A novel method to analyse DART TOFMS spectra based on Convolutional Neural Networks: a case study on methanol extracts of wool fibres from endangered camelids

Jahanbanifard, M.; Price, E.; González Benito, A.; Raggi, L.A.; Javanmardi, S.; Lens, F.; ... ; Verbeek, F.J.

International Journal of Mass Spectrometry 489 (2023) 117050

Contents lists available at ScienceDirect

International Journal of Mass Spectrometry

journal homepage: www.elsevier.com/locate/ijms

A novel method to analyse DART TOFMS spectra based on Convolutional Neural Networks: A case study on methanol extracts of wool fibres from endangered camelids

Mehrdad Jahanbanifard [a, b, *], Erin Price [c], Benito A. González [d], Luis A. Raggi [e], Shima Javanmardi [a], Frederic Lens [b, f], Barbara Gravendeel [b, h], Edgard Espinoza [g], Fons J. Verbeek [a]

[a] Leiden Institute of Advanced Computer Science (LIACS), Leiden University, Niels Bohrweg 1, 2333, CA, Leiden, the Netherlands
[b] Naturalis Biodiversity Center, Darwinweg 2, 2333, CR, Leiden, the Netherlands
[c] USFS International Programs Wood Identification and Screening Center, 1490 E. Main St., Ashland, OR, 97520, USA
[d] Faculty of Forest Sciences and Nature Conservation, Universidad de Chile, Chile
[e] Faculty of Veterinary Sciences, Animal Sciences Department, Universidad de Chile, Chile
[f] Leiden University, Institute of Biology Leiden, Plant Sciences, Sylviusweg 72, 2333, BE, Leiden, the Netherlands
[g] National Fish and Wildlife Forensic Laboratory, 1490 E. Main St., Ashland, OR, 97520, USA
[h] Radboud Institute for Biological and Environmental Sciences, Heyendaalseweg 135, 6500 GL Nijmegen, The Netherlands

ARTICLE INFO

Article history:
Received 12 December 2022
Received in revised form
26 March 2023
Accepted 28 March 2023
Available online 4 April 2023

Keywords:
Classification
Chemical fingerprint
Deep learning
Transfer learning

ABSTRACT

Monitoring the illegal trade of wool fibres of wild vicuña (*Vicugna vicugna*) and guanaco (*Lama guanicoe*) is highly desirable. The high market value of fleece from these camelid species poses a threat to their wild populations. A previous study showed that direct analysis in real time time-of-flight mass spectrometry (DART-TOFMS) effectively identifies wool fibres to species. Producing high-resolution data in a short period of time makes DART-TOFMS a reliable identification tool, even though data analysis can still be improved. The present study proposes a novel data analysing pipeline based on Convolutional Neural Networks (CNN), applicable to any kind of DART-TOF MS data. We tested our proposed method on keratin fibres of four camelid species (*Vicugna vicugna*: n = 19; *Vicugna pacos*: n = 20; *Lama guanicoe*: n = 20, and *Lama glama*: n = 20). Analyses showed that selecting 512 ions with the highest relative intensity provides the best resolution and yields 100% accuracy for species identification.

1. Introduction

South American camelids species (*Lama* spp. and *Vicugna* spp.) are worldwide renowned for their fine wool; two wild species in particular, vicuña (*Vicugna vicugna*) and guanaco (*Lama guanicoe*) produce some of the finest wool fibres in the world [1]. The high market value of the raw fibre of wild camelids has been an economic incentive for developing sustainable use of the species for the benefit of local communities [2]. Currently, both species of wild camelids are listed under Appendices of the Convention on International Trade in Endangered Species (CITES) for the international trade of individuals and their products. All *guanaco* species and the vicuña populations of Argentina, Bolivia, Chile and Peru are listed in Appendix II. A small population of vicuña located in the geographic regions of Antofagasta and Atacama in Chile is listed under Appendix I [3].

An earlier study showed that ambient ionisation using direct analysis in real time time-of-flight mass spectrometry (DART-TOFMS) is an accurate tool for distinguishing wool fibres of camelid species with high accuracy [1]. DART-TOFMS is a technique that allows fast acquisition of chemical profiles of almost any material, and its ease of use is an advantage over other chemical mass spectrometry methods [4]. This method has been used in various forensic science studies, such as drug analysis [5], wood identification [6] and rhinoceros keratin identification [7].

Despite the ever-growing application of DART-TOFMS for identification purposes, approaches for data analysis have not kept pace

* Corresponding author. Leiden Institute of Advanced Computer Science (LIACS), Leiden University, Niels Bohrweg 1, 2333, CA, Leiden, the Netherlands.
E-mail address: m.jahanbanifard@liacs.leidenuniv.nl (M. Jahanbanifard).

https://doi.org/10.1016/j.ijms.2023.117050

with its growing demand. The conventional approach to data analysis for classification typically involves the use of dimensionality reduction and feature extraction techniques, such as principal component analysis (PCA), in conjunction with a classifier such as the support vector machine (SVM) [4]. Other classification methods such as random forest, as demonstrated by Finch et al. [8] and Deklerck et al. [9], as well as kernel discriminant analysis (KDA), as shown by Paredes-Villanueva et al. [10], have also been utilised.

Although Principal Component Analysis (PCA) has demonstrated efficacy in the context of smaller datasets, its applicability diminishes when addressing larger datasets comprising tens of thousands of samples. As an alternative, Convolutional Neural Networks (CNNs) and heatmap analysis of peak patterns, as opposed to individual data points, offer a more suitable approach for analysing extensive datasets.

The advantages of CNNs and deep learning methodologies in general extend beyond their applicability to large datasets with intricate patterns. Firstly, CNNs can autonomously learn features from raw spectra, eliminating the necessity for manual feature extraction or dimensionality reduction. Secondly, CNNs can identify nonlinear and complex patterns within spectra that may remain undetected using a conventional method. Thirdly, CNNs can more effectively manage noisy and variable spectra compared to conventional classifiers by employing convolutional filters and pooling layers. Consequently, deep learning methods have garnered increasing popularity and are favoured over traditional approaches [19,22,24].

CNNs are a class of algorithms inspired by the structure of the visual cortex that assign weights and biases to several objects to recognise them in the images [11–13].

Early use of CNNs was in modern computer vision and it was introduced by LeCun et al., in 1998. Using the backpropagation approach, they trained the convolution kernel coefficients directly from images of hand-written numbers [14]. Since then, several convolutional networks with different structures have been introduced such as VGG [15], ResNet (residual network) [16], Inception [17] and DenseNet [18]. The structure of a CNN consists of convolutional layers followed by pooling layers and fully connected layers. The primary layers extract low-level features such as colours and edges, while the deeper layers in the network extract high-level features such as texture [19]. Pooling layers minimise data dimensions by merging the outputs of neuron clusters of one layer into a single neuron in the following layer to avoid an exponential rise in the number of parameters [20]. Fully connected layers are usually the final layers in which all the inputs from one layer connect to all the activation units of the following layers. They compile the features extracted by previous layers to generate the final output of the model [21].

This study presents a DART-TOFMS data analysis pipeline powered by CNNs. We tested our data set with three well-known networks, ResNet-50, Inception-v3 and DenseNet201, and six classifiers, namely softmax, linear discriminant analysis (LDA), linear support vector machine (LSVM), cubic SVM (CSVM), quadratic SVM (QSVM) and k-nearest neighbour (KNN) to classify wool fibres of four camelid species based on the deep features extracted by CNNs (Fig. 1). The selected networks are mainly introduced for the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), a competition to assess algorithms (usually CNNs) for object recognition and image classification tasks at a large scale [19,22,23]. The details of the employed models will be discussed in the next section.

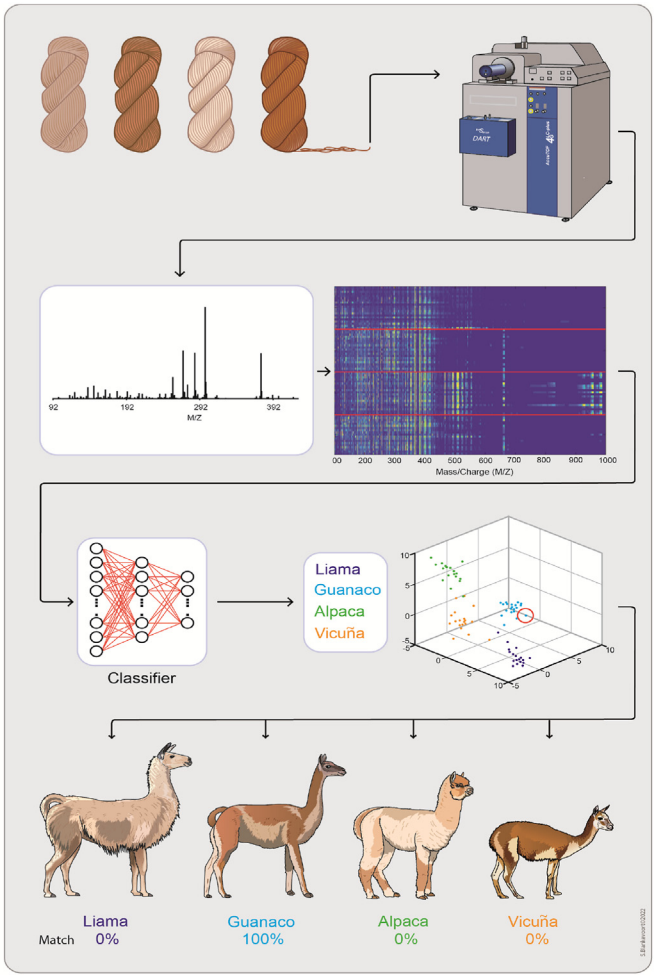

Fig. 1. Visual illustration of the pipeline. Methanol extracts of wool fibres from four camelid species are analysed with a DART TOFMS, which generates a unique chemical peak pattern. This pattern is converted into heatmaps and fed into a CNN for classification.

2. Material and methods

2.1. Wool fibre sampling

Fibre samples were collected from wild guanaco (*Lama guanicoe*, n = 20) and vicuña (*Vicugna vicugna*, n = 19) carcasses found in Chile [1]. The domestic alpaca (*Vicugna pacos*, n = 20) were collected from a private ranch in White City, Oregon, USA [1], and domestic llama (*Lama glama*, n = 20) samples were collected from herds living in San Pedro de Atacama, Chile and added to the data set.

2.2. DART-TOFMS data acquisition

Detailed instrument settings and sample preparation can be found in Ref. [1]. In order to reduce potential variation and keep sample sizes consistent, untreated wool fibre samples containing both underfur and guard hair fibres were pressed into felt-like discs using a 13 mm FTIR pellet die (Specac Ltd, Orpington, UK) and a laboratory press (8 tons of pressure; Fred Carver Inc., Wabash In,

USA). Approximately one-quarter of each disc (~41 mg) was placed in a borosilicate test tube with 1 mL of MeOH and extracted for 30 min. Each sample was vortexed for 10s upon immersion in MeOH and again immediately prior to spectra collection. A single spectrum per sample was collected from each extraction in positive-ion mode by dipping the closed end of a capillary tube into the extract and holding the tube in front of the DART-TOFMS. Spectra were collected over a range of 100–1000 *m/z* and calibrated using poly (ethylene glycol) 600 (Ultra Scientific, Kingstown, RI, USA) which was sampled using a capillary tube as described above. Single total ion chromatograms representing a single sample underwent ambient background subtraction and averaging prior to being exported as centroided text files using msAxel (version 1.05.2, JEOL Ltd.).

2.3. Data processing pipeline

Once the DART-TOFMS data were acquired, each spectrum was calibrated, averaged and the background subtracted before being stored in text files by msAxel (version 1.0.5.2; JEOL Ltd). The first column represents the mass-charge ratio (*m/z*), and the second column contains absolute intensity values. To enable comparison between spectra, we converted the intensity of each spectrum to a relative intensity and normalized it to a 0–100 range. Mass-to-charge ratio values were rounded to AccuTOF 4G's resolution of 5 milli mass units (mmu). A ranking approach was utilised to identify high peaks in our analysis. All peaks were sorted according to their intensity and the top n peaks were selected for further investigation (e.g., top 16, 128, 512, and 1024) to find the optimum number of ions for classification. Finally, each spectrum was fed into the msheatmap function as present in the Bioinformatics toolbox of MATLAB to generate a heatmap file.

2.4. Deep neural networks training

For the feature extraction phase, ResNet50, Inception-V3 and DenseNet201 were used. These networks benefit from improved architecture over the traditional network models. Their architecture provides them with the capability of overcoming the challenge of increasing network depths without falling into the vanishing or exploding gradient problem [19]. In conventionally structured models like AlexNet [12] and VGGNet [15], the inputs, i.e. the images, are processed by several convolutional layers followed by fully connected layers to extract high-level features. The network tries to calculate the gradients of the front layers in an *n*-layer network by multiplying the gradient value of each *n* layer. The gradients can become zero (vanished) in multiplying small values or become overly large (exploded) in multiplying large values. A summary of the architectures of ResNet50, Inception-V3 and DenseNet201 will be given in the next section.

2.4.1. Resnet-50

By introducing the residual block concept in 2015, ResNet pushed the depth of convolutional networks to their limits and mitigated the vanishing gradient problem, which often occurs in deep neural networks that use gradient-based learning methods and backpropagation. ResNet benefits from a structure called shortcut or skip-connection to fit the input from the previous layer to the next layer without modifying it. Instead of learning the signal representation directly, skip-connection provides a deeper network by learning the residual representation functions [16]. Stacking residual blocks throughout ResNet's network allows it to have more than 150 layers. Despite its deep structure, ResNet models have far fewer trainable parameters than conventional methods. For instance, ResNet-50 has around 23 million trainable parameters and a top 5 error of 3.57% on ImageNet, while Alex net has around 60 million parameters and a top 5 error of 15.3%. Fig. 2 illustrates the architecture of ResNet-50.

ResNet-50 consists of four stages with a total of 50 layers including convolutional, max pooling and average pooling layers.

2.4.2. Inception-V3

Inception v3 was developed using GoogleNet, which uses different kernel sizes (1×1, 3×3, 5×5) to extract feature maps at different scales. Inception v3 replaces the 5×5 kernels of Inception v1 with two 3×3 kernels and replaces the $n \times n$ symmetric convolution of Inception v1 with $1 \times n$ and $n \times 1$ asymmetric convolutions to extract feature maps in different scales and stack them. It is, therefore, possible to reduce the number of parameters by increasing the depth of the network. This architecture allows the model to extract more features in total [17]. Inception's improved module reduces computational complexity and improves key feature extraction. The model of Inception v3 is shown in Fig. 3.

There are three layers in each convolutional module: the convolutional layer, the batch norm layer, and the rectified linear unit (ReLU) layer. It includes four layers of 3×3 convolution modules, two layers of 3×3 Max pooling layers, and one layer of 3×3 convolution modules.

2.4.3. DenseNet201

DenseNet was proposed by Huang et al., in 2017 [18]. DenseNet is a model built similarly to ResNet, with the main difference being that each layer of the model is fed-forward to the next to maximize the flow of information between the layers of the network. In other words, the layers in DenseNet connect to every other layer feed-forwardly and the outputs of each layer act as input for all successive layers. A layer obtains additional inputs and passes its feature maps to the next layer based on inputs received from previous layers. With such an architecture, DenseNet comes with several impressive advantages. These include the ability to mitigate the vanishing gradient problem, strengthen feature propagation, improve feature reuse, and substantially reduce the number of parameters required in the algorithm. As a result, the network is lighter and more compact, leading to higher computational and memory efficiency. Fig. 4 depicts the DenseNet-201 architecture.

DenseNet-201 consists of four blocks of convolutional layers, three transition layers, and one fully connected layer followed by classification layer.

2.4.4. Fine-tuning, parameters, training and validation

The application of the transfer learning approach benefited from the pre-trained network by adjusting its parameters typical to our data set. This procedure is also known as fine-tuning and is faster than training from scratch as a pre-trained network already has established weights. The pre-trained weights result from learning over a data set (usually ImageNet) and help the network train the features faster [24]. To realise this, the last fully connected layer of networks was modified to contain the same number of outputs according to the number of classes we used for classification. To ensure that the results were robust, we ran each algorithm three times and reported the average of the validation accuracy.

The following parameters were set for all the networks: optimiser = SGDM, epoch = 500, batch size = 64, learning rate = 0.001. The output of each fully connected layer of networks was fed into six different classifiers (Softmax, LDA, LSVM, CSVM, QSVM and KNN), and the performance of each was reported separately. A 5-fold cross-validation approach was taken to protect the models against overfitting. 5-fold cross-validation divides the whole data set into five separated groups, trains the model with four groups, and tests it over the remaining set (validation set). This

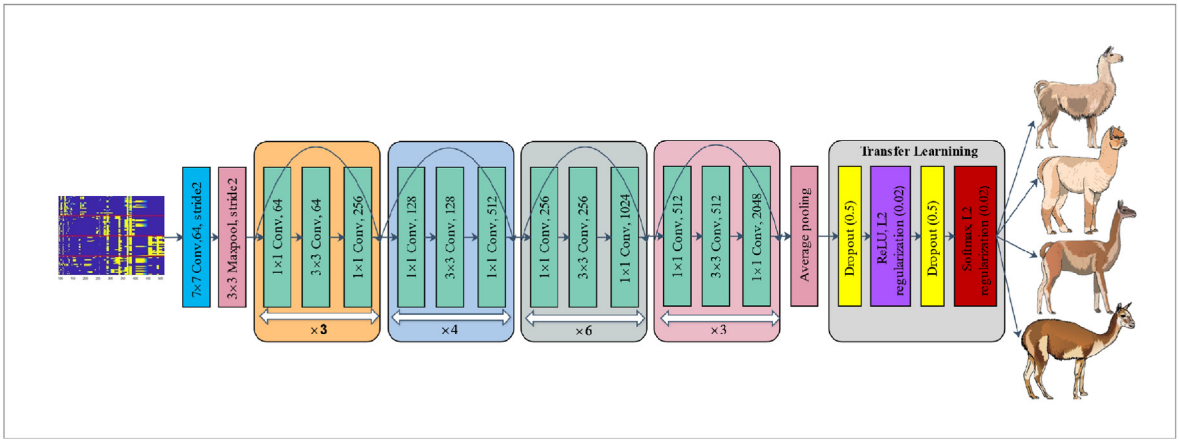

Fig. 2. ResNet-50 architecture.

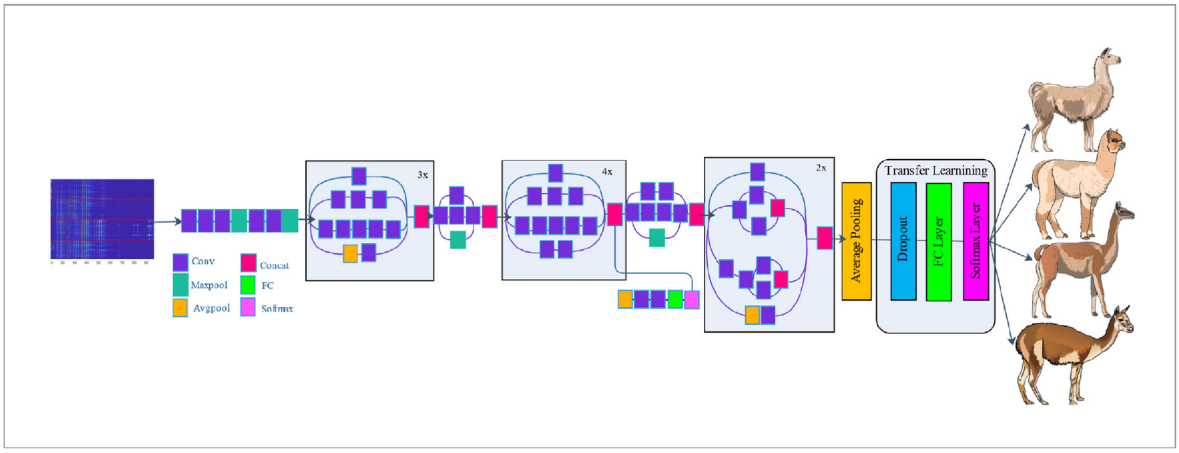

Fig. 3. Inception-V3 architecture.

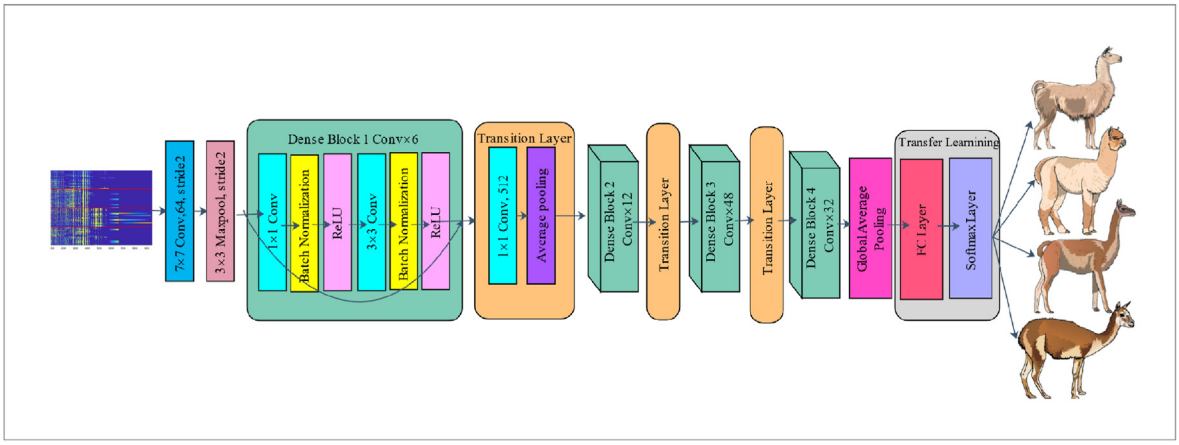

Fig. 4. DenseNet-201 architecture.

procedure was repeated until all the groups had been considered for validation once. The performance of each model was assessed based on the validation accuracy, which is the percentage of correctly classified test data by the trained model. All the algorithms were implemented in the MATLAB Deep Learning Toolbox (MATLAB R2021a, MathWorks Inc.).

3. Results

3.1. Peak pattern analyses

The DART-TOFMS peak patterns of 79 camelid wool fibre samples, summarised in four different peak numbers (16, 128, 512 and

1024) are illustrated in Fig. 5.

The concentration of high-intensity ions is in the [100–500 m/z] area, especially [250–450 m/z]. However, *V. pacos* has relatively high-intensity ions around the [900–1000 m/z] area, making it unique. As the peak number increases, the contrast of the relative intensity decreases, while the resolution of the peak pattern heatmaps also increases. There is a trade-off between all these factors, which is in line with the results (Table 1), showing that too low or too high peak numbers affect the classification results.

3.2. Classification results based on the features extracted by ResNet50

The classification performance of five classifiers was evaluated (Table 1).

The results of our study showed that the LSVM classifier achieved the highest accuracy (98.7%), followed by the QSVM (97.5%), CSVM (96.2%), KNN (94.9%), LDA (90.3%), and softmax (84%) classifiers. The differences in performance between the classifiers can be attributed to various factors, such as their underlying algorithms, hyperparameters, and training procedures. For example, the LSVM and QSVM classifiers are both support vector machine classifiers, but they use different kernel functions or parameters that can affect their performance. Similarly, KNN performs differently based on the number of neighbours chosen or the distance metric used. Overall, the results suggest that the choice of classifier can have a significant impact on the accuracy of image classification.

In our study, almost all the classifiers achieved high accuracy using the extracted features from ResNet50. There are several factors that may have contributed to the success of ResNet50 in our task. Firstly, ResNet50 is a deep convolutional neural network with 50 layers, allowing it to capture more complex features from input images. Secondly, ResNet50 uses skip connections to alleviate the vanishing gradient problem that can occur in very deep networks, making it easier for the network to learn useful features. Additionally, ResNet50 has been pre-trained on large-scale datasets, such as ImageNet, enabling it to learn generic features that can be transferred to other tasks. All together, these factors provide insight why ResNet50 performed well in our study.

Table 1
Classification results of different classifiers applied to the extracted features by ResNet50.

Data sets	Classifiers Accuracy %					
	Softmax	LDA	LSVM	QSVM	CSVM	KNN
16 Peaks	72.2	92.4	91.1	91.1	91.1	89.9
128 Peaks	72.2	92.4	89.9	86.1	86.1	87.3
512 Peaks	93	100	100	98.7	100	100
1024 Peaks	84	90.3	98.7	97.5	96.2	94.9

LDA outperformed every other classifier by a narrow margin. The analyses show that increasing the number of peaks from 16 to 512 leads to an increase in the accuracy of the classification. However, going up to 1024 peaks and beyond, decreases the accuracy of the classification as most of the spectra have less than 1000 ions and further increasing the peak number does not increase the resolution anymore.

3.3. Classification results based on the features extracted by InceptionV3

The classification results of the extracted features by InceptionV3 are summarised in Table 2.

LDA again showed the best performance amongst the methods applied and 512 remained the optimum peak number for the most accurate classification. Our results showed that InceptionV3

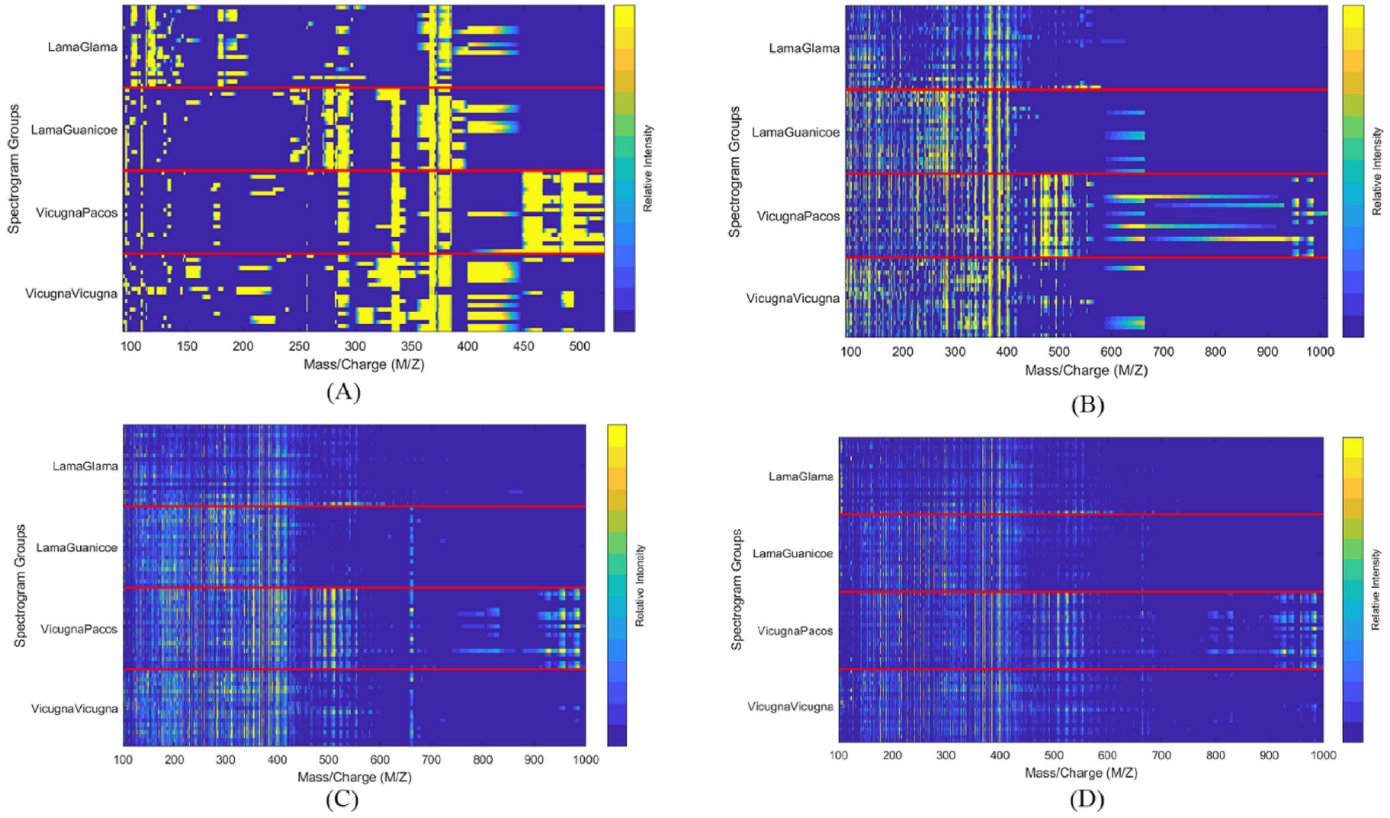

Fig. 5. Peak patterns of four selected data sets derived from DART-TOF Mass Spectra of Camelid wool fibres: 16-peak heatmaps (A), 128-peak heatmaps (B), 512-peak heatmaps (C) and 1024-peak heatmaps (D). The colour bars show the relative intensity from 0 (dark blue) to 1 (yellow).

Table 2
Classification results of different classifiers applied to the extracted features by InceptionV3.

Data sets	Classifiers Accuracy %					
	Softmax	LDA	LSVM	QSVM	CSVM	KNN
16 Peaks	77.7	92.4	89.9	87.3	83.5	86.1
128 Peaks	68	89.9	86.1	83.5	79.9	87.3
512 Peaks	79.1	93.7	93.7	94.9	92.4	92.4
1024 Peaks	63.9	88.6	89.9	89.9	89.9	89.9

achieved an average accuracy of 86.24 ± 2% for all classifications, which is lower than the performance achieved by ResNet50. One possible explanation for this difference in performance is the architecture of the two models. InceptionV3 uses a different type of convolutional layer called the inception module, which has multiple branches for processing different scales of features in parallel. While this can be beneficial for capturing a wider range of features, it can also lead to increased complexity and may require more data to train effectively. Additionally, InceptionV3 has fewer parameters than ResNet50, which may limit its ability to capture fine-grained details in the input images.

Furthermore, the performance of the different classifiers varied for InceptionV3 compared to ResNet50. For example, the softmax classifier achieved a lower accuracy of 63.9% for InceptionV3, compared to 84% for ResNet50. This may be due to the fact that softmax classifiers typically work better with highly separable features, which may not be the case for InceptionV3 features. On the other hand, the LDA, LSVM, QSVM, CSVM, and KNN classifiers achieved higher accuracies for InceptionV3 compared to softmax, indicating that these classifiers are more suitable for handling the more complex features extracted by InceptionV3. In conclusion, our results suggest that the choice of feature extraction method can have a significant impact on the performance of classifiers in image classification tasks. While ResNet50 may have certain advantages for our specific dataset, InceptionV3 can still provide useful features for classification with appropriate classifiers.

3.4. Classification results based on the features extracted by DenseNet201

DenseNet201 is the third CNN that we applied as a feature extractor (Table 3).

The classification results show that LDA performance is slightly better than the other models. In line with what we observed with the other two networks applied, 512 peaks provided the highest resolution for the classification task. DenseNet201 marginally outperformed ResNet50 based on the average accuracy of all models (91.6 ± 0.7%). One possible explanation for this improved performance is the architecture of DenseNet201, which incorporates densely connected blocks to improve information flow throughout the network. This can lead to better feature representation and more efficient use of parameters, which may be particularly beneficial for our specific dataset.

Additionally, by using the extracted features from DenseNet201, the softmax classifier achieved an accuracy of 80.5%, which is lower than for ResNet50, but higher than for InceptionV3. This may be due to the fact that DenseNet201 has a larger number of parameters than InceptionV3, which can allow it to capture more fine-grained details in the input images. On the other hand, the LDA, LSVM, QSVM, CSVM, and KNN classifiers achieved higher accuracies for DenseNet201 compared to both ResNet50 and InceptionV3, indicating that these classifiers may be better suited for handling the features extracted by DenseNet201.

Table 3
Classification results of different classifiers applied to the extracted features by DenseNet201.

Data sets	Classifiers Accuracy %					
	Softmax	LDA	LSVM	QSVM	CSVM	KNN
16 Peaks	77.7	92.4	92.4	92.4	88.6	91.1
128 Peaks	81.9	94.9	94.9	92.4	92.4	93.7
512 Peaks	90.2	98.7	97.5	94.9	96.2	94.9
1024 Peaks	80.5	97.5	97.5	96.2	94.9	98.7

In conclusion, our study highlights the importance of selecting appropriate feature extraction methods and classifiers for achieving high accuracy in image classification tasks. DenseNet201, with its unique architecture and ability to capture fine-grained details, is promising in achieving high accuracy in classification tasks. Future studies can further explore the potential of this and other feature extraction methods in different datasets and applications.

4. Discussion and conclusion

This study contributes to specifying the taxonomic origin of wool fibres from four camelid species. Our results discriminate with a high level of accuracy all four species of South American camelid using keratin analysis combined with AI. Previous studies on molecular differentiation of species have relied on genetic, genomic, and proteomic analyses, which are indicative of the intricate nature of the group and its evolutionary background [25].

From phylogenetic studies based on mitochondrial DNA only two lineages, *Lama* and *Vicugna* [26], could be identified. Mutation associated to colour genes relate to the domestication process have been successful for the differentiation between wild camelids from the domestic ones [27]. On the other hand, extensive genomic analysis comparing the four species convincingly supported the hypothesis that the alpaca was domesticated from the vicuña and that the llama was domesticated from the guanaco. Interestingly, high level of introgression from llama in alpaca has been detected [28]. This might affect all parameters for production as present in current domestic South American camelids, such as the fibre quality and fibre composition. In this way, proteomic studies of fibres from the four species and archaeological fibres revealed 7 peptides specific to South American camelids. Additionally, 5 new taxonomic peptides found in keratins discriminate only the two wild species, guanaco and vicuña [29].

Here, results support again the previous study using DART-TOFMS on fibre from three species that discriminate with high level of accuracy [1], but now among all South American camelid species. Having a fast and accurate analysing pipeline together with a high-throughput identification tool such as DART-TOFMS is vital for law enforcement and wildlife trade regulations to identify and protect species, especially endangered species like vicuña and guanaco [1].

Price et al. [1] applied discriminant analysis of principal components (DAPC) to spectra collected via DART-TOFMS from three camelids: *V. pacos*, *V. vicugna*, and *L. guanicoe*. Leave-One-Out Cross-Validation (LOOCV) of the model resulted in a score of 93.4%. Model accuracy was tested with spectra from seven specimens that were not included in the training model, with an assignment probability over 98% [1]. Our study included spectra from *L. glama* and benefited from three well-known convolutional neural networks (ResNet50, InceptionV3 and DenseNet201) coupled with five linear classifiers (softmax, LSVM, CSVM, QSVM and KNN) to analyse the spectra. The features extracted by the DenseNet201 network achieved a higher accuracy on average. Moreover, ResNet50 could even provide 100% separation based on the 512-peak heatmaps (Fig. 6).

Subtle differences in keratins and the process of keratinization modulated by evolutionary trajectory of each species or differences

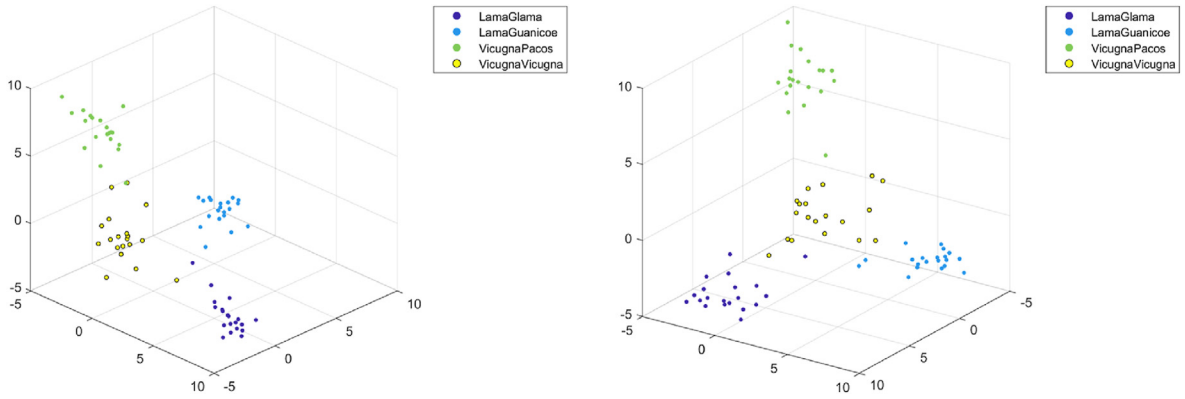

Fig. 6. Two views from the 3D distribution of the feature space of the 512-peak data set as extracted by ResNet50.

on nutritional physiology could explain the differences of keratin composition among species. The structure, function and evolution of keratins, keratin granules and keratin filaments are subject to selective pressure because of their interactions with their environment, adapting gradually to new conditions [30], which could explain the pattern among South American camelids.

Another complementary explanation can be attributed to differences in the type, quality, and quantity of forage resources. For vicuñas and guanacos, this should be considered at the local scale. For alpacas and llamas, the supplementation of feed provided by controlled production systems can also result in differences in the nutritional composition of the diet. These differences in diet may potentially affect the primary nutrients required for the metabolism of fibre production.

Although several studies are focused on the effect of nutrition over productive parameters in South American camelids such as fibre diameter and length, and fleece weight [31–33], the effect of nutrition on the intimate structure of the hair and wool have not been addressed. For example, wool follicle development is affected by short-term sub-maintenance nutrition in sheep [34]. Alternatively, it is also possible that in South American camelids ruminal metabolism mediated by microorganisms provides the essential metabolites for keratin production independent of diet quality. Therefore, specific studies are necessary to assess these or another plausible hypothesis.

The proposed method is suitable for batch data analyses and appropriate for fitting in an automated DART-TOFMS data analysing pipeline. As researchers, the authors acknowledge that CNN models are widely known for requiring large training sets, which could be challenging to prepare such datasets in the field of wildlife forensic and may not perform optimally when applied to high-dimensional data with limited sample sizes. However, in this study, different CNN models and classifiers were investigated on various datasets to ensure that the results are reliable, reproducible, and not overfitted.

The application of DART-TOFMS and AI for discriminating fibre keratins in South American camelids can also be useful for archaeological studies where ancient textiles are found. This analysis approach can then be applied to differentiate fibres from wild species when trafficking is detected. All considered, the combination of DART-TOFMS with CNN's is a robust analysis technique that provides a lot of new possibilities for fibre keratin research.

Legal note

The findings and conclusions in this article are those of the authors and do not represent the views of the US Fish & Wildlife or the US Forest Service.

Author statement

Mehrdad Jahanbanifard: Conceptualization, Methodology, Software, Validation, Formal Analyses, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization.

Erin Price: Conceptualization, Methodology, Validation, Writing - Original Draft, Writing - Review & Editing.

Benito A. González Resources, Writing - Original Draft, Writing - Review & Editing.

Luis A. Raggi Resources, Writing - Original Draft, Writing - Review & Editing.

Shima Javanmardi: Writing - Original Draft, Writing - Review & Editing, Visualization.

Frederic Lens Writing - Review & Editing.

Barbara Gravendeel Writing - Review & Editing, Visualization.

Edgard Espinoza: Validation, Resources, Data Curation, Writing - Review & Editing, Supervision.

Fons J. Verbeek Writing - Review & Editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgements

We thank Bas Blankevoort for making the infographic, Ernesto Perez for field sampling of llama herds in Chile and Ms. Kristina Kish and Mr. Cristobal Barros for providing the funding which supported the logistical collection of the camelid samples in extremely remote areas of S. America. This study was partially supported by the Plant. ID Innovative Training Network that received funding from the European Union's Horizon 2020 research and innovation program under Marie Skłodowska-Curie grant agreement No 765000.

References

[1] E. Price, D. Larrabure, B. Gonzales, P. McClure, E. Espinoza, Forensic identification of the keratin fibers of South American camelids by ambient ionization mass spectrometry: vicuña, alpaca and guanaco, Rapid Commun. Mass Spectrom. 34 (23) (2020) e8916, https://doi.org/10.1002/rcm.8916. Mar.

[2] D. Castilla, et al., Enteric viral infections among domesticated South American

camelids: first detection of mammalian orthoreovirus in camelids, Animals (Basel) 11 (5) (2021), https://doi.org/10.3390/ani11051455. Mar.
[3] R.B. Centro N. Patagónico, et al., "{IUCN} red list of threatened species," {IUCN} red list of threatened species [Online]. Available: https://www.iucnredlist.org/species/11186/18540211, 2016.
[4] E. Sisco, T.P. Forbes, Forensic applications of {DART}-{MS}: a review of recent literature, Forensic Chem. 22 (2021), 100294, https://doi.org/10.1016/j.forc.2020.100294. Mar.
[5] R. Lian, et al., Rapid screening of abused drugs by direct analysis in real time (DART) coupled to time-of-flight mass spectrometry (TOF-MS) combined with ion mobility spectrometry (IMS), Forensic Sci. Int. 279 (2017) 268–280, https://doi.org/10.1016/j.forsciint.2017.07.010. Mar.
[6] C. Lancaster, E. Espinoza, Analysis of select Dalbergia and trade timber using direct analysis in real time and time-of-flight mass spectrometry for CITES enforcement, Rapid Commun. Mass Spectrom. 26 (9) (2012) 1147–1156, https://doi.org/10.1002/rcm.6215. Mar.
[7] E.R. Price, P.J. McClure, R.L. Jacobs, E.O. Espinoza, Identification of rhinoceros keratin using direct analysis in real time time-of-flight mass spectrometry and multivariate statistical analysis, Rapid Commun. Mass Spectrom. 32 (24) (2018) 2106–2112, https://doi.org/10.1002/rcm.8285. Mar.
[8] K. Finch, E. Espinoza, F.A. Jones, R. Cronn, Source identification of western Oregon douglas-fir wood cores using mass spectrometry and random forest classification, Appl. Plant Sci. 5 (5) (2017), 1600158, https://doi.org/10.3732/apps.1600158. Mar.
[9] V. Deklerck, et al., A protocol for automated timber species identification using metabolome profiling, Wood Sci. Technol. 53 (4) (2019) 953–965, https://doi.org/10.1007/s00226-019-01111-1. Jul.
[10] K. Paredes-Villanueva, E. Espinoza, J. Ottenburghs, M.G. Sterken, F. Bongers, P.A. Zuidema, Chemical differentiation of Bolivian Cedrela species as a tool to trace illegal timber trade, Forestry: Int. J. Financ. Res. 91 (5) (2018) 603–613, https://doi.org/10.1093/forestry/cpy019. Dec.
[11] D.H. Hubel, T.N. Wiesel, Receptive fields and functional architecture of monkey striate cortex, J. Physiol. 195 (1) (1968) 215–243, https://doi.org/10.1113/jphysiol.1968.sp008455. Mar.
[12] A. Krizhevsky, I. Sutskever, G.E. Hinton, {ImageNet} classification with deep convolutional neural networks, Commun. ACM 60 (6) (2012) 84–90, https://doi.org/10.1145/3065386. Mar.
[13] M. V Valueva, N.N. Nagornov, P.A. Lyakhov, G. V Valuev, N.I. Chervyakov, Application of the residue number system to reduce hardware costs of the convolutional neural network implementation, Math. Comput. Simulat. 177 (2020) 232–243, https://doi.org/10.1016/j.matcom.2020.04.031. Mar.
[14] Y. LeCun, et al., Backpropagation applied to handwritten zip code recognition, Neural Comput. 1 (4) (1989) 541–551, https://doi.org/10.1162/neco.1989.1.4.541. Mar.
[15] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, 2014 {arXiv} preprint {arXiv}:1409.1556.
[16] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: {IEEE} Conference on Computer Vision and Pattern Recognition ({CVPR}), 2016, pp. 770–778, https://doi.org/10.1109/{CVPR}.2016.90. Mar.
[17] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Inception v3 Rethinking the inception architecture for computer vision, in: Proceedings of 2016 {IEEE} Conference on Computer Vision and Pattern Recognition ({CVPR}), 2016, pp. 2818–2826, https://doi.org/10.1109/{CVPR}.2016.308. Mar.
[18] G. Huang, Z. Liu, L. van der Maaten, K.Q. Weinberger, {DenseNet} Densely connected convolutional networks, in: 2017 {IEEE} Conference on Computer Vision and Pattern Recognition ({CVPR}), 2017, pp. 2261–2269, https://doi.org/10.1109/{CVPR}.2017.243. Mar.
[19] S.-H.M. Ashtiani, S. Javanmardi, M. Jahanbanifard, A. Martynenko, F.J. Verbeek, Detection of mulberry ripeness stages using deep learning models, IEEE Access (2021) 1, https://doi.org/10.1109/{ACCEß}.2021.3096550.
[20] D. Ciresan, U. Meier, J. Schmidhuber, Multi-column deep neural networks for image classification, in: 2012 {IEEE} Conference on Computer Vision and Pattern Recognition, 2012, pp. 3642–3649, https://doi.org/10.1109/{CVPR}.2012.6248110. Mar.
[21] P. Li, B. Wang, L. Zhang, Virtual fully-connected layer: training a large-scale face recognition dataset with limited computational Resources, in: 2021 {IEEE}/{CVF} Conference on Computer Vision and Pattern Recognition ({CVPR}), 2021, pp. 13310–13319, https://doi.org/10.1109/{CVPR46437}.2021.01311. Mar.
[22] F. Lens, et al., Computer-assisted timber identification based on features extracted from microscopic wood sections, IAWA J. (2020) 1–21, https://doi.org/10.1163/22941932-bja10029. Mar.
[23] S. Javanmardi, S.-H.M. Ashtiani, F.J. Verbeek, A. Martynenko, Computer-vision classification of corn seed varieties using deep convolutional neural network, J. Stored Prod. Res. 92 (2021), 101800, https://doi.org/10.1016/j.jspr.2021.101800. Mar.
[24] E.C. Too, L. Yujian, S. Njuki, L. Yingchun, A comparative study of fine-tuning deep learning models for plant disease identification, Comput. Electron. Agric. 161 (2019) 272–279, https://doi.org/10.1016/j.compag.2018.03.032. Mar.
[25] Molecular evolution of the family Camelidae: a mitochondrial DNA study, Proc. R. Soc. Lond. B Biol. Sci. 256 (1345) (1994) 1–6, https://doi.org/10.1098/rspb.1994.0041. Apr.
[26] J.C. Marín, et al., Sistemática, taxonomía y domesticación de alpacas y llamas: nueva evidencia cromosómica y molecular, Rev. Chil. Hist. Nat. 80 (2) (2007), https://doi.org/10.4067/S0716-078X2007000200001. Jun.
[27] B.A. González, A.M. Agapito, F. Novoa-Muñoz, J. Vianna, W.E. Johnson, J.C. Marín, Utility of genetic variation in coat color genes to distinguish wild, domestic and hybrid South American camelids for forensic and judicial applications, Forensic Sci Int Genet 45 (2020), 102226, https://doi.org/10.1016/j.fsigen.2019.102226. Mar.
[28] R. Fan, et al., Genomic analysis of the domestication and post-Spanish conquest evolution of the llama and alpaca, Genome Biol. 21 (1) (2020) 159, https://doi.org/10.1186/s13059-020-02080-6. Dec.
[29] C. Azémard, et al., Untangling the fibre ball: proteomic characterization of South American camelid hair fibres by untargeted multivariate analysis and molecular networking, J. Proteonomics 231 (2021), 104040, https://doi.org/10.1016/j.jprot.2020.104040. Jan.
[30] H.H. Bragulla, D.G. Homberger, Structure and functions of keratin proteins in simple, stratified, keratinized and cornified epithelia, J. Anat. 214 (4) (2009) 516–559, https://doi.org/10.1111/j.1469-7580.2009.01066.x. Apr.
[31] J.C. Crossley, C.G. Borroni, A.S. Raggi, Correlation between mean fibre diameter and total follicle density in alpacas of differing age and colour in the Parinacota province of the Chilean high plain, J. Appl. Anim. Res. 42 (1) (2014) 27–31, https://doi.org/10.1080/09712119.2013.795899. Jan.
[32] K.A. Jakes, S. Shim, A. Thompson, A pilot study of the effects of diet on Huacaya and Suri alpaca fibre, J. Camel Pract. Res. 13 (2006) 185–192. Mar.
[33] B.A. McGregor, Comparative productivity and grazing behaviour of Huacaya alpacas and Peppin Merino sheep grazed on annual pastures, Small Rumin. Res. 44 (3) (2002) 219–232, https://doi.org/10.1016/S0921-4488(02)00050-0. Jun.
[34] X. Lv, et al., Effect of nutritional restriction on the hair follicles development and skin transcriptome of Chinese merino sheep, Animals 10 (6) (2020) 1058, https://doi.org/10.3390/ani10061058. Jun.