

Using Fine-Tuned CNNs to Grade Collectible Coins

Chad Stachowicz

PhD Advisors: Dr. Ahmed Helmy & Dr. Siddharth Krishnan

2025

Abstract

This paper explores the viability of automated coin grading using convolutional neural networks (CNNs). A benchmarking dataset of professionally graded United States coins is curated and a deep-learning model is fine-tuned to predict a coin's grade on the Sheldon scale. The concept of a *step-grade* is introduced as a more realistic and robust evaluation metric, accounting for the inherent subjectivity of professional grading. The model achieves an accuracy of $86.3\% \pm 1$ step-grade and $96.4\% \pm 2$ step-grades, demonstrating that modern CNNs can provide practical assistance in numismatic grading tasks.

1 Introduction

Growing up, my father was a coin dealer and ever since I can remember, collecting coins was a part of my life. A coin to me is a physical token of a moment in time. Minted with its year on it, it is a constant reminder of the history around us and a way to walk back in time in your mind. People collect coins not only for their history but also for their propensity to hold and gain in value over time. There are many factors that contribute to a coin's value. How old the coin is, its relative rarity and survival rate, its condition or grade, its variety, eye appeal, and other factors all contribute to the value of a coin [1]. This paper will focus on using Artificial intelligence (AI), specifically convolutional neural networks (CNNs), to determine the coin's grade on the Sheldon scale.

The Sheldon scale is a 70-point numeric system with 1 being the lowest and 70 being the highest possible measure to quantify a coin's state of preservation and overall quality [2]. All coins used or graded in this paper were produced in the United States between 1793 and 2025, which is the entirety of the non-colonial or territorial minted coins in the USA.

This research explores the viability of automated AI grading by fine-tuning a CNN based off of ConvNeXt as well as a benchmarking dataset other researchers can use for similar research. This paper offers the following contributions not yet seen in the numismatic industry using AI:

- The first public dataset of US Coins for coin grading (to our knowledge), which will be made public once we figure out the appropriate manner to do so.
- A standard scale for benchmarking numismatic AI grading that is representative of the reality of grading being subjective even between professional graders.
- The *step-grade* as a more accurate measure than the raw Sheldon grade when building models for coins.
- A state-of-the-art model that achieves $86.3\% \pm 1$ step-grade and $96.4\% \pm 2$ step-grades of accuracy against a generalized dataset of US coins.

It's notable that much of the research time was spent on exploring modeling ideas, procuring the dataset and tuning hyper-parameters to optimize the model; however, AI wrote most of the code. AI-assisted coding tools (e.g., Cursor with Claude Opus 4.5) drove most of the monotony of code-writing that normally would have made a project of this scale impossible in the given time-frame. Not only did this research include training the model, it also included building an inference package and getting real-world confirmation from coin experts the grades were usefully accurate. The inference webpage can be found here [4].

2 Problem and Prior Work

In the professional certification realm, there are three top-tier companies specializing in the grading of coins: NGC, PCGS, and CACG. However, there are many others over the years that have existed and continue to exist with less reputability. There are also thousands of coin dealers and tens of thousands or more serious collectors in the US capable of grading coins although their skill in it may widely fluctuate.

Grading coins, however, is more of an art than a science and there is a great deal of subjectivity in the field as to what a coin's grade is and the reasoning behind that. The same coin will often be sent between different grading companies in the hope of receiving a higher or better grade, as coins are worth more money in the collectibles market when they are graded higher. This subjectivity in grading is one of the main drivers of activity in the collectors market, but it is also what makes grading coins difficult. New collectors or family members of collectors are often taken advantage of by dealers when bringing old collections to coin shops or dealers because they are not experts in condition, grading, or even understanding what they have. The goal in this research is to bring a level of objectivity to a market that is heavily subjective in nature.

Prior work has been done in this field before, with the most complete work being done by Chen, who used image processing techniques such as Sobel edge detection to create gradient images in order to find marks and significant elements of the coin [3]. This method compares gradient images to a computed "perfect coin" of a coin series by comparing 18 sub-regions across the coin. Chen does this on both the obverse and reverse of the coin and then feeds all these data points into a multilayer perceptron (MLP) that predicts the grade.

Chen's paper focuses on five series and provides metrics for accuracy for a single coin type, Franklin halves, for only Mint State grades and achieves approximately 95.1% accuracy within ± 1 grade. This metric may sound impressive; however, Franklin halves are typically a lower-valued coin, and thus the vast majority that are graded are Mint State and high-end Mint State at that, leaving the algorithm relatively few distinct grade choices to distinguish compared to a model that needs to do the entire Sheldon scale. Chen also trained a model on all series and grades and achieved only 91.3% accuracy within ± 4 grades. The range of ± 4 is too large to be useful in practice and represents a significant amount of error.

3 Background on Coin Grading

Coin grading on the Sheldon scale is a 1–70 scale that describes how well preserved a coin is from its original minting [2]. Coins that are graded 1–59 are considered *circulated coins*, or coins that have made it into the pockets of people and used as money. Sheldon grades 60–70 are reserved for Mint State coins, which are also called *uncirculated* coins by many. These are coins that never made it into general circulation and were pulled for or by collectors either from mint sets issued by the government, rolls of coins fresh from minting, or other similar sources.

Although grades are defined from 1–70, in practice they are not all used. Grades have monikers that are used for grading which are Poor 1 (P01), Fair 2 (FR02), About Good 3 (AG03), Good 4 (G04), Good 6 (G06), Very Good 8 (VG08), Very Good 10 (VG10), Fine 12 (F12), Fine 15 (F15), Very Fine 20 (VF20), Very Fine 25 (VF25), Very Fine 30 (VF30), Very Fine 35 (VF35), Extremely Fine 40 (XF40), Extremely

Fine 45 (XF45), About Uncirculated 50 (AU50), About Uncirculated 53 (AU53), About Uncirculated 55 (AU55), About Uncirculated 58 (AU58), Mint State 60 (MS60), Mint State 61 (MS61), Mint State 62 (MS62), Mint State 63 (MS63), Mint State 64 (MS64), Mint State 65 (MS65), Mint State 66 (MS66), Mint State 67 (MS67), Mint State 68 (MS68), Mint State 69 (MS69), and Mint State 70 (MS70), with the final two digits corresponding to the Sheldon grade.

For example, a VF23 is not an option you can receive from any of the grading companies; such a coin will either receive a VF20 or a VF25. We use this discrete scale for the grades we detect and predict in our model. For the purposes of measuring accuracy, we define a *step-grade*, which is one step on the common grading scale that professional graders use. For example, MS63 $\rightarrow$ MS64 is a step-grade, as well as VF20 $\rightarrow$ VF25.



Figure 1: Side-by-side Sheldon examples

There are further nuances that this paper does not directly address. Grading coins by their Sheldon value is not the only challenge; there are also questions of authenticity and un-gradable or “details”-graded coins that companies and dealers consider, which our models do not currently handle. A “details”-graded coin is one that has been modified or damaged. Examples of these issues include coins that have been cleaned, severely scratched, cut or hacked apart, mounted in jewelry, dug out of the ground, or damaged by corrosion. All of these are considered serious negatives when valuing a coin. This research looks only at straight-graded coins graded by NGC, PCGS, and CACG.

4 Methods

The first major problem we faced is that, to our knowledge, there is no public dataset for US coins with grading information available. We found datasets for world coins and US coins that could be used for classifying different currencies, but not for grading them. Grading coins also requires that very small scratched and fine marks be seen, so high-resolution, high-quality photos must be used.

The best sources for graded coins with high-resolution imaging that we could identify were auction sites and PCGS and CACG professional imaging. We collected a large number of photos (on the order of tens of thousands) of many series of US coinage and split these into a 70/15/15 training, validation, and testing split. These coins were from 1793–2025 and were represented by many major series across all denominations. The series used include Indian Head cents, Buffalo nickels, Mercury dimes, two-cent pieces, three-cent pieces, Walking Liberty half dollars, twenty-cent pieces, Standing Liberty quarters, Flying Eagle cents, large cents, Peace dollars, half cents, and half dimes. This variety was chosen to promote generalization, as the Sheldon scale can be used on any coin type and the goal was to build a generalized model across all US coinage.

For pre-processing images prior to training, Hough circle detection was used to isolate the coin, resize

it to 512 pixels, then place it onto an all-white background. Minor color and hue augmentation as well as random rotations of up to 5 degrees on the coins help the model generalize to different coin positions and imaging variations.

The obverse (front of the coin) and reverse of the coin were treated as separate 512×512 images and run them each through their own CNN backbone. The features from both CNNs are then fused together before classification through a single regression head. An ordinal regression formulation with mean squared error (MSE) as the training loss and mean absolute error (MAE) as a validation metric is used. Our per-coin error metric is

$$|\text{actual step-grade} - \text{predicted step-grade}| \quad (1)$$

The training loop itself uses a two-stage fine-tuning process. In the first stage, the backbone is frozen and only the new classifier head is trained for 5 epochs. In the second stage, the backbone is unfrozen and the entire model is trained for an additional 25 epochs. Batches of 16 images with a learning rate of 0.001 was used to train the models. The research experimented with fine-tuning both ResNet-50 and ConvNeXt as backbones; the results were quite similar. Early stopping during training of our models was employed based on the best MAE validation score. The training loop used a learning rate scheduler configured with a patience of 3 and a factor of 0.5 to reduce the learning rate when validation performance plateaued.

Lastly, the grading company was added as a feature that is fused as an embedded dimension with the outputs of the dual CNN before our classifier head. This is motivated by the fact that CACG is known as a very strict grader. We also assigned a bias of +0.5 to the grade of any CACG coins to represent this strictness. This translates to that if a CACG coin was graded MS65, 65.5 was the score used for training.

5 Results

Three major methods were tested to see which performed better. VGG-19 was trained from scratch and Resnet50 and ConvNeXt were fine tuned with the two stage process defined in methods. Fine-tuning ConvNeXt performed the best of the three methods achieving $86.3\% \pm 1$ step-grade and $96.4\% \pm 2$ setp-grades of accuracy in 39 epochs taking slightly longer than a day to train on a Mac Studio M4 Ultra. Fine-tuning Resnet-50 also did quite well achieving $82.8\% \pm 1$ step-grade and $95.5\% \pm 2$ step-grades. The results we achieved from training VGG-19 from scratch was $62.1\% \pm 1$ step-grade and $74.4\% \pm 2$. Both the ConvNeXt model and the Resnet-50 models are available on the inference demo site in the references.



Figure 2: ConveNeXt Training vs Validation MAE

Scientific stats are great, but what matters more than the opinion of machines, is the opinion of numis-

matists. Once this research was complete, numismatic experts were asked to try using the model. Here are some real pieces of feedback. “That grading app is very interesting. I tested it on a raw Morgan under low light and it got it right!”, said Rick Snow, a renowned expert in numismatics and author of many numismatic books[5]. Interestingly, these models had no Morgan dollars in their training sets, showing great generalization. “Got my [18]72 spot on” and “Nailed my 1856”, said my neighbor and copper enthusiast, Jonathan who was the first person to try the model after it was built and whose coins weren’t part of the test or training set. These pieces of feedback from members of the numismatic community that have much more experience than AI has at grading coins, weighs as heavy as numbers in a field that has traditionally been subjective in nature.

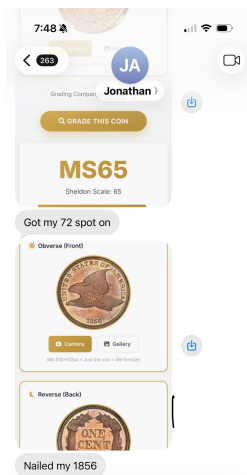


Figure 3: Feedback on AI Coin Grading Model

6 Conclusion

The research as a whole was certainly a success but there are many limitations in our work that need to be addressed in future works. Authenticity is a major part of grading companies’ purview and we didn’t work on that problem at all. A fake coin is worth nothing. Coins are not all the same once they’ve gone through the hands of society, many of them have different degrees of consequential damage which make them what the market calls a ‘details’ graded coin. These are worth much less money as they are considered damaged or altered. Examples of these issues are coins that may have been cleaned, severely scratched, cut or hacked apart, put into jewelry, dug out of the ground, have corrosion or a multitude of other issues that are considered serious negatives when valuing a coin. This research only focused on assigning a Sheldon Grade to the coin. Picking out these damaged coins and noting why they are damaged is just as important as assigning it a Sheldon grade overall. There are also many additional beneficial designations such as ‘Full Bell Lines’ or ‘FBL’ on Franklin Halves and ‘Full Steps’ or ‘FS’ on Jefferson Nickels which this model does not detect which significantly increase the value of a coin.

There are also limitations in not being able to have the coin in hand. Photos of coins can be bad, blurry, improper resolution or litany of other problems. Photos can make it hard to convey the accurate grade of a coin. Eye appeal also matters when grading coins and they regularly look much different in hand so photos can only do so much justice. Especially when it comes to luster which is the reflection of light based on a slight microscopic difference in surface metal created during the minting of a coin. This can be seen pretty decently on video but not in an image since it takes movement of the coin against light to produce. Future research might consider grading from a video of the coin instead of just images.

To make a truly great model, 100's of thousands of images need to be gathered and filtered through to make sure bad images, blurry images or poorly/wrongly graded don't make it into the training and validation sets. Research is also warranted in using larger images to help show the details of coins better than 512x512 images, however compute resources and time-frame for this project limited testing this theory out.

Although there are many places for improvement in the process, this research shows that AI Coin grading can be viable given more time, effort and focus from researchers. This was far more successful that it was expected to be on the onset of the project.

Acknowledgments

First, we thank the professional grading services and auction houses whose high-quality imaging were essential in constructing our dataset, and we acknowledge the assistance of modern AI coding tools that accelerated the implementation of our experimental pipeline. Next we thank our instructors in our machine learning class which without assigning this project, I never would have done.

References

- [1] "An Economic Model of the Coin Collecting Industry," available online at ScienceDirect, <https://www.sciencedirect.com/science/article/pii/S016726819700084X>, accessed 2025.
- [2] M. Dickie, "An Economic Model of the Coin Collecting Industry," *Journal of Political Economy*, vol. 102, no. 2, pp. 215–230, 1994.
- [3] J. Chen, "Development of an automated coin grading system: Integrating image pre-processing, feature extraction, and ML modeling," M.S. thesis, Virginia Tech, Blacksburg, VA, Dec. 2024. [Online]. Available: <https://vtechworks.lib.vt.edu/items/7ced86e8-4534-406c-9b75-6922b516d7bc>
- [4] Chad Stachowicz (2025), *Numisking Coin Grader*, <https://cstachowicz--numisking-coin-grader-fastapi-app.modal.run/>.
- [5] Eagle Eye Rare Coins, "About us," *Indiancent.com*. Available: <https://www.indiancent.com/about-us-eerc>. Accessed: Dec. 11, 2025.