Microchemical Journal 206 (2024) 111417

Contents lists available at ScienceDirect

Microchemical Journal

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

Serum species identification using mid-infrared and far-infrared spectroscopy combined with neural network algorithms

Xinghao Huang [a,1], Qiliang Wang [a,1], Mao Mao [b,1], Ningyi Wang [a], Jiamin Zhang [a], Xu Wu [a], Xueting Dai [b], Zhengan Tian [b,*], Yan Peng [a,*]

[a] Terahertz Technology Innovation Research Institute, Terahertz Spectrum and Imaging Technology Cooperative Innovation Center, Shanghai Key Lab of Modern Optical System, University of Shanghai for Science and Technology, Shanghai 200093, China
[b] Shanghai International Travel Health Care Center (Shanghai Customs Port Clinic), Shanghai, China

ARTICLE INFO

Keywords:
Serum species identification
Infrared spectroscopy
Convolutional Neural Network
Feature correlation analysis

ABSTRACT

Serum is a widely used biological fluid containing rich biological information, commonly employed in clinical diagnosis and medical treatment. However, since human serum contains genetic information of races and involves national information security, accurate identification of species serum is required for customs import and export trade. In this study, we propose a species serum identification method combining mid-infrared and far-infrared spectroscopy with neural network algorithms. By collecting spectra of 147 serum samples from 8 species and conducting preliminary analysis on specific spectral ranges, differences between spectra of certain species were identified. Subsequently, a Convolutional Neural Network (CNN) model was constructed for species identification and classification, achieving recognition accuracy of 95.00 % for human and non-human serums (binary classification) and 8 species of serums (octal classification). Through correlation analysis, the corresponding relationships between species serums and their spectral characteristic regions, as well as relevant molecular groups and chemical bonds, were clarified. This reveals the underlying mechanism for species recognition based on serum spectra. Furthermore, by nesting multiple segmented progressive sub-models and optimizing the weight ratios of spectral features, the accuracy of species recognition was improved to 98.95 % while the computational complexity was greatly reduced. These results provide a reliable and efficient method for rapid differentiation of species serums, with important implications for the identification of other biological fluids.

1. Introduction

Blood is a common and informative form of biological evidence. It plays a crucial role in analytical chemistry [1], forensic analysis [2–6], customs quarantine [7–9] and wildlife conservation [10,11]. Serum is obtained after centrifugation of whole blood, and it contains a wealth of biomolecules (lipids, proteins, carbohydrates, nucleic acids) [12] and important biological information for donors [13,14]. It can be used to diagnose diseases by detecting changes in biomarkers in the serum. For example, in 2023, the study by Zhenning Jin et al. showed that both phosphocreatine and cyclic guanosine monophosphate may serve as predictive biomarkers of coronary heart disease in patients with type 2 diabetes mellitus [15]. Therefore, serum is widely used in disease diagnosis. However, it contains proteins, hormones, and other substances that carry racial genetic information and are involved in national information security. Consequently, customs are rigorous in discriminating serum between humans and nonhumans, as well as between different species during import and export. Currently, conventional methods used for blood discrimination between species include Mass Spectrometry (MS) [16] and High-Performance Liquid Chromatography (HPLC) [17]. In 2013, Heyi Yang et al. demonstrated that MS can be used for body fluid identification [16], which used Matrix-assisted laser desorption/ionization (MALDI) mass spectrometry (MS) to identify various biomarkers present in blood (hemoglobin), saliva (amylase), semen (semenogelin). In 1990, Inoue et al. used reverse-phase HPLC to identify fresh blood from humans and 28 animal species by analysing characteristic chromatograms and Heme peaks [17]. Due to its high resolution and sensitivity, HPLC has been applied in

* Corresponding author.
E-mail addresses: tianzhenganciq@163.com (Z. Tian), py@usst.edu.cn (Y. Peng).
[1] Equal contributions.

https://doi.org/10.1016/j.microc.2024.111417
Received 9 May 2024; Received in revised form 12 July 2024; Accepted 12 August 2024
Available online 20 August 2024
0026-265X/© 2024 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

blood analysis and proven to be a reliable method. However, both methods require the addition of organic solvents such as acetonitrile before testing, which can be sample-destructive, time-consuming and expensive [18]. Therefore, customs are still seeking a rapid, efficient and non-destructive detection method to accurately discriminate between different species of blood.

Vibrational spectroscopy is a rapid, non-destructive method for the identification of vibrations and rotational frequency resonance absorption of chemical bonds, groups, etc. Therefore, this technique enables qualitative and quantitative analysis of trace samples [19,20]. Infrared spectroscopy has been widely used as an analytical method in vibrational spectroscopy. [21–24]. It can provide information on chemical composition of samples and distinguishing different samples quantitatively and qualitatively [25]. In biological sample research, it has been successfully applied in disease diagnosis (such as cancer) [26,27] and identification of body fluids (such as blood and saliva) [28]. In 2008, De Wael et al. differentiated human and animal blood by Raman and mid-infrared spectroscopy, but it was not feasible to differentiate blood spectra from different species visually [29]. In subsequent studies, researchers differentiated blood vibrational spectra through advanced statistical analysis. In 2009, Virkler et al. successfully differentiated blood samples from humans, cat, and dog using Raman spectroscopy data and applying Principal Component Analysis (PCA) model. Separation between the species within a PCA model surpassed a 99 % confidence interval [30]. In 2019, Shan Huang et al. established a 1D Convolutional Neural Network (CNN) model to further expand the number of animal categories. The accuracy rate reached 97.33 % for distinguishing the blood Raman spectra of 20 animal species [31]. In 2015, Mistek-Morabito et al. successfully differentiated humans, cat, and dog using Partial Least Squares Discriminant Analysis (PLS-DA) with a 100 % accuracy rate [3]. Furthermore, in 2020, they expanded the number of blood species to 12 with 99.6 % accuracy [32].

Currently, there are a few studies on serum discrimination. For example, in 2024, Yuchen et al. used a novel probe C1 combining benzothiazole and spiropyran fragments for the specific detection of human serum albumin (HSA), achieving fluorescence differentiation between HSA and bovine serum albumin(BSA) [33]. In 2017, Fan, Q. et al. invented a method to discriminate male and female animal serum samples of the same species of mammals by near-infrared spectroscopy [34]. However, current studies on serum species identification are not comprehensive, covering only a limited number of species, and the test methods are complex. What's more, as the import and export of serum products in customs accounts for more than 57 % of the total blood products, distinguishing serum is crucial. In this study, we utilized a Convolutional Neural Network (CNN) model combined with mid-infrared and far-infrared spectroscopy for species identification based on serum samples. Initially, mid-infrared and far-infrared spectra (4000–30 cm^{-1}) of serum from different species were collected using a Fourier Transform Infrared Spectrometer (FT-IR). Subsequently, preliminary spectral analysis was conducted to identify partial differentiation criteria between species. Then a CNN model with convolutional layers, pooling layers, and fully connected layers was constructed and trained. In the data set, there are poultry and livestock, which ensuring the applicability, robustness and specificity of the CNN model. Afterwards, by nesting multiple segmented progressive sub-models and optimizing the weight ratios of spectral features, the accuracy was further improved. These findings demonstrate that this method provides a reliable and efficient approach for species serum differentiation.

2. Materials and methods

2.1. Serum samples

In this study, the serum samples used were provided by Technical Center for Animal Plant and Food Inspection and Quarantine, Shanghai Customs. The serum samples used in the experiments have been approved by the relevant ethical committees and detailed documentations are in the Supplementary material. And these experiments were conducted in accordance with established ethical guidelines and informed consent was obtained from the patients (or relatives/guardians) in compliance with all regulations. In total, 24 human serum samples and 123 animal serum samples were used for differentiating between human and animal blood. The animal species included horse, pig, goat, cow, chicken, cat, and dog. Within these species, there were serum samples from 12, 36, 4, 17, 18, 18, and 18 different individuals, respectively. All sample sources were licensed and met safety quarantine standards. To maintain sample integrity, all samples were stored frozen at −20°C.

2.2. Experimental equipment

The experiment utilized a VERTEX 70v FT-IR, capable of collecting a full spectrum from 6000 cm^{-1} to 30 cm^{-1} in the mid to far-infrared range. The signal to noise ratio (SNR) was larger than 50000:1, the wave number accuracy is up to 0.005 cm^{-1}, and the resolution is better than 0.4 cm^{-1}. Spectra were collected in a vacuum environment to minimise the effect of water vapour on the experiments. Samples were vortexed and dried using a vacuum centrifuge concentrator (Eppendorf Concentrator plus).

2.3. Experimental methods

In the experiment, each serum sample was defrosted, then vortexed for 5 min using a vacuum centrifugal concentrator to ensure sample uniformity. Subsequently, 10 μL of the sample solution was deposited onto a high-resistance silicon wafer evenly by using a pipette. The sample was then dried using the vacuum centrifuge concentrator to prevent interference from infrared spectral peaks caused by water [35]. After the serum sample evaporated, a uniform thin layer formed on the silicon wafer. The silicon wafer with the sample film was placed into the VERTEX 70v to collect spectral data. Spectra were collected in the range of 4000–30 cm^{-1}, with a resolution of 4 cm^{-1}. 5 spectra were tested for each species sample, with a total of 735 spectral data. And 64 scans were performed for each spectrum.

2.4. Data pre-processing

All data processing methods were implemented based on data analysis algorithms built with python. Due to noise in the sidebands, the spectral range of 4000–100 cm^{-1} was used. Baseline correction was applied to all collected spectra, and normalized to the 0–1 range based on the extreme value normalization method. The five spectra obtained from repeated testing of each donor were averaged to generate an average spectrum and error bar for each sample.

2.5. Analytical method

2.5.1. Multilinear de-baseline

In order to solve the problem of the presence of baselines in the collected spectral data affecting the subsequent detection, we constructed a python based de-baseline program. Due to the special characteristics of biological samples' spectra, the baseline is not fundamental linear, so the linear baseline cannot remove all external influences well at some positions, and considering the existence of broad absorption peaks near 3700–4700 cm^{-1}, polynomial fitting of the baseline will lead to the reduction of the peak heights here and thus affect the subsequent detection. Therefore, we propose a multilinear baseline removal method, which generates a multilinear baseline by generating a linear baseline between the spectral data points corresponding to the determined frequency points and then combining them to form a folded baseline, and then subtracting the baseline from the spectral data through a program to obtain the spectral data after the removal of the

baseline. This method avoids the insufficient or excessive changes to the absorption peak waveforms caused by ordinary linear baselines and polynomial fitting baselines, and can remove baselines well without causing loss of spectral information.

2.5.2. Extreme value normalization

In order to address the problem of different magnitudes and units between features, which may lead to certain features dominating the model training process, we employed the min–max normalization technique. This normalization method scales all features to a fixed range, typically [0, 1] or [−1, 1]. This not only removes the difference in magnitude between features, but also helps to learn the relationships between features in a more balanced and efficient way.

Extreme value normalization involves subtracting the minimum value from data points and then dividing by the data range (maximum value minus minimum value). For an input data matrix M, min–max normalization can be represented as follows:

$$M = \frac{M - M_{\min}}{M_{\max} - M_{\min}} \tag{1}$$

where Mmin and Mmax represent the minimum and maximum values of the data matrix M respectively.

The normalization operation performed here effectively avoids the bias caused by the quantitative differences in the different features of the data during the training of the neural network. This normalization step helps to ensure that each attribute of the data is scaled to a similar range, helping the model to better learn the correlation between features and improve model performance and generalization.

2.5.3. Moving average filter

Moving average filtering is based on the statistical law, the continuous sampling data is viewed as a queue with a fixed length of N. After a new measurement, the first data of the above queue is removed, the rest of the N-1 data are sequentially shifted forward and new sampling data is inserted as the tail of the new queue; then arithmetic operations are performed on this queue and the result is done as the result of this measurement. For an N-point queue, suppose the input is x and the output is y its calculation is as follows:

After the sliding average filtering can be very good to reduce the high-frequency component in the spectral data, effectively reducing the additional high-frequency noise on the spectral curve due to the environmental noise during the detection process.

$$y(n) = \frac{1}{N} \sum_{k=n-N+1}^{n} x(k) \tag{2}$$

After the sliding average filtering can be very good to reduce the high-frequency component in the spectral data, effectively reducing the additional high-frequency noise on the spectral curve due to the environmental noise during the detection process.

2.5.4. Convolutional Neural Network

Convolutional Neural Network (CNN) has excellent performing in various fields such as data classification, image recognition and natural language processing, making them an indispensable tool in data analysis [36–38]. A key feature of CNN is their ability to learn and extract features from input data [39]. When applied to a large amount of spectral data, CNN can effectively discern feature differences between different categories, making them suitable for serum spectral classification. Given that the spectral data of samples are a set of one-dimensional data containing amplitude, frequency, and other information, we use a nested approach based on a 1-dimensional CNN architecture with multiple sub-models in a segmented progression to classify the serum types, taking into account the amount of computation.

3. Results and discussion

3.1. Spectroscopy analysis

The serum spectra of 8 different species are shown in Fig. 1. Each spectra in Fig. 1 represents an individual within a species. It can be observed that they are very similar to each other, with main peaks appearing at the same frequencies. Absorption peaks are present in the lipid (3000–2800 cm^{-1}), protein (1700–1500 cm^{-1}), nucleic acid (1250–1000 cm^{-1}), and carbohydrate (1000–800 cm^{-1}) regions [40]. The representative band assignments of molecular vibrations for the spectra of serum are contained in Table 1. These results are due to the similarity in serum composition among species, but the concentration of components varies. For example, the differences between human, cat, and dog serum are mainly attributed to concentrations of glucose, ascorbic acid, enzymes and hormone [3].

In the partially enlarged image shown in Fig. 1, spectral differences between species are evident. Peaks are observed around 1735 cm^{-1} (the shoulder of the Amide I band [40]) in chicken, cat, human, and dog spectra, while absent in horse, pig, cow, and goat spectra. Thus, based on the presence or absence of this peak, the spectra of these 8 species can be preliminarily classified into two major categories: Class 1 [Fig. 1(a)-Fig. 1(d)] and Class 2 [Fig. 1(e)-Fig. 1(h)].

In the spectra of Class 1 [Fig. 2(a)-Fig. 2(d)], analysis was conducted by extracting the 1185–1069 and 956–820 cm^{-1} ranges. The relative peak heights and errors were obtained by averaging the spectra of all samples within the same species. The relative peak heights at 1081 cm^{-1} (glucose C-O symmetric stretching [44]) and 1120 cm^{-1} (C-N symmetric stretching [44]) for horse, pig, cow, and goat were 1.808 ± 0.027, 1.473 ± 0.021, 1.449 ± 0.067, and 1.147 ± 0.073, respectively. Subsequently, we obtained Fig. 2(e) based on the above values. From this figure, it can be concluded that horse and goat are distinguishable from pig and cow, and that horse and goat are distinguishable from each other. Further analysis of the relative peak heights of pig and cow at 931 and 832 cm^{-1} revealed values of 0.276 ± 0.004 and 0.420 ± 0.008. Then Fig. 2(f) indicates a clear distinction in the relative peak height range of pig and cow in 956–820 cm^{-1} frequency range. Therefore, we can effectively distinguish four species in Class 1 combining the 1185–1069 and 956–820 cm^{-1} frequency ranges.

In the spectra of Class 2 [Fig. 3(a) Fig. 3(d)], analysis was conducted by extracting the 1190–1010 and 250–141 cm^{-1} ranges. It was observed that there were significant differences between chicken and other species within the two extracted ranges. The difference in the 1190–1010 cm^{-1} range is most likely due to the higher glucose concentration in chicken serum (averaging 170 mg/100 mL) compared to the other three species (averaging not more than 100 mg/100 mL) [3]. After further analysis, the relative peak heights near 1168 cm^{-1} (C-O vibration [43]) and 1080 cm^{-1} (glucose C-O symmetric stretching [44]) for cat, human, and dog were 1.262 ± 0.042, 0.782 ± 0.028, and 1.151 ± 0.034, respectively. Fig. 3(e) indicates that humans can be distinguished from cat and dog. Further analysis of the relative peak heights of cat and dog in 213–164 cm^{-1} range showed values of 1.160 ± 0.014 and 1.048 ± 0.016, respectively. According to Fig. 3(f), there is a significant difference between the relative peak heights of cat and dog in the range, and thus they can be distinguished from each other. Therefore, we can effectively distinguish the four species in Class 2 by combining the 1190–1010 and 250–141 cm^{-1} frequency ranges.

The above analysis provided a preliminary discussion based on the amplitude differences of relative peak heights. However, considering the possibility of extreme values in samples due to individual heterogeneity, we also introduced a Convolutional Neural Network (CNN) model for further analysis.

3.2. Algorithmic identification of species serum

Based on the above spectral differences between different species, we

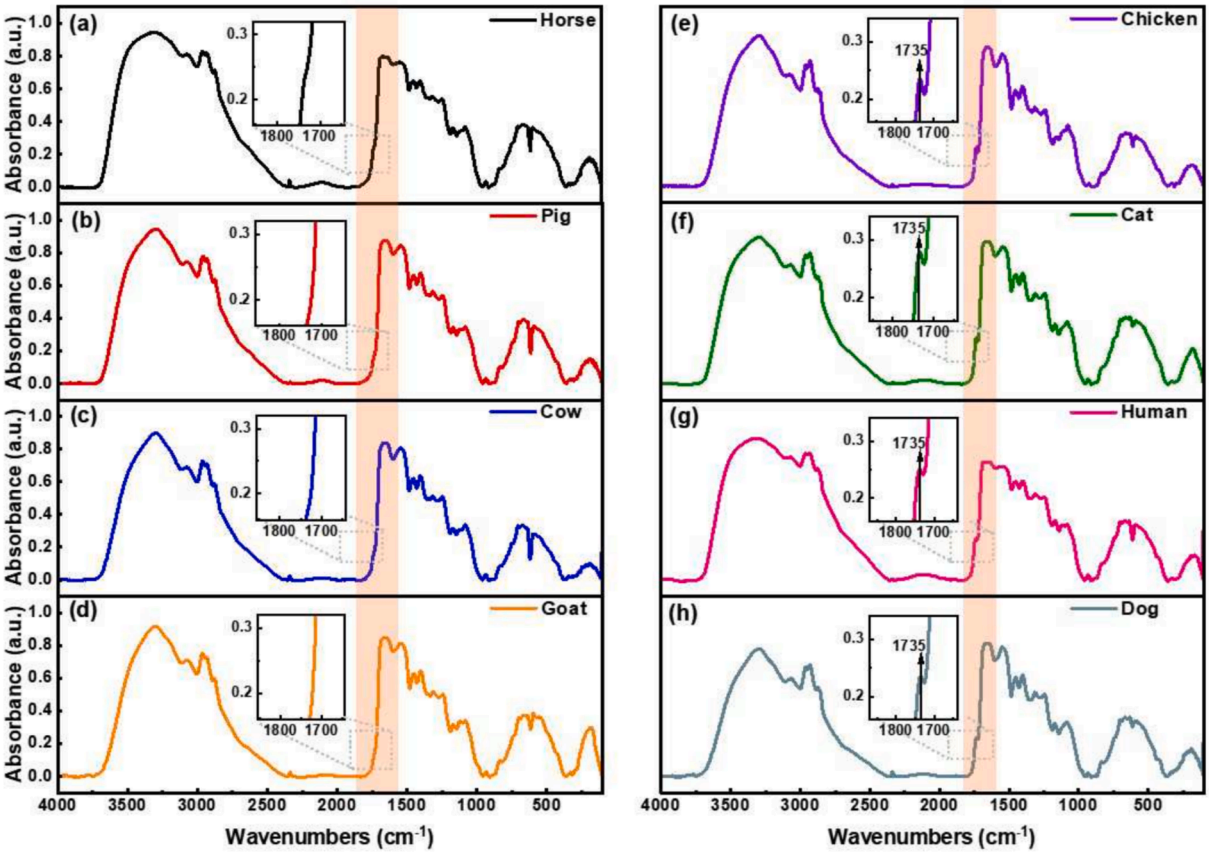

Fig. 1. Mid-infrared and far-infrared spectra of serum from different species. The spectra correspond to the following species: (a) Horse, (b) Pig, (c) Cow, (d) Goat, (e) Chicken, (f) Cat, (g) Human, (h) Dog. The spectra are zoomed in the range of 1855–1645 cm^{-1}, and based on the presence or absence of absorption peak at 1735 cm^{-1}, the species are classified into Class 1 ((a)–(d)) and Class 2 ((e)–(h)).

Table 1
Assignment of the infrared bands in human blood.

Wavenumbers (cm^{-1})	Assignment
3500–3200	Water and hydroxyl (O–H stretching) [40,41]
2959	Lipids (asymmetric stretching of CH_3) [42–44]
2931	Lipids (asymmetric stretching of CH_2) [40]
2872	Lipids (symmetric stretching of CH_3) [42]
1700–1600	Amide I (C═O stretching) [40,41]
1560–1500	Amide II (N–H bending, C–N stretching) [41,45,46]
1239	Amide III (C–N stretching) [41,47]
1082	Glucose (C–O stretching) [44]
699	Amide IV (C–H bending) [46]

try to use artificial neural network models to extract and enhance the relevant information in the spectra for training in order to further improve the speed and accuracy of recognition.

3.2.1. CNN algorithm model

In this section we use a traditional convolutional neural network (CNN) model structure consisting of two convolutional layers, two pooling layers, and one fully connected layer, as shown in the structure in Fig. 4(a). The convolutional layers utilize kernels to perform element-wise multiplication and summation on the input data. This process aims to extract local features and capture spatial relationships and specific patterns [48,49]. Given that our input data is one-dimensional spectra, we designed both convolutional layers according to a 1D CNN architecture. These layers use a convolutional kernel size of 3 and a stride of 1. To reduce data dimensionality and the number of parameters while preserving important features, we opted for max pooling layers. The parameters of these two max pooling layers are normalized with a kernel size and stride of 2.

The fully connected layer enables the neural network to understand the relationships and weights between the features. In the output layer, we employ the SoftMax activation function to map the final features to their corresponding output categories. We chose the Adaptive Moment Estimation optimizer (also known as the Adam optimizer) as the model's optimization method. The Adam optimizer, with its adaptive learning rate and dynamic momentum, effectively accelerates model convergence and yields superior results, thereby reducing the need for hyperparameter tuning.

The above model was trained under the conditions of 200 epoch, batch-size of 64 per epoch, and learning-rate of 0.001. 19 groups of 95 sample spectra were randomly selected from the above 147 groups of serum sample spectra as the validation set for verification, and the rest were used as the training set. Considering that the sample spectra testing conditions are not ideal, there may be small differences in the same set of samples, which can be regarded as the differences between different individuals of the same species, the validation set can be expanded from 19 groups to 95 groups to improve the accuracy. The final classification results are displayed in Fig. 4(a) and (c). The model's direct classification accuracy reached 95 %, with a few cat samples misclassified as humans. Upon analysis, this may be due to an excess of redundant information in the spectra, which could submerge the potential subtle differences between human and cat blood. Therefore, in the following, we performed a correlation analysis of the sample spectra to reduce the weight of the redundant information, and we modified the architecture of the neural network model and proposed a nested method of constructing multiple sub-models with segmental progression to increase the accuracy and reduce the amount of computation.

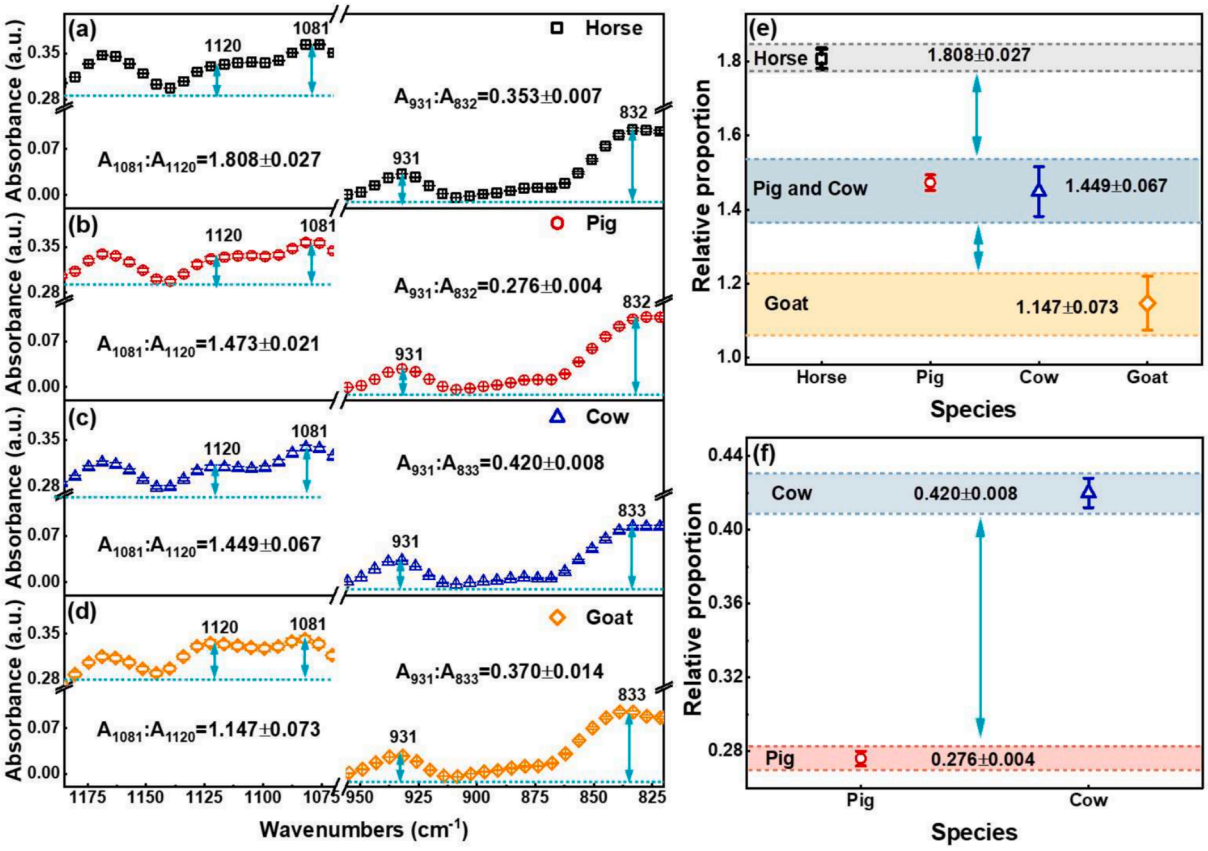

Fig. 2. The spectra of species in Class 1 and the relative peak height ranges corresponding to two frequency ranges. The spectral range of Figures (a)–(d): 1185–1069 and 956–820 cm^{-1}. Figure (e) shows the relative peak height range in the 1185–1069 cm^{-1} frequency range, and Figure (f) shows the relative peak height range in the 956–820 cm^{-1} frequency range.

3.2.2. Feature correlation analysis

The observed feature peaks in the range of 4000–100 cm^{-1} contain equivalent information content, such as frequency, amplitude, relative ratio, peak area, and half-peak full width. However, not all absorption peaks are effective for neural networks to learn spectral features. The introduction of excessive nonessential spectral information not only makes the model bulky and costly to train, but also reduces the percentage of effective information due to too much redundant information, thereby affecting the model's accuracy.

To address this, we integrated a variance-based feature correlation analysis, which assigns higher weights to features with higher relevance and lower weights to useless information. The method aims to analyze the correlation between each spectral feature of the model and the category it belongs to, the specific implementation method is to calculate the correlation between each column of the spectral features and the category it belongs to respectively. The higher the correlation, the more important the feature is for the neural network to learn the spectral features. We constructed a correlation calculation program based on python, and referenced the f_classif function in the sklearn package to calculate the ANOVA f-value. The larger the f-value, the stronger the correlation between the corresponding feature and the result. And we quantified the number of different species in the classification result, and traversed the correlation between the quantized value of each feature to obtain the correlation between each band and the quantized value. In that way, we obtained the correlation between each feature and the quantized value to obtain the importance of each waveband for the result of the identification of the serum importance of each band to the identification results.

In Fig. 5, we plotted the correlation between all the columns and the categories they belong to. A breakpoint in the range of 2400–2000 cm^{-1} is added to ignore the effect of carbon dioxide (2349 cm^{-1}) on the correlation analysis. Comparison of the experimental spectra reveals high correlation features and obvious absorption peaks near 1739 cm^{-1}, 1024 cm^{-1}, 942 cm^{-1}, and 182 cm^{-1}.

The peak near 1739 cm^{-1} can be considered as the shoulder peak of amide I, which is essentially caused by the absorption of the stretching vibration of the C═O bond in the peptide chain constituting the backbone of the protein structure [40]. The absorption in the bands from 1750 cm^{-1} to 1600 cm^{-1} are all generated by amide I. The ratio of the absorption intensities at different frequency points reflects the ratio of the different secondary structures of proteins present in the samples. Because the protein structure in serum of different species must be different, and the proportion of the secondary structure of proteins is also different, so the shapes of the absorption peaks have differences between different kinds of samples. We found that the infrared spectra of the eight animal serum samples in this study have obvious differences at 1739 cm^{-1}, which can be used as a basis for classification. 1024 cm^{-1} is the absorption produced by amino acids [46]. Similar to the analysis of protein structure, there are differences in the types and amounts of amino acids in the serum of different species. Each amino acid corresponds to one or more frequency points, and the intensity of the absorption reflects the content of that amino acid, which in turn leads to differences in the absorption spectra. 942 cm^{-1} is the absorption due to vibrations outside the O-H bond. This absorption can be present in a wide range of molecules, but it is certain that the intensity of the absorption peaks here varies from species to species, so this can be used as a basis for classification. The far infrared absorption near 182 cm^{-1} is caused by resonance absorption due to molecular vibrations and rotations, as well as lattice absorption. The specific molecules that cause the absorption peaks here have not been reported, but based on Fig. 3, we

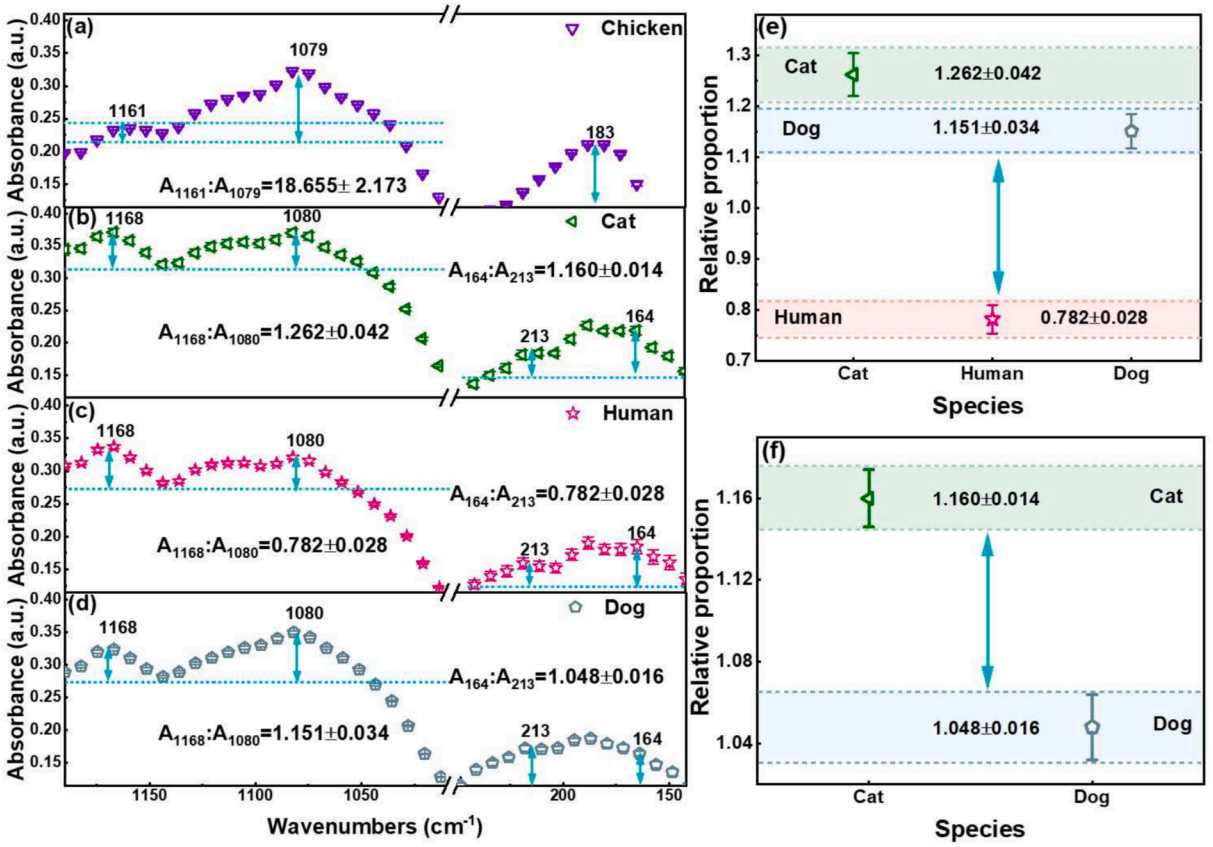

Fig. 3. The spectra of species in Class 2 and the relative peak height ranges corresponding to two frequency ranges. The spectral range of Figures (a)–(d): 1190–1010 and 250–141 cm^{-1}. Figure (e) shows the relative peak height range in the 1190–1010 cm^{-1} frequency range, and Figure (f) shows the relative peak height range in the 250–141 cm^{-1} frequency range.

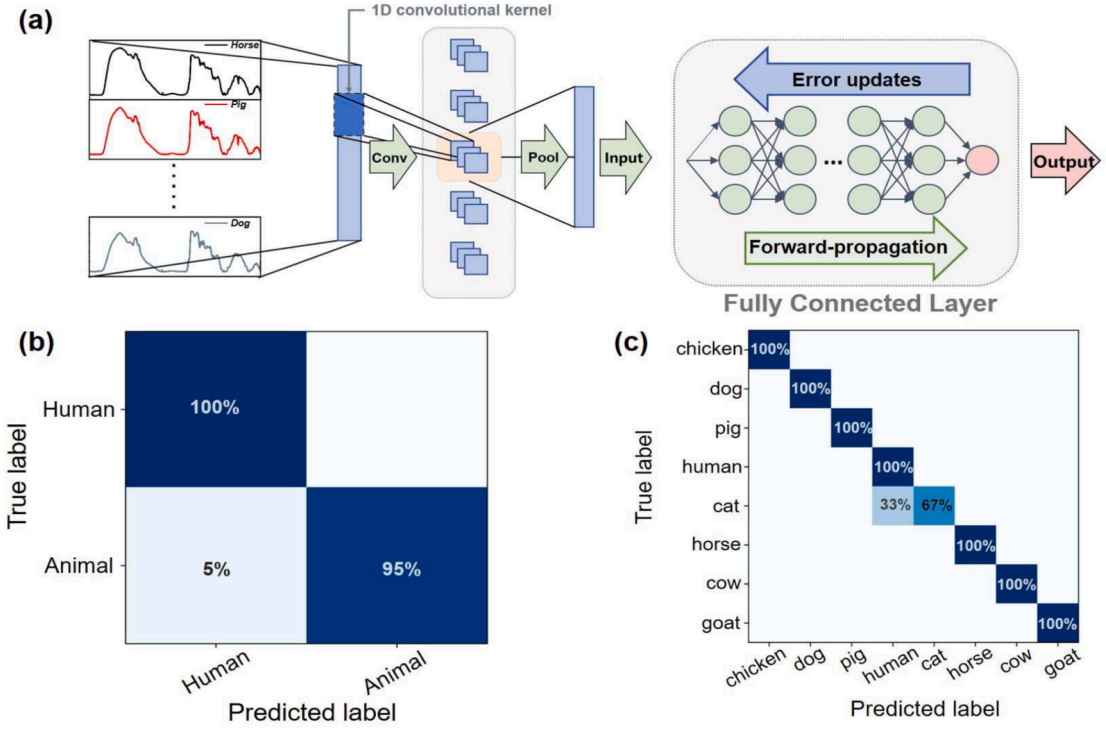

Fig. 4. (a) Schematic of the CNN network model structure. (b) Confusion matrix of human and non-human binary classification results. (c) Confusion matrix of all-species eight classification results.

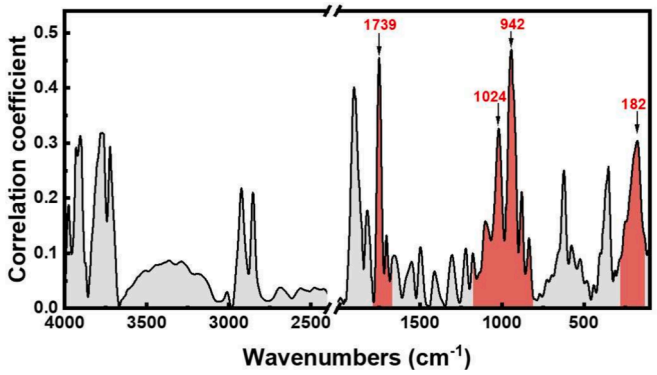

Fig. 5. Correlation analysis plot, with bands of lower importance or with no obvious absorption peaks shown in grey, and bands of higher importance for which corresponding absorption peaks can be found shown in red.

can see that there is a significant difference between chicken, cat, and dog in this band. Therefore, it can be used as a basis for distinguishing chicken from other species.

The above analysis is the potential deep mechanism that different species can be distinguished, and we improved the algorithm based on the above analysis. We focused on analysing the bands in the spectrum that have strong correlation with the classification results, which further improves the correct classification rate.

3.2.3. *Structure of the algorithm after feature correlation analysis*

From spectral and correlation analyses, we discerned that the feature recognition information of various serum species may distribute in different bands. Consequently, we employed a nested approach to construct multiple sub-models in a segmented progression. The original CNN model was divided into three sub-models, each focusing on recognizing different bands. We refer to the improved algorithm as the CNNs algorithm model. Fig. 6(a) schematically illustrates the structure of the nested CNNs algorithm model.

The first sub-model classifies the eight species into two main classes based on the feature peaks at 1735 cm^{-1}. These are then further divided by two sub-models. Sub-model A identifies the characteristic peaks at 1801–1701 cm^{-1} to separate the two main classes. Sub-model B identifies the 1201–900 cm^{-1} band, subdividing the first main category into four. Sub-model C identifies the 250–150 cm^{-1} and 1201–1000 cm^{-1} bands, dividing the second main category into four. Based on the classification results of sub-model A, it is decided whether sub-model B or sub-model C will further classify the samples, ultimately realizing the eight classifications of all species. This strategy of segmented progressive classification by constructing multiple sub-models can effectively enhance the accuracy rate. Moreover, the computation amount and the time required to train the CNN model are significantly reduced as a single sub-model only needs to recognize specific bands.

Given the different recognition bands of different models and different classification tasks, we trained the three sub-models with different hyper-parameters (using different Epoch, Batch Size, and

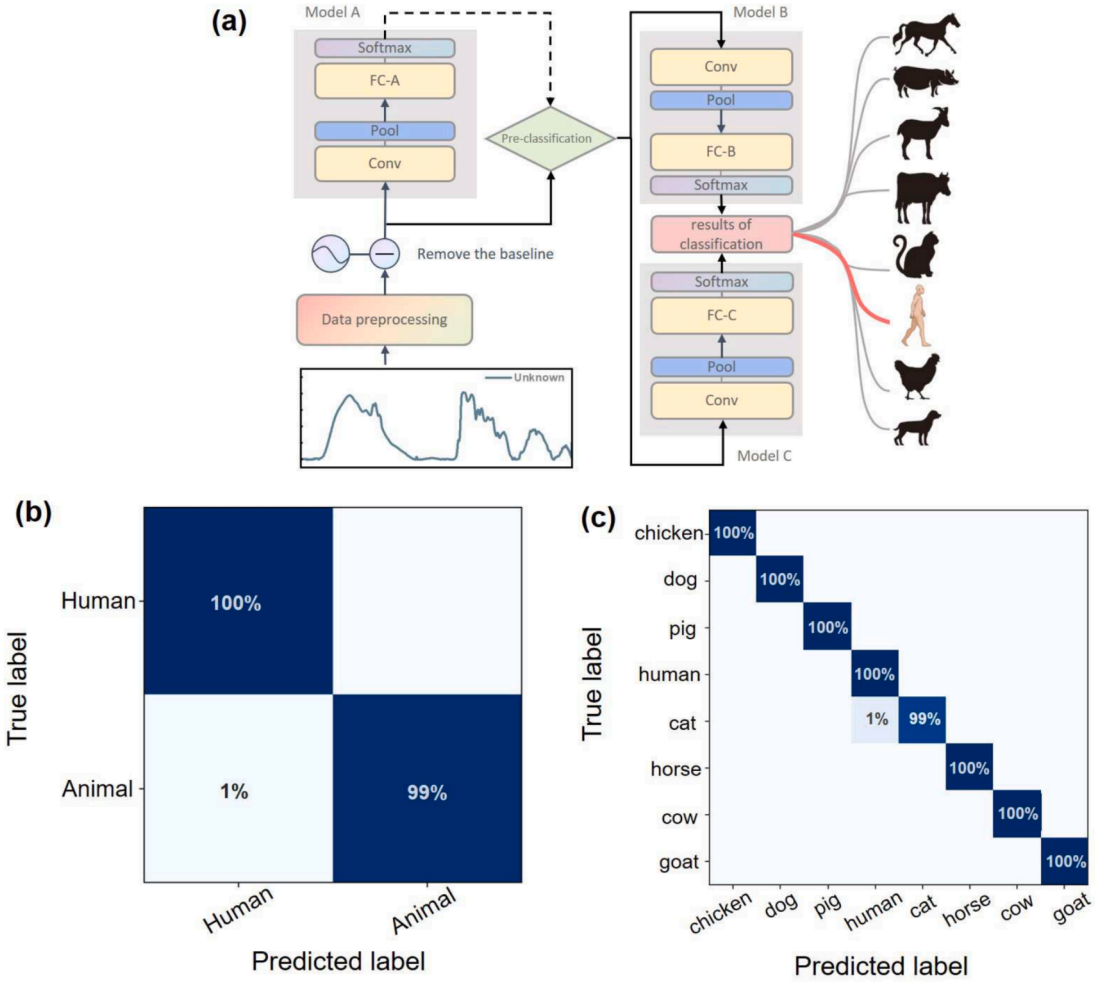

Fig. 6. (a) Schematic model of CNNs algorithm. (b) Confusion matrix of human and non-human binary classification results. (c) Confusion matrix of all-species eight classification results.

Learning Rate as shown in Table 2). After several attempts, we determined the training parameters with better performance as shown in Table 2.

3.2.4. Binary classification of human and non-human serum samples

The experimentally measured spectral data was categorized into two groups: human and non-human. The dataset comprised a total of 735 pieces of spectral data. Of these, 640 were utilized as a training set to train the aforementioned CNNs algorithmic model, while the remaining 95 were used as a validation set for external validation.

The model's binary classification performance on the validation dataset is depicted in Fig. 6(b). Out of the 95 validation set samples, one sample, which was actually from a non-human, was incorrectly classified as a human serum. The accuracy is 98.94 %. The overall result meets the practical requirements for distinguishing between human and non-human serum. Following this, we extended the model's application to the eightfold classification of human and animal serum.

3.2.5. Octal classification of human and animal serum samples

The spectral data, based on the sample sources, are classified into eight categories: human, cat, dog, chicken, pig, goat, cow, and horse. We trained a CNNs eight-classification recognition model using these spectral data. The division of the training and validation sets was performed in the same manner as the previously mentioned binary classification. Fig. 6(c) presents the eight-classification results of the CNNs model for the spectral validation dataset. Out of the 95 samples, one was misclassified, resulting in an accuracy of 98.94 %. It's noteworthy that the accuracy of the eight-classification model did not decline compared to the binary classification model. This suggests that our testing method can classify the serum spectra of the remaining species, with the exception of potential confusion between human and cat blood. Given the relative similarity in substance content and concentration between human and cat serum, and the existence of individual differences, the spectra of human and cat serum did not achieve 100 % classification. Future work could involve expanding the test bands and identifying new differentiation points for discrimination.

4. Conclusion

In this study, we introduce a method for identifying serum from different species by integrating mid- and far-infrared spectroscopy with neural network algorithms. We gathered spectral data from 147 serum samples from horse, pig, goat, cow, chicken, cat, dog, and human. By selecting specific intervals, we were able to identify some differences in the spectra between species for preliminary analysis. Subsequently, we built an artificial neural network model to train the dataset. This model achieved recognition accuracies of 95.00 % for both human and non-human serum (binary classification) and serum from eight species (octal classification). We further analysed the correlation between the species' serum and their spectral features in the model. Additional information, such as frequency, amplitude, relative peak height, peak area, and half-peak full width, was extracted from the spectra. We also clarified the molecular groups and chemical bonds corresponding to some of the spectral feature regions with high correlation. By constructing a nested approach with multiple sub-models in a segmented progression, we optimized the weight ratio between spectral features. This increased the accuracy of species identification to 98.94 %, while significantly reducing the computational volume. This study enables rapid, non-destructive, and accurate recognition and differentiation of serum between different species. It provides a valuable reference for studies aiming to identify other species of body fluids.

CRediT authorship contribution statement

Xinghao Huang: Writing – original draft, Investigation, Formal analysis. **Qiliang Wang:** Software, Writing – original draft, Formal-analysis. **Mao Mao:** Resources. **Ningyi Wang:** Investigation. **Jiamin Zhang:** Investigation. **Xu Wu:** Methodology. **Xueting Dai:** Resources. **Zhengan Tian:** Funding acquisition, Project administration. **Yan Peng:** Funding acquisition, Project administration, Supervision, Writing – review & editing.

Table 2
Training hyperparameters for three different models.

	Epoch	Batch Size	Learning Rate
Model A	40	48	0.0001
Model B	120	32	0.0001
Model C	300	48	0.0001

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

This work was supported in part by the National Key Research and Development Program of China (2022YFA1404004); National Natural Science Foundation of China (61988102, 62335012); General Administration of Customs Project (2023HK066).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.microc.2024.111417.

References

[1] G. Mclaughlin, K.C. Doty, I.K. Lednev, Raman spectroscopy of blood for species identification, Anal. Chem. (2014).
[2] G. McLaughlin, K.C. Doty, I.K. Lednev, Discrimination of human and animal blood traces via raman spectroscopy, For. Sci. Int. 238 (2014) 91–95.
[3] E. Mistek, I.K. Lednev, Identification of species' blood by attenuated total reflection (atr) fourier transform infrared (ft-ir) spectroscopy, Anal. Bioanal. Chem. 407 (2015) 7435–7442.
[4] P. Bai, J. Wang, H. Yin, Y. Tian, W. Yao, J. Gao, Discrimination of human and nonhuman blood by raman spectroscopy and partial least squares discriminant analysis, Anal. Lett. 50 (2017) 379–388.
[5] J. Fujihara, Y. Fujita, T. Yamamoto, N. Nishimoto, K. Kimura-Kataoka, S. Kurata, Y. Takinami, T. Yasuda, H. Takeshita, Blood identification and discrimination between human and nonhuman blood using portable raman spectroscopy, Int. J. Legal Med. 131 (2017) 319–322.
[6] T. Das, A. Harshey, K. Nigam, V.K. Yadav, A. Srivastava, Analytical approaches for bloodstain aging by vibrational spectroscopy: current trends and future perspectives, Microchem. J. 158 (2020) 105278.
[7] S.C. Renner, D. Neumann, M. Burkart, U. Feit, P. Giere, A. Groeger, A. Paulsch, C. Paulsch, M. Sterz, K. Vohland, Import and export of biological samples from tropical countries-considerations and guidelines for research teams, Organ. Divers. Evol. 12 (2012) 81–98.
[8] W. Xiong, W. Jian, L. Peng-xi, Z. Ting-Ting, Identification of animal whole blood based on near infrared transmission spectroscopy, Spectrosc. Spectral Anal. 36 (2016) 80–83.
[9] Y. Liu, Z. Wang, Z. Zhou, T. Xiong, Analysis and comparison of machine learning methods for blood identification using single-cell laser tweezer raman spectroscopy, Spectrochim. Acta Part A 277 (2022).
[10] L. Zhang, S. Zhang, M. Sun, Z. Wang, H. Li, Y. Li, G. Li, L. Lin, Blood species identification using near-infrared diffuse transmitted spectra and pls-da method, Infrared Phys. Technol. 76 (2016) 587–591.
[11] D.L. Dalton, A. Kotze, Dna barcoding as a tool for species identification in three forensic wildlife cases in south africa, Forensic Sci. Int. 207 (2011) e51–e54.
[12] M. Ge, Y. Wang, T. Wu, H. Li, C. Yang, T. Chen, H. Feng, D. Xu, J. Yao, Serum-based raman spectroscopic diagnosis of blast-induced brain injury in a rat model, Biomed. Opt. Express 14 (2023) 3622–3634.

[13] X. Zheng, G. Wu, G. Lv, L. Yin, X. Lv, Rapid discrimination of hepatic echinococcosis patients' serum using vibrational spectroscopy combined with support vector machines, Photodiagnosis Photodyn. Ther. 40 (2022).
[14] S.S. Panikar, N. Banu, E.-R. Escobar, G.-R. García, J. Cervantes-Martínez, T.-C. Villegas, P. Salas, E. De la Rosa, Stealth modified bottom up sers substrates for label-free therapeutic drug monitoring of doxorubicin in blood serum, Talanta 218 (2020) 121138.
[15] Z. Jin, W. Hu, Y. Yang, Serum metabolomic analysis revealed potential metabolite biomarkers for diabetes mellitus with coronary heart disease, Anal. Methods 15 (2023) 3432–3438.
[16] H. Yang, B. Zhou, H. Deng, M. Prinz, D. Siegel, Body fluid identification by mass spectrometry, Int. J. Legal Med. 127 (2013) 1065–1077.
[17] H. Inoue, F. Takabe, O. Takenaka, M. Iwasa, Y. Maeno, Species identification of blood and bloodstains by high-performance liquid chromatography, Int. J. Legal Med. 104 (1990) 9.
[18] L. Zhang, H. Ding, L. Lin, Y. Wang, X. Guo, H. Tian, Transmission versus reflection spectroscopy for discrimination of human and nonhuman blood, Infrared Phys. Technol. 99 (2019) 1–4.
[19] E. Gebel, Species in a snap: Raman analysis of blood, Anal. Chem. 81 (2009) 7862.
[20] K. Olbrich, Z. Setkowicz, K. Kawon, M. Czyzycki, N. Janik-Olchawa, I. Carlomagno, G. Aquilanti, J. Chwiej, Vibrational spectroscopy methods for investigation of the animal models of glioblastoma multiforme, Spectrochim. Acta Part A 303 (2023) 123230.
[21] E. Manzano, N. Navas, R. Checa-Moreno, L. Rodriguez-Simón, L. Capitán-Vallvey, Preliminary study of uv ageing process of proteinaceous paint binder by ft-ir and principal component analysis, Talanta 77 (2009) 1724–1731.
[22] H. Zhu, Y. Wang, H. Liang, Q. Chen, P. Zhao, J. Tao, Identification of portulaca oleracea l. From different sources using gc–ms and ft-ir spectroscopy, Talanta 81 (2010) 129–135.
[23] X. Li, P. Zeng, X. Wu, X. Yang, J. Lin, P. Liu, Y. Wang, Y. Diao, Resd-net: a model for rapid prediction of antioxidant activity in gentian root using ft-ir spectroscopy, Spectrochim. Acta Part A 310 (2024) 123848.
[24] G. Qin, A. Zhang, F. Chen, H. Man, H. Liu, M. Li, Z. Jia, Identification of appetite suppressants through fourier transform infrared spectroscopy and filtered spectral feature extraction, Microchem. J. 197 (2024) 109843.
[25] E. Mistek-Morabito, I.K. Lednev, Discrimination of menstrual and peripheral blood traces using attenuated total reflection fourier transform-infrared (atr ft-ir) spectroscopy and chemometrics for forensic purposes, Anal. Bioanal. Chem. 413 (2021) 2513–2522.
[26] M. Khanmohammadi, M.A. Ansari, A.B. Garmarudi, G. Hassanzadeh, G. Garoosi, Cancer diagnosis by discrimination between normal and malignant human blood samples using attenuated total reflectance-fourier transform infrared spectroscopy, Cancer Invest. 25 (2007) 397–404.
[27] M.J. Baker, J. Trevisan, P. Bassan, R. Bhargava, H.J. Butler, K.M. Dorling, P. R. Fielden, S.W. Fogarty, N.J. Fullwood, K.A. Heys, C. Hughes, P. Lasch, P. L. Martin-Hirsch, B. Obinaju, G.D. Sockalingum, J. Sule-Suso, R.J. Strong, M. J. Walsh, B.R. Wood, P. Gardner, F.L. Martin, Using fourier transform ir spectroscopy to analyze biological materials, Nat. Protocols 9 (2014) 1771–1791.
[28] G. Wang, W. Cai, H. Wu, C. Yang, K. Yu, R. Liu, X. Wei, H. Lin, Q. Sun, Z. Wang, Identification of human and non-human bloodstains on rough carriers based on atr-ftir and chemometrics, Microchem. J. 180 (2022) 107620.
[29] K. De Wael, L. Lepot, F. Gason, B. Gilbert, In search of blood – detection of minute particles using spectroscopic methods, Forensic Sci. Int. 180 (2008) 37–42.
[30] K. Virkler, I.K. Lednev, Blood species identification for forensic purposes using raman spectroscopy combined with advanced statistical analysis, Anal. Chem. 81 (2009) 7773–7777.
[31] S. Huang, P. Wang, Y. Tian, P. Bai, D. Chen, C. Wang, J. Chen, Z. Liu, J. Zheng, W. Yao, J. Li, J. Gao, Blood species identification based on deep learning analysis of raman spectra, Biomed. Opt. Express 10 (2019) 6129–6144.
[32] E. Mistek-Morabito, I.K. Lednev, Discrimination between human and animal blood by attenuated total reflection fourier transform-infrared spectroscopy, Commun. Chem. 3 (2020).
[33] Q. Fan, S. Li, R. Wu, Q. Xie, Determining sex of mammal by collecting male and female animal serum samples of mammal, performing e.g. near-infrared spectroscopy, selecting spectral range, establishing and validating sex discrimination model, and performing spectroscopy, CN106918571-A; CN106918571-B (2017).
[34] Y. Huyan, X. Nan, H. Li, S. Sun, Y. Xu, A novel fa1-targeting fluorescent probe for specific discrimination and identification of human serum albumin from bovine serum albumin††electronic supplementary information (esi) available: materials, instruments, methods, synthesis, supporting figures, scheme and tables. see, Chem. Commun. 60 (2024) 3810–3813, https://doi.org/10.1039/d4cc00407h.
[35] M. Salmain, A. Vessieres, P. Brossier, G. Jaouen, Use of fourier transform infrared spectroscopy for the simultaneous quantitative detection of metal carbonyl tracers suitable for multilabel immunoassays, Anal. Biochem. 208 (1993) 117–120.
[36] Y. Lu, Y. Cao, X. Tang, N. Hu, Z. Wang, P. Xu, Z. Hua, Y. Wang, Y. Su, Y. Guo, Deep learning-assisted mass spectrometry imaging for preliminary screening and pre-classification of psychoactive substances, Talanta 272 (2024) 125757.
[37] K.N. Basri, F. Yazid, M.N. Mohd Zain, Z. Md Yusof, R. Abdul Rani, A.S. Zoolfakar, Artificial neural network and convolutional neural network for prediction of dental caries, Spectrochim. Acta Part A 312 (2024) 124063.
[38] Y. Liu, H. Pu, Q. Li, D.-W. Sun, Discrimination of pericarpium citri reticulatae in different years using terahertz time-domain spectroscopy combined with convolutional neural network, Spectrochim. Acta Part A 286 (2023) 122035.
[39] J. Li, Z. Yang, Y. Zhao, K. Yu, Hsi combined with cnn model detection of heavy metal cu stress levels in apple rootstocks, Microchem. J. 194 (2023) 109306.
[40] S. Olsztynska-Janus, K. Szymborska-Malek, M. Gasior-Glogowska, T. Walski, M. Komorowska, W. Witkiewicz, C. Pezowicz, M. Kobielarz, S. Szotek, Spectroscopic techniques in the study of human tissues and their components. part i: Ir spectroscopy, Acta Bioeng. Biomech. 14 (2012) 101–115.
[41] C.-M. Orphanou, L. Walton-Williams, H. Mountain, J. Cassella, The detection and discrimination of human body fluids using atr ft-ir spectroscopy (vol 252, pg e10, 2015), Forensic Sci. Int. 261 (2016) 82.
[42] S. Gunasekaran, D. Uthra, Ftir and uv-visible spectral study on normal and jaundice blood samples, Asian J. Chem. 20 (2008) 5695–5703.
[43] K.M. Elkins, Rapid presumptive "fingerprinting" of body fluids and materials by atr ft-ir spectroscopy, J. Forensic Sci. 56 (2011) 1580–1587.
[44] Z. Movasaghi, S. Rehman, I.U. Rehman, Fourier transform infrared (ftir) spectroscopy of biological tissues, Appl. Spectrosc. Rev. 43 (2) (2008).
[45] K. Gajjar, J. Trevisan, G. Owens, P.J. Keating, N.J. Wood, H.F. Stringfellow, P. L. Martin-Hirsch, F.L. Martin, Fourier-transform infrared spectroscopy coupled with a classification machine for the analysis of blood plasma or serum: a novel diagnostic approach for ovarian cancer, Analyst 138 (2013) 3917–3926.
[46] N. Kanagathara, M. Thirunavukkarasu, C.E. Jeyanthi, P. Shenbagarajan, Ftir and uv-visible spectral study on normal blood samples, Int. J. Pharm. Biol. Sci. 1 (2011) 74–81.
[47] A. Takamura, K. Watanabe, T. Akutsu, T. Ozawa, Soft and robust identification of body fluid using fourier transform infrared spectroscopy and chemometric strategies for forensic analysis, Scientific Rep. 8 (2018).
[48] X. Cao, L. Zhang, Z. Wu, Z. Ling, J. Li, K. Guo, Quantitative analysis modeling for the chemcam spectral data based on laser-induced breakdown spectroscopy using convolutional neural network, Plasma Sci. Technol. 22 (2020).
[49] H. Gu, S. Wang, S. Hu, X. Wu, Q. Li, R. Zhang, J. Zhang, W. Zhang, Y. Peng, Identification of panax notoginseng origin using terahertz precision spectroscopy and neural network algorithm, 2024.