided exceptional coins with depictions of busts were often misclassified by the pre-trained CNN as well as the baseline CNN. There were three such coin types which were all often misclassified. The aGrad-CAM activation maps show that the CNNs would pick up on the hair, nose, mouth, or the legend[4] on the coins to classify them as observe instead of reverse. The major problem for the pre-trained CNN was that two coin types, depicted almost identical busts of Julius Caesar, where one coin depicted the image on the observe side and the other coin depicted the image on the reverse side. The images of these coins can be seen in Figure 12. Grad-CAM activation maps of these coins where they were misclassified highlighted the legend of the coins that had the writing "CAESAR" on them. The specimens that were classified correctly highlighted the small flower depiction on the observe coins. Figure 13 shows the similarities between two different coins that have the depiction of Julius Ceasar on its observe side and reverse side. This further proves the capability of the CNN classifiers on identifying ancient coins even with highly similar samples of coins.

[4] The legend is the writing on a coin that provides information about the production of the coin.

Figure 12: Reverse-sided coins that were misclassified due to exceptional depictions (baseline CNN).

Figure 13: The similarities in activation maps of the misclassified bust of Ceasar on the left (coin 488/1-2) correctly classified bust of Ceasar on the right (coin 480/5a-b) (baseline CNN).

Evaluation of the misclassified exceptional reverse coins unveiled a mistake with the original RRC-60 dataset. Most probably because the observe and the reverse sides of the coin type 488/1-2 with the bust of Marcus Antonius and the bust of Julius Caesar on each side look very similar to each other, some images of the observe side of the coin were mixed in with the images of the reverse side of the coin in the original dataset. This may have caused further misclassification errors on the CNNs.

The results of the machine learning classifiers random forest and SVM can be found in Table 4 with the contrast-enhanced data and in Table 5 with the original data. According to the results random forest classifier consistently outperformed the SVM. This is probably caused by a mixture of factors: The ensembled nature of the random forest is effective at reducing overfitting. It can adapt to different decision boundaries which helps while classifying ancient coins with high inter-class variations.

Comparison between the two feature extraction methods shows that the HOG transformation was better at preserving information that is dicriminative to the observe and the reverse sides of the ancient coins than PCA. All of the classifications that were performed with HOG transformation achieved better results as the ratio of components increased from %25 to %50. Both the random forest and the SVM algorithms that were trained with both the original dataset and the CLAHE dataset performed the best with HOG fetures at %50.

Table 4: The accuracy, precision, recall, and AUC scores of Random Forest and SVM trained with HOG features and PCA features from the CLAHE dataset (contrast-enhanced images).

Models	Dataset (%)	Accuracy	Precision	Recall	AUC
Random Forest	HOG (%25)	0.885	0.920	0.844	0.946
	HOG (%50)	**0.902**	0.929	0.871	0.957
	PCA (%75)	0.780	0.811	0.732	0.871
	PCA (%95)	0.773	0.791	0.745	0.859
SVM	HOG (%25)	0.855	0.855	0.857	0.935
	HOG (%50)	**0.872**	0.867	0.880	0.947
	PCA (%75)	0.780	0.786	0.772	0.856
	PCA (%95)	0.697	0.696	0.705	0.772

The effect of contrast enhancement with CLAHE was tested on both the random forest and the SVM. The results showed mixed correlations between accuracy scores and contrast enhancements depending on the feature extraction method used. The HOG feature extraction was affected slightly positively by contrast enhancement. Considering that this feature extraction method relies heavily on information about the edges and orientations in an image, a more significant effect was expected. However, Pca was affected negatively by the contrast-enhanced dataset. Contrast enhancement caused the PCA to require significantly more components to reach the same amounts of explained variance as the original images. This shows that the dataset was less compact and concise after contrast enhancement which resulted in a decrease in the overall performance of the machine learning classifiers.

Confusion matrices of the classification results, seen in Appendix C (page 30), indicate that the random forest classifiers that were trained with the HOG features were better at the classification of reverse sides of the coins than the observe sides of the coins which are supported by the precision and the recall scores. The rest of the classifications showed no significant difference between classifying the observe and the reverse sides of the ancient coins.

5 DISCUSSION

There were two main research questions and 4 sub-research questions that were introduced in the introduction. The main research questions are regarding the performance of machine learning classifiers and deep learning classifiers. The sub-questions are regarding different pre-processing methods that could increase the performance of the classifiers. The results

Table 5: The accuracy, precision, recall, and AUC scores of Random Forest and SVM trained with HOG features and PCA features from the original dataset (without contrast-enhanced images).

Models	Dataset (%)	Accuracy	Precision	Recall	AUC
Random Forest	HOG (%25)	0.878	0.915	0.834	0.940
	HOG (%50)	**0.888**	0.917	0.855	0.949
	PCA (%75)	0.819	0.824	0.814	0.899
	PCA (%95)	0.814	0.803	0.836	0.894
SVM	HOG (%25)	0.855	0.854	0.860	0.931
	HOG (%50)	**0.866**	0.868	0.865	0.936
	PCA (%75)	0.823	0.839	0.802	0.897
	PCA (%95)	0.761	0.757	0.773	0.832

of this research were insightful in answering the research questions that were proposed in the introduction section. The machine learning classifiers performed at a very high level and the random forest classifier performed better than the SVM. However, the CNNs performed even better than the machine learning algorithms and the pre-trained model achieved the best accuracy scores. The contrast enhancement on images was helpful when used in combination with HOG features and harmful when used in combination with PCA. It had no significant effect on the already high accuracy of CNNs. Feature extraction methods were successful at increasing the accuracy of the machine learning classifiers. HOG features proved to be a better choice than PCA with every trial for coin classification. Performing augmentation to increase the dataset was not a reliable method and makes the prediction of the CNNs harder by adding unnecessary variance. Finally, an increase in image size consistently resulted in higher accuracy with CNNs.

The pre-trained CNNs showed great potential at being used as a reliable option for the first step of the classification of ancient coins. It was very successful at classifying the observe sides from the reverse sides of ancient coins. A deeper investigation of the pre-trained CNNs results led to a demonstration of real-world applications of these methods. Pre-trained CNN was successful at detecting labeling errors for a coin type that had 6 images of its observe side on the wrong folder with the images of its reverse side. This and further specific classifications of ancient coins when combined together can make way for a very efficient and very accurate cumulative classifier in the future.

There have been other researches on the classification of ancient coins that compared machine learning and CNN methods, see Kim and Pavlovic (2017). The results of this research align with the findings from those research in that the CNNs performed better overall when compared with the machine learning algorithms. However, our results achieved higher accuracy scores in general, because classification was done between only two classes. Also similar to other research, the CNN models had difficulties while classifying ancient coins that suffered from excessive wear Arandjelović (2010).

The low accuracy rates of both the machine learning and the deep learning models in classifying worn-out ancient coins were the most prominent limitation of this research. The implementation of contrast enhancement methods, augmentation methods, and feature extraction methods were all unsuccessful at overcoming the challenges of classifying noisy images of ancient worn coins. Another limitation was the inclusion of exceptional-sided coins that are mentioned during the description of the dataset. These coins were misclassified more often compared to normal coins.

One of the aims of this research was to propose a method that would be useful for numismatists in their day-to-day classification of ancient coins. The methods that were proposed, however, would fail to adapt to natural usage for this task due to practicality limitations. A model that segments and classifies ancient coins in images of multiple coins that are put together would increase the practicality for daily uses immensely. Another improvement would be to successfully classify ancient coins that have part of their surfaces covered (in a pile) with a classifier model that can figure out the class of a coin from partial/fractional features of the coin. These implementations would overcome some limitations that were caused by the specificity of the dataset that was used to train the models.

6 CONCLUSION

This research proposed a sub-method for ancient coin classification that can be utilized in various applications. Efficient and accurate classification of the observe and the reverse sides of ancient coins has many benefits ranging from assisting numismatists with classifying ancient coins to validating the ancient coin datasets that are being used for research purposes. The proposed methods made use of feature extraction techniques, contrast enhancement and augmentation which resulted in various degrees of success. HOG features were more beneficial for the accuracy of machine learning classifiers compared to PCA features but both affected had a positive impact. CLAHE method had mixed effects on the accuracy of machine learning classifiers but increased the accuracy of CNNs significantly. Augmentation of the training data by flipping or rotating images resulted in a decrease in the performance of both machine learning and deep learning classifiers. All of the proposed machine learning and deep learning methods achieved high accuracy scores. The pre-trained CNN was compared with the baseline CNN and the machine learning classifiers and achieved very promising results with extremely high accuracy that shows the effectiveness of training with high amounts of data and the feasible reach of the goals of this research.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... Zheng, X. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems.* Retrieved from https://www.tensorflow.org/ (Software available from tensorflow.org)

Anwar, H., Anwar, S., Zambanini, S., & Porikli, F. (1908). Coinnet: Deep ancient roman republican coin classification via feature fusion and attention. arxiv 2019. *arXiv preprint cs.CV/1908.09428*.

Arandjelović, O. (2010). Automatic attribution of ancient roman imperial coins. In *2010 ieee computer society conference on computer vision and pattern recognition* (pp. 1728–1734).

Aslan, S., Vascon, S., & Pelillo, M. (2020). Two sides of the same coin: Improved ancient coin classification using graph transduction games. *Pattern Recognition Letters*, *131*, 158–165.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Chollet, F., et al. (2015). *Keras.* https://keras.io.

Cooper, J., & Arandjelović, O. (2020). Understanding ancient coin images. In *Recent advances in big data and deep learning: Proceedings of the inns big data and deep learning conference innsbddl2019, held at sestri levante, genova, italy 16-18 april 2019* (pp. 330–340).

Dadi, H. S., & Pillutla, G. M. (2016). Improved face recognition rate using hog features and svm classifier. *IOSR Journal of Electronics and Communication Engineering*, *11*(04), 34–44.

Dalara, A., Sindhu, C., & Vasanth, R. (2022). Entity recognition in indian sculpture using clahe and machine learning. In *2022 first international conference on electrical, electronics, information and communication technologies (iceeict)* (pp. 1–12).

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... Oliphant, T. E. (2020, September). Array programming with NumPy. *Nature*, *585*(7825), 357–362. Retrieved from https://doi.org/10.1038/s41586-020-2649-2 doi: 10.1038/s41586-020-2649-2

Huber, R., Ramoser, H., Mayer, K., Penz, H., & Rubik, M. (2005). Classification of coins using an eigenspace approach. *Pattern Recognition Letters*, *26*(1), 61–75.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. doi: 10.1109/MCSE.2007.55

Kim, J., & Pavlovic, V. (2017). Discovering characteristic landmarks on ancient coins using convolutional networks. *Journal of Electronic Imaging*, *26*(1), 011018–011018.

Kiourt, C., & Evangelidis, V. (2021). Ancoins: Image-based automated identification of ancient coins through transfer learning approaches. In *Pattern recognition. icpr international workshops and challenges: Virtual event, january 10-15, 2021, proceedings, part vii* (pp. 54–67).

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, *60*, 91–110.

Ma, Y., & Arandjelović, O. (2020). Classification of ancient roman coins by denomination using colour, a forgotten feature in automatic ancient coin analysis. *Sci*, *2*(2), 37.

McKinney, W., et al. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th python in science conference* (Vol. 445, pp. 51–56).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*.

Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., Greer, T., ... Zuiderveld, K. (1987). Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, *39*(3), 355–368.

Reisert, M., Ronneberger, O., & Burkhardt, H. (2006). An efficient gradient based registration technique for coin recognition. In *Proc. of the muscle cis coin competition workshop, berlin, germany* (pp. 19–31).

Schlag, I., & Arandjelovic, O. (2017). Ancient roman coin recognition in the wild using deep learning based recognition of artistically depicted face profiles. In *Proceedings of the ieee international conference on computer vision workshops* (pp. 2898–2906).

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the ieee international conference on computer vision* (pp. 618–626).

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2818–2826).

Van Der Maaten, L., & Postma, E. (2006). *Towards automatic coin classification*. na.

Zaharieva, M., Kampel, M., & Zambanini, S. (2007). Image based recognition of ancient coins. In *Computer analysis of images and patterns: 12th international conference, caip 2007, vienna, austria, august 27-29, 2007. proceedings 12* (pp. 547–554).

Zambanini, S., & Kampel, M. (2013). Coarse-to-fine correspondence search for classifying ancient coins. In *Computer vision-accv 2012 workshops: Accv 2012 international workshops, daejeon, korea, november 5-6, 2012, revised selected papers, part ii 11* (pp. 25–36).

Zambanini, S., Kavelar, A., & Kampel, M. (2014). Classifying ancient coins by local feature matching and pairwise geometric consistency evaluation. In *2014 22nd international conference on pattern recognition* (pp. 3032–3037).

APPENDIX A

The Figure 14.

APPENDIX B

The Figure 15.

APPENDIX C

The Figure 16.

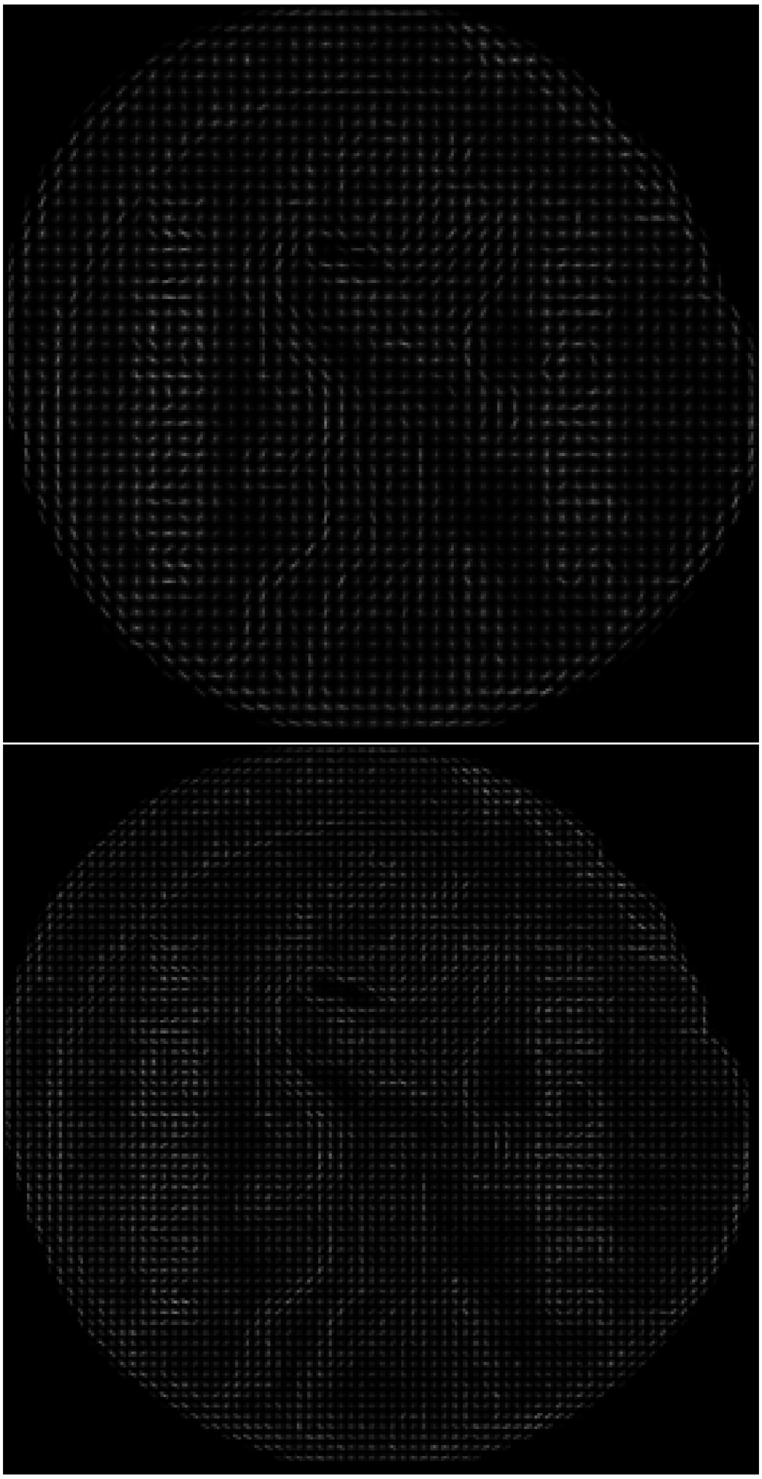

Figure 14: The higher resolution images of sample coins after HOG transformation 20 percent and 50 percent respectively.

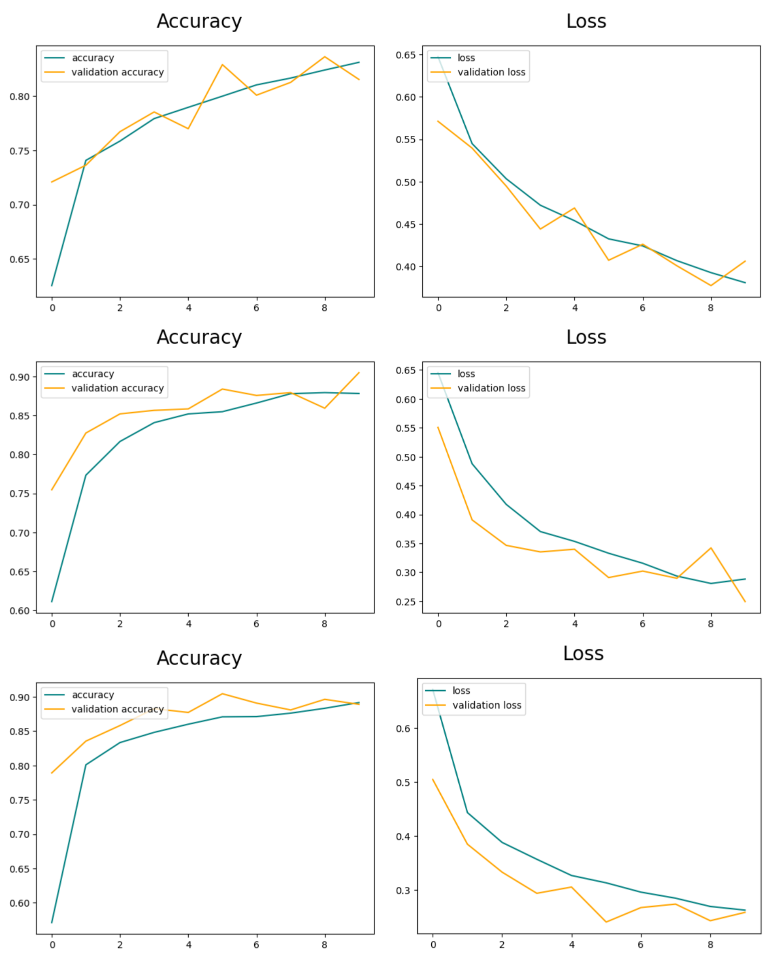

Figure 15: Training and validation accuracy and loss of the baseline CNNs that were trained with augmented data.

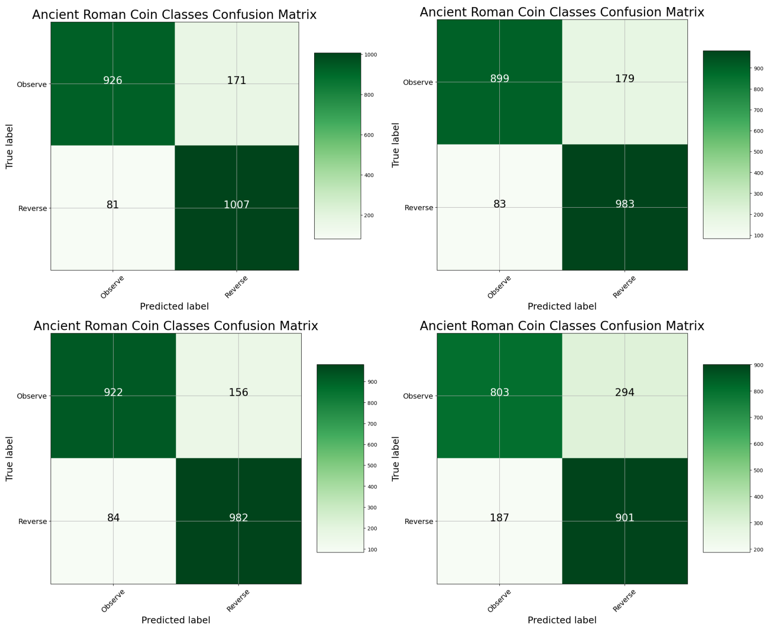

Figure 16: Confusion matrices for the classification results of random forest with HOG features with both the original and the CLAHE dataset.