University of Dundee

Identification of bullets fired from air guns using machine and deep learning methods

Mookiah, Muthu Rama Krishnan; Puch-Solis, Roberto; Nic Daeid, Niamh

[Link to publication in Discovery Research Portal](#)

Citation for published version (APA):
Mookiah, M. R. K., Puch-Solis, R., & Nic Daeid, N. (2023). Identification of bullets fired from air guns using machine and deep learning methods. *Forensic Science International*, *349*, [111734]. https://doi.org/10.1016/j.forsciint.2023.111734

Identification of bullets fired from air guns using machine and deep learning methods

Muthu Rama Krishnan Mookiah *, Roberto Puch-Solis, Niamh Nic Daeid

Leverhulme Research Centre for Forensic Science, School of Science and Engineering, University of Dundee, Nethergate Dundee DD1 4HN, Scotland, UK

ARTICLE INFO

ABSTRACT

Ballistics (the linkage of bullets and cartridge cases to weapons) is a common type of evidence encountered in criminal cases around the world. The interest lies in determining whether two bullets were fired using the same firearm. This paper proposes an automated method to classify bullets from surface topography and Land Engraved Area (LEA) images of the fired pellets using machine and deep learning methods. The curvature of the surface topography was removed using loess fit and features were extracted using Empirical Mode Decomposition (EMD) followed by various entropy measures. The informative features were identified using minimum Redundancy Maximum Relevance (mRMR), finally the classification was performed using Support Vector Machines (SVM), Decision Tree (DT) and Random Forest (RF) classifiers. The results revealed a good predictive performance. In addition, the deep learning model DenseNet121 was used to classify the LEA images. DenseNet121 provided a higher predictive performance than SVM, DT and RF classifiers. Moreover, the Grad-CAM technique was used to visualise the discriminative regions in the LEA images. These results suggest that the proposed deep learning method can be used to expedite the linkage of projectiles to firearms and assist in ballistic examinations.

In this work, the bullets that were compared were air pellets fired from both air rifles and a high velocity air pistol. Air guns were used to collect the data because they were more accessible than other firearms and could be used as a proxy, delivering comparable LEAs. The methods developed here can be used as a proof-of-concept and are easily expandable to bullet and cartridge case identification from any weapon.

1. Introduction

Exploring linkages between fired bullets and/or cartridge cases recovered from a scene with each other or with exemplar bullets/cartridge cases from a known (recovered) weapon can be very beneficial to an investigation where firearms have featured. The barrel of a gun leaves distinctive striations, or contour variations, in a fired bullet. If a bullet is not destroyed and still bears these striations, then they can be used to assess whether two bullets were fired by the same firearm and, if a firearm is recovered, explore whether the bullet was fired by that particular firearm. The sections of the bullet that bears the striations are called the engraved area. The section of the engraved area that makes high and low contact points inside the barrel are called Land Engraved Area (LEA) and Groove Engraved Area (GEA) respectively [1], Fig. 1.

Traditionally, bullet signatures are analysed and compared by an expert using a comparison microscope [3,4]. This assessment is subjective and time consuming [3,5,6]. More recently, machine learning is seeing a rapid development in image comparison problems outside of forensic science but is increasingly attracting attention within the forensic science field.

The evaluation of bullet signatures is based on 3D scans of the LEA. The GEA tend to be less discriminative [1]. Some authors use the whole image while others use the topography of a representative line segment of the LEA. In either case, in a second step, features are extracted to which classification algorithms could be applied.

Banno et al. [7] developed a neural network to identify the similarity between two bullet striation signatures. The signals were obtained from striation images of unidentified bullets and reference bullets. The network was able to match the unidentified and reference bullets, however the model was developed using only ten images and the final matching decision was made by the forensic expert. Changmai et al. [8] used the whole image, and to focus on the striations, the image was segmented (i.e. the selection of areas of interest in the image) using a fuzzy c-means algorithm followed by preprocessing (contrast enhancing, noise reduction and smoothing). Then shape features were extracted from the segments, and a K-nearest neighbour (kNN) classifier was applied. Vanderplas et al. [9] used the

---

* Corresponding author.
E-mail address: mrkmookiah@dundee.ac.uk (M.R.K. Mookiah).

Nomenclature			
EMD	Empirical Mode Decomposition.	CMS	Consecutive Matching Striae.
IMF	Intrinsic Mode Function.	PCA	Principal Component Analysis.
SD	Standard Deviation.	LDA	Linear Discriminant Analysis.
AM	Amplitude Modulated.	kNN	K-nearest neighbour.
FM	Frequency Modulated.	FFT	Fast Fourier Transform.
mRMR	minimum Redundancy Maximum Relevance.	ROI	Region of Interest.
SVM	Support Vector Machines.	FIS	Feature Importance Score.
SVM-Q	Support Vector Machines-Quadratic.	AUC	Area Under the Receiver Operating Characteristic Curve.
DT	Decision Tree.	LR	Likelihood Ratio.
LEA	Land Engraved Area.	GEA	Groove Engraved Area.
		RF	Random Forest.

whole LEA image and extracted their 3D topography which were then classified with a Random Forest (RF) score, cross-correlation and Consecutive Matching Striae (CMS), applied to same-source and different-source bullets. Dutta et al. [10] developed a deep learning neural network (fully convolutional network) to extract striations from the fired bullets specifically for segmentation rather than classification. The Residual U-net and Inception U-net architectures were used for the segmentation of striations [11]. The performance of the two models were compared. The residual U-net delivered better training accuracy than the Inception U-net. Pisantanaroj et al. [12] used the whole image and proposed a deep learning framework for firearm identification where feature extraction and classification was performed simultaneously using three different deep neural networks: DenseNet121, ResNet50 and Xception. The model was trained with 718 bullets from eight different firearms. Hamzah [13] used striation signatures followed by the extraction of surface topography line segment and the application of Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) for classification. Morrison et al. [14] used images of cartridge cases from which they extracted the Zernike-moment based features. The dimension of the data was reduced with PCA and LDA was used as a probabilistic approach to calculate a Likelihood Ratio (LR).

In this paper we explore machine learning algorithms on surface topography line segments within the LEAs, and also deep learning algorithms using automatic feature extraction directly from the images. The extraction of features from the surface topography is achieved using Empirical Mode Decomposition (EMD). We also use feature ranking by applying minimum Redundancy Maximum Relevance (mRMR) to select the most discriminant features. Classification is performed with Support Vector Machines (SVM), a Decision Tree (DT) and a RF. To our knowledge, EMD and mRMR have not been used before in this context. For the deep learning method, the process starts with preprocessing the LEA image to segment the pellet region and select the Region of Interest (ROI). Image augmentation is
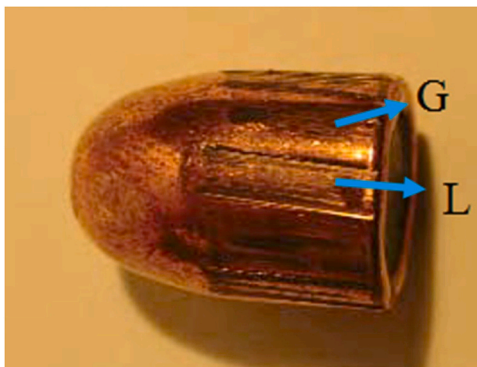
performed to increase the training samples and DenseNet121 is used to discriminate the sample set of pellets fired by two air rifles (Edgar and Baikal) and one air pistol (HW). Performance of the classifiers are evaluated using five-fold cross validation. Our classification method is similar to Pisantanaroj et al. [12], however, here we have chosen smaller number of training parameters. We also used the Grad-CAM technique [15] to visualise a heatmap of the influential features in the LEA to make our deep learning algorithm transparent. In addition, we compare the results from the two methods, one using SVM, DT and RF on LEA topography, and the other using deep learning on the whole image, both applied to the same dataset.

2. Materials and methods

2.1. Dataset

A dataset of 3D scans of pellets reported in [16] was used in this study. Two air rifles and one air pistol were used to generate the data. The air rifles (an Edgar Brother Model 35 and a Baikal 90042234–35) both had 12 right rifling with a break barrel design and spring piston mechanisms, Fig. 2(a) and (b). These air rifles were a number of years old and there were no records available of the number of air pellets the weapons may have fired. The breech of the rifles was opened in order to feed the pellet into the barrel and one pellet was loaded into the chamber at a time. The air pistol was a Weihrauch model HW45 which had a 12 right rifling and a spring piston [17], Fig. 2(c). It had two velocity settings depending on how the cocking mechanism is used. In this work, the high velocity setting was used. In all cases (both rifles and the air pistol) the ammunition used was RWS® superdome 4.5 mm (0.177 cal) round nosed unjacketed pellets [18], Fig. 2(d).

Each weapon was fired multiple times and the fired pellets collected. In total 50 pellets were fired using the air pistol and 5 pellets each using the two air rifles. All the fired pellets were scanned using an Alicona® infinite focus microscope [19] across each of the 12 LEAs. In each case the pellet was placed on a platform of the microscope that moved along the x, y and z axes in order to enable measurement of the 3D surface topography. The Alicona microscope software displayed the LEA image on a monitor and the primary profile of the LEA was selected following the ISO4287 standard [20]. The Alicona software was then used to transform the LEA striation image into surface topography [13]. Only LEA surfaces with clear striations were considered for the analysis. A total of 38, 34 and 600 LEA regions were imaged from the pellets recovered from the Baikal air rifle, the Edgar air rifle and the HW air pistol respectively.

2.2. Preprocessing and ROI Selection

Preprocessing was performed to remove background noise from the original LEA images, Fig. 3(a). The quality map, Fig. 3(b), generated by the Alicona microscope was used to obtain a segmentation mask, Fig. 3(c). Multilevel thresholding using Otsu's method [21] was applied
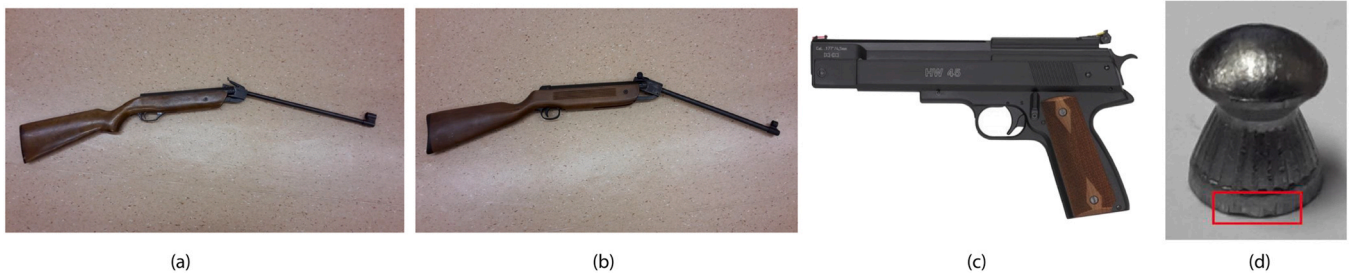

Fig. 1. Image of a fired bullet; G denotes GEA and L denotes LEA (Image courtesy of Chen et al. [2]).

Fig. 2. (a) Baikal air rifle, (b) Edgar air rifle, (c) HW air pistol, and (d) Air pellet (red rectangle represents pellet bearing striation marks).
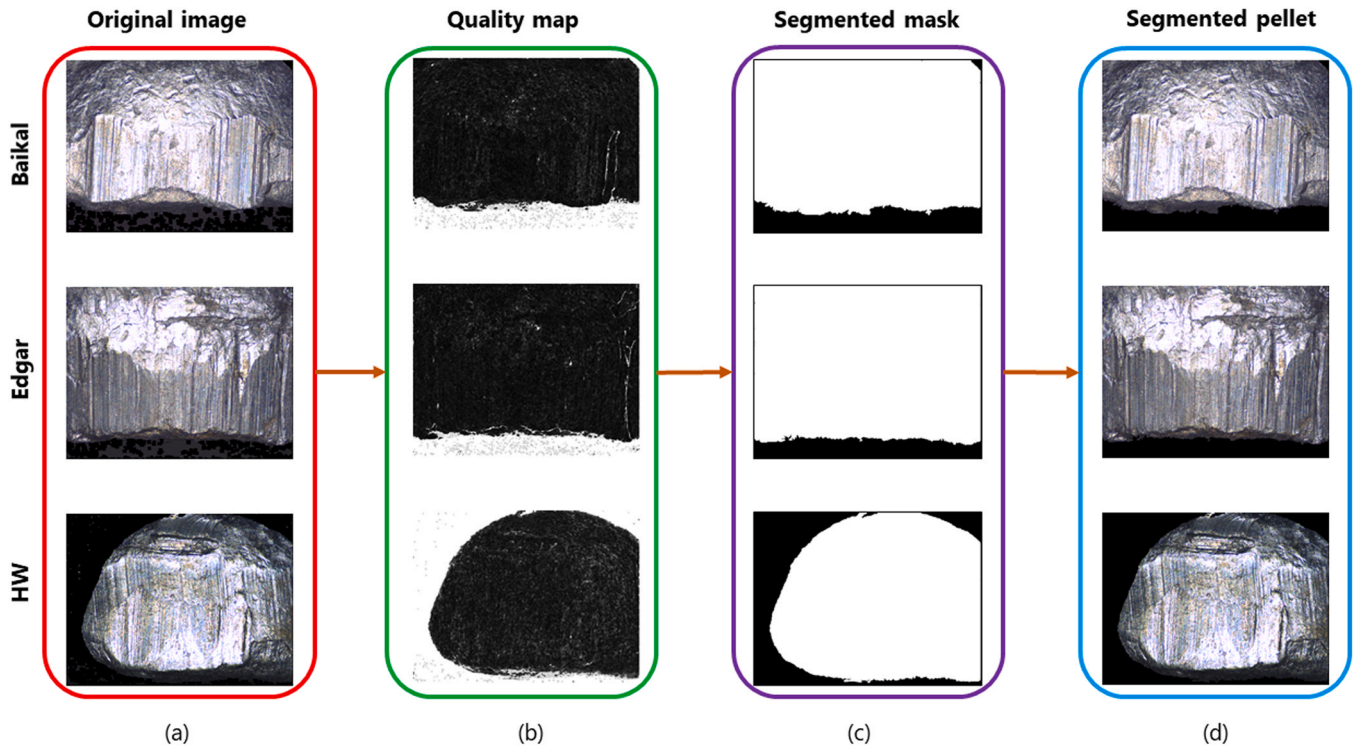

Fig. 3. Block diagram of preprocessing and ROI extraction.

on the quality map to identify the threshold to separate image foreground and background. Further, binarisation was performed using threshold and morphological operations were applied to remove noise after binarisation [22]. Finally, the segmented mask was used to extract the LEA region, Fig. 3(d), from the original image. Note that the background regions in the segmented images are black, while the same regions in the original image contain grey patches which are features of the background and not features of the pellet images.

The preprocessed images were used to select the ROI for deep learning experiments. The ROI was selected from where the LEA surface topography was recorded. The red box was drawn between the LEA shoulders to select a single LEA or ROI. The deep learning algorithm uses the whole ROI and therefore areas that were deformed during the pellet impact are not considered part of the ROI. In the first row of Fig. 4, the red rectangle indicates the selected ROI and the corresponding cropped region is shown in the second row. E.g., the ROI of Edgar pellet does not consider the deformed area in the right top corner of the pellet image.

2.3. Preprocessing of LEA Surface Topography

The curvature of LEA surface topography on the line segment was removed using locally estimated scatterplot smoothing (loess)

[23,24]. In a nutshell, for a given point p^* in the scatterplot, a subset of points near p^* are selected and a linear regression is fitted with this subset of points. The predicted value associated with p^* is calculated using this local regression. This is done for all points. A smoothing parameter controls the number of points that are selected near p^*. The fewer points are selected near p^*, the more the predicted curve follows the points in the scatterplot.

Examples of the loess fit for the images of pellets fired from each of the three firearms are shown in Fig. 5(a), (b) and (c). In each figure, the first row shows the LEAs and line segments, the second row shows the loess fit (in red) on the surface topography line segment (in blue), and the third row shows the residual signature. Hereafter, the residual signature is referred to as *striation signature*.

2.4. Striation signature decomposition

EMD is an adaptive and data driven method used for analysing nonlinear and non-stationary signals [25]. It decomposes the signal into Amplitude Modulated (AM) and Frequency Modulated (FM) signal components, called Intrinsic Mode Function (IMF) [26,27]. EMD generates a set of IMFs, $\{IMF_i: i = 1, 2, …, N\}$, and a residual signal $r(t)$, for a given striation signature $x(t)$ which can be expressed as:
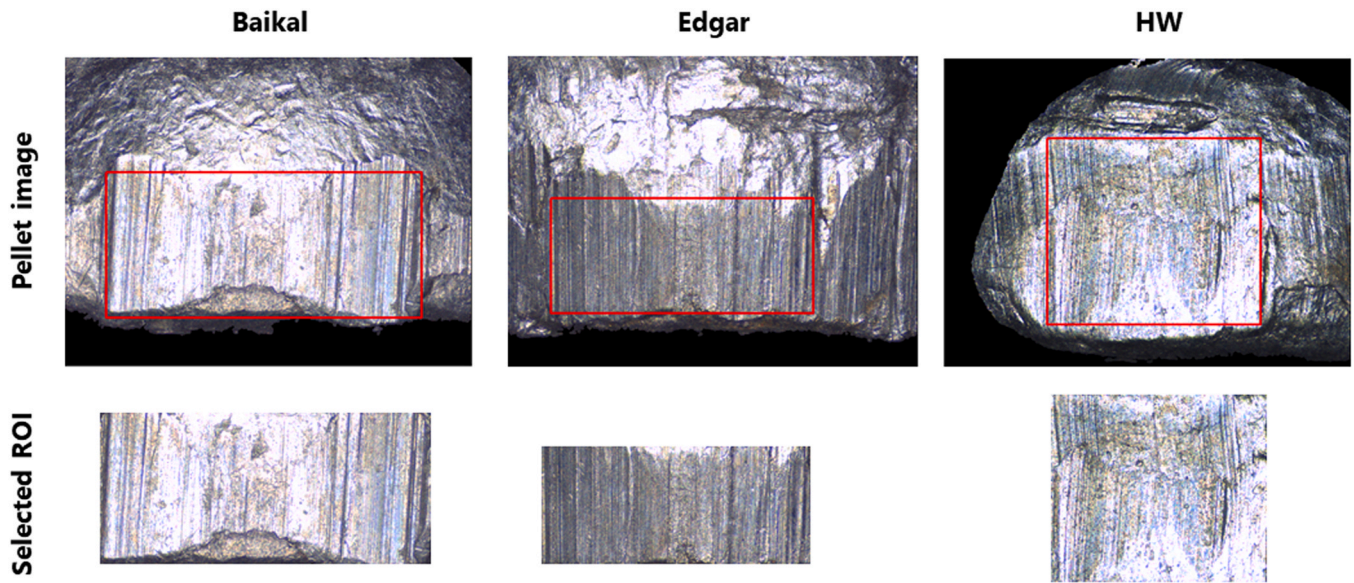
Baikal **Edgar** **HW**

Fig. 4. ROI extraction, first row: original LEA region with selected ROI (highlighted with red rectangle), second row: extracted ROI.

$$x(t) = \sum_{i=1}^{N} IMF_i(t) + r(t) \tag{1}$$

The IMFs are extracted using a sifting algorithm given below [25].

(i) Find location of all maxima and minima of $x(t)$.
(ii) Cubic spline interpolation is used to connect all maxima (minima) to obtain an upper (lower) signal envelop $x_u(t)$ $(x_l(t))$.
(iii) The mean envelope, $m_1(t)$, between the upper and lower envelopes is calculated as $m_1(t) = [x_l(t) + x_u(t)]/2$.
(iv) Subtract the local mean envelope from the original signal to obtain the AM and FM oscillation $s_1(t) = x(t) - m_1(t)$
(v) Verify if $s_1(t)$ satisfies the conditions of an IMF [25,26,28,29],

   (a) The number of extrema and zero-crossings of $s_1(t)$ are the same, or differ by exactly one, with $x(t)$, and
   (b) The upper envelope is a symmetric reflection of the lower envelope with respect to $m_1(t)$

   If $s_1(t)$ satisfies the conditions above, then it is the first IMF of $x(t)$, else repeat steps (i) to (v) on $s_1(t)$ until the newly calculated $s_1(t)$ satisfies the conditions of an IMF.

The successive IMFs were extracted by repeating steps (ii)–(vi) recursively to the residue signal $r(t) = x(t) - s_1(t)$ until the residual signal $r(t)$ became a monotonic function where any more IMFs could be extracted [26]. This amplitude and instantaneous frequency combination is used for time-frequency analysis [29]. The IMFs captured the peaks and valleys of striations on the LEA from the surface topography profile [30]. The first five IMFs and residual of the pellet signatures are shown in Fig. 6.

2.5. Feature extraction using EMD

Skewness, Kurtosis, Energy, and entropy measures: Approximate, Shannon, Kapur, Renyi, and Yager, were extracted as features on the first four IMFs. The rest of the IMFs had low-frequency responses and largely contained noise. They are briefly described below, where IMF is used to represent one of the four IMFs.

Skewness and Kurtosis provides information about shape and distribution [31,32] of an IMF. They are calculated using the equations:

$$Skewness = \frac{1}{N} \frac{\sum_{i=1}^{N} \{IMF(t_i) - m\}^3}{\sigma^3} \tag{2}$$

$$Kurtosis = \frac{1}{N} \frac{\sum_{i=1}^{N} \{IMF(t_i) - m\}^4}{\sigma^4} \tag{3}$$

where N is number of data points, and m and σ are the mean and standard deviation of $\{IMF(t_i): i = 1, 2, ..., N\}$.

Energy is the amount of information present in the IMF [33] and can be calculated as follows,

$$Energy = \sum_{i=1}^{N} \{IMF(t_i)\}^2 \tag{4}$$

Approximate entropy is the logarithmic likelihood and is used to measure data regularity [34–36].

$$Approximate\ Entropy = \phi_m(r) - \phi_{m+1}(r) \tag{5}$$

where $\phi_m(r) = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} \ln[IMF_i^m(r)]$, and $IMF_i^m(r)$ is a correlation integral; m is series of patterns of length; r is a fixed parameters, here r is to be 0.2 times the standard deviation of the data [36].

Shannon entropy can be used to measure the average information present in an IMF and is calculated using the equation,

$$Shannon\ Entropy = -\sum_{j=1}^{M} p(y_j) \log_2[p(y_j)] \tag{6}$$

where $\{y_j: j = 1, ..., M\}$ and $\{p(y_j): j = 1, ..., M\}$ are the bins and frequencies of a histogram of $\{IMF(t_i): i = 1, ..., N\}$.

Renyi entropy is a generalised entropy [37–40], computed using the equation,

$$Renyi\ Entropy = \frac{1}{1-\alpha} \log_2 \sum_{i=1}^{N} IMF(t_i)^\alpha, \quad \text{for } \alpha \neq 1, \alpha > 0, \tag{7}$$
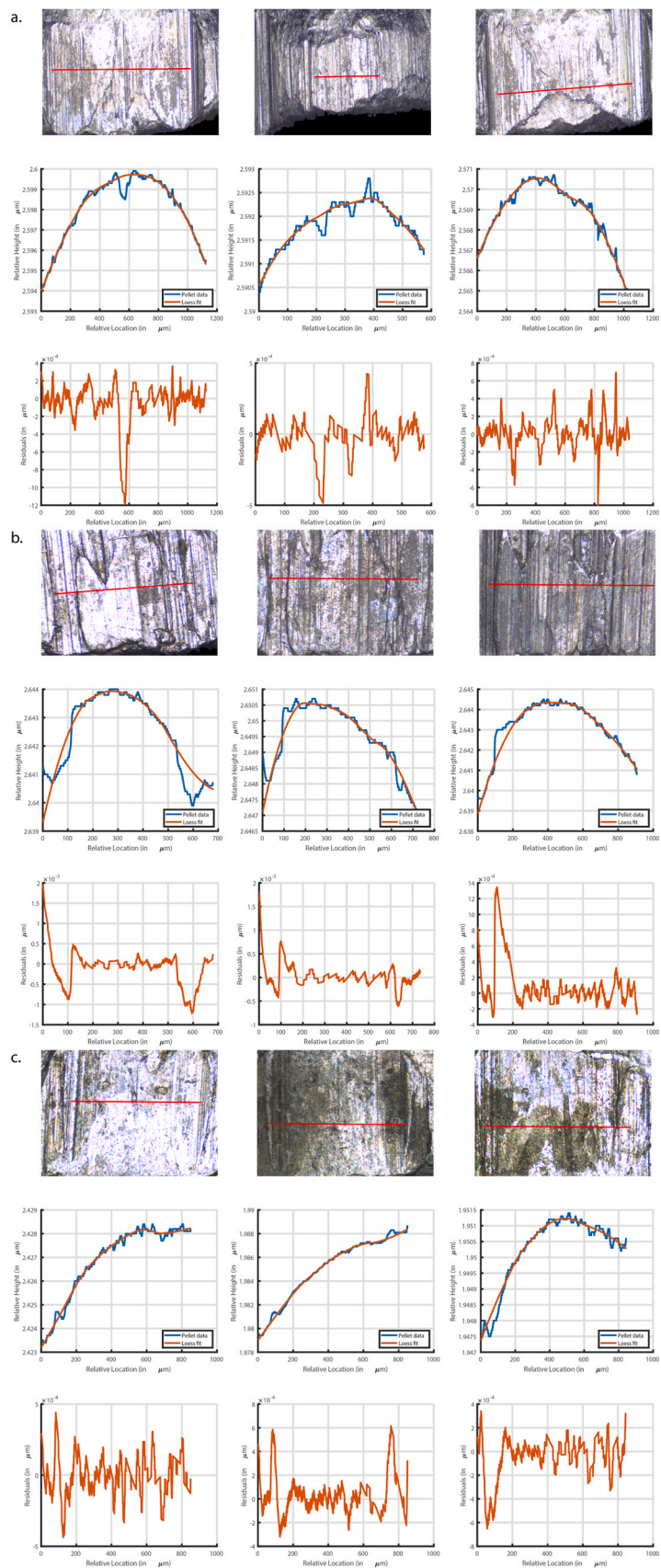
Fig. 5. Preprocessing of surface topography recorded from (a) Baikal air rifle, (b) Edgar air rifle and (c) HW air pistol.
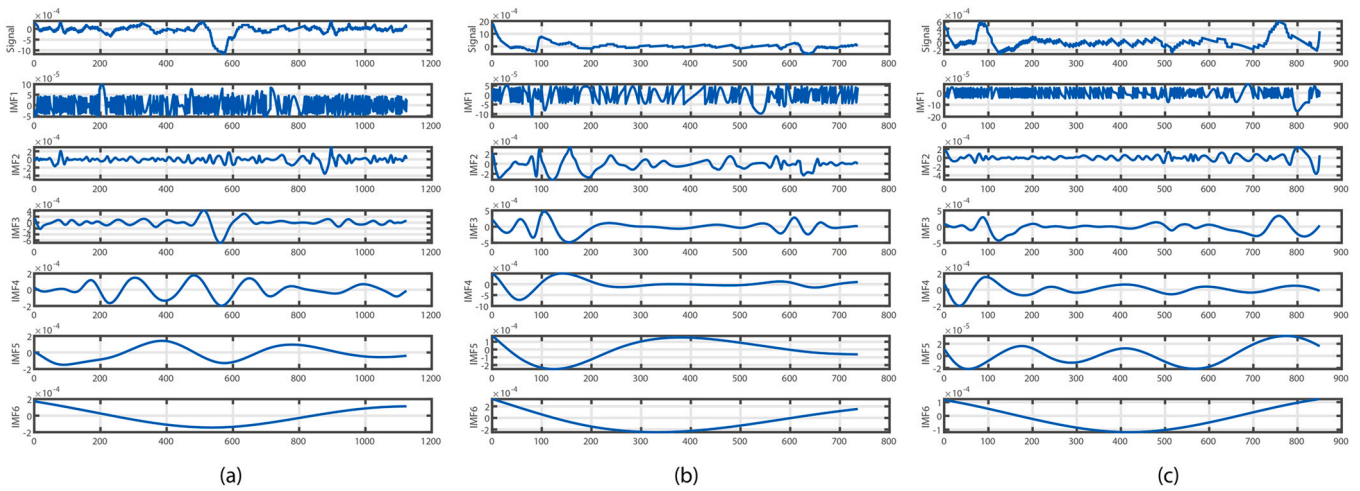
Fig. 6. IMFs of surface topography for the pellets fired from (a) Baikal, (b) Edgar and (c) HW using EMD.

where α is the diversity index. Here $\alpha = 3$, taken from previous experiments which provided a high classification performance [37].

Kapur entropy is a generalised version of Renyi entropy [37–40] and it can be computed using the equation,

$$\text{Kapur Entropy} = \frac{1}{1-\alpha}\left[\log_2 \sum_{i=1}^{N} IMF(t_i)^\alpha - \log_2 \sum_{i=1}^{N} IMF(t_i)^\beta\right], \quad (8)$$

for $\alpha \neq 1$, $\beta > 0$, $\alpha + \beta - 1 > 0$. Diversity indices α and β are set here to $\alpha = 0.5$ and $\beta = 0.7$, selected based on previous experiments which provided a high classification performance [37].

Yager entropy is computed as follows [37],

$$\text{Yager Entropy} = 1 - \frac{\sum_{i=1}^{N} |2IMF(t_i) - 1|}{|N|} \quad (9)$$

2.6. Ranking of features across the IMFs

Eight features were calculated for each of the four IMFs, amounting to a total of 32 features. The features associated with a particular IMF were denoted with a numerical suffix, e.g. Energy1, Energy2, Energy3 and Energy4 are the Energy features associated with IMF1 to IMF4. The features across IMFs were ranked using mRMR [41]. This method selects features that are highly relevant to the output (LEAs of the pellets produced by the Baikal, Edgar air rifles and the HW air pistol) with low correlation between the features [41]. Feature relevance was calculated using the F-statistic and redundancy was computed using Pearson correlation coefficient. These two criteria were combined to obtain a Feature Importance Score (FIS). The features were ranked by descending order of FIS, Fig. 7. The last ten features revealed a low FIS value and the classification was undertaken with and without the last ten features to ascertain whether these features were informative.

2.7. Machine and deep learning methods

The classification across the three air weapons were performed using machine learning (SVM, DT and RF) and deep learning (DenseNet121) methods [42–44].

SVM is a statistical learning method, that constructs an optimal hyperplane to discriminate the air weapon classes using a structural risk minimisation algorithm [42]. In this work, multi-class classification was performed using one-vs-all SVM. This splits multi-class classification into multiple binary classifiers, where each model provides a probability score and a maximum class score, which were used to predict the air weapon class [45]. A similar approach was adapted to the DT classifier [43]. The DT split the complex solution into several simpler solutions using the nodes root, internal and leaf, deriving the final decision using the training data [46]. RF is an ensemble method [47], it combines predictions from other smaller DT models. Multiple DTs were created using different random subsets of the training data. AdaBoost [48] ensemble aggregation with five decision splits and two hundred ensemble learning cycles were used to obtain better classification performance.

The algorithms from sections 2.2 to 2.6 and machine learning algorithms in section 2.7 were implemented using MATLAB version 2020a from Statistics and Predictive Maintenance toolboxes.

DenseNet is one of the convolutional neural networks, where each layer is connected with every other layer in a feed-forward nature. The feature maps are concatenated to get the maximum information between the layers while also reducing the vanishing-gradient problem [44]. This network consists of dense blocks and transition layers, where each dense block has two convolution units and the transition layers has batch normalisation, activation (ReLU), convolution and pooling, to reduce the number of channels by half from previous layers. In this work, we adapted the DenseNet121, which has 121 layers. The final fully connected layer was replaced with global average pooling, batch normalisation, dropout and two fully connected layers followed by another set of batch normalisation and dropout, then a final fully connected layer (Fig. 8). The final classification was performed using the softmax classifier [44]. The block diagram of the deep learning approach used for the air gun pellet classification is shown in Fig. 8.

Both machine and deep learning models were evaluated using fivefold cross validation. In the DenseNet121, the dataset was divided into training, validation and testing. Image augmentation with horizontal flip, vertical flip, shear (range = 0.2) and zoom (range = 0.2) were performed. The LEA images were resized into a standard size of 512 × 512 × 3. The deep network was trained to classify pellets fired using the Baikal and Edgar air rifles and the HW air pistol. Initially, Imagenet weights were initialised using the Glorot uniform technique. The last eight layers of the network was trained for 100 epochs using categorical cross-entropy loss and all other layers were frozen to avoid over-fitting. ADAM optimisation was used to back-propagate the model error with 16 batch updates and a learning rate of 0.01. Network weights were saved at the minimum validation loss.
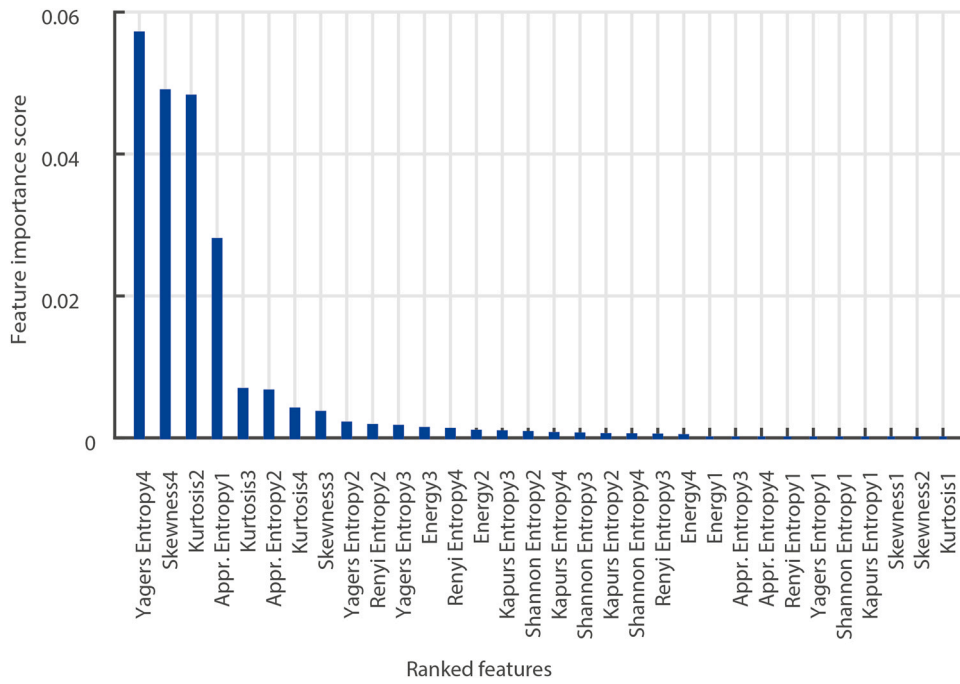
Fig. 7. EMD feature ranking using mRMR.

Training parameters (number of patches, batch size, learning rate and epochs) were chosen experimentally to avoid over-fitting. The trained model has 1.5 million trainable parameters and was implemented in Python 3 using Keras 2.2.4. The model ran on an Intel Core i7–8700 CPU with an NVIDIA TITAN Xp GPU. The training performance of DenseNet121 for all five folds are shown in Fig. A1. The accuracy and loss plots demonstrate that the model does not either over-fit or under-fit.

The classification performance of machine and deep learning models were evaluated using measures that take values between 0 and 1, where larger values are associated with better performance. There are two groups of measures, one group focus on the performance of a specific class, e.g. whether HW pistol pellets were classified correctly or not, and consist of precision, recall, F1-score and Area Under the Receiver Operating Characteristic Curve (AUC). The other group measures the performance of the classifier for all three classes together and consists of micro and macro averages.

3. Results

3.1. Machine learning

Probability density plots of the twelve most discriminating features are shown in Fig. 9. A probability density plot shows the
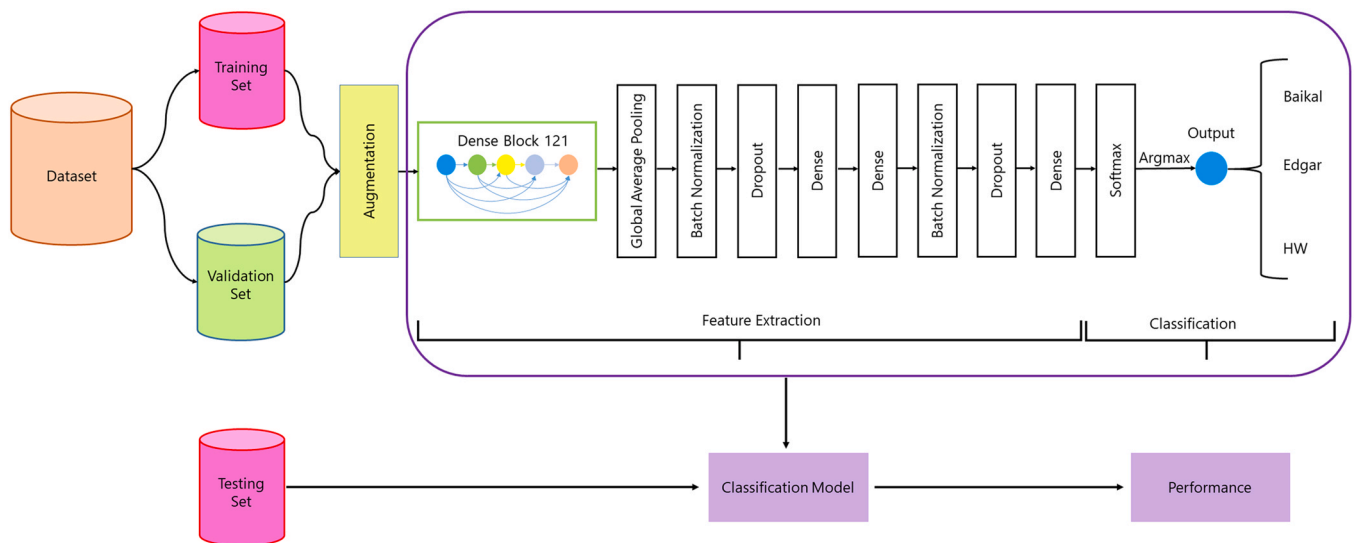

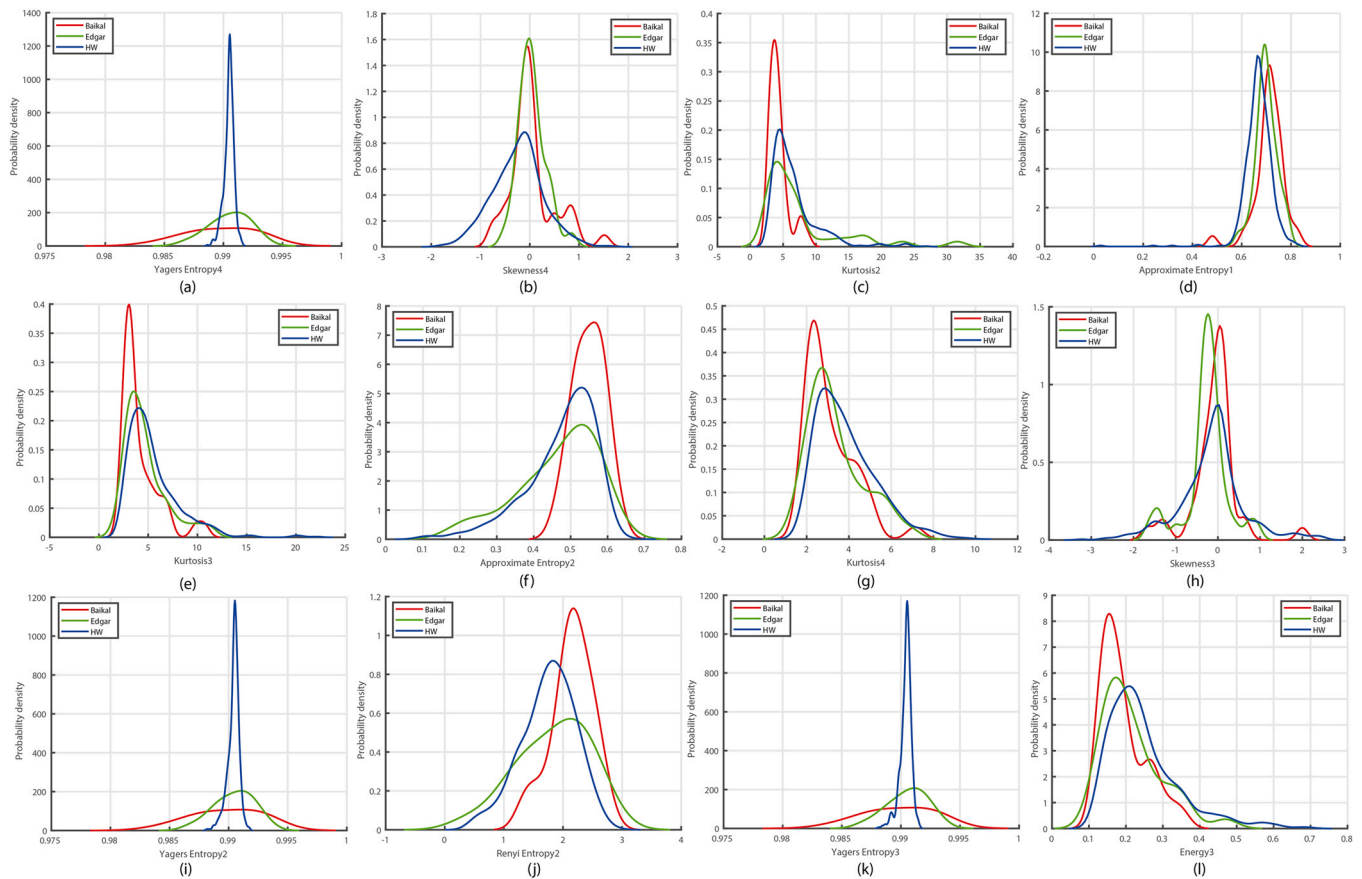**Fig. 8.** Block diagram of the deep learning approach.

Fig. 9. Probability density plots of ranked features.

densities, in the y-axis, of the values of a feature, in the x-axis. A density, in the statistical sense, is not a probability but a non-negative number that encodes how likely this value is compared with other values. Thus, the most likely value has the highest density value, and an unlikely value has a density close to zero. The variation range, or range of likely values, are the values where the densities are greater than zero. E.g. in Fig. 9(b) the variation range of the *Skewness4* feature for HW pistol pellets is between about -2 and 1.5. In the plots, we can see that the variation ranges of the pellets fired by the Baikal rifle and the HW pistol tend to be more separated than the variation ranges of the Edgar rifle and the HW pistol. E.g., in Fig. 9(a) the variation range of Baikal rifle contains a larger interval of values that are not in the variation range of HW pistol than the variation range of Edgar rifle. This contributes to better discrimination of pellets from the Baikal rifle and the HW pistol. The variation ranges of the pellets from the Edgar rifle and HW pistol share more values.

The ranked features were sequentially fed to SVM, DT and RF classifiers. The classification was performed using (i) the first 22 ranked features and (ii) all ranked features. The performance measures for both experiments are summarised in Figs. 10 and 11, where micro and macro average denote the overall performance of the classifiers for all three classes. Numerical values are reported in Appendix A (Tables A1 and A2).

The results of the first 22 ranked features using SVM-Quadratic revealed that the classification performance for the HW air pistol was good (F1-Score: 0.96 and AUC: 0.82), while the performance for the Baikal air rifle was less promising in comparison to the HW air pistol (F1-Score: 0.63 and AUC: 0.80). The performance of the Edgar

air rifle was poor (F1-Score: 0.34 and AUC: 0.66). Similar trends were observed for DT classifier with a slightly poorer classification performance, whereas the RF classifier with Adaboost ensemble aggregation provides a slightly higher F1-score for the Edgar air rifle (F1-Score: 0.35) and HW air pistol (F1-Score: 0.97) but a slightly worse score for the Baikal rifle, Fig. 10 and Table A1.

The results of all features revealed that the classification performance for the Baikal (F1-Score: 0.59 and AUC: 0.79) and Edgar (F1-Score: 0.33 and AUC: 0.66) air rifles was better using SVM-Q, while for the HW air pistol (F1-Score: 0.98 and AUC: 0.81) the classification performance was better using RF. However, DT provides a slightly higher AUC for the Edgar air rifle (AUC: 0.27) and HW air pistol (AUC: 0.86) compared to RF. The trends in classification performance with all features and the first 22 ranked features are similar, however, the results also revealed that adding features reduced the classification performance, Fig. 11 and Table A2.

3.2. Deep learning

The results of the deep learning approach are summarised in Fig. 12 and the numerical values are reported in Appendix A (Table A3). The pellet classification performance was considerably improved when compared to the machine learning methods. The best classification performance of the DenseNet121 returned F1-scores of 0.97, 0.78 and 0.99 for pellets fired from the Baikal air rifle, the Edgar air rifle and the HW air pistol respectively. The F1-score for the pellets fired from the Edgar air rifle has a lower value compared to the Baikal air rifle and the HW air pistol, however it was notably
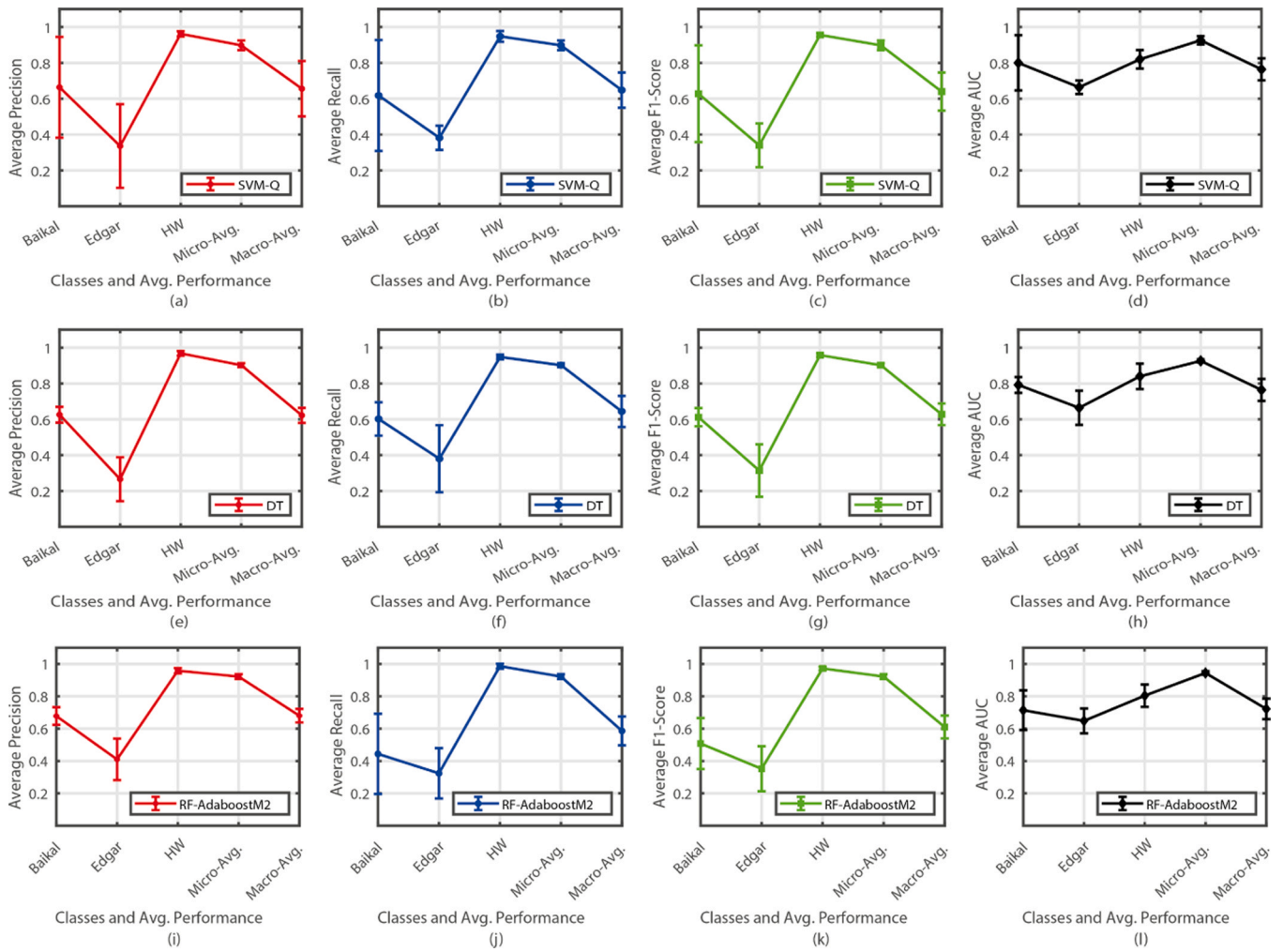
Fig. 10. Performance measures of Support Vector Machines-Quadratic (SVM-Q), DT and RF using the first 22 ranked features.

improved when compared to the machine learning methods (Figs. 10 to 12 and Tables A1 to A3).

The Grad-CAM [15] output provided a visual explanation of the model. Fig. 13(c), (f) and (i) revealed the discriminating areas. The red, yellow, green and blue signify the degree of discrimination of the features from the highest (red) to the lowest (blue). The intensity of the color represents the degree of importance: a brighter red color indicates a higher importance for that region, while a darker blue color indicates a lower importance [15].

4. Discussion and conclusion

Striation marks observed on fired bullets are commonly used to generate a bullet signature to facilitate the linkage of bullets together if they have been fired from the same firearm or link to bullets to specific firearms [49]. In this study we compared the classification performance of surface topography features using machine learning methods (SVM, DT and RF), and an automatic classification of LEA images using deep learning.

For the machine learning methods, the results revealed that the RF provided the highest classification performance for pellets fired from the HW air pistol (22 features, average F1-score: 0.97 and all features, F1-score: 0.98). However, DT provided the highest AUC (22 features, average AUC: 0.84 and all features, AUC: 0.86), Tables A1

and A2. AUC often provides misleading results for imbalanced datasets, whereas F1-score is able to measure performance objectively when the class balance is skewed [50]. The classification performance for the pellets fired by both air rifles was lower on all features. In general, The classification performance using the first 22 ranked features is better than using all features. The selected features not only reduce the dimension of the data but also improve the classification performance. Including low discriminating features reduced the classification performance. The features of the pellets fired by the HW air pistol was more separable from those produced by the air rifles, which can be inferred from the probability density plots, Fig. 9. One of the reasons is that the sample size for the HW pistol (600 scans) is much larger than those of the rifles (38 and 34 scans for the Baikal and Edgar rifles), Appendix A.

The proposed deep learning model, DenseNet121, provided the highest average F1-scores of 0.97, 0.78 and 0.99 for classifications of the air pellets fired from the Baikal and Edgar air rifles and the HW air pistol respectively, Table A3. The average AUC to classify the three groups of fired pellets are considerably higher (≥0.99) as compared to the machine learning methods, Table A1 to Table A3. The advantages of using DenseNet121 are (i) enabling the transfer of learning from pre-trained networks (trained using large datasets), and (ii) addressing the vanishing gradient problem which strengthens feature propagation and reduces the number of
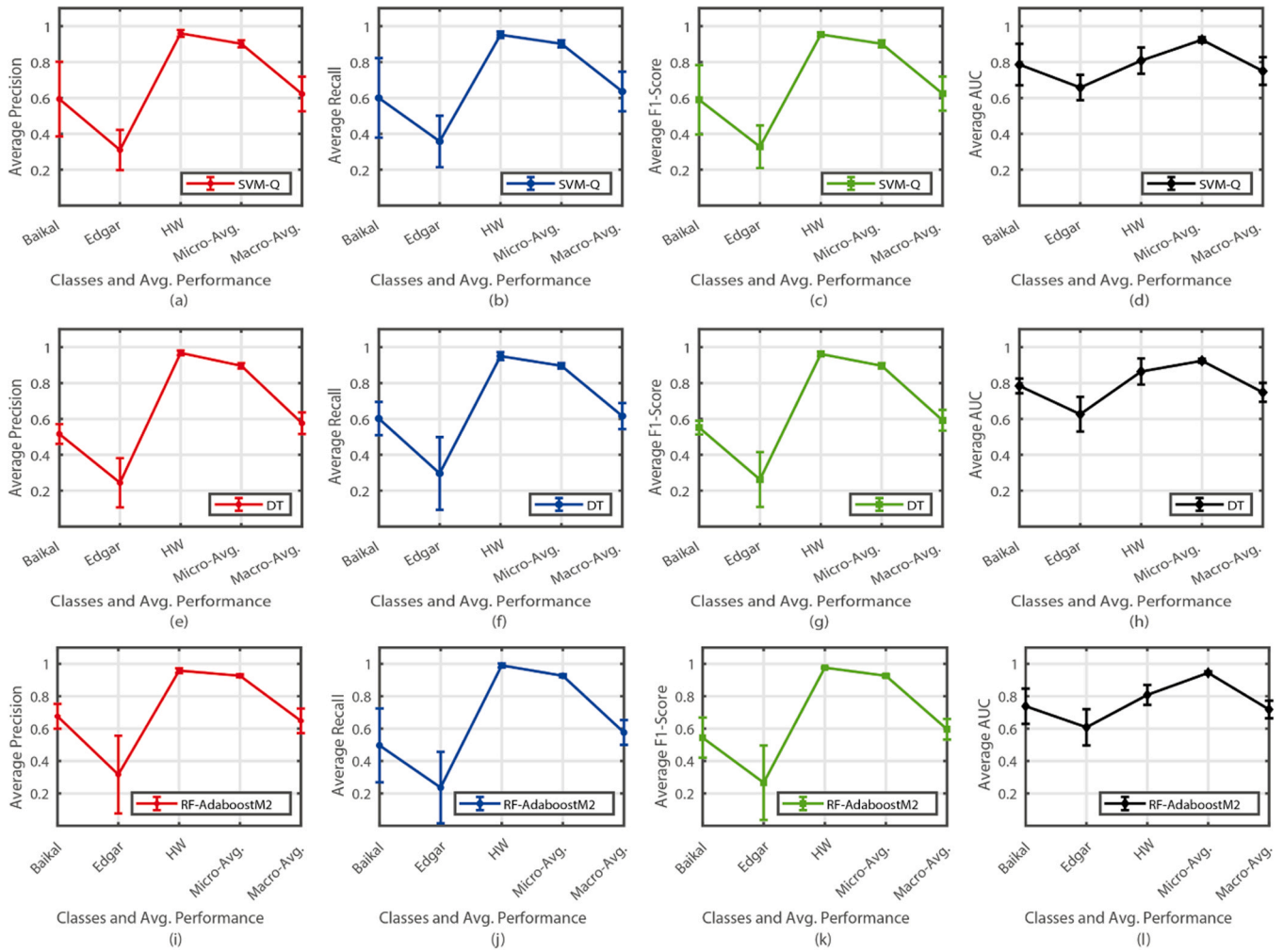
Fig. 11. Performance measures of SVM-Q, DT and RF using all ranked features.
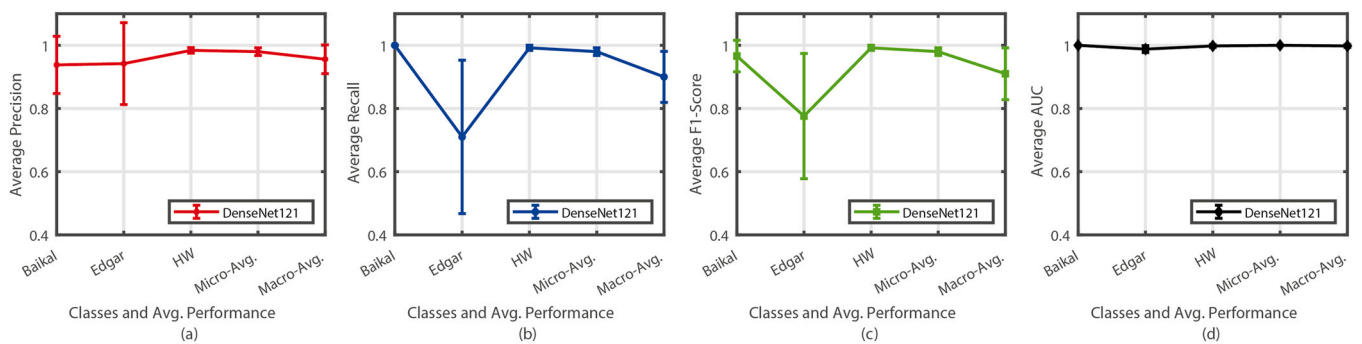

Fig. 12. Performance measures of DenseNet121.

parameters [44]. Hence, it provided better classification performance than machine learning methods.

Further advantages of the proposed machine learning method are that it did not require signal alignment (of the LEA topography from the line segment), while for the deep learning approach, both signal alignment and feature extraction is not needed. Moreover, the proposed deep learning model required less training parameters (1.5 million) compared to other models reported in the literature [12]. Limitations of the proposed approach are: (i) limited dataset for

the proof of concept study, i.e., 38, 34 and 600 LEA regions for the Baikal and Edgar air rifles and the HW air pistol respectively. Unbalanced sample size affects classification performance. A smaller sample size results in poor classification performance. E.g., the SVM with the quadratic kernel using 22 ranked features resulted in an F1-score of 0.63, 0.34 and 0.96 for Baikal, Edgar rifles and HW pistol, respectively, Table A1. A similar trend was observed for DenseNet121, Table A3, (ii) the machine learning approach reported lower performance, and (iii) the experiment was conducted using air
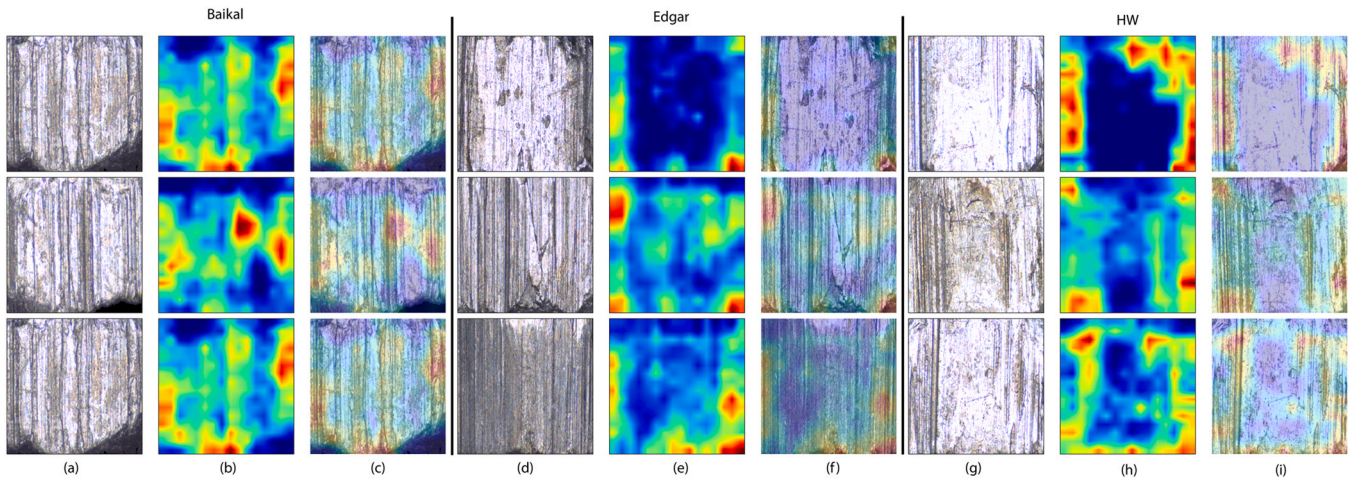
Fig. 13. Grad-CAM visualization: (a), (d) and (g) Original LEA images; (b), (e) and (h) Class discriminative regions; and (c), (f) and (i) Class discriminative regions superimposed with original LEA image.

pellets as a pilot study and would be expected to transfer to bullets fired from other firearms.

In summary, the efficiency of firearm classification was considerably improved using DenseNet121, Table A3, and supported by Grad-CAM visualisation, Fig. 13. This method can be easily transferred to similar applications such as bullet comparison [51]. The methods presented here provide a contribution to the body of research on the application of machine learning and deep learning to forensic science, suggesting a viable deep learning approach to linking bullets to firearms.

Ethics statement

There were no ethical requirements for the collection and analysis of the data.

CRediT authorship contribution statement

Muthu Rama Krishnan Mookiah: Writing – original draft preparation, Methodology, Conceptualisation. **Roberto Puch-Solis**: Writing – original draft preparation, Conceptualisation, Supervision. **Niamh Nic Daeid**: Conceptualisation, Supervision, Funding acquisition, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

Acknowledgements

Appendix A. SVM, DT, RF and DenseNet121 classification performance

Tables A1, A2, A3, Figure A1.

Table A1
SVM, DT and RF performance measures using 22 ranked features.

Classes	Precision	Recall	F1-Score	AUC
Support Vector Machine-Quadratic				
Baikal	0.66 ± 0.28	0.62 ± 0.31	0.63 ± 0.27	0.80 ± 0.15
Edgar	0.34 ± 0.23	0.38 ± 0.07	0.34 ± 0.12	0.66 ± 0.04
HW	0.96 ± 0.02	0.95 ± 0.03	0.96 ± 0.01	0.82 ± 0.05
Micro Average	0.90 ± 0.03	0.90 ± 0.03	0.90 ± 0.03	0.93 ± 0.02
Micro Average	0.66 ± 0.15	0.65 ± 0.10	0.64 ± 0.11	0.76 ± 0.06
Decision Tree				
Baikal	0.63 ± 0.04	0.60 ± 0.09	0.61 ± 0.05	0.79 ± 0.04
Edgar	0.27 ± 0.12	0.38 ± 0.19	0.31 ± 0.15	0.66 ± 0.10
HW	0.97 ± 0.01	0.95 ± 0.01	0.96 ± 0.01	0.84 ± 0.07
Micro Average	0.90 ± 0.01	0.90 ± 0.01	0.90 ± 0.01	0.93 ± 0.01
Macro Average	0.62 ± 0.04	0.64 ± 0.09	0.63 ± 0.06	0.76 ± 0.06
Random Forest with AdaBoost Ensemble Aggregation				
Baikal	0.68 ± 0.06	0.44 ± 0.25	0.51 ± 0.16	0.71 ± 0.12
Edgar	0.41 ± 0.13	0.32 ± 0.16	0.35 ± 0.14	0.65 ± 0.08
HW	0.96 ± 0.02	0.99 ± 0.02	0.97 ± 0.01	0.80 ± 0.07
Micro Average	0.92 ± 0.02	0.92 ± 0.02	0.92 ± 0.02	0.94 ± 0.01
Macro Average	0.68 ± 0.04	0.59 ± 0.09	0.61 ± 0.06	0.72 ± 0.06

Table A2
SVM, DT and RF performance measures using all ranked features.

Classes	Precision	Recall	F1-Score	AUC
Support Vector Machine-Quadratic				
Baikal	0.59 ± 0.21	0.60 ± 0.22	0.59 ± 0.19	0.79 ± 0.12
Edgar	0.31 ± 0.11	0.36 ± 0.14	0.33 ± 0.12	0.66 ± 0.07
HW	0.96 ± 0.02	0.95 ± 0.02	0.95 ± 0.01	0.81 ± 0.07
Micro Average	0.90 ± 0.02	0.90 ± 0.02	0.90 ± 0.02	0.92 ± 0.02
Macro Average	0.62 ± 0.10	0.64 ± 0.11	0.62 ± 0.10	0.75 ± 0.08
Decision Tree				
Baikal	0.52 ± 0.06	0.60 ± 0.09	0.55 ± 0.04	0.78 ± 0.04
Edgar	0.24 ± 0.14	0.30 ± 0.20	0.26 ± 0.15	0.63 ± 0.10
HW	0.97 ± 0.01	0.95 ± 0.02	0.96 ± 0.01	0.86 ± 0.07
Micro Average	0.90 ± 0.02	0.90 ± 0.02	0.90 ± 0.02	0.92 ± 0.01
Macro Average	0.58 ± 0.06	0.62 ± 0.07	0.59 ± 0.06	0.75 ± 0.05
Random Forest with AdaBoost Ensemble Aggregation				
Baikal	0.68 ± 0.08	0.50 ± 0.23	0.54 ± 0.12	0.74 ± 0.11
Edgar	0.32 ± 0.24	0.24 ± 0.22	0.27 ± 0.23	0.61 ± 0.11
HW	0.96 ± 0.02	0.99 ± 0.01	0.98 ± 0.01	0.81 ± 0.06
Micro Average	0.93 ± 0.01	0.93 ± 0.01	0.93 ± 0.01	0.94 ± 0.01
Macro Average	0.65 ± 0.08	0.58 ± 0.08	0.60 ± 0.06	0.72 ± 0.05

Table A3
DenseNet121 performance measures.

Classes	Precision	Recall	F1-Score	AUC
Baikal	0.94 ± 0.09	1.00 ± 0.00	0.97 ± 0.05	1.00 ± 0.00
Edgar	0.94 ± 0.13	0.71 ± 0.24	0.78 ± 0.20	0.99 ± 0.01
HW	0.98 ± 0.01	0.99 ± 0.01	0.99 ± 0.01	0.99 ± 0.01
Micro Average	0.98 ± 0.01	0.98 ± 0.01	0.98 ± 0.01	1.00 ± 0.00
Macro Average	0.97 ± 0.05	0.90 ± 0.08	0.91 ± 0.08	0.99 ± 0.01
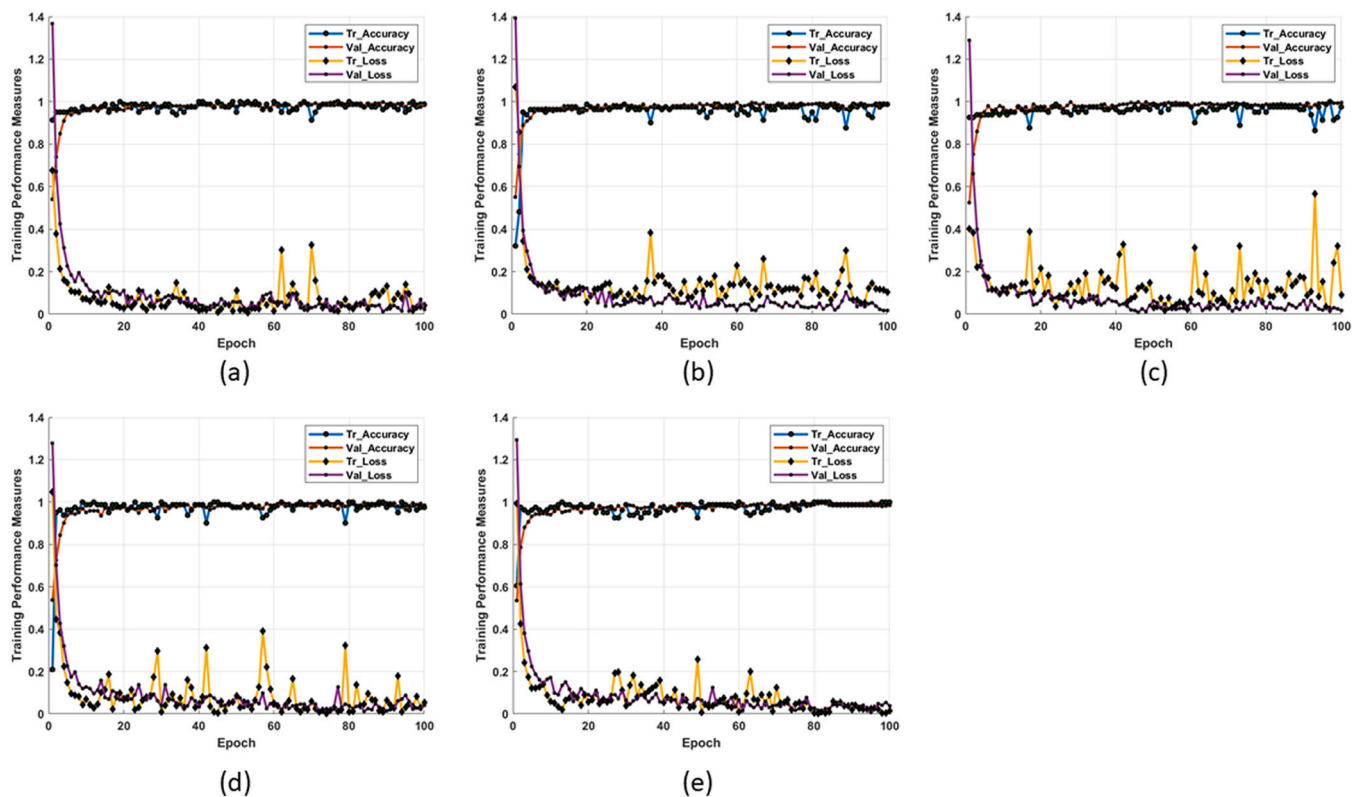
Fig. A1. Training and validation performance of DenseNet121 using 5-fold cross validation.

B. Effect of smaller sample size on classification performance

To study the effect of a smaller sample size, we selected 5 pellets from each of the air rifles and pistol. A total of 34, 34 and 34 LEA regions were considered from the Baikal air rifle, the Edgar air rifle and the HW air pistol respectively. A similar experiment from section 2.2 to section 2.7 was repeated and the results are shown in Tables B1, B2 and B3.

The results revealed that the smaller sample size reduced the classification performance for both machine and deep learning methods (Tables B1, B2 and B3) compared to the results reported in Tables A1, A2 and A3.

Table B1
SVM, DT and RF classifier performance measures using 22 ranked features.

Classes	Precision	Recall	F1-Score	AUC
Support Vector Machine-Quadratic				
Baikal	0.69 ± 0.21	0.53 ± 0.27	0.57 ± 0.21	0.70 ± 0.11
Edgar	0.47 ± 0.13	0.56 ± 0.13	0.51 ± 0.13	0.62 ± 0.12
HW	0.75 ± 0.16	0.71 ± 0.15	0.71 ± 0.05	0.78 ± 0.04
Micro Average	0.60 ± 0.09	0.60 ± 0.09	0.60 ± 0.09	0.70 ± 0.07
Macro Average	0.64 ± 0.11	0.60 ± 0.09	0.60 ± 0.11	0.70 ± 0.07
Decision Tree				
Baikal	0.62 ± 0.10	0.76 ± 0.14	0.68 ± 0.07	0.76 ± 0.06
Edgar	0.48 ± 0.28	0.47 ± 0.32	0.46 ± 0.28	0.61 ± 0.18
HW	0.71 ± 0.19	0.53 ± 0.10	0.59 ± 0.08	0.70 ± 0.05
Micro Average	0.59 ± 0.10	0.59 ± 0.10	0.59 ± 0.10	0.69 ± 0.08
Macro Average	0.60 ± 0.12	0.59 ± 0.10	0.58 ± 0.11	0.69 ± 0.08
Random Forest with AdaBoost Ensemble Aggregation				
Baikal	0.58 ± 0.06	0.65 ± 0.12	0.61 ± 0.06	0.71 ± 0.05
Edgar	0.51 ± 0.14	0.56 ± 0.10	0.52 ± 0.07	0.64 ± 0.06
HW	0.83 ± 0.23	0.58 ± 0.08	0.68 ± 0.12	0.76 ± 0.08
Micro Average	0.60 ± 0.06	0.60 ± 0.06	0.60 ± 0.06	0.70 ± 0.04
Macro Average	0.65 ± 0.08	0.60 ± 0.06	0.60 ± 0.06	0.70 ± 0.04

Table B2
SVM, DT and RF classifier performance measures using all ranked features.

Classes	Precision	Recall	F1-Score	AUC
Support Vector Machine-Quadratic				
Baikal	0.56 ± 0.10	0.53 ± 0.21	0.52 ± 0.12	0.66 ± 0.08
Edgar	0.46 ± 0.17	0.44 ± 0.18	0.45 ± 0.17	0.59 ± 0.10
HW	0.63 ± 0.15	0.65 ± 0.15	0.63 ± 0.12	0.72 ± 0.10
Micro Average	0.54 ± 0.08	0.54 ± 0.08	0.54 ± 0.08	0.66 ± 0.06
Macro Average	0.55 ± 0.10	0.54 ± 0.08	0.53 ± 0.09	0.66 ± 0.06
Decision Tree				
Baikal	0.54 ± 0.14	0.57 ± 0.22	0.54 ± 0.17	0.66 ± 0.12
Edgar	0.47 ± 0.21	0.56 ± 0.31	0.50 ± 0.25	0.62 ± 0.21
HW	0.79 ± 0.21	0.56 ± 0.14	0.64 ± 0.13	0.73 ± 0.09
Micro Average	0.56 ± 0.14	0.56 ± 0.14	0.56 ± 0.14	0.67 ± 0.11
Macro Average	0.60 ± 0.13	0.56 ± 0.14	0.56 ± 0.13	0.67 ± 0.11
Random Forest with AdaBoost Ensemble Aggregation				
Baikal	0.64 ± 0.05	0.61 ± 0.20	0.61 ± 0.11	0.72 ± 0.08
Edgar	0.51 ± 0.03	0.56 ± 0.10	0.53 ± 0.05	0.65 ± 0.03
HW	0.82 ± 0.18	0.74 ± 0.16	0.76 ± 0.08	0.82 ± 0.06
Micro Average	0.64 ± 0.06	0.64 ± 0.06	0.64 ± 0.06	0.73 ± 0.04
Macro Average	0.66 ± 0.08	0.64 ± 0.06	0.63 ± 0.07	0.73 ± 0.04

Table B3
DenseNet121 performance measures.

Classes	Precision	Recall	F1-Score	AUC
Baikal	0.57 ± 0.53	0.34 ± 0.44	0.38 ± 0.43	0.94 ± 0.04
Edgar	0.41 ± 0.26	0.57 ± 0.44	0.43 ± 0.27	0.77 ± 0.20
HW	0.67 ± 0.21	0.83 ± 0.31	0.67 ± 0.15	0.91 ± 0.08
Micro Average	0.58 ± 0.10	0.58 ± 0.10	0.58 ± 0.10	0.76 ± 0.07
Macro Average	0.55 ± 0.20	0.58 ± 0.10	0.49 ± 0.09	0.90 ± 0.08

References

[1] K. Rice, U. Genschel, H. Hofmann, A robust approach to automatically locating grooves in 3d bullet land scans, J. Forensic Sci. 65 (3) (2020) 775–783.

[2] Z. Chen, J. Song, J.A. Soons, R.M. Thompson, X. Zhao, Pilot study on deformed bullet correlation, Forensic Sci. Int. 306 (2020) 110098.

[3] Z. Chen, W. Chu, J.A. Soons, R.M. Thompson, J. Song, X. Zhao, Fired bullet signature correlation using the congruent matching profile segments (cmps) method, Forensic Sci. Int. 305 (2019) 109964.

[4] T.V. Vorburger, J. Song, N. Petraco, Topography measurements and applications in ballistics and tool mark identifications, Surf. Topogr. Metrol. Prop. 4 (1) (2015) 013002.

[5] M. Heizmann, Strategies for the automated recognition of marks in forensic science, in: Investigative Image Processing II, Vol. 4709, International Society for Optics and Photonics, 2002, 68–79.

[6] F.P. León, Automated comparison of firearm bullets, Forensic Sci. Int. 156 (1) (2006) 40–50.

[7] A. Banno, Estimation of bullet striation similarity using neural networks, J. Forensic Sci. 49 (3) (2004) JFS2002361-5.

[8] P. Changmai, K. Bora, R. Suresh, N. Deb, L.B. Mahanta, On the study of automated identification of firearms through associated striations, in: 31st International symposium on ballistics, 2019, 1192–1203.

[9] S. Vanderplas, M. Nally, T. Klep, C. Cadevall, H. Hofmann, Comparison of three similarity scores for bullet lea matching, Forensic Sci. Int. 308 (2020) 110167.

[10] S.K. Dutta, S. Saikia, A. Barman, R. Roy, K. Bora, L.B. Mahanta, R. Suresh, Study on enhanced deep learning approaches for value-added identification and segmentation of striation marks in bullets for precise firearm classification, Appl. Soft Comput. 112 (2021) 107789.

[11] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, 234–241.

[12] P. Pisantanaroj, P. Tanpisuth, P. Sinchavanwat, S. Phasuk, P. Phienphanich, P. Jangtawee, K. Yakoompai, M. Donphoongpi, S. Ekgasit, C. Tantibundhit, Automated firearm classification from bullet markings using deep learning, IEEE Access 8 (2020) 78236–78251.

[13] N.H. Hamzah, Development of an objective method for the comparison of fired projectiles using an air pistol as a template, Forensic Sci. Int. 264 (2016) 106–112.

[14] N. Basu, R.S. Bolton-King, G.S. Morrison, Forensic comparison of fired cartridge cases: Feature-extraction methods for feature-based calculation of likelihood ratios, Forensic Sci. Int.: Synerg. 5 (2022) 100272, , https://doi.org/10.1016/j.fsisyn.2022.100272 ⟨https://www.sciencedirect.com/science/article/pii/S2589871×22000572⟩.

[15] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, 618–626.

[16] N.H. Hamzah, Development of an objective method for the comparison of fired projectiles using an air pistol as a template, Thesis (ph.d.), Centre for Forensic Science, (2017). 10.48730/j2cw-1h15.

[17] S.W. H. Weihrauch, Weihrauch sport air pistols(May 2022). ⟨https://www.weihrauch-sport.de/luftpistolen⟩.

[18] V.J. DiMaio, Gunshot Wounds: Practical Aspects of Firearms, Ballistics, and Forensic Techniques, CRC Press,, 2015.

[19] Alicona, Alicona infinite focus(May 2022). ⟨https://www.alicona.co.uk/home/products/infinitefocus.html⟩.

[20] K. ISO4287, Geometrical product specifications (gps)-surface texture: Profile method-terms, Definitions and Surface Texture Parameters.1997-04.

[21] N. Otsu, A threshold selection method from gray-level histograms, IEEE Trans. Syst. Man Cybern. 9 (1) (1979) 62–66.

[22] R.C. Gonzalez, Digital image processing, Pearson Education india, 2009.

[23] E. Hare, H. Hofmann, A. Carriquiry, Automatic matching of bullet land impressions, Ann. Appl. Stat. (2017) 2332–2356.

[24] W.S. Cleveland, Robust locally weighted regression and smoothing scatterplots, J. Am. Stat. Assoc. 74 (368) (1979) 829–836.

[25] N.E. Huang, Z. Shen, S.R. Long, M.C. Wu, H.H. Shih, Q. Zheng, N.-C. Yen, C.C. Tung, H.H. Liu, The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis, in: Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, Vol. 454, The Royal Society, 1998, 903–995.

[26] M.R.K. Mookiah, U.R. Acharya, H. Fujita, J.E. Koh, J.H. Tan, C.K. Chua, S.V. Bhandary, K. Noronha, A. Laude, L. Tong, Automated detection of age-related macular degeneration using empirical mode decomposition, Knowl. -Based Syst. 89 (2015) 654–668.

[27] R.B. Pachori, P. Avinash, K. Shashank, R. Sharma, U.R. Acharya, Application of empirical mode decomposition for analysis of normal and diabetic rr-interval signals, Expert Syst. Appl. 42 (9) (2015) 4567–4581.

[28] S. Pal, M. Mitra, Empirical mode decomposition based ecg enhancement and qrs detection, Comput. Biol. Med. 42 (1) (2012) 83–92.

[29] T. Tanaka, D.P. Mandic, Complex empirical mode decomposition, IEEE Signal Process. Lett. 14 (2) (2007) 101–104.

[30] S. Bigdeli, H. Danandeh, M.E. Moghaddam, A correlation based bullet identification method using empirical mode decomposition, Forensic Sci. Int. 278 (2017) 351–360.

[31] R.M. Haralick, K. Shanmugam, I.H. Dinstein, Textural features for image classification, IEEE Trans. Syst. Man Cybern. 6 (1973) 610–621.

[32] F. Albregtsen, et al., Statistical texture measures computed from gray level coocurrence matrices, Image processing laboratory, department of informatics, university of oslo 5(5). 2008.

[33] M.R.K. Mookiah, U.R. Acharya, C.M. Lim, A. Petznick, J.S. Suri, Data mining technique for automated diagnosis of glaucoma using higher order spectra and wavelet energy features, Knowl. Based Syst. 33 (2012) 73–82.

[34] S.M. Pincus, Approximate entropy as a measure of system complexity, Proc. Natl. Acad. Sci. 88 (6) (1991) 2297–2301.

[35] U.R. Acharya, F. Molinari, S.V. Sree, S. Chattopadhyay, K.-H. Ng, J.S. Suri, Automated diagnosis of epileptic eeg using entropies, Biomed. Signal Process. Control 7 (4) (2012) 401–408.

[36] A. Delgado-Bonal, A. Marshak, Approximate entropy and sample entropy: A comprehensive tutorial, Entropy 21 (6) (2019) 541.

[37] M.R.K. Mookiah, U.R. Acharya, J.E. Koh, V. Chandran, C.K. Chua, J.H. Tan, C.M. Lim, E. Ng, K. Noronha, L. Tong, et al., Automated diagnosis of age-related macular degeneration using greyscale features from digital fundus images, Comput. Biol. Med. 53 (2014) 55–64.

[38] J. Karmeshu, Entropy Measures, Maximum Entropy Principle and Emerging Applications, 119 Springer Science & Business Media,, 2003.

[39] V.P. Singh, Entropy Theory and Its Application in Environmental and Water Engineering, John Wiley & Sons,, 2013.

[40] Q. Hu, D. Yu, Entropies of fuzzy indiscernibility relation and its operations, Int. J. Uncertain. Fuzziness Knowl. -Based Syst. 12 (05) (2004) 575–589.

[41] M. Radovic, M. Ghalwash, N. Filipovic, Z. Obradovic, Minimum redundancy maximum relevance feature selection approach for temporal gene expression data, BMC Bioinforma. 18 (1) (2017) 1–14.

[42] V.N. Vapnik, An overview of statistical learning theory, IEEE Trans. Neural Netw. 10 (5) (1999) 988–999.

[43] S.R. Safavian, D. Landgrebe, A survey of decision tree classifier methodology, IEEE Trans. Syst. Man Cybern. 21 (3) (1991) 660–674.

[44] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, 4700–4708.

[45] E.L. Allwein, R.E. Schapire, Y. Singer, Reducing multiclass to binary: A unifying approach for margin classifiers, J. Mach. Learn. Res. 1 (Dec) (2000) 113–141.

[46] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Routledge,, 2017.

[47] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

[48] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. Syst. Sci. 55 (1) (1997) 119–139.

[49] W. Chu, J. Song, T.V. Vorburger, R. Thompson, R. Silver, Selecting valid correlation areas for automated bullet identification system based on striation detection, J. Res. Natl. Inst. Stand. Technol. 116 (3) (2011) 647.

[50] T. Fawcett, An introduction to roc analysis, Pattern Recognit. Lett. 27 (8) (2006) 861–874.

[51] F.P. León, Automated comparison of firearm bullets, Forensic Sci. Int. 156 (1) (2006) 40–50.