Highlights

Firearm Brand Classification Using Deep Learning on Cartridge Case Images

- Firearm brand classification across a wide spectrum of brands
- Large-scale dataset comprising over one million samples
- Practically deployable solution achieving approximately 92% accuracy

Firearm Brand Classification Using Deep Learning on Cartridge Case Images

ARTICLE INFO

Keywords:
Ballistic examination
Firearms identification
Deep learning
Firearms brand classification

ABSTRACT

When a firearm is discharged, it leaves characteristic marks on the cartridge case, which are analyzed in forensic ballistics to identify the firearm. Conventional ballistic examination systems rely on high-quality images of cartridge cases and bullets, scanning databases to generate ranked candidate lists based on similarity scores. However, these systems often overlook the distinctive signatures of the firearm brand, which could refine search spaces and improve identification accuracy. In this study, we propose a deep learning-based approach leveraging normalized height maps and shape index transformation of cartridge cases for firearm brand classification. Using the BALISTIKA system, we generated high-resolution surface representations from over 350,000 cartridge cases representing the most populous 21 firearm brands, representing 97% of firearms encountered in criminal cases in Türkiye, including handcrafted firearms and converted blank pistols (CBPs). By oversampling the minority classes in the dataset using rotated samples, we expanded it to over a million samples and mitigated class imbalance. We evaluated both traditional machine learning (SVM, Random Forest) and deep learning models (ResNet, Vision Transformer), with deep learning approaches achieving superior performance of up to 92% accuracy. These findings demonstrate that automated firearm brand classification enables forensic examiners to confidently prioritize cartridge cases from the same brand during ballistic comparisons. This approach is expected to substantially reduce examination time and enhance the efficiency of forensic investigations.

1. Introduction

Ballistic examination seeks to identify the firearm used to discharge a cartridge case or bullet recovered from a crime scene. This process involves analyzing the surface markings on the evidence, as firearms leave distinctive impressions on the cartridge during firing. Given the microscopic nature of these marks, they are typically examined under high magnification. The differentiation and matching of these markings with other samples is a highly specialized task, requiring ballistic expertise. To support this analysis, ballistic experts also utilize advanced ballistic imaging and comparison systems alongside traditional microscopic examinations.

Modern ballistic examination systems, such as BALISTIKA [1], IBIS [2], EvoFinder [3], and ARSENAL [4], operate by analyzing high-resolution images of cartridge cases and bullets, and provide forensic experts with a list of potential matches based on the likelihood that the evidence was fired from a specific firearm. During database searches, all evidence is compared using features extracted through traditional image processing techniques. While these systems offer recommendations, they are not capable of automatically identifying the firearm without expert validation. One reason for this is the complexity of the firearm-cartridge matching problem, where current matching accuracy levels remain inadequate for operational deployment. Additionally, erroneous firearm-cartridge matches could lead to false accusations, potentially affecting individuals' lives unjustly. Therefore, there remains hesitation in fully trusting intelligent systems for ballistic firearm identification.

On the other hand, deep learning-based retrieval systems [5–36] are now widely employed across various domains and have demonstrated significantly more reliable and successful outcomes compared to traditional methods. Over the past decade, with advancements in processor and GPU capabilities, CNN-based image processing methods have replaced older approaches in numerous projects due to their high accuracy, flexibility, and more generalized problem-solving capacity. With the introduction of transformers [37] into the field of image processing (e.g., Vision Transformers [38]), new attention-based approaches have also begun to gain traction in practical applications including classification, clustering, and matching tasks.

Although contemporary ballistic examination systems do not offer fully automated matching, they can leverage deep learning techniques to narrow down the ballistic data space for the desired query in tasks such as classification and clustering. This, in turn, can enhance the ballistic examination process by providing more accurate results in a shorter time frame. One possible first step in this regard could be the development of a system capable of distinguishing firearm brands. Currently, there are over a hundred firearm brands used in various calibers of ammunition. Modern ballistic examination systems do not analyze the unique "signature" of firearm brands, often overlooking this detail. Such signatures could either mislead the matching process by being mistaken for the unique marks of a specific firearm or, conversely, aid in the process by eliminating certain firearm brands and thereby narrowing the search space for more accurate matching. Despite its potential for improvement, the primary challenge in this area is the lack of a sufficiently diverse ballistic data pool. Existing studies [39–42] have been conducted with too few brands and samples, rendering them impractical for real-world applications.

In this study, we evaluate both classical machine learning algorithms and deep learning models to classify firearm brands based on signatures found on cartridge case surfaces, aiming to develop a solution with practical applicability for forensic applications. To this end, we utilized the BALISTIKA system to generate high-quality height maps representing cartridge case surfaces from over 350,000 samples spanning 89 different brands. We used the height maps both as raw feature maps as well as by converting them to a descriptive feature called the shape index [43], which encodes the curvature information at every point on the image space. Among the 89 different brands, we specifically focus on the most popular 21, which account for 97% of the firearm brand population involved in criminal cases, including handmade firearms and modified blank guns with replaced barrels. By oversampling the minority classes, we expanded the dataset to over one million samples and trained Random Forest [44], Support Vector Machine [45], ResNet [46], and Vision Transformer [38] models. We achieved approximately 92% accuracy with the best model, which is highly valuable for practical applications.

Figure 1: Samples of cartridge case and bullet evidence: the *right* shows the bullet, and the *left* shows the cartridge case. The photograph is provided by the National Institute of Standards and Technology (NIST).

Figure 2: *Left:* Marks on the cartridge case caused by factory production, such as the headstamp marks of the Mechanical and Chemical Industry Corporation (MKE) of Türkiye. *Right:* Marks on cartridge cases fired from firearms of the same brand group can be misinterpreted as originating from the same firearm.

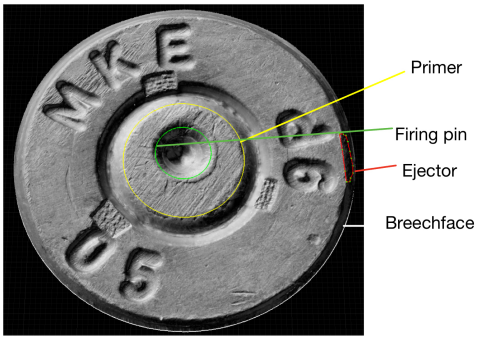

Figure 3: Marked regions on the surface of a cartridge case: breech face marks, primer marks, firing pin impression, and ejector marks.

2. Background

The aim of ballistic examination is to identify the firearm from which a cartridge, found at a crime scene, was fired. To achieve this, investigators first obtain the suspect firearm, conduct test firings, and compare the resulting cartridge cases with those in the database. If a match is found, the matching cartridge cases are considered to have been fired from the same firearm.

2.1. Ballistic Marks

A firearm operates by igniting the gunpowder within a cartridge when the firing pin strikes the primer, propelling the bullet toward its target. As the bullet is discharged, the now-empty cartridge case is first pulled from the chamber by the extractor and then forcefully ejected from the firearm by the ejector, clearing the chamber for the next cartridge to be loaded (Fig. 1).

In a firearm mechanism, the firing pin, breech face, and ejector are the main components that come into contact with the cartridge case, while the barrel comes into contact with the bullet. Each of these parts leaves distinct, characteristic marks on either the cartridge case or the bullet. Breech face impressions, firing pin impressions, and ejector marks are found on the cartridge case, while rifling marks from the barrel's grooves are imprinted on the bullet.

In addition to the marks left by the firearm, cartridges may also exhibit marks from the manufacturing process. The primary challenge lies in distinguishing the forensically valuable marks made by the firearm from those originating during production. The headstamp, for instance, although irrelevant for firearm identification, may physically obstruct the firearm from imprinting its own characteristic marks on that region. A third category of marks can be considered brand-related marks. Firearms from certain brands or groups may leave similar marks on cartridges, making them appear as if they were fired from the same weapon (Fig. 2). While such marks may confuse the identification of the specific firearm, they can also help narrow down the pool of firearms that need to be examined.

In this study, we aimed to distinguish brand-related marks on the cartridge cases. The cartridge case examination involves analyzing the marks found in four defined regions: breech face marks, primer marks, firing pin marks, and ejector marks (Fig. 3). Among these, the breech face is typically the region where the fewest characteristic marks from the firearm are found, though this can vary depending on the firearm's brand. Additionally, the headstamp of the cartridge is located in this region and should be disregarded during the matching process since it is not caused by the firearm. The marks on the primer and firing pin result from the impact of the firing pin strike, with the most distinctive marks often found in the firing pin region. Ejector marks form when the cartridge is expelled from the chamber by the firearm's ejector mechanism. Although they are not always present or prominent, when distinct, these marks provide highly discriminative features for matching.

3. Related work

The most fundamental ballistic examination relies on the pairwise comparison of ballistic evidence under a microscope. However, given the growing population, the corresponding increase in crime rates, and the demand for swift responses from the legal system, it is no longer feasible to compare all archived evidence microscopically for a single case in the current forensic context. To address this challenge, computer-assisted ballistic identification systems

have been developed. Among the most widely used systems globally are BALISTIKA [1], IBIS [2], ARSENAL [4], and EvoFinder [3].

IBIS [2] captures images using a high-resolution camera under ring lighting. The legacy IBIS system produces 2D images and performs comparisons based on these images. Although newer IBIS systems can produce 3D images, they continue to perform similarity comparisons using 2D images to maintain compatibility with older systems. The reference sample is compared with other evidence in the archive, and the system presents the operator with a ranked list sorted by similarity to the reference.

BALISTIKA [1] similarly provides a recommendation list but conducts comparisons based on 3D images generated through photometric stereo [47–58]. Various image processing techniques are applied to these 3D images to extract multiple features. Each feature yields an individual similarity score. The system then compiles these scores to generate a cumulative score, which is used to rank the evidence in the recommendation list.

In addition to these commercially utilized systems, there are several academic studies exploring forensic firearm identification using deep learning techniques. Most of these studies focus on bullet identification. The study [59] proposes a firearm brands classifier from bullet markings using a panoramic photograph taken by a mobile phone. The pretrained DenseNet121 [60], ResNet50 [46] and Xception [61] classification models are further trained with 718 bullet images to classify eight firearm brands. The lowest class accuracy among the brands is around 91%. Despite the promising accuracy, the limited scope of only eight brands substantially restricts its practical applicability in real-world forensic investigations.

The study [39] proposes a bullet classifier to distinguish among three firearm brands. The performance of various machine learning methods (Support Vector Machines [45], Decision Tree [62], Random Forest [44]) and a deep learning method (DenseNet121 [60]) is compared using a dataset composed of features extracted from 50 fired pellets. These features are derived from the curvature of the surface topology of bullet images using Empirical Mode Decomposition (EMD) [63]. The study identifies DenseNet121 as the best-performing classifier, achieving F1 scores of 0.97 for the ’Baikal’ class, 0.78 for the ’Edgar’ class, and 0.99 for the ’HW’ class.

Another study [40] aims to extract bullet markings by employing a U-net segmentation model [64]. Instead of focusing on firearm identification or improving evidence matching performance, the study aims to assist ballistic experts by determining the true bullet markings, specifically striation marks, with around 88% accuracy. Thus, the effects of the predicted bullet markings on overall matching performance are not evaluated.

A recent study [41] focuses on cartridge cases and aims to automatically predict two Regions of Interest (ROIs), corresponding to breech face and firing pin marks. Similar to the approach in [40], they take advantage of a U-Net segmentation model, trained on 1,195 samples of cartridge case images obtained from the IBIS system. Ground truth for the training process is derived from boundaries annotated by a ballistic expert. The model achieves an IoU of 95.6%

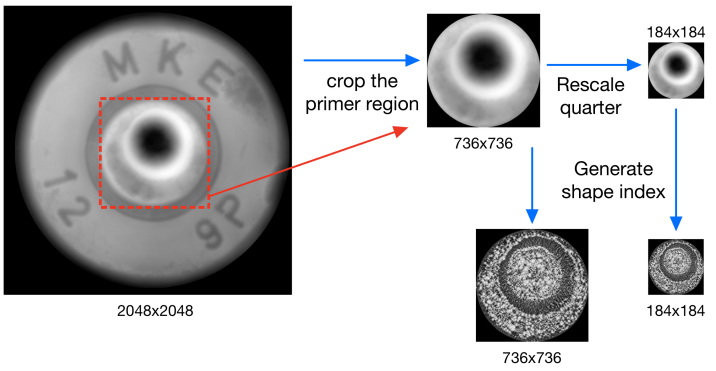

Figure 4: This figure illustrates the process of generating a shape index [43] from a height map image. The original height map images of cartridge cases have a resolution of 2048×2048 pixels. From these images, the primer region is cropped based on the breech face location, resulting in images of 736×736 pixels. To avoid excessively large input sizes during training with shape index images, we further reduced the resolution of the cropped height maps to a quarter of their size. Shape index images were generated from both the original and downsampled height maps for use in different training configurations.

for breech face regions and 95.9% for firing pin areas. While the results are promising, the effect of these predicted regions on the overall matching performance of the system has not been explored.

In addition to studies that aim to assist ballistic experts, some studies focus on making firearm-evidence matching automatically. One study proposes a Siamese model to decide whether two cartridge cases are fired from the same firearm or not [42]. The model is trained with the images of firing pin pairs fired from 300 firearms. The firing pin images are taken from the IBIS system. The results show the true positive rate is 87% with the 0.5 threshold. However, under this configuration, the model produces 3,339 false positives. For such a system to have practical forensic value, the false positive rate should ideally be zero. When the decision threshold is adjusted to achieve an almost zero false positive rate, the true positive rate drops significantly to only 12%. Thus, the study makes a valuable contribution to the field, but is not practical, yet.

4. Method

4.1. Data Preparation

We utilized the data generated by the BALISTIKA system [1], developed by TUBITAK Space Technologies Research Institute in Türkiye. BALISTIKA is a ballistic examination system that captures 3D images of both cartridge cases and bullets, and provides ballistic experts with a ranked recommendation list by comparing evidence across an extensive historical database. The 3D images are generated using the photometric stereo method, which reconstructs surface geometry from 16 images captured under different lighting directions by a high-quality camera. In this study, we used the height map images produced by the photometric stereo process.

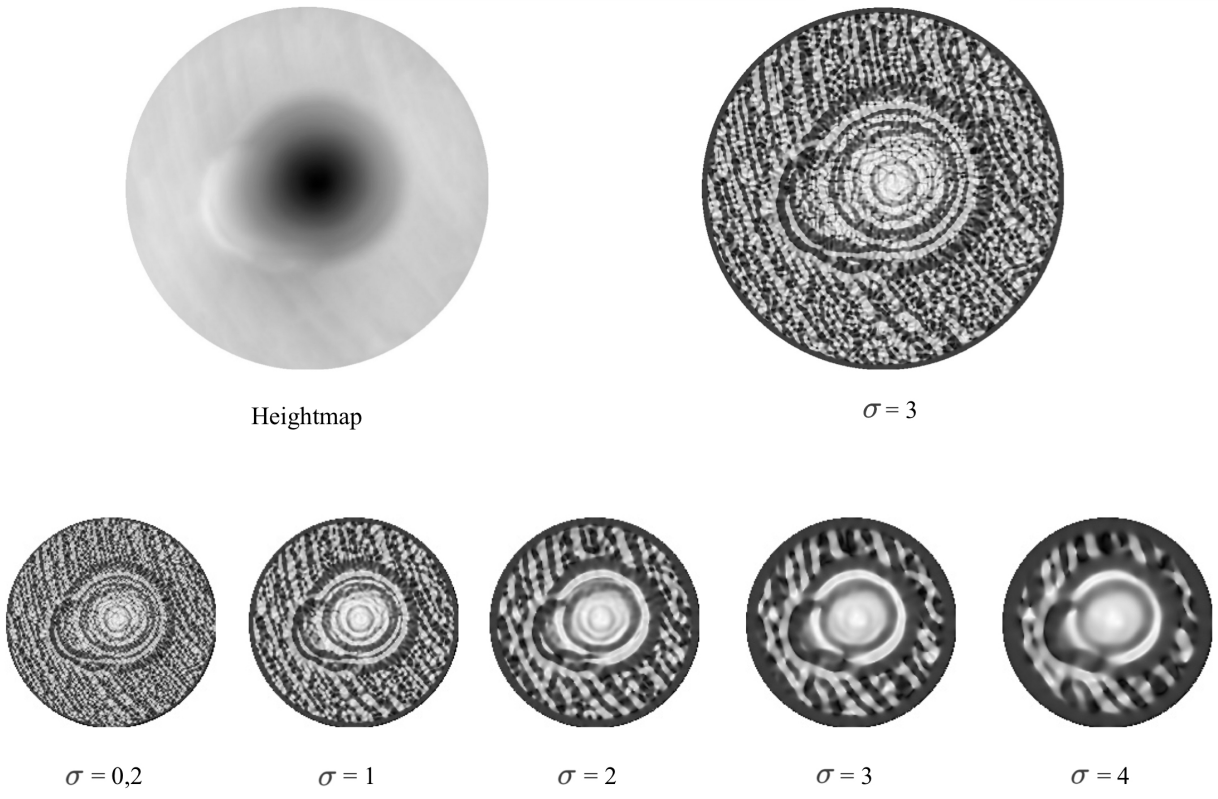

Figure 5: A height map of a cartridge case and the corresponding shape index images generated with different σ values. The shape index image with the larger dimension (top-right) is derived from the original, non-downsampled height map, while the smaller images (bottom) are generated from downsampled, reduced-resolution height maps.

Using the height maps generated by the BALISTIKA system, we collected 362,659 samples from 89 different firearm brands, covering all calibers supported by the system, such as 9×19mm and 7×65mm. Of these 89 brands, 87 are factory-manufactured original firearm brands, one represents handmade firearms, and another group includes modified firearms in which blank-firing pistols have been converted into functional firearms by replacing their barrels. The dataset images cover the breech face, primer, firing pin, and ejector regions within each cartridge case (Fig. 3). In this study, we used only the primer and firing pin regions for training.

Height map images, originally generated for 3D rendering purposes, are particularly challenging to interpret visually due to their pixel values being clustered within a narrow range, resulting in low-contrast representations that obscure the fine ballistic markings of interest. Consequently, distinguishing characteristic firearm impressions with the naked eye becomes nearly impossible. Moreover, without appropriate normalization, these clustered pixel intensities fail to provide meaningful input distributions for model training.

To prepare the data for effective model learning, we evaluated two preprocessing approaches: (1) normalizing the raw height maps to distribute pixel intensities across the usable dynamic range, and (2) applying the Shape Index

transformation [43], which maps curvature-based surface features into a standardized range, thereby enhancing local geometrical details. Our motivation for using the Shape Index was its prior use on ballistic data, where it has been shown to enhance discriminative features and improve performance [65].

The Shape Index is a classical technique in image processing that describes the local shape characteristics of surfaces by mapping curvature information to values between -1 and 1 (positive for convex, negative for concave, and near zero for flat regions). This transformation facilitates human visual interpretation, increases input discriminability, and reduces model training time. Given the eigenvalues λ_1 and λ_2 of the Hessian matrix ($\lambda_1 \geq \lambda_2$), the shape index is calculated as follows:

$$s = \frac{2}{\pi} \arctan \frac{\lambda_2 + \lambda_1}{\lambda_2 - \lambda_1} \tag{1}$$

The Gaussian filter used to extract the Hessian matrix takes σ value as a parameter. This parameter thus acts as a hyperparameter during shape index generation. While creating the shape index images, we explored the optimal window size for the region that would be most suitable for brand classification. This step was critical for determining the optimal scale at which distinctive brand-related features are most salient within the dataset.

The original height map images of cartridge cases have a resolution of 2048×2048 pixels. Because the breech face contains factory-originated marks irrelevant to our analysis, we focused on the primer area by cropping it, resulting in images of 736×736 pixels (ShapeIndex_N configuration in Table 1). To reduce input size during training, these cropped images were further downsampled to a quarter of their original size, resulting in 184×184 pixels (ShapeIndex_S configuration in Table 1). In our experiments, we used these height maps both to train deep learning models and to generate shape index transformations (Fig. 4).

We evaluated the effect of different σ values $(0.2, 1, 2, 3, 4)$ in shape index generation (Fig. 5). Smaller σ values preserve finer details crucial for distinguishing subtle features, whereas larger values emphasize broader features through smoothing. Experiments with ResNet-18 in ShapeIndex_S set showed that a σ value of 3 was near-optimal for shape index generation from down-sampled height maps, resulting in the highest accuracy for ResNet-based brand classification (Fig. 6). Therefore, σ=3 was adopted for all shape index-based experiments throughout this study.

The complete cartridge case dataset comprises 89 firearm brands. Analysis of the brand distribution reveals significant variation in sample sizes across brands. While some brands have as many as 30,000 samples, the majority (68 brands) contain fewer than 1,000 samples, and 24 of these have fewer than 5 samples each. Beyond creating substantial class imbalance, the 24 brands with fewer than 5 samples contain insufficient data for meaningful model training and generalization. Notably, the 21 most populous brands each include over 1,000 samples and together account

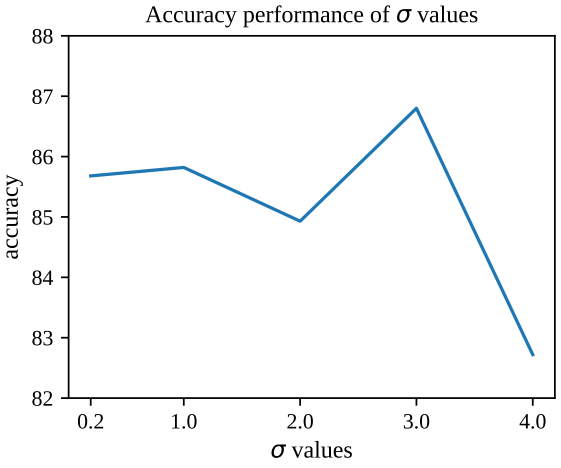

Figure 6: Effect of σ value used for the computation of shape index for the down-sampled height maps on the accuracy of the ResNet model. Lower σ values capture finer details, which are crucial for distinguishing subtle features in the data. However, as σ increases, the shape index emphasizes broader features by smoothing out smaller details or noise. The best accuracy was obtained for $\sigma = 3$.

for 97% of the entire dataset (Fig. 7). Consequently, these 21 brands—representing 351,626 out of the total 362,659 samples—were selected as the primary focus for some of our experiments.

To address class imbalance, we employ two strategies: focal loss [66], which addresses the imbalance problem by reducing the relative loss contribution of easy examples and thus focusing training on difficult and underrepresented samples, and oversampling of minority classes. Unlike standard cross-entropy loss, which treats all samples equally, focal loss incorporates a modulating factor to focus learning on difficult or misclassified samples, along with class weighting factors to address class imbalance. In our case, the primary motivation for using focal loss was not only class balancing, but also encouraging the model to prioritize harder examples during training. As for oversampling, given the rotation-invariant nature of cartridge case images, we oversampled minority classes for these 21 brands by generating rotated versions of existing samples, ensuring each class contained at least 50,000 samples. Overall, this oversampling strategy expanded the training set to exceed a million samples (Heightmap_BAL and ShapeIndex_BAL indicate the oversampled datasets in Table 1). Prior to training, the dataset was normalized using its calculated mean and standard deviation to ensure input compatibility with the models.

4.2. Models

We trained our normalized height map and shape index datasets using several learning models, including Support Vector Machines (SVM) [45], Random Forest [44], ResNets [46], and Vision Transformers (ViT) [38].

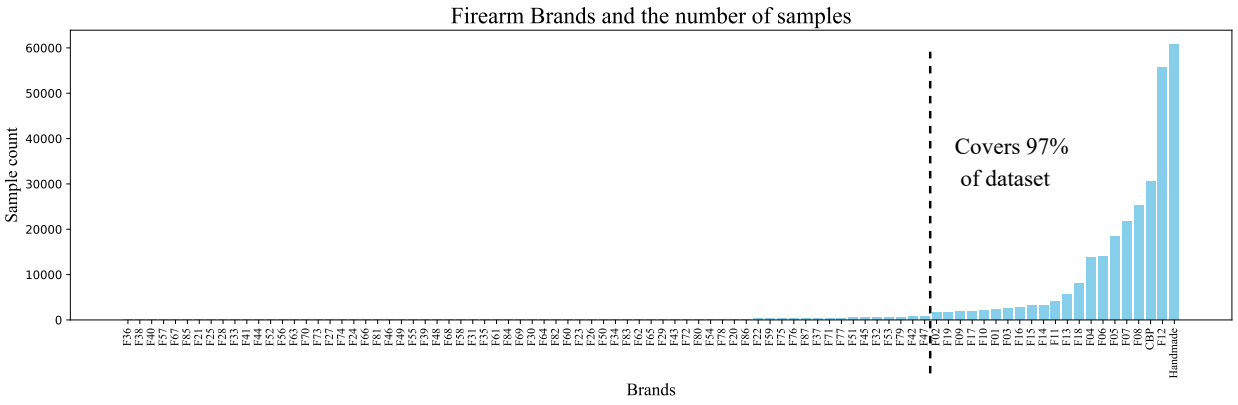

Figure 7: Complete cartridge case dataset is comprised of 89 brands. The anonymized brand names and their sample counts are provided. While some popular brands include approximately 30,000 samples, 68 brands have fewer than 1,000 samples. Notably, these 21 brands represent 97% of the firearm brand population associated with criminal cases in Türkiye, including handmade firearms and converted blank pistols (CBPs).

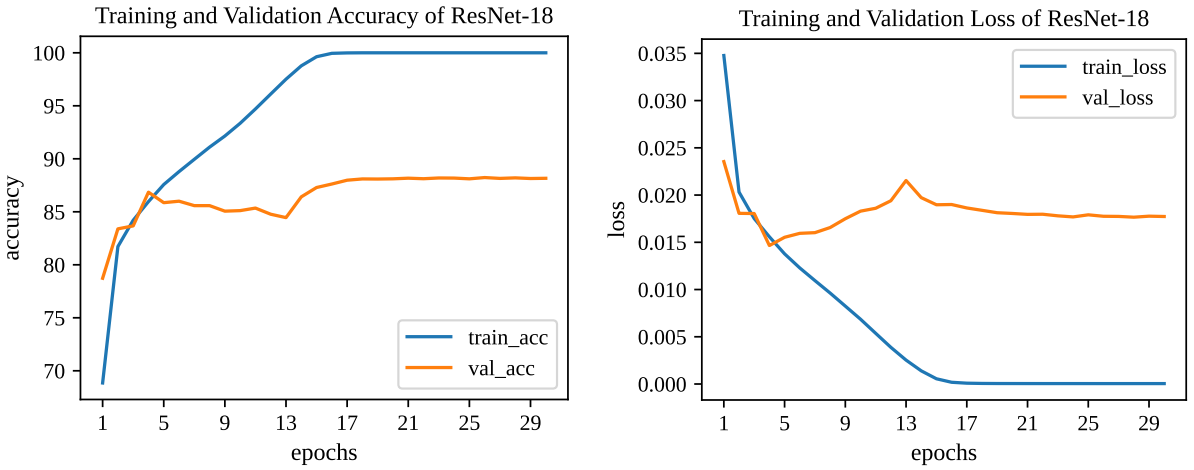

Figure 8: Accuracy and loss graphs of ResNet-18 model in training.

4.2.1. Machine Learning Models

Support Vector Machine (SVM) [45], as popular margin-based classifier, is trained using height map and shape index features extracted from ResNet-18 pretrained on ImageNet, considering the 21 most populous firearm brands. In Table 1, this model is indicated as 'ResNet-18 (ImageNet, FE) + SVM', where 'FE' denotes that ResNet-18 is employed as a feature extractor for the SVM classifier. Features were extracted from the original full-resolution shape index images (ShapeIndex_N configuration) and height map images (HeightMap configuration). A similar configuration was applied to Random Forest classifiers [44], a widely used ensemble learning method, indicated as 'ResNet-18 (ImageNet, FE) + RF' in Table 1.

4.2.2. CNN Based Models

For CNN-based deep learning approaches, we employed ResNet-18 [46] and ResNet-50 models pretrained on ImageNet [67], and fully fine-tuned them on the normalized height map and shape index datasets for 30 epochs using focal loss. To increase the contribution of minority classes to the loss, the alpha parameter of focal loss was set according to the inverse class frequencies. The gamma parameter was evaluated on the first six epochs using values of 1, 2, and 3, and the value that yielded the lowest validation loss ($\gamma = 2$) was then used for the full training. Models were optimized using stochastic gradient descent (SGD) with a learning rate of 0.005 and a momentum of 0.9. A learning rate warm-up was applied during the first five epochs to stabilize training.

Fig. 8 show the training accuracy and loss metrics for ResNet-18 trained with ShapeIndex_N configuration. As shown in the figures, the validation loss reaches its minimum value at epoch 4. After this point, the warm-up phase concludes, and the loss begins to increase. Around epoch 13, the loss begins to decrease again, accompanied by a slight improvement in accuracy. Between approximately epochs 16 and 18, both loss and accuracy reach a plateau. The final model was selected at epoch 18, as it lies within the saturation range and has the lowest validation loss within this interval.

4.2.3. Transformers

In addition to CNN-based models, we explored transformer architectures [37] for firearm brand classification. We hypothesized that the self-attention mechanism inherent in transformers could allow the model to learn the signature of the firing pin left on the cartridge case, which could be distributed across the image, more robustly. For training and testing, we employed a similar methodology to that used for CNN-based models.

Unlike CNN-based models, we employed the AdamW optimizer [68], widely adopted for transformer architectures due to its decoupled and explicit weight decay regularization. During the 50-epoch training, we introduced a "warm-up" phase in the first 5 epochs, where the learning rate was gradually increased. The overall learning rate was set to 0.0001, and the weight decay was configured as 0.05.

The training and validation accuracy and loss metrics are presented in Fig. 9. Upon examining these figures, an unusual trend is observed: as the number of epochs increases, both the validation loss and accuracy rise simultaneously. While loss captures fine-grained changes in prediction confidence, an increase in loss does not necessarily indicate incorrect predictions. For instance, the model may still correctly predict the class with the highest probability, even if the confidence in that prediction decreases, leading to higher loss values.

Another possible explanation for this trend is the significant class imbalance in the dataset. If the model begins to overlook underrepresented classes, this could result in an increase in loss due to the model making more errors in these minority classes. At the same time, accuracy might increase because the model performs better on the dominant

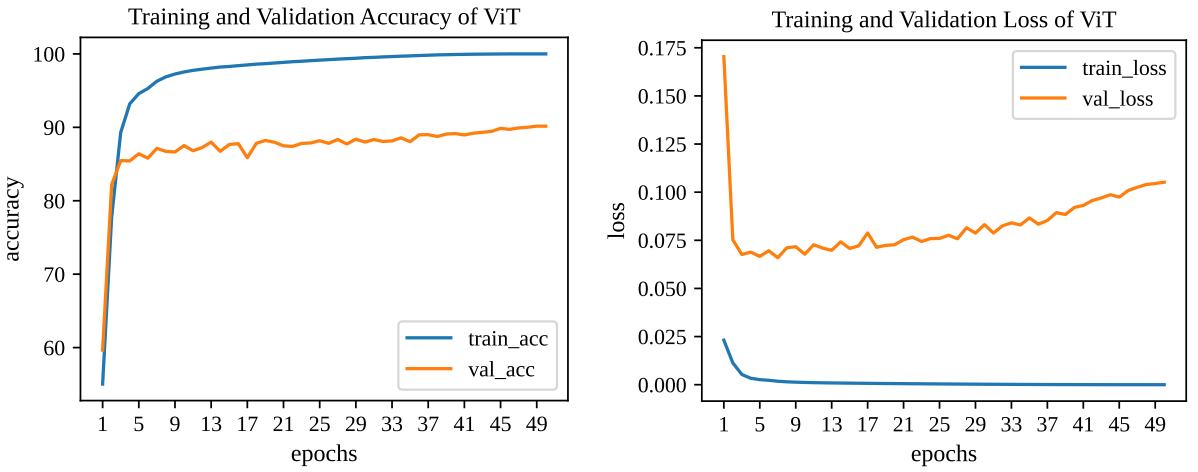

Figure 9: Accuracy and loss graphs of ViT model in training.

classes. As no notable improvement in accuracy was observed beyond approximately epoch 36, and validation loss increased significantly, we selected the model with the lowest validation loss near this epoch as the final model.

5. Evaluation

In this section, we evaluate the classification performance of different models and preprocessing approaches used in this study. The results are reported in terms of accuracy, calculated as the ratio of correct predictions to the total number of queries (Equation 2), to provide a clear measure of each model's effectiveness. A summary of the accuracy outcomes for all models and dataset configurations is presented in Table 1.

$$acc(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(y_i == \hat{y}_i) \tag{2}$$

The experiments conducted with various models revealed distinct performance differences across approaches. Among the classical machine learning models, Support Vector Machines (SVM) achieved their highest accuracy of 65.3% on the 21-brand subset when trained using features extracted from the height maps, while Random Forest classifiers reached a comparable accuracy of 64.1% under the same configuration. Overall, the accuracies of both models remained below 66%, highlighting the limited discriminative capacity of these approaches for the subtle brand-related patterns in the dataset. Both models showed slightly better results with height maps than with shape index images, possibly because height maps retain intensity variations that align more closely with the low-level filters learned from ImageNet, whereas the shape index transformation increases contrast and emphasizes fine-scale details, which may diverge from the distributions expected by the pretrained feature extractor and reduce feature quality.

Table 1
This table presents the test accuracy results of the models for the 21 most populous brands. The last row shows the accuracy obtained for all 89 brands.

Model	Configuration	Accuracy (%)
ResNet-18 (ImageNet, FE) + SVM	ShapeIndex_N	56.1
ResNet-18 (ImageNet, FE) + SVM	HeightMap	65.3
ResNet-18 (ImageNet, FE) + RF	ShapeIndex_N	63.3
ResNet-18 (ImageNet, FE) + RF	HeightMap	64.1
ResNet-18	ShapeIndex_S	86.8
ResNet-18	HeightMap	88.0
ResNet-18	ShapeIndex_N	88.2
ResNet-18	HeightMap_BAL	89.7
ResNet-18	ShapeIndex_BAL	**90.8**
ResNet-50	ShapeIndex_S	87.3
ResNet-50	HeightMap_BAL	**91.2**
ResNet-50	ShapeIndex_BAL	**91.6**
ViT	HeightMap_BAL	**90.0**
ViT	ShapeIndex_BAL	**90.2**
ResNet-18	HeightMap (89 brands)	86.0

In contrast, deep learning models demonstrated substantially higher performance. ResNet-based architectures achieved up to 86% test accuracy on the normalized height map dataset covering 89 brands. Their performance further improved with the use of high-resolution inputs and class balancing strategies. In particular, oversampling minority classes (Heightmap_BAL and ShapeIndex_BAL configurations) effectively addressed class imbalance in the 21-brand subset and yielded the highest accuracies across all configurations with ResNet-18 (89.7% and 90.8%, respectively). Moreover, preserving the original resolution of the height maps during the shape index transformation (ShapeIndex_N) retained fine ballistic markings more effectively than downsampled inputs (ShapeIndex_S), enhancing discriminability (88.2% vs. 86.8%). The shape index transformation, although a non-linear mapping derived from height maps, further enhanced surface details and slightly outperformed normalized height maps. Among all models tested, ResNet-50 achieved the highest accuracy, approximately 92%, when trained on the balanced ShapeIndex_BAL dataset with focal loss.

Similarly, Vision Transformer models achieved an accuracy of around 91% on the test set, demonstrating comparable performance to the best CNN-based configurations. These results suggest that both transformer and convolutional architectures can effectively capture brand-specific ballistic features and provide highly promising solutions for firearm brand classification tasks.

5.1. Class Topology and Difficulty Analysis

For per-class performance analysis, two testing approaches were employed. In the first approach, the query count per class was adjusted to reflect real-world occurrence statistics, resulting in an imbalanced distribution. In the second approach, the number of samples per class was balanced to ensure uniform representation.

We report the model's precision, recall, and F1-score metrics for the 21 populous firearm brands. Table 2 and Table 3 summarize these results along with their average performances using micro, macro, and weighted averaging methods. Micro averaging computes metrics globally by aggregating True Positives, False Positives, and False Negatives across all classes, making it suitable for balanced datasets. Macro averaging calculates metrics independently for each class and takes their arithmetic mean, treating all classes equally. Weighted averaging, on the other hand, adjusts for class imbalance by weighting each class's metrics according to its support. For information security purposes, brand names in the results have been anonymized. The terms "handmade" and "CBP" refer to handcrafted firearms and converted blank pistols, respectively.

The ResNet-50 and Vision Transformer models that were evaluated exhibited higher performance on the imbalanced test set compared to the balanced test set. For ResNet, the micro-averaged F1 score increased from 0.84 (balanced) to 0.91 (imbalanced), while for ViT, it increased from 0.80 to 0.90. These results are shown in Table 2 and Table 3, respectively. Although such an outcome can be expected given the dominance of majority classes in the imbalanced setting, it was nonetheless surprising in our context, as the oversampled training configuration was intended to mitigate class imbalance effects and yield comparable performance across both test types. This trend may be attributed to models learning majority classes more effectively, leading to improved accuracy when these classes dominate the test distribution. This may rather be explained by the fact that rotational oversampling of minority classes in the balanced test set increased the number of inherently challenging examples within these classes. Conversely, when minority classes are oversampled for the balanced test set, their difficult or outlier examples are replicated along with the easier ones, whereas majority classes, due to their larger sample sizes, inherently have a lower proportion of outliers. While the exact underlying cause warrants further investigation, these results suggest that the models are well-adapted to real-world class distributions and underscore the importance of carefully considering dataset structure when evaluating operational performance in forensic applications.

As a further investigation, we analyzed the relationships between per-class accuracy and two key variables: sample count and intra-class distance. Pearson and Spearman correlation analyses were conducted for each comparison to evaluate both linear and monotonic trends.

The results showed no strong correlation between class sample count and accuracy (Fig. 10), suggesting that the model effectively addressed the class imbalance problem and was able to learn even from underrepresented classes. In other words, the number of samples per class did not emerge as the dominant factor determining model performance.

Similarly, no significant relationship was observed between intra-class distance and accuracy (Fig. 11). This indicates that intra-class distance alone does not sufficiently explain the difficulty of classifying a particular class. Direct measurements of intra-class variation showed that most classes are well-centered with comparable variances. Additionally, analysis of inter-class distances, with an average value of 0.2272 (Table 5), revealed that the centers of

Table 2
The classification results of firearm brands on the test set of ShapeIndex_BAL using ResNet-50, reflecting the imbalanced real-world distribution of firearms. Brands with F1 score greater than 0.9 are shown in bold.

Brands	#Train	#Test	Precision	Recall	F1 Score
CBP	61120	3821	0.92	0.92	**0.92**
F01	51635	279	0.77	0.66	0.71
F02	47520	204	0.87	0.69	0.77
F03	60559	316	0.78	0.55	0.65
F04	55200	1683	0.98	0.95	**0.96**
F05	55077	2253	0.79	0.86	0.82
F06	55916	1861	0.85	0.83	0.84
F07	65394	2794	0.94	0.94	**0.94**
F08	50444	3149	0.96	0.96	**0.96**
F09	54360	232	0.79	0.81	0.80
F10	49080	262	0.95	0.92	**0.94**
F11	53040	503	0.96	0.97	**0.97**
F12	55683	6896	0.94	0.96	**0.95**
F13	51066	714	0.83	0.82	0.82
F14	51552	402	0.83	0.68	0.75
F15	50576	446	0.82	0.79	0.80
F16	51813	321	0.83	0.87	0.85
F17	51660	218	0.93	0.89	**0.91**
F18	56882	980	0.83	0.77	0.80
F19	50760	220	0.77	0.65	0.70
Handmade	60862	7627	0.94	0.95	**0.95**
	ViT Micro avg.		0.90	0.90	0.90
	ViT Macro avg.		0.85	0.80	0.82
	ViT Weighted avg.		0.90	0.90	0.90
	ResNet Micro avg.		0.91	0.91	0.91
	ResNet Macro avg.		0.87	0.83	0.85
	ResNet Weighted avg.		0.91	0.91	0.91

certain classes are located quite close to each other in the feature space, making them inherently harder to distinguish (Table 6).

Confusion matrix analyses showed that frequently confused class pairs often involved one majority and one minority class (Table 4). To further investigate whether these confusions stemmed from class imbalance or intrinsic data characteristics, we systematically analyzed the frequently confused brand pairs and examined their inter-class distances in the feature space. These pairs exhibited uniformly low inter-class distances without distinctive separation patterns, suggesting genuine morphological similarities rather than model limitations.

To understand these similarities from a forensic perspective, we consulted ballistics experts. They explained that many of the frequently confused brand pairs involve domestic reproductions of foreign firearm designs (e.g., F02−F08, F03−F05, F19−F12) or brand families containing sub-models with distinct characteristic marks, some of which resemble those of other brands (e.g., F14−F12). Additionally, several poorly performing classes were noted to contain

Table 3
The classification results of firearm brands on balanced test set of ShapeIndex_BAL using ResNet-50 model. Brands with F1 score greater than 0.9 are shown in bold.

Brands	#Train	#Test	Precision	Recall	F1 Score
CBP	61120	200	0.81	0.94	0.87
F01	51635	200	0.92	0.65	0.76
F02	47520	200	0.98	0.69	0.81
F03	60559	200	0.96	0.56	0.70
F04	55200	200	0.96	0.93	**0.95**
F05	55077	200	0.53	0.85	0.66
F06	55916	200	0.71	0.85	0.78
F07	65394	200	0.83	0.94	0.88
F08	50444	200	0.87	0.96	**0.91**
F09	54360	200	0.96	0.81	0.88
F10	49080	200	0.98	0.94	**0.96**
F11	53040	200	0.99	0.98	**0.98**
F12	55683	200	0.64	0.95	0.76
F13	51066	200	0.83	0.84	0.83
F14	51552	200	0.90	0.69	0.79
F15	50576	200	0.92	0.80	0.85
F16	51813	200	0.96	0.88	**0.92**
F17	51660	200	0.97	0.90	**0.93**
F18	56882	200	0.84	0.78	0.81
F19	50760	200	0.96	0.66	0.78
Handmade	60862	200	0.66	0.96	0.78
	ViT Micro avg.		0.80	0.80	0.80
	ViT Macro avg.		0.84	0.80	0.82
	ViT Weighted avg.		0.84	0.80	0.80
	ResNet Micro avg.		0.84	0.84	0.84
	ResNet Macro avg.		0.87	0.84	0.85
	ResNet Weighted avg.		0.87	0.84	0.84

highly heterogeneous models lacking consistent brand-specific morphological features, further complicating accurate classification.

Overall, these findings indicate that the observed confusions are not merely artifacts of class imbalance but rather stem from genuine morphological similarities and intra-class heterogeneity inherent to firearm manufacturing practices.

5.2. Model Explainability Analysis

To interpret and visualize the decision-making process of the trained models, we employed two widely used model explainability techniques: Grad-CAM [69] and Occlusion sensitivity [70]. Grad-CAM generates class-discriminative heatmaps by computing the gradients of target classes flowing into the final convolutional layers, thereby highlighting important regions contributing to the prediction. In contrast, occlusion sensitivity systematically occludes parts of the input image to measure changes in output probability, revealing which regions are most critical for the model's

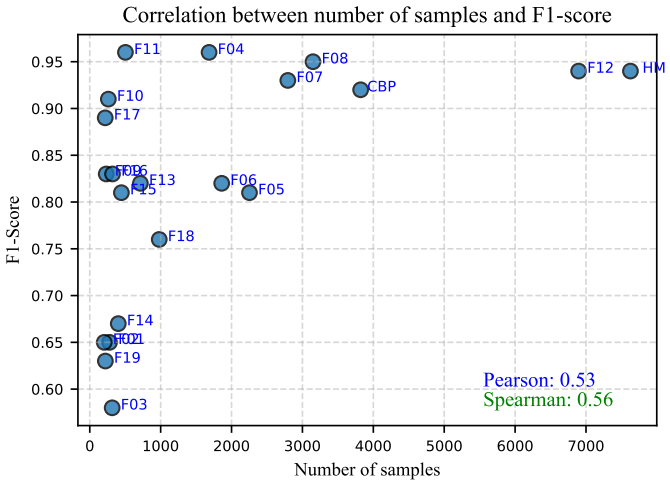

Figure 10: Pearson and Spearman correlation coefficients for the relationship between sample count and F1 score.

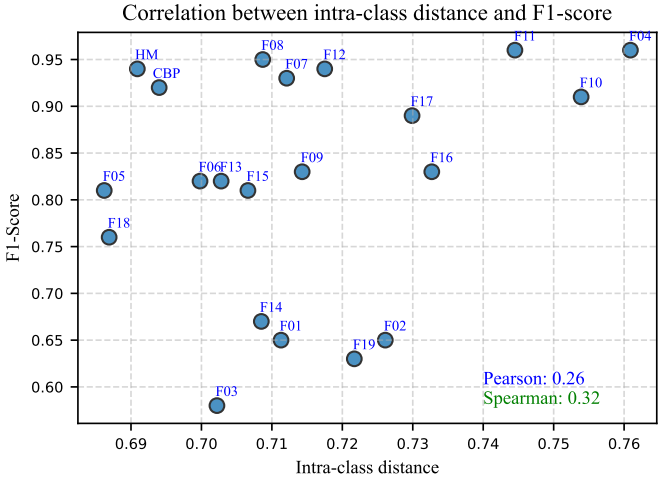

Figure 11: Pearson and Spearman correlation coefficients for the relationship between intra-class distance and F1 score.

decision. Together, these methods provide complementary insights into the spatial features utilized by the models in firearm brand classification.

As seen in Fig. 12, model visualizations were examined using the ResNet-18 model trained with the highest input resolution (ShapeIndex_N configuration). Grad-CAM was not applied to the Vision Transformer model, as gradient-based visualizations like Grad-CAM are incompatible with ViT's architecture due to its patch-based processing and lack of convolutional layers.

Grad-CAM produced reasonable localization maps highlighting the general primer area but did not focus on fine detailed marks or thin impressions. This limitation is likely due to the resolution of the final convolutional layer output;

Table 4
Confusion matrix shows the ViT model's classification results on the test set that reflects the imbalanced real-world distribution of firearms. Rows represent the ground truth classes, while columns represent the predicted classes, and thus the diagonals indicate correct predictions. Although the matrix is normalized such that each row sums to 100, values less than 0.5 are rounded down to zero, which may cause row sums to appear slightly less than 100. The diagonals correspond to recall, and the ratio of each diagonal value to its respective column sum corresponds to precision.

	CBP	F01	F02	F03	**F04**	F05	F06	F07	**F08**	F09	F10	**F11**	**F12**	F13	F14	F15	F16	F17	F18	F19	**HM**
CBP	**93**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5
F01	0	**62**	1	1	0	**8**	4	2	1	1	1	0	0	0	6	**8**	2	0	0	0	2
F02	0	1	**55**	0	0	3	5	2	**18**	1	0	0	1	0	**9**	0	1	1	0	0	0
F03	0	1	0	**51**	0	**26**	2	2	3	1	0	0	1	0	**8**	0	0	1	1	1	3
F04	0	0	0	0	**95**	0	0	0	2	2	0	0	0	0	0	0	0	0	1	0	0
F05	0	1	0	2	0	**83**	2	1	3	0	0	0	2	0	3	1	0	1	0	0	2
F06	0	0	0	0	0	3	**81**	2	0	1	0	0	1	0	6	1	1	0	0	0	2
F07	0	0	0	0	0	1	1	**92**	0	0	0	0	0	0	2	0	0	0	0	0	1
F08	0	0	0	0	0	0	0	1	**96**	0	0	1	0	0	0	0	0	0	0	0	0
F09	0	0	1	0	2	3	2	3	0	**78**	0	0	0	6	0	0	1	0	0	3	0
F10	1	0	1	0	1	0	0	1	1	0	**89**	2	0	2	0	0	0	0	0	0	0
F11	0	0	0	0	0	1	1	2	0	0	0	**95**	0	0	0	0	0	0	0	0	0
F12	0	0	0	1	1	1	0	0	0	0	0	0	**96**	0	0	0	0	0	1	0	0
F13	0	0	0	2	2	1	3	2	0	0	1	0	4	**81**	0	0	0	0	2	1	1
F14	0	0	0	6	6	3	3	3	0	0	1	0	**10**	1	**60**	0	0	0	3	0	2
F15	1	0	0	7	1	0	7	0	0	0	1	0	6	0	0	**75**	1	0	0	0	0
F16	0	0	1	1	1	0	**8**	4	0	0	1	0	0	0	0	0	**83**	0	0	0	0
F17	0	1	3	1	2	0	2	3	0	0	1	0	0	0	0	0	0	**86**	0	0	0
F18	0	0	0	6	6	2	1	1	0	0	1	0	5	1	1	0	0	0	**75**	0	1
F19	0	0	0	2	4	2	1	3	0	0	4	0	**22**	2	1	1	1	0	3	**51**	2
HM	3	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**95**

despite using the highest available input size, the feature maps remained relatively coarse, and upscaling these low-resolution maps to the input image size resulted in blurred, broadly highlighted regions.

In contrast, occlusion sensitivity analysis yielded highly detailed and precise maps, effectively capturing fine morphological features across the primer surface. These results appeared sufficiently precise to allow direct forensic interpretation. Consequently, occlusion sensitivity heatmaps were overlaid onto the BALISTIKA render outputs to facilitate expert evaluation of model focus areas in a realistic forensic visualization setting.

For this analysis, we first computed a custom difficulty metric incorporating prediction confidence, correctness, loss value, and each sample's distance to its class center in the feature space. Using this metric, we identified the ten easiest and ten hardest samples for each of the 21 most prevalent brands. From these subsets, we randomly selected two easy and two hard examples per brand to generate the occlusion maps. The maps were produced with low patch and stride sizes, resulting in high-resolution heatmaps that clearly revealed the image regions contributing most to predictions (Figs. 13, 14, 15).

The ballistics expert assessments confirmed that the model consistently focused on relevant discriminative features, as easy examples typically exhibited characteristic marks representative of their firearm brand classes. Their evaluation

Table 5
Inter-class distance matrix for the 21-brand set. Distance values are presented in $-\log_{10}(x)$ format to improve readability. The mean inter-class distance across all brand pairs is 0.2272.

	F02	F03	F04	F05	F06	F07	HM	F08	F09	F10	CBP	F11	F12	F13	F14	F15	F16	F17	F18	F19
F01	3.2	3.4	2.6	3.0	3.3	4.5	3.1	3.9	3.9	2.7	3.1	2.8	3.5	3.4	3.9	3.7	3.0	3.1	3.0	3.3
F02		3.0	2.8	2.8	2.9	3.2	2.8	3.1	3.3	2.9	2.8	3.1	3.4	3.0	3.1	3.0	3.5	3.8	2.8	3.7
F03			2.6	3.2	4.0	3.4	3.3	3.5	3.3	2.6	3.4	2.7	3.2	4.6	3.6	3.7	2.9	2.9	3.2	3.1
F04				2.5	2.5	2.6	2.5	2.6	2.6	3.5	2.5	3.1	2.7	2.6	2.6	2.6	2.9	2.8	2.5	2.7
F05					3.2	3.0	3.7	3.0	2.9	2.5	3.5	2.6	2.9	3.2	3.0	3.1	2.7	2.7	4.5	2.8
F06						3.3	3.4	3.4	3.2	2.6	3.6	2.7	3.1	3.9	3.4	3.5	2.8	2.9	3.3	3.0
F07							3.0	3.8	4.0	2.7	3.1	2.8	3.6	3.4	3.8	3.6	3.0	3.1	3.0	3.4
HM								3.1	3.0	2.5	3.9	2.6	2.9	3.3	3.1	3.2	2.7	2.8	3.8	2.9
F08									3.6	2.7	3.2	2.8	3.4	3.6	5.2	4.0	3.0	3.0	3.0	3.2
F09										2.7	3.1	2.8	3.8	3.3	3.6	3.5	3.1	3.1	2.9	3.5
F10											2.6	3.3	2.8	2.6	2.7	2.6	3.0	2.9	2.5	2.8
CBP												2.6	3.0	3.4	3.2	3.3	2.8	2.8	3.5	2.9
F11													2.9	2.7	2.8	2.8	3.3	3.2	2.6	3.0
F12														3.2	3.4	3.3	3.1	3.2	2.9	3.7
F13															3.6	3.8	2.9	2.9	3.2	3.1
F14																4.1	3.0	3.0	3.0	3.2
F15																	2.9	3.0	3.1	3.2
F16																		3.9	2.7	3.3
F17																			2.7	3.4
F18																				2.8

of the hard examples provided further insights into model performance by revealing two distinct patterns: some hard examples were actually mislabelled in the dataset, with the model's seemingly incorrect predictions being correct upon expert review, while others genuinely lacked clear brand-specific features, reflecting true forensic ambiguity rather than model failure. Overall, these findings demonstrate that the model's decision-making aligns well with forensic expertise, focusing on meaningful image regions that carry brand-identifying information.

5.3. Implications for Operational Deployment and Forensic Application

To evaluate the practical applicability of our approach, we conducted extensive testing of the ResNet-18 model integrated within the BALISTIKA third-generation comparison system. The evaluation covered three common calibers (7.65×17mm, 9×19mm and 7.62×39mm) and involved 427 cartridge case queries against a large database consisting of 944,837 cartridge cases.

By incorporating the model's high-confidence predictions (confidence >99%) as a ranking enhancement strategy, we observed an approximate 5% improvement in top-100 matching accuracy. This improvement demonstrates the model's substantial practical value in effectively narrowing down candidate pools within large-scale forensic databases, thereby contributing to a reduction in expert workload and accelerating case processing.

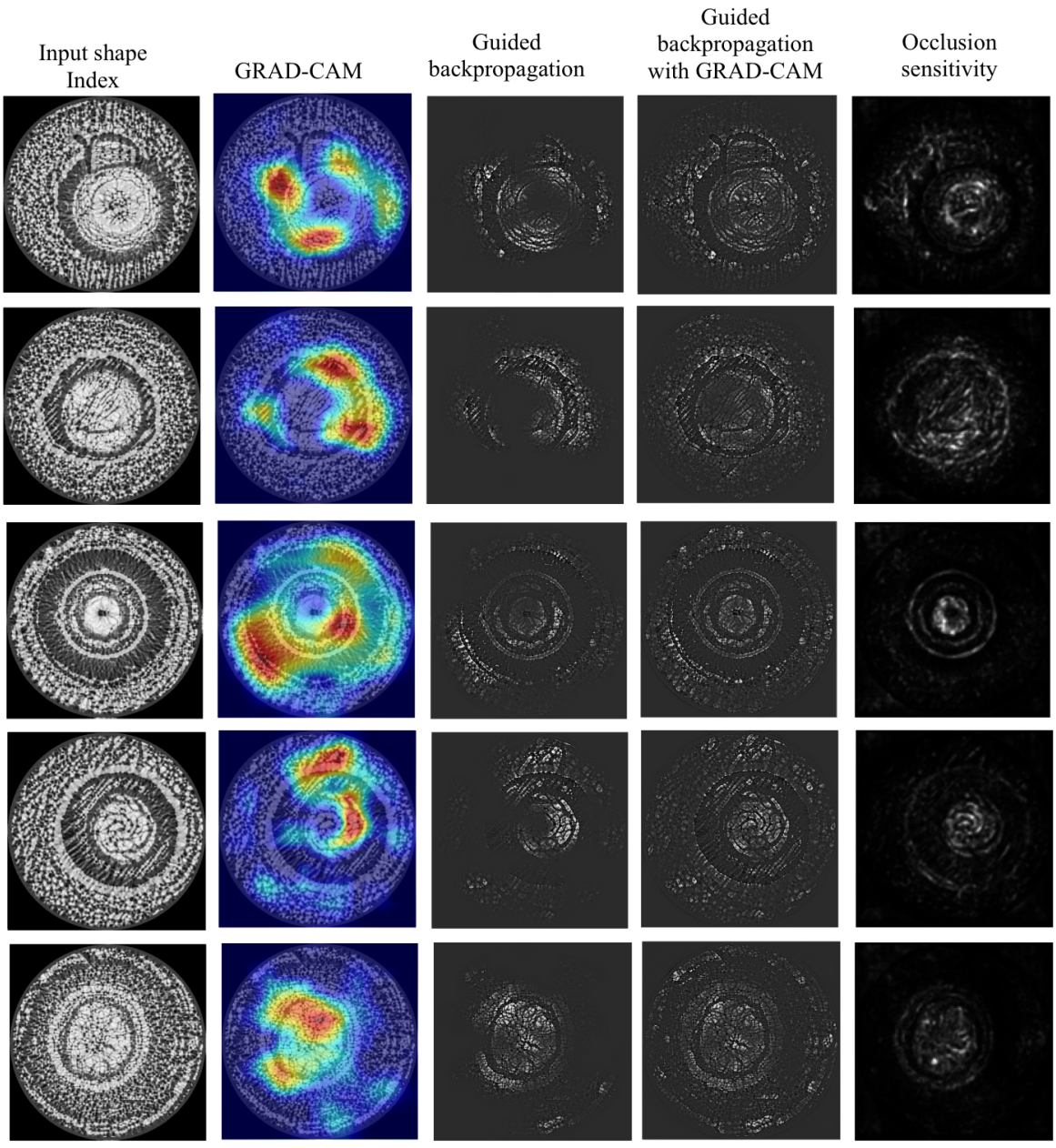

Figure 12: The figure presents the Grad-CAM, Guided Backpropagation combined with Grad-CAM, and Occlusion Sensitivity outputs of the ResNet-50 model. The leftmost image shows the shape index input provided to the model.

6. Conclusion

In this study, we conducted firearm brand classification on cartridge case images using the BALISTIKA dataset. Although our experiments included 89 firearm brands, we primarily focused on the 21 most prevalent brands, which collectively account for 97% of all samples. We trained models directly on normalized height map images as well as their corresponding shape index transformations to emphasize fine markings. Our analyses revealed that the

Table 6
Intra-class distances of the brands computed using features extracted from the ResNet-18 model with the Heightmap_BAL configuration. Brands are ordered in ascending order based on their mean intra-class distances.

Brands	Mean	Median	std	max
F05	0.6862	0.6867	0.0276	0.8081
F18	0.6869	0.6875	0.0274	0.7852
HM	0.6909	0.6910	0.0257	0.7951
CBP	0.6940	0.6947	0.0242	0.7874
F06	0.6998	0.6997	0.0318	0.8176
F03	0.7022	0.7036	0.0242	0.7943
F13	0.7028	0.7039	0.0259	0.8126
F15	0.7066	0.7065	0.0280	0.8028
F14	0.7085	0.7099	0.0247	0.7859
F08	0.7087	0.7096	0.0245	0.7997
F01	0.7113	0.7127	0.0263	0.7903
F07	0.7121	0.7130	0.0251	0.8021
F09	0.7143	0.7145	0.0274	0.8147
F12	0.7175	0.7180	0.0274	0.8259
F19	0.7217	0.7231	0.0247	0.7864
F02	0.7261	0.7278	0.0271	0.7951
F17	0.7299	0.7306	0.0282	0.8192
F16	0.7327	0.7330	0.0273	0.8169
F11	0.7445	0.7458	0.0233	0.8132
F10	0.7539	0.7539	0.0313	0.8458
F04	0.7609	0.7633	0.0311	0.8420

height map images exhibited strong pixel value clustering, and applying normalization significantly improved model performance by enhancing representational learning. Similarly, shape index transformations inherently provide a form of normalization by converting raw height values into relative curvature-based features, emphasizing meaningful local variations for the classification task.

To capture surface features at different scales, we generated shape index data with varying σ values and found that selecting an appropriate σ was crucial to emphasize distinctive marks relevant to firearm brand differentiation. Using these prepared datasets, we trained Support Vector Machine, Random Forest, ResNet, and Vision Transformer models with optimized configurations.

The results showed that deep learning models substantially outperformed traditional machine learning approaches. Moreover, shape index images generated from original height maps achieved slightly higher accuracy (88.2%) compared to normalized height maps (88.0%) for brand classification with ResNets. To address the significant class imbalance in the dataset, we employed focal loss and oversampled the minority classes by generating rotated versions of existing samples. This balancing strategy further improved model performance, with the ResNet-50 achieving the highest accuracy of 91.6%, indicating strong potential for practical forensic applications.

Finally, to interpret model decisions and identify the regions most influential in classification, we employed Grad-CAM and occlusion sensitivity visualization techniques. While Grad-CAM produced inherently low-resolution

saliency maps, occlusion sensitivity yielded promising results by accurately highlighting critical regions in high resolution. Evaluations conducted with ballistics experts confirmed that the models consistently focused on the correct forensic features and produced reliable predictions. Furthermore, the model successfully identified mislabeled samples within the dataset, demonstrating both its practical applicability for real-world forensic analysis and its potential to improve data quality in operational deployments.

Overall, this study demonstrates a solution capable of achieving practically applicable accuracy for firearm brand classification across a wide range of brands. Although the number of firearm brands encountered in forensic casework may increase over time, and our focus on the 21 most common brands does not fully cover the entire dataset, these brands account for the overwhelming majority of real-world cases. In forensic practice, the ability to distinguish among such frequently encountered brands significantly narrows down the pool of possible matches, thereby reducing the workload for ballistic experts. The proposed solution enables efficient brand-level narrowing within datasets containing millions of samples, dramatically accelerating the comparison process and enhancing operational efficiency in forensic investigations.

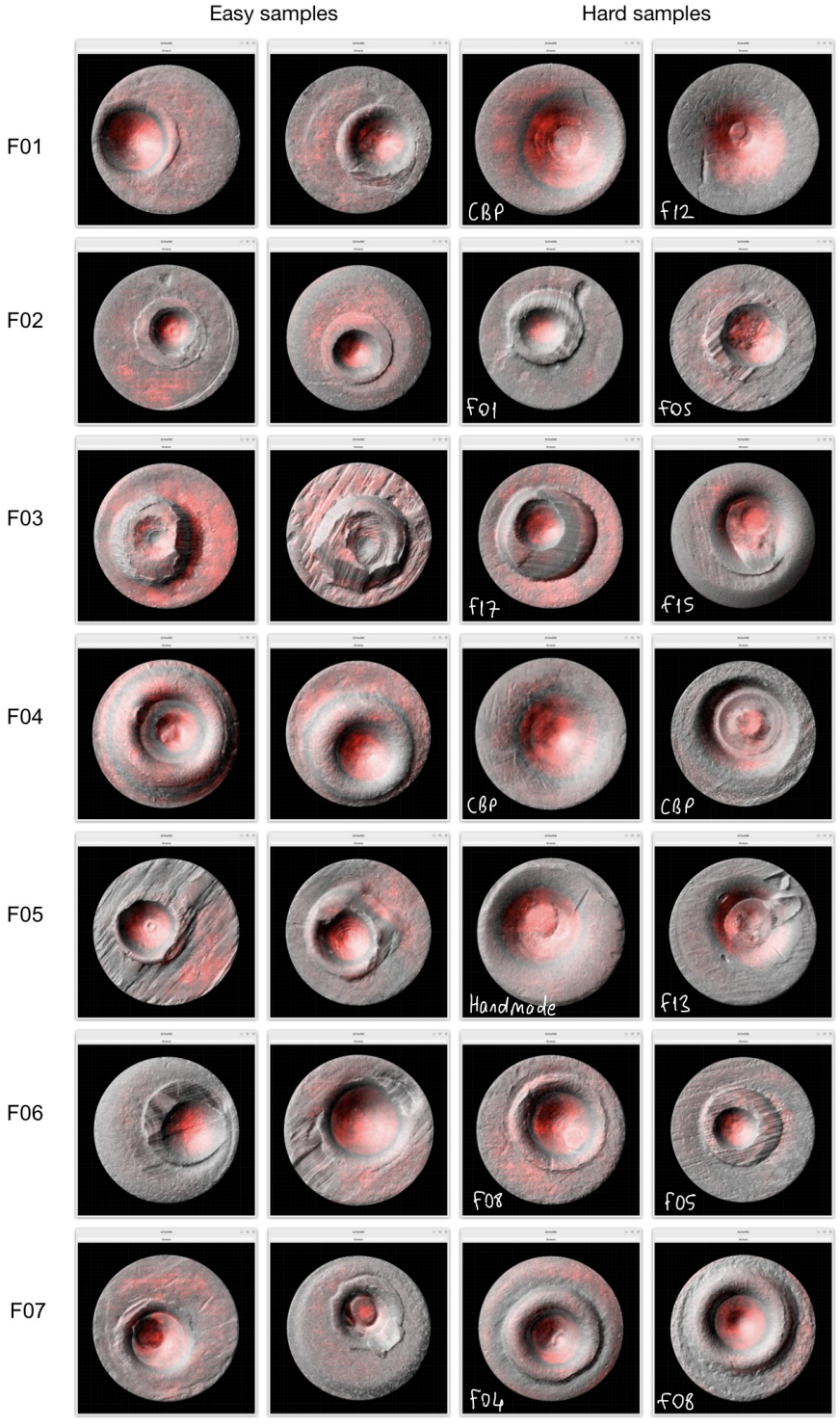

Figure 13: Occlusion sensitivity maps of the most popular firearm brands using the ResNet-18 model. The two images on the left show easy samples, while the two on the right are considered hard by the model. The model correctly predicted the easy samples but misclassified the hard ones. Its predictions for the hard samples are labeled in the bottom-left corner of these images.

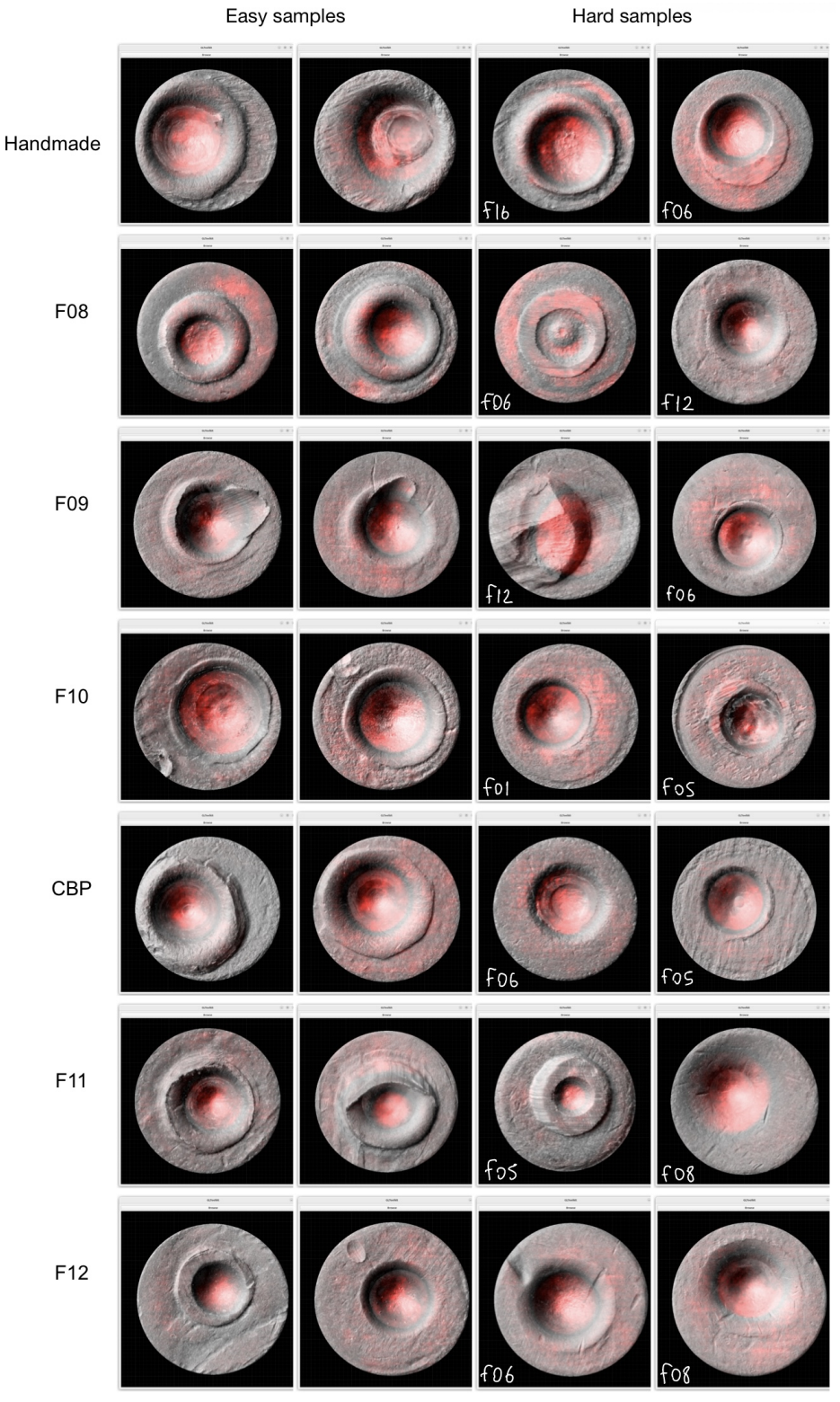

Figure 14: Occlusion sensitivity maps of the most popular firearm brands using the ResNet-18 model. The two images on the left show easy samples, while the two on the right are considered hard by the model. The model correctly predicted the easy samples but misclassified the hard ones. Its predictions for the hard samples are labeled in the bottom-left corner of these images.

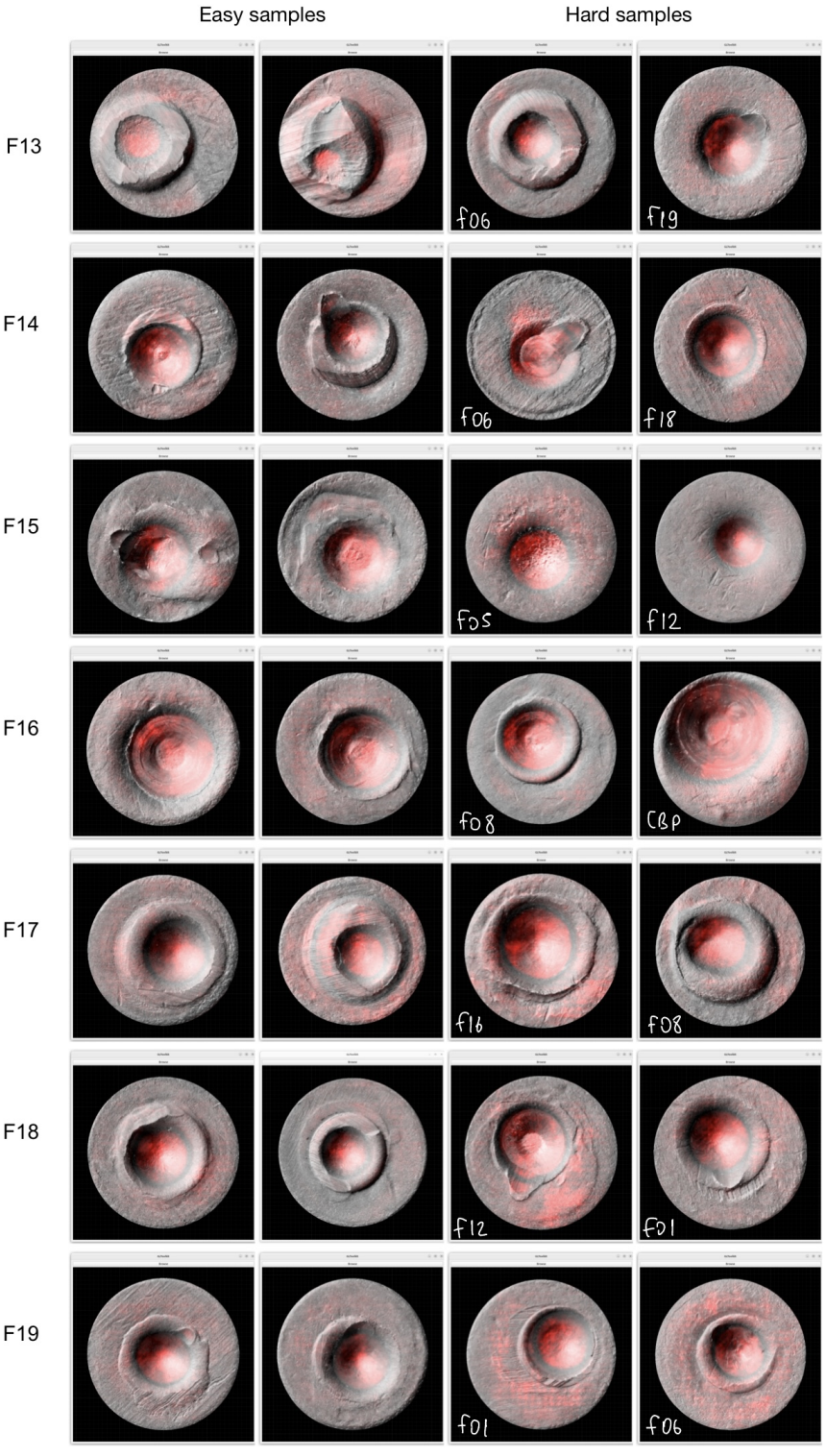

Figure 15: Occlusion sensitivity maps of the most popular firearm brands using the ResNet-18 model. The two images on the left show easy samples, while the two on the right are considered hard by the model. The model correctly predicted the easy samples but misclassified the hard ones. Its predictions for the hard samples are labeled in the bottom-left corner of these images.

References

[1] "Balistika." https://balistika.com.tr/. Accessed: 19 Nov. 2023.

[2] "The ibis solution – integrated ballistic identification system." https://www.ultra-forensictechnology.com/en/products-and-services/firearm-and-tool-mark-identification-ibis/ibis-solution-overview/#:~:text=IBIS%20uses%20specialized%203D%20microscopy,seamlessly%20within%20an%20integrated%20network. Accessed: 19 Nov. 2023.

[3] "Evofinder automated ballistics identification." http://evofinder.com/. Accessed: 19 Nov. 2023.

[4] "Arsenal automated ballistics identification system." https://papillonsystems.com/products/programs/arsenal/. Accessed: 19 Nov. 2023.

[5] S. R. Dubey, "A decade survey of content based image retrieval using deep learning," *Computer Vision and Pattern Recognition (CVPR)*, 2020.

[6] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," *Computer Vision and Pattern Recognition (CVPR)*, 2014.

[7] H. Lai, Y. Pan, Y. Liu, and S. Yan, "Simultaneous feature learning and hash coding with deep neural networks," *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[8] R. Zhang, L. Lin, R. Zhang, W. Zuo, and L. Zhang, "Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification," *IEEE Transactions on Image Processing*, 2015.

[9] A. Gordo, J. Almazan, J. Revaud, and D. Larlus, "End-to-end learning of deep visual representations for image retrieval," *International Journal of Computer Vision (IJCV)*, 2017.

[10] H. Liu, R. Wang, S. Shan, and X. Chen, "Deep supervised hashing for fast image retrieval," *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[11] W.-J. Li, S. Wang, and W.-C. Kang, "Feature learning based deep supervised hashing with pairwise labels," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2016.

[12] Y. Cao, B. Liu, M. Long, and J. Wang, "Hashgan: Deep learning to hash with pair conditional wasserstein gan," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[13] J. Li, W. W. Ng, X. Tian, S. Kwong, and H. Wang, "Weighted multideep ranking supervised hashing for efficient image retrieval," *International Journal of Machine Learning and Computing (IJMLC)*, 2019.

[14] S. Huang, Y. Xiong, Y. Zhang, and J. Wang, "Unsupervised triplet hashing for fast image retrieval," in *Thematic Workshops of ACM Multimedia*, 2017.

[15] K. G. Dizaji, F. Zheng, N. Sadoughi, Y. Yang, C. Deng, and H. Huang, "Unsupervised deep generative adversarial hashing network," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[16] Y. Gu, H. Zhang, Z. Zhang, and Q. Ye, "Unsupervised deep triplet hashing with pseudo triplets for scalable image retrieval," *IEEE Multimedia Tools and Applications*, 2019.

[17] Y. Shen, L. Liu, and L. Shao, "Unsupervised binary representation learning with deep variational networks," *International Journal of Computer Vision (IJCV)*, 2019.

[18] J. Zhang and Y. Peng, "Ssdh: Semi-supervised deep hashing for large scale image retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 1, pp. 212–225, 2017.

[19] S. Zhang, J. Li, and B. Zhang, "Pairwise teacher-student network for semi-supervised hashing," in *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[20] Y. Shen, J. Qin, J. Chen, M. Yu, L. Liu, F. Zhu, F. Shen, and L. Shao, "Auto-encoding twin-bottleneck hashing," in *Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2827, 2020.

[21] E.-J. Ong, S. Husain, and M. Bober, "Siamese network of deep fisher-vector descriptors for image retrieval," *arXiv preprint arXiv:1702.00338*, 2017.

[22] A. Jose, S. Yan, and I. Heisterklaus, "Binary hashing using siamese neural networks," in *IEEE International Conference on Image Processing (ICIP)*, 2017.

[23] A. Pandey, A. Mishra, V. K. Verma, A. Mittal, and H. Murthy, "Stacked adversarial network for zero-shot sketch based image retrieval," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020.

[24] V. K. BG, G. Carneiro, and I. Reid, "Learning local image descriptors with deep siamese and triplet convolutional networks by minimising global loss functions," in *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[25] J. Song, T. He, L. Gao, X. Xu, A. Hanjalic, and H. T. Shen, "Unified binary generative adversarial network for image retrieval and compression," *International Journal of Computer Vision (IJCV)*, 2020.

[26] L.-W. Ge, J. Zhang, Y. Xia, P. Chen, B. Wang, and C.-H. Zheng, "Deep spatial attention hashing network for image retrieval," *Journal of Visual Communication and Image Representation (JVCIR)*, 2019.

[27] Z. Chen, J. Lin, Z. Wang, V. Chandrasekhar, and W. Lin, "Beyond ranking loss: Deep holographic networks for multi-label video search," in *IEEE International Conference on Image Processing (ICIP)*, 2019.

[28] A. Qayyum, S. M. Anwar, M. Awais, and M. Majid, "Medical image retrieval using deep convolutional neural network," *Neurocomputing*, 2017.

[29] S. R. Dubey, S. K. Roy, S. Chakraborty, S. Mukherjee, and B. B. Chaudhuri, "Local bit-plane decoded convolutional neural network features for biomedical image retrieval," *Neural Computing and Applications*, 2019.

[30] Y. Chen and X. Lu, "Deep discrete hashing with pairwise correlation learning," *Neurocomputing*, 2020.

[31] X.-S. Wei, J.-H. Luo, J. Wu, and Z.-H. Zhou, "Selective convolutional descriptor aggregation for fine-grained image retrieval," *IEEE Transactions on Image Processing (TIP)*, 2017.

[32] Q. Qin, L. Huang, and Z. Wei, "Deep multilevel similarity hashing with fine-grained features for multi-label image retrieval," *Neurocomputing*, 2020.

[33] J. Zhang and Y. Peng, "Query-adaptive image retrieval by deep weighted hashing," *IEEE Transactions on Multimedia (TMM)*, 2018.

[34] X. Zeng, Y. Zhang, X. Wang, K. Chen, D. Li, and W. Yang, "Fine-grained image retrieval via piecewise cross entropy loss," *Image and Vision Computing (IVC)*, 2020.

[35] R. Xia, Y. Pan, H. Lai, C. Liu, and S. Yan, "Supervised hashing for image retrieval via image representation learning," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2014.

[36] X. Yan, L. Zhang, and W.-J. Li, "Semi-supervised deep hashing with a bipartite graph," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017.

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[38] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.

[39] M. R. K. Mookiah, R. Puch-Solis, and N. N. Daeid, "Identification of bullets fired from air guns using machine learning and deep learning," *the Forensic Science International*, vol. 349, 2023.

[40] S. K. Dutta, S. Saikia, A. Barman, R. Roy, K. Bora, L. B. Mahanta, and R. Suresh, "Study on enhanced deep learning approaches for value-added identification and segmentation of striation marks in bullets for precise firearm classification," *Applied Soft Computing*, vol. 112, p. 107789, 2021.

[41] M.-E. L. Bouthillier, L. Hrynkiw, A. Beauchamp, L. Duong, and S. Ratte, "Automated detection of regions of interest in cartridge case images using deep learning," *the Journal of Forensic Sciences (JFS)*, June 2023.

[42] P. Giverts, K. Sorokina, and V. Fedorenko, "Examination of the possibility to use siamese networks for the comparison of firing pin marks," *the Journal of Forensic Sciences (JFS)*, September 2022.

[43] J. J. Koenderink and A. J. van Doorn, "Surface shape and curvature scales," *Image and Vision Computing*, 1992, 10, 557-564. DOI:10.1016/0262-8856(92)90076-F.

[44] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[45] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[47] U. S. Buyukatalay, o. Birgul, "3b gericatım icin golge ve parlama bolgesi saptamalı dongulu fotometrik stereo," *SIU 2009 IEEE 17. Sinyal Isleme ve Iletisim Uygulamaları Kurultayı*, Nisan 2009.

[48] A. O. ozturk, o. Birgul, E.Esen, and U. Sakarya, "Fotometrik stereo yonteminde fotoğraflardaki gurultulerin gericatılmıs kovan yukseklik haritasına etkileri," *ICIP 2009 IEEE International Conference on Image Processing*, November 2009.

[49] S. Buyukatalay, o. Birgul, and U.Halıcı, "Isık kaynakları seciminin ve yuzey ozelliklerinin fotometrik stereo hatasına etkisi," *Sinyal Isleme ve Iletisim Uygulamaları Kurultayı*, Nisan 2010.

[50] U. Sakarya, S. Es, U. Leloğlu, E. Tunalı, and o. Birgul, "A case study of an automated firearms identification system: Balistika 2010," *BIT's 1st Annual World Congress of Forensics (WCF-2010)*, October 21-23, 2010.

[51] H. Tarakcı, S. Arslan, E. Gurbuz, s. Tekdal, and A. Yılmaz, "Bpel kullanarak adli balistik bilisimde servis temelli mimarinin uygulanması," *BAsARIM2010 2. Ulusal Yuksek Basarımlı ve Grid Hesaplama Konferansı*, 2010.

[52] A. Yılmaz, H. Tarakcı, and S. Arslan, "Ballon: An ontology for forensic ballistics domain," *KEOD 2010 International Conference on Knowledge Engineering and Ontology Development*, October 2010.

[53] A. O. Ozturk, S. Buyukatalay, and O. Birgul, "Tek kamera ve tek yapısal isık kaynağı kullanarak mermi kovanı yukseklik haritasının olusturulması," *SIU 2010 IEEE 18. Sinyal Isleme ve Iletisim Uygulamaları Kurultayı*, Nisan 2010.

[54] M. Turkmenoğlu, O. Sengul, and L. B. Yalcıner, "Optik goruntuleme sistemleri icin mtf olcumleri," *Fotonik 2010*, Ekim 2010.

[55] A. Sayar, F. Tetiker, E. Acar, B. O. Acar, and U. Sakarya, "Sift ozniteliği kullanarak mermi cekirdeği esleme," *IEEE 19. Sinyal Isleme, Iletisim ve Uygulamaları Kurultayı*, Nisan 2011.

[56] U. Leloglu, "Characterisation of tool marks on cartridge cases by combining multiple images," *IET Image Processing*, vol.6, no.7, pp.854-862, October 2012.

[57] U. Sakarya, O. Topcu, U. Leloğlu, M. Soysal, and E. Tunalı, "Automated region segmentation on cartridge case base," *Forensic Science International*, vol. 222, Issues 1–3, pp. 277-287, 10 October 2012.

[58] U. Sakarya, U. M. Leloglu, and E. Tunalı, "Three-dimensional surface reconstruction for cartridge cases using photometric stereo," *Forensic Science International*, 2008.

[59] P. Pisantanaroj, P. Tanpisuth, P. Sinchavanwat, S. Phasuk, P. Phienphanich, P. Jangtawee, K. Yakoompai, M. Donphoongpi, S. Ekgasit, and C. Tantibundhit, "Automated firearm classification from bullet markings using deep learning," *IEEE Access*, vol. 8, pp. 78236–78251, 2020.

[60] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2016.

[61] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.

[62] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[63] G. Rilling, P. Flandrin, P. Goncalves, *et al.*, "On empirical mode decomposition and its algorithms," in *IEEE-EURASIP workshop on nonlinear signal and image processing*, vol. 3, pp. 8–11, Citeseer, 2003.

[64] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015.

[65] U. Sakarya, U. M. Leloğlu, O. Topçu, H. S. Arkan, M. Z. Kadıoğlu, and S. Çanga, "Shape index and polynomial coefficient based pattern analysis and comparison method for cartridge cases and bullets in forensic science." WO Patent WO2013182871A1, Dec. 2013.

[66] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2999–3007, 2017.

[67] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[68] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.

[69] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, p. 336–359, Oct. 2019.

[70] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*, pp. 818–833, Springer, 2014.