Artificial Intelligence in the Life Sciences 3 (2023) 100071

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

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

Deep metric learning for the classification of MALDI-TOF spectral signatures from multiple species of neotropical disease vectors

Fernando Merchan[a,f], Kenji Contreras[a], Rolando A. Gittens[b,f], Jose R. Loaiza[b,c,d,f], Javier E. Sanchez-Galan[b,e,f,*]

[a] Facultad de Ingeniería de Eléctrica, Universidad Tecnológica de Panamá, Campus Victor Levi Sasso, Panama, Panama
[b] Centro de Biodiversidad y Descubrimiento de Drogas, Instituto de Investigaciones Científicas y Servicios de Alta Tecnología (INDICASAT AIP), City of Knowledge, Panama, 0843-01103, Panama
[c] Smithsonian Tropical Research Institute, Panama, Panama
[d] Programa Centroamericano de Maestría en Entomología, Universidad de Panama, Panama, Panama
[e] Facultad de Ingeniería de Sistemas Computacionales, Universidad Tecnologica de Panama, Campus Victor Levi Sasso, Panama, Panama
[f] Sistema Nacional de Investigación (SNI), Secretaria Nacional de Ciencia, Tecnologia e Innovacion (SENACYT), Panama, Panama

ARTICLE INFO

Keywords:
MALDI-TOF
Machine learning
Metric learning
Culicidae
Ixodidae
Molecular taxonomy

ABSTRACT

Deep Learning techniques have significant advantages for mass spectral classification, such as parallelized signal correction and feature extraction. Deep Metric Learning models combine Metric Learning to determine the degree of similarity or difference between a set of mass spectra with the generalization power of Deep Learning to improve feature extraction even further. The two most popular of these models combine multiple neural networks with identical architectures and are commonly called Siamese (SNN) and Triplet Neural Networks (TNN). Herein, using both SNNs and TNNs, we intended to taxonomically categorize two sets of previously-validated mass spectra that corresponded to 30 species of Neotropical arthropods in the Culicidae and Ixodidae families, some of which are disease vectors. The effectiveness of SNNs and TNNs to correctly classify 826 spectra from 12 mosquito species and 310 spectra from 18 species of hard ticks was highly effective, with both algorithms performing with minimal average loss during cross-validation. SNNs produced accuracy rates for ticks and mosquitoes of 91.22% and 94.46%, respectively, while accuracy rates of 93% and 99% were obtained with TNNs. Our results indicate that Deep Metric Learning is a practical machine learning tool for quickly and precisely classifying MALDI-TOF-generated mass spectra of Neotropical and public-health-relevant arthropod species.

1. Introduction

At the forefront of public health concerns is the global threat from emerging vector-borne zoonotic diseases, which is increasing with the continuous geographical expansion of humans and vector species, and the spread of novel pathogens [1]. Despite historical efforts to prevent the spillover of pathogens from wildlife into people, the risk of emerging vector-borne diseases is still high in many regions around the world, particularly in developing countries such as Panama where large socio-economically vulnerable indigenous communities reside [2].

Several vector-borne pathogens present in Panama, including mosquito-transmitted *Plasmodium* (e.g., malaria) and tick-transmitted pathogenic *Rickettsia* (e.g., Rocky Mountain spotted fever), are only mitigated by the elimination and monitoring of vector populations [3,4]. However, most studies on arthropod species in Panama have concentrated on a few species, and the vast majority of species remain unstudied to this day [5,6]. This is in part due to the difficulties encountered when trying to lab-raise rare species, as well as the lack of accurate, reliable, and cost-effective species identification methods other than the traditional morphological identification techniques. Moreover, cryptic species complexes are a group of taxa with similar morphologies that belong to different evolutionary units and are easily confused with one another when recognized using morphological characters [2,7].

Accordingly, developing novel identification approaches for arthropods that can accurately and rapidly catalog the large biodiversity of arthropods that exist, is a topic of great interest in Panama [8,9], where several vector-borne diseases are still common [10,11]. With the largest number of described species of mosquitoes and hard ticks in Central America, Panama offers an unique opportunity to assess the usefulness of novel techniques to taxonomically characterize this elevated biodiversity. Previous work in Panama has established the presence of 265

* Corresponding author.
E-mail addresses: fernando.merchan@utp.ac.pa (F. Merchan), javier.sanchezgalan@utp.ac.pa (J.E. Sanchez-Galan).

https://doi.org/10.1016/j.ailsci.2023.100071
Received 29 December 2022; Received in revised form 2 April 2023; Accepted 3 April 2023
Available online 6 April 2023

mosquito species and 23 Ixodid tick species that can parasitize wildlife and humans [2,7].

Mass Spectrometry (MS) is a high-performance analytical tool that can be employed to characterize and identify complex biological samples. There exist different techniques to acquire spectral signals. One of particular interest in this work is known as Matrix-Assisted Laser Desorption/Ionization - Time of Flight or MALDI-TOF, which is considered a two-dimensional method geared to understand complex molecular samples via ionization and later time-of-flight discrimination [12,13]. Besides "true" proteomics, an exciting and growing application of MALDI-TOF MS is the use of protein profiles, or "mass fingerprints", to identify species of microorganisms [14], and more recently of arthropod specimens, from eggs all the way to adults [15–18]

A typical MALDI-TOF mass fingerprint experiment generates a large number of spectral signals that must be pre-processed before machine learning algorithms can automatically classify and identify arthropod species [19,20]. Mass spectral features allow for the identification of distinct biomarkers that characterize different biological species. For instance, MALDI-TOF has been previously used in Panama by our team, for species identification in two of the most important taxonomic families of arthropods that contain vectors of disease: *Culicidae* (i.e., mosquitoes) [2,16,21] and *Ixodidae* (i.e., hard ticks) [7].

Loaiza et al. [2] and Gittens et al. [7] used Principal Component Analysis (PCA) [22] and Linear Discriminant Analysis (LDA) [23] to categorize Panama's mosquitoes and ticks mass spectra, respectively, acquired from samples stored under various conditions (e.g., dry-frozen or in alcohol at room temperature). Both of these algorithms focus on transforming the input data unto an orthogonal sub-space in which new features allow to discard irrelevant information, improving classification efficiency. This type of machine learning framework is preceded by a pre-processing stage consisting of signal smoothing (Savitzky-Golay filter) and baseline correction to remove noise and improve performance classification. Another approach by López-Fernández et al. [24] featured a software platform named Mass-Up which implemented unsupervised learning algorithms together with PCA to facilitate the pre-processing and analysis of MALDI spectra. However, these types of algorithms may not be efficient or robust enough to continuously learn and reliably classify mass spectra from various arthropod families, collected and stored with different techniques and processed and analyzed with different parameters and equipment.

Another branch of machine learning that shows promise for mass spectra analysis and classification is Deep Learning, which comprises all the techniques and algorithms related to Artificial Neural Networks (ANN). The inception of these algorithms was inspired by the functioning of neurons in the human brain. These models carry out an adaptive learning process through multiple nonlinear mathematical transformations based on multivariable calculus and tensor algebra [25].

Deep Metric Learning is a technique that combines Deep Learning elements with the concept of Metric Learning (MeL), in which a model attempts to learn features from objects (such as signals and images) portrayed as numerical vectors in terms of their distance and location in a multidimensional space. This concept implies that two vectors in a neighboring space will have similar features, while distances between them indicate varying degrees of dissimilarity [26]. A classic example of MeL algorithms corresponds to K-Nearest Neighbors (K-NN) [27], which performs the classification process based on a distance metric (e.g., Euclidean, angular [cosine], Manhattan) to a reference sample and the k neighboring samples that are in the nearby vector space.

To understand the differences between Deep Metric Learning and Deep Learning-based methods, one may consider the analogy of a person (i.e., an intelligent model) learning to discriminate between different objects in a given data set. One approach to this task is for the person to carefully study the defining characteristics of each class of objects (i.e., Deep Learning). Alternatively, the person learns to determine the similarities between objects regardless of their predetermined label. Furthermore, one can illustrate both methodologies by describing how a machine learning model learns to classify an object (Fig. 1). For example, classical neural networks learn a feature space or representation (a set of all possible numerical values of a class), while Deep Metric Learning methods seek to generate an *embedding*, an alternative characterization of the original data projected into a latent space (i.e., a low-dimensional space generated to preserve the most prominent features) where similar objects appear closer [26].

Table 1
Anopheline mosquito species data set obtained from Loaiza et al. [2].

ID	Name	Abreviation	Spectral Count
1	*Anopheles albimanus*	Anabm	119
2	*Anopheles apicimacula*	Anapcml	110
3	*Anopheles aquasalis*	Anasls	56
4	*Anopheles darlingi*	Andalg	40
5	*Anopheles malefactor*	Anmaft	39
6	*Anopheles nuneztovari*	Annvi	192
7	*Anopheles pseudopunctipennis*	Anowd	45
8	*Anopheles punctimacula*	Anpsdt	81
9	*Anopheles strodei*	Anptml	49
10	*Anopheles triannulatus*	Anstro	26
11	*Anopheles neivai*	Antria	24
12	*Chagasia bathana*	Chaba	45

Some mass fingerprinting applications related to Deep Learning include the work of Rakotonirina et al. [28], which employed Convolutional Neural Network (CNN) models to detect the presence of the *Wolbachia* bacteria in the mosquito *Aedes aegypti*. The study featured a voting scheme in which the MALDI spectra dataset was partitioned into 15 parts, each of which was modelled by a similar neural network, and the final inference was computed by considering each model prediction. Moreover, computer vision has been used to test the effectiveness of Transfer Learning models to improve the identification of triatomine bugs that transmit Chagas disease to humans across the Americas [29]. In recent years with the advent of machine learning oriented frameworks utilizing Deep Learning [30] models, new approaches have been implemented in the context of arthropod detection and classification using CNN [31]. However, to our knowledge no previous studies have applied Deep Metric Learning for the rapid identification of arthropods from the biodiverse Neotropical region. In the work presented here, advanced feature extraction techniques based on Deep Learning and CNN, purposely designed to work with mass spectra without any pre-processing, were used to improve the classification tasks of Neotropical mosquito and tick species, some of which are important disease vectors.

2. Materials and methods

2.1. MALDI-TOF MS data sets

Two previously published data sets of MALDI-TOF spectra from Neotropical arthropods were used, belonging to the taxonomic families *Culicidae* (*i.e.*, mosquitoes) [2] and *Ixodidae* (*i.e.*, hard ticks) [7]. All the methods for sample collection in the field, sample processing (*i.e.*, protein extraction from arthropod legs), MALDI-TOF analyis and spectra selection were described in detail in the cited studies. In brief, we compiled samples from 12 distinct biological species of mosquitoes (Table 1) and 18 distinct biological species of ticks (Table 2), each species with a different number of specimens and, thus, a different number of spectra collected. All the samples were pre-identified using morphological keys. Mosquitoes were stored dry-frozen before MS analysis, while ticks had been stored for several years in 70% ethanol. For spectra selection into the database, three technical replicates were collected from each specimen, and each spectra had to meet a minimum criterion of at least one peak with an intensity of 3500 arbitrary units (a.u.) for mosquitoes, and 1000 a.u. for ticks (Fig. 2).

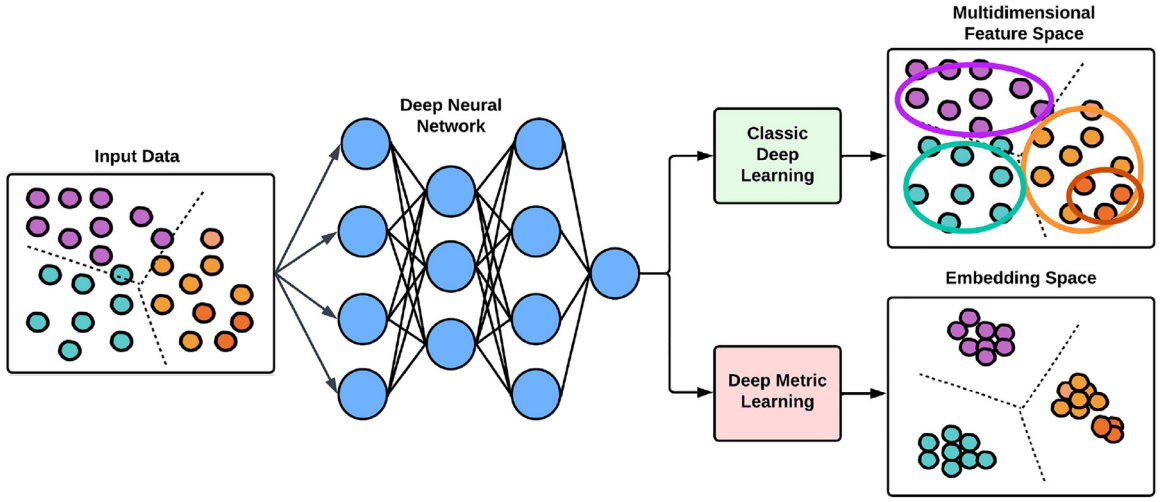


Fig. 1. Graphical representation comparing traditional feature learning methods with metric learning.

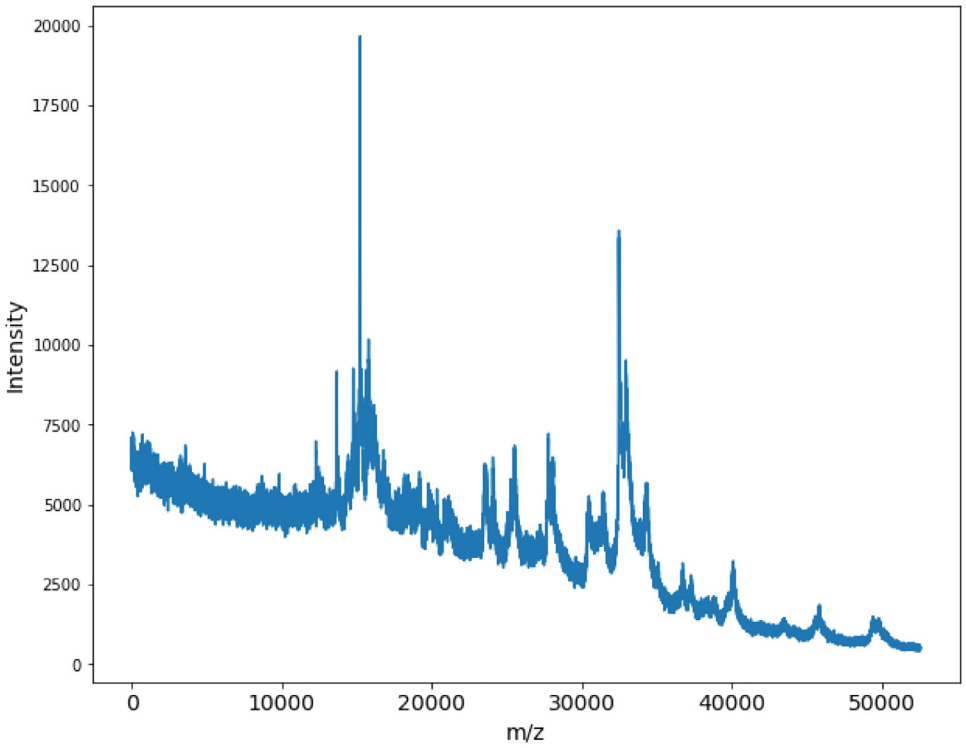


Fig. 2. Example of a raw spectrum of tick species *Amblyomma naponense*, collected as intensity vs. m/z (mass over charge), with a molecule size varying from 0 to 50,000 Da.

2.2. Machine learning architecture: Convolutional neural networks (CNN)

CNN have been the most commonly employed architectures in recent years for image recognition, classification, and detection tasks. This class of neural networks has three main types of layers that perform the operations of convolution, *pooling* (for feature extraction) and classification by means of an ANN. These architectures are made up of modules or blocks that are stacked one after the other depending on the complexity of the model [30].

The network architecture utilized in this study (Fig. 3) draws inspiration from the classic convolutional architecture, LeNet [32]. The design includes parameters and an activation function called Leaky ReLu [33], which mitigates the vanishing gradient issue that can make the neural network struggle to learn new data due to very small changes in parameter optimization during training.

Additionally, the last layer, which produces embeddings, includes $\mathcal{L}^2$ normalization [34], which aims to maximize the distance between positive (similar) and negative (different) samples [35]. Additionally, to address the challenge of unbalanced classes, we applied the technique of class weights [36]. This technique penalizes the misclassification of minority classes within the data set.

The final architecture comprises three convolutional blocks, each with one-dimensional convolution layers. The first block has 32 filters, the second has 16 filters, and the third has 8 filters. Batch normalization [37] is applied to each convolutional block. The dense layers consist of 32 and 16 neurons and incorporate dropout. Finally, the optimizer algorithm used is Adaptive Moment Estimation (ADAM) [38].

2.2.1. Siamese neural networks (SNN)

The *SNN* consist of a pair of interconnected neural networks with identical parameters that are trained in parallel, whose unified architecture and learning is governed by the contrastive loss function

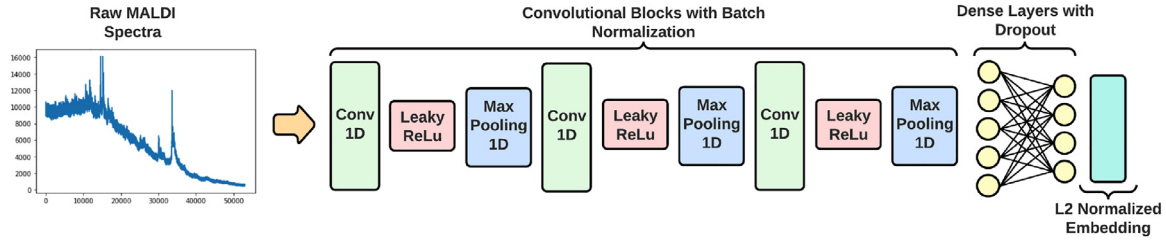


Fig. 3. Diagram of the unified architecture used for the analysis of both data sets.

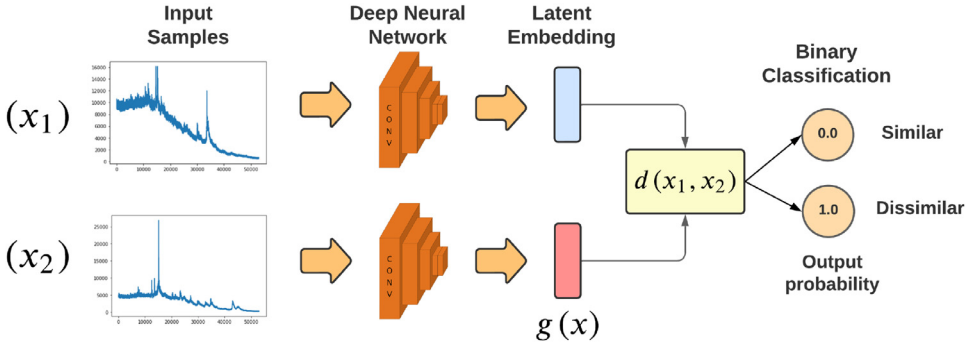


Fig. 4. Graphical representation of the unified architecture of a Siamese neural network.

Table 2
Ixodid tick species data set obtained from Gittens et al. [7].

ID	Name	Abreviation	Spectral Count
1	*Amblyomma cajenense*	Amb Caj	6
2	*Amblyomma calcarattum*	Amb Cal	15
3	*Amblyomma dissimile*	Amb Dis	9
4	*Amblyomma geayi*	Amb Gea	12
5	*Amblyomma nodosum*	Amb Nod	10
6	*Amblyomma oblongoguttatum*	Amb Obl	8
7	*Amblyomma ovale*	Amb Ova	11
8	*Amblyomma pecarium*	Amb Pec	11
9	*Amblyomma saberanae*	Amb Sab	11
10	*Amblyomma varium*	Amb Var	9
11	*Ammblyomman naponense*	Amb Nap	9
12	*Ammblyomma tapirellum*	Amb Tap	56
13	*Dermacentor nitens*	Derm Nit	60
14	*Haemaphysalis juxtackochi*	Hae Jux	6
15	*Ixodes affinis*	Ix Aff	9
16	*Ixodes boliviensis*	Ix Bol	11
17	*Rhipicephalus microplus*	Rhi Mic	50
18	*Rhipicephalus sanguineus*	Rhi San	6

[39] (Fig. 4), given by Eq. (1):

$$CLoss = (1-y)\frac{1}{2}(d) + (y)\frac{1}{2}\max\{0, \alpha - d\} \tag{1}$$

where,

$$d(x_1, x_2) = \left\| g(x_1) - g(x_2) \right\| \tag{2}$$

The contrastive loss function [39] (Eq. (1)) aims to optimize the performance of the classification model by measuring the similarity between pairs of samples, denoted by a binary label y,where $y = 0$ denotes similarity between samples and $y = 1$ denotes dissimilarity. The function takes in two input samples, x_1 and x_2, and calculates the Euclidean distance between their respective embeddings (*i.e.*, mapping functions $g(x_1)$ and $g(x_2)$). Then, the difference between embeddings is computed and normalized to form the distance measure, $d(x_1, x_2)$.

During training, the parameters of $g(x_n)$ are adjusted to minimize the loss (error rate) by reducing the distance measure between similar samples while maximizing it between dissimilar samples. This is, the function optimizes an embedding space where similar spectra are in the same embedding neighborhood while dissimilar ones are distant. This is completed by setting a threshold value (α) such that the distance between similar samples is smaller than α, while the distance between dissimilar samples is larger than α.

This structure is similar to that of the binary cross-entropy loss function, which is commonly used in binary classification problems [40]. The main advantage of this loss function is that it allows for a multi-class problem to be translated into a binary classification problem. This enables the model to perform flexible comparisons, such as comparing one class to many by making pairs consisting of one species and many others. This flexibility is particularly useful in problems where there are many classes and comparisons between classes are important for the analysis, such as in the case of arthropod species.

2.2.2. Triplet neural networks (TNN)

The (*TNN*) are considered an extension of SNN because of the unified architecture with identical parameters and the learning technique that attempts to train a model to minimize intra-class distance and maximize inter-class distance between classes of objects [41] (Fig. 5). However, the Triple Loss function (Eq. (3)) is used to learn how to generate an embedding space $g(x)$ by computing the distance between triplets of samples.

$$TLoss = \sum_{a,p,n y_p = y_a, y_n \neq y_a} max\left\{d(x_a, x_p) + \alpha - d(x_a, x_n), 0\right\}_+ \tag{3}$$

A triplet is formulated based on an anchor x_a, a positive x_p and a negative x_n sample, each with their respective labels (y_a, y_p, y_n), under the condition that $y_p = y_a$ and $y_n \neq y_a$, where the Euclidean distance (d) is calculated based on the *embeddings* generated by the neural networks (Fig. 5).

Unlike SNNs that can be trained directly with random pairs of samples generated during the training, TNNs require a stage of triplet selection, since training a network with all possible triplets from a data set represents a high and inefficient cubic computational complexity $\mathcal{O}(n)^3$, and not all triplets contribute to the efficient learning of the network.

A way to understand the necessity of selecting the right triplets to feed to a neural network is to assume that a model, like a human being, learns better to identify diverse objects from studying hard examples where the different classes have similar or near identical features. According to this assumption and the work of Schroff et al. [41], one of the most efficient methods is to generate triplets during the training cycle (epoch) and only employ the hard triplets generated from the mini batch of samples at each part of the cycle, also known as Online Triplet Mining

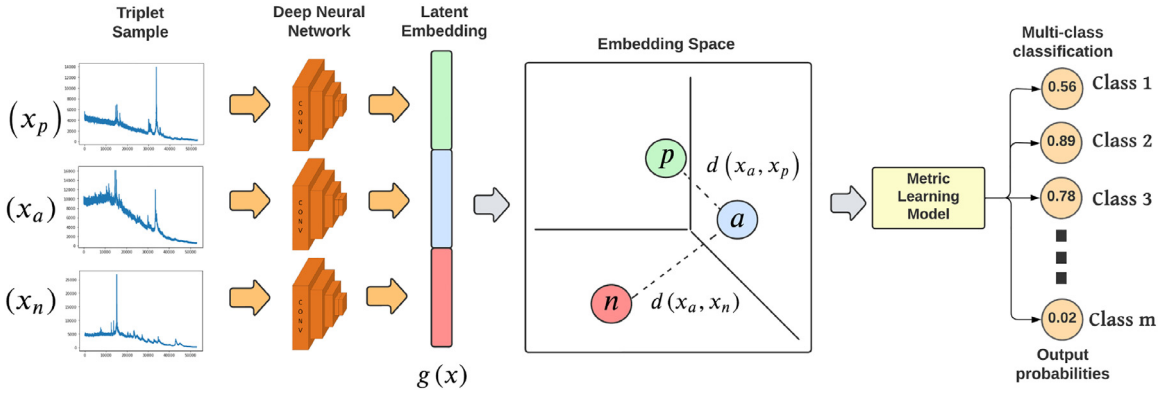
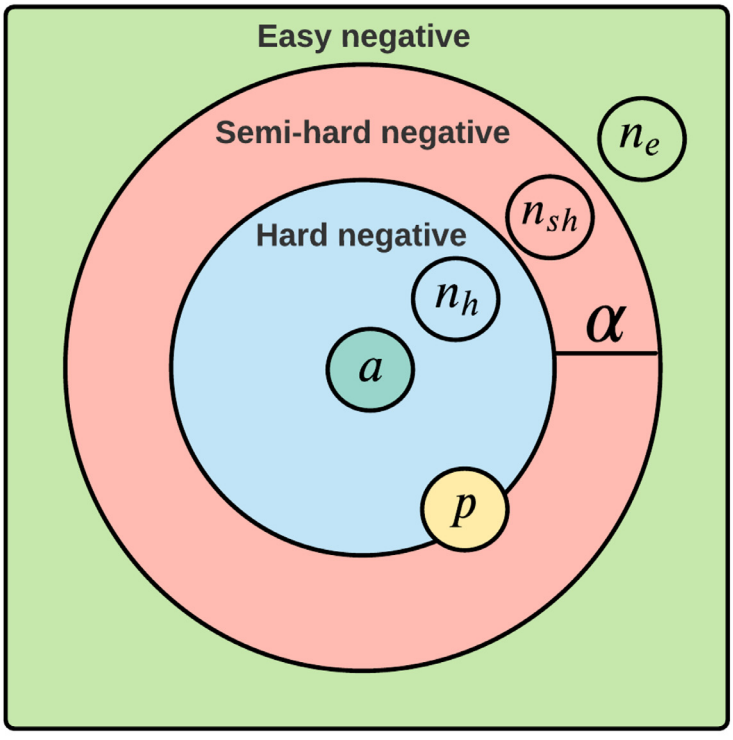


Fig. 5. Graphical representation of the unified architecture of a triple neural network.



$$n_e : d(a,p) + \alpha < d(a,n)$$
$$n_{sh} : d(a,p) < d(a,n) < d(a,p) + \alpha$$
$$n_h : d(a,n) < d(a,p)$$

Fig. 6. Description of the types of triplets based on the parameter α.

[26]. The selection criteria also depends on the threshold parameter α (Fig. 6).

Moreover, it is necessary to remark that TNNs are used as a method of feature extraction, unlike SNNs where both feature extraction and classification are performed by the same architecture. The neural networks will have their parameters optimized during training to produce a new embedding space where another metric learning model (in this case, K-NN) will perform the multi-class classification stage.

2.3. Data analysis: Experimental settings

Data sets used in this study are described in Tables 1 and 2 with a total of 826 mosquito and 310 tick spectra, respectively. The distribution of samples for model training and testing was 80%/20% in the case of mosquitoes, as had been the case with our previously published machine learning algorithms [2,7], but had to be adjusted for ticks to 70%/30% as will be described ahead. As explained in Section 2.2, the usage of CNN allowed to train models with input raw spectra without the need for significant pre-processing steps. Therefore, the only step before training was to take n first features of each spectra sample (according to the smallest spectrum of the data set), yielding numerical arrays of $n = 52906$ and $n = 52570$ features for ticks and mosquitoes respectively.

Due to the special training scheme that SNN architectures imply (Section 2.2.1), pairs of samples were first generated randomly, maintaining a similar distribution of samples between the different classes to avoid biases due to sample imbalance (Table 3). After performing

Table 3
Distribution of spectra for training and validation with *SNN*.

	Sample	# Training Pairs	# Testing Pairs
Mosquitoes	826	578	248
Ticks	310	432	186

Table 4
Distribution of spectra for training and validation with *TNN*.

	Sample	Training (%)	Testing (%)
Mosquitoes	826	80%	20%
Ticks	310	70%	30%

Table 5
Generic Confusion Matrix.

		Predicted	
		Positive	Negative
Actual	**Positive**	TP	FN
	Negative	FP	TN

manual parameter experimentation, we selected the threshold value of $\alpha = 0.55$ for both ticks and mosquitoes.

For the TNN, the α was set to 0.4, and it should be noted that the best performance for TNN training was achieved using an Online Triplet Mining scheme with Hard Negative Triplets (Fig. 6). Furthermore, this prompted an adjustment to 70%/30% train/testing data splitting for the tick's data set, which had more classes with fewer samples per class compared to mosquitoes (Table 4). Ultimately, this adjustment was necessary to avoid computational errors produced during the validation stage after each training epoch, where the mining algorithm failed to find triplets that fulfill the similarity criteria described in Section 2.2.2.

Furthermore, in order to avoid over-fitting, the SNN models were trained with *k-fold* cross validations, with $k = 5$, for 30 epochs with a *batch size* of 16 samples. The TNN were trained with a *batch size* of 24 samples, using a K-Shuffle Split validation scheme (also known as Monte Carlo Cross Validation) for 20 iterations and 32 epochs in each one. This validation scheme produces better results when dealing with data sets containing limited samples per class [42].

2.4. Evaluation metrics

Error metrics were calculated for each model and data set combination. A confusion matrix listing True Negatives (TN), True Positives (TP), False Negatives (FN) and False Positives (FP) was also used to represent the total number and proportion of samples that were correctly or incorrectly predicted (Table 5).

Moreover, using the items of the confusion matrix, other error metrics such as Accuracy, Precision, Recall and F1-score (Eqs. 4–7, respectively) were computed and later compared for each model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{7}$$

In general, the accuracy corresponds to the number of samples that were correctly predicted in their particular class (*i.e.*, species) among the total number of samples used in the test.

Table 6
Resulting confusion matrix for the SNN model for the mosquito data set.

		Predicted		
		Positive	Negative	Total
Actual	**Positive**	235 – 94.8%	13 – 5.2%	248
	Negative	28 – 11.3%	220 – 88.7%	248
Total		263	233	496

Table 7
Resulting confusion matrix for the SNN model for the ticks data set.

		Predicted		
		Positive	Negative	Total
Actual	**Positive**	83 – 89.2%	10 – 10.9%	93
	Negative	6 – 6.5%	87 – 93.5%	93
Total		89	97	186

2.5. Software and hardware specifications

The models developed in this work have been implemented using the Python programming language, with open-source software libraries such as: Tensorflow [43], Scipy [44] and Scikit-learn [45]. Deep Learning models were trained on the Google Colab [46] online platform with a Tesla T4 GPU with 16GB VRAM.

Additionally, to illustrate the particular learning process of the models based on *Deep Metric Learning*, embedding visualizations were created using the dimension reduction algorithm called *pacmap* [47].

3. Results

3.1. Classification with siamese neural networks (SNN)

SNN models exhibited good performance to classify mosquitoes (Fig. 7) and ticks (Fig. 8) mass spectra. The training loss and accuracy of SNN with the mosquito data set improved considerably after the first cross-validation (Fig. 7A and C, Fold 1 in blue), which was the only iteration that clearly separated from the rest. Similarly, the loss and accuracy average for the validation stage using the remaining 20% of the data set also seemed to converge from Fold 2 to 5 (Fig. 7B and D) for mosquitoes and Fig. 8-B, D for ticks. The jagged and oscillating curves in these validation graphs could be signs of slight over-fitting, with Fold 3 (green line) exhibiting the greatest instability. However, after Fold 4 and Fold 5 the model seems to have reached a stable steady state, meaning that it successfully learned to distinguish from existing species with a final loss and accuracy average of 0.0392 and 91.22%, respectively.

The results of the SNN model performance with the tick data set was very similar to the mosquitoes, except for the fact that the model seemed to require Fold 3 or more to reach peak accuracy in training and validation (Fig. 8B and C). In addition, training and validation curves improved "slower" but had less instability than mosquito curves, although Fold 4 and 5 also offered the best performance, with a final loss of 0.0185 and accuracy average of 94.46%.

Confusion matrices were computed for each model, and are shown in Table 6 for mosquitoes and in Table 7 for ticks. In general, the SNN model for mosquitoes had around 95% of True Positives (TP) and 89% for True Negatives (TN), while the numbers were reversed for ticks, with 89% for TP and almost 94% for TN. Error metrics were also calculated for the SNN models and are shown in Table 8. Evaluating all the parameters considered in the error metrics, the model was able to achieve 91% in all of them, varying only on the third decimal.

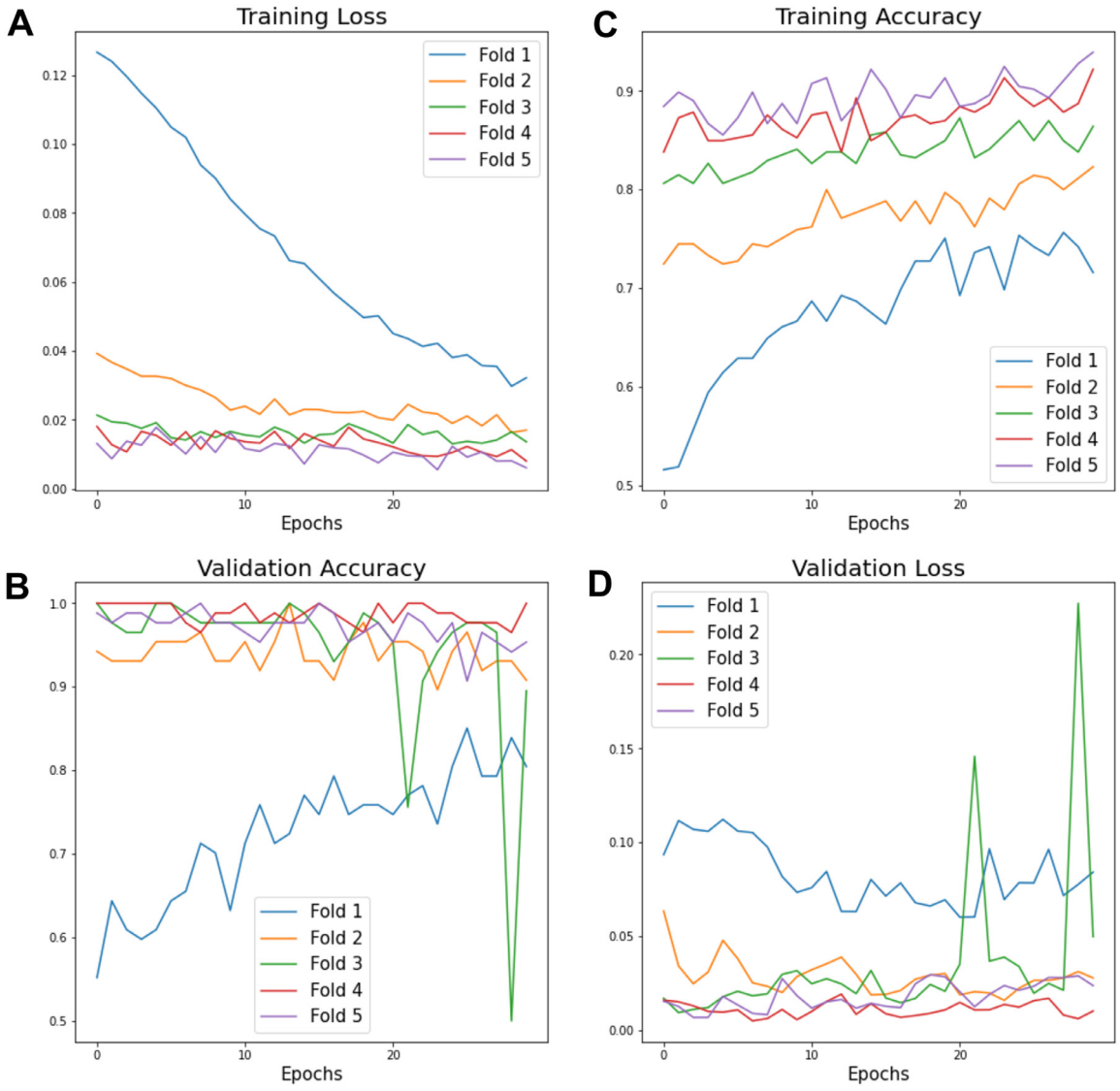


Fig. 7. Loss and Accuracy curves shown for Training (A, C) and Testing Validation (B, D) using the *SNN* model for the mosquitos data set).

Table 8
Resulting Metrics for the SNN model using both sets of spectra.

	Accuracy	Precision	Recall	F1-Score
Mosquitoes	0.9188	0.9188	0.9117	0.9172
Ticks	0.9147	0.9139	0.9139	0.9139

Table 9
Error metrics for the *TNN* model for both data sets.

	Accuracy	Precision	Recall	F1-Score
Mosquitoes	**0.9979**	0.9979	0.9965	**0.9971**
Ticks	0.9325	0.8980	0.9355	0.9144

3.2. Classification with triplet neural networks (TNN)

TNN models were very effective at classifying mass spectra data sets and outperformed SNN models. At the end of the entire training and validation cycle, the minimum validation loss reached by the TNN models was 0.00001 for mosquitoes and 0.0160 for ticks. Furthermore, the error metrics of accuracy and *F1-Score* obtained for both *knn* models with their respective test sets reached up to 99.7% for both parameters in the mosquito data set, and 93.2% and 91.4%, respectively, for ticks (Table 9).

The Confusion matrices shown in Fig. 9, explore which species are easier to to predict for both mosquitoes and ticks. In the case of mosquitoes, the matrix showed correct predictions for all except one species at a 100%, failing only in a sample of the *Anabm* (*Anopheles albimanus*) that was misclassified as *Anowd* (*Anopheles oswaldoi*). For ticks the results were similar, but with slightly more misclassifications, such as a sample from *Amb Caj* (*Amblyomma cajennense*) species wrongly classified as *Amb Obl* (*Amblyomma oblongoguttatum*), a sample from *Derm Nit* (*Dermacentor nitens*) misclassified as *Rhi Mic* (*Rhipicephalus microplus*) species, another sample of *Hae Jux* (*Haemaphysalis juxtackochi*) misclassified as *Derm Nit* (*Dermacentor nitens*), and finally a sample of *Rhi San* (*Rhipicephalus sanguineus*) classified mistakenly within the *Amb Gea* (*Amblyomma geayi*).

3.3. Triplet neural networks embedding visualizations

The embedding generated by TNN was graphed in two dimensions for a simple visualization of mosquitoes (Fig. 10) and ticks (Fig. 11) data

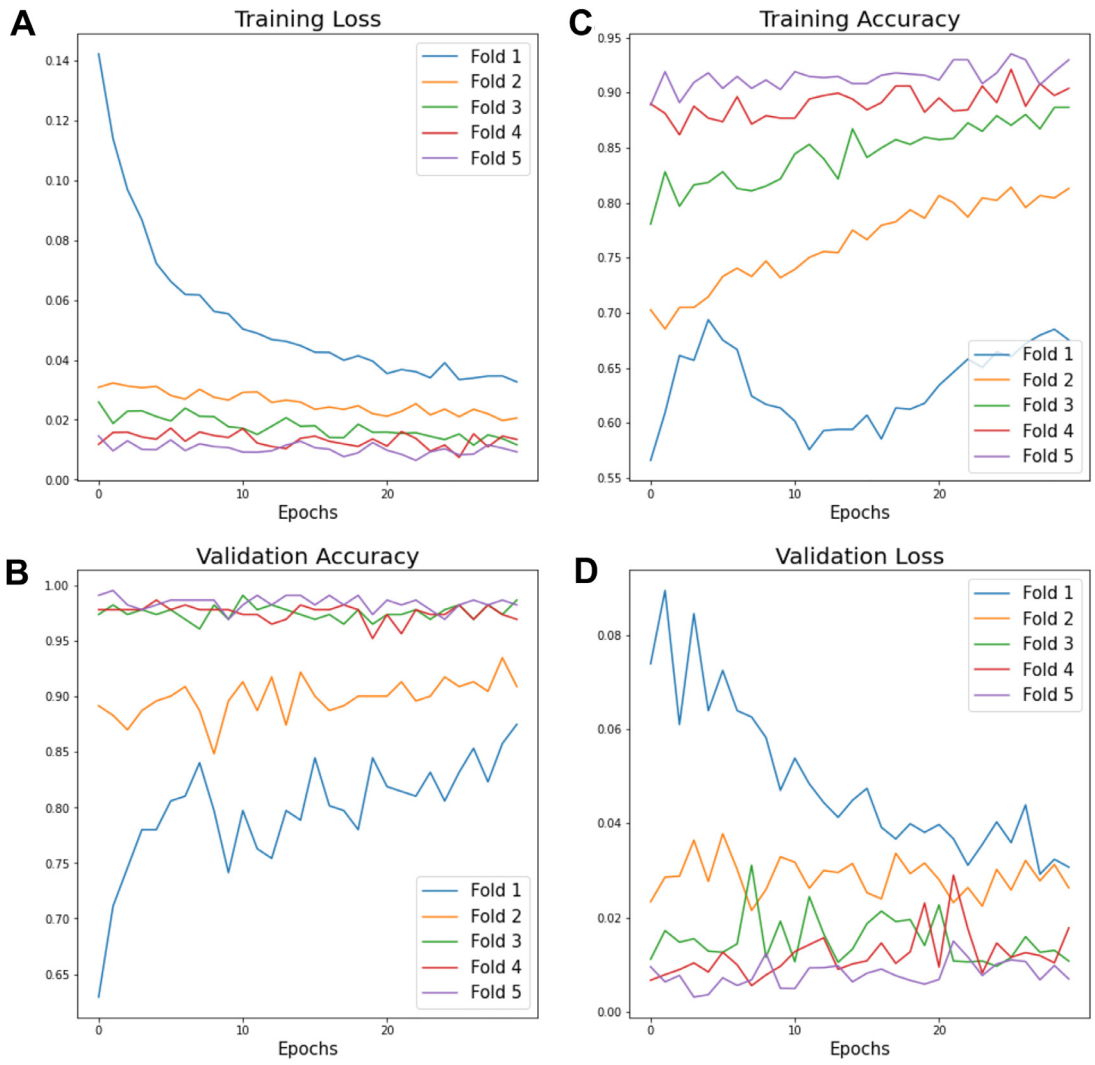


Fig. 8. Loss and Accuracy curves shown for Training (A, C) and Testing Validation (B, D) using the *SNN* model for the ticks data set).

sets. These graphs provide a visual representation of how the model learns to separate and/or join samples depending on their degree of similarity.

For example, the embedding prior to TNN training clearly demonstrates that samples of one species of mosquito (Fig. 10A) or ticks (Fig. 11A) can be found anywhere on the map, whereas after the network is trained, samples of the same species are closely clustered together and species with similar spectral characteristics form neighborhoods or clusters where each point represents one spectrum (Figs. 10B and 11 B).

4. Discussion

In our previous work establishing these self-curated data sets and analyzing them with PCA and LDA machine learning algorithms [2,7], we compared the smoothed and baseline-corrected spectra generated from unidentified field-collected arthropod samples with the mean spectra from a subset of the same field-collected samples that had already been identified using conventional morphological methods. The main contribution of these studies was to show a way around working without a commercial reference library of protein spectra profiles, using what we called the "self-curated reference library" concept. Because we are not using high-quality, lab-reared specimens to train the algorithm, this methodology required such pre-processing of the spectra to provided an accurate, and unbiased method to identify mosquito and tick species, with global positive identification rates of 93.3% for mosquitoes and 94.2.0% for ticks.

Here, we broadened the scope of data analysis by applying two Deep Metric Learning techniques to the same two MALDI-TOF data sets. Our idea of using Deep Metric Learning comes from the fact that these networks have a collective architecture that eliminates the need for spectral pre-processing steps as well as classification algorithms to discriminate the taxonomic identity of different arthropod species. The ability to rapidly and effectively classify arthropod species from distantly related species is one of the benefits of using these types of networks, which have proven to be highly efficient to analyze complex mass spectrometry data sets from two of the most important arthropod groupings in the Neotropics.

Roughly 1150 MALDI-TOF spectra from 30 arthropod species, some of which are vectors of disease (*i.e.*, 826 spectra from 12 mosquito species and 310 spectra from 18 tick species), were successfully classified using both SNN and TNN networks. For the SNN, the validation step at the end of each iteration with a subset of 20% of the spectra yielded an average loss and accuracy of 0.0392 and 91.22% for ticks and an average loss and accuracy of 0.0185 and 94.46% for mosquitoes, respectively. In comparison, the TNN validation step at the end of each iteration with 20% of the spectra yielded an average loss and accuracy of 0.00001 and 99% for mosquitoes, and an average loss and accuracy of 0.0160 and 93% for ticks, respectively. Even if the TNN model only improved the identification success rate for mosquitoes (*i.e.*, from 93.3%

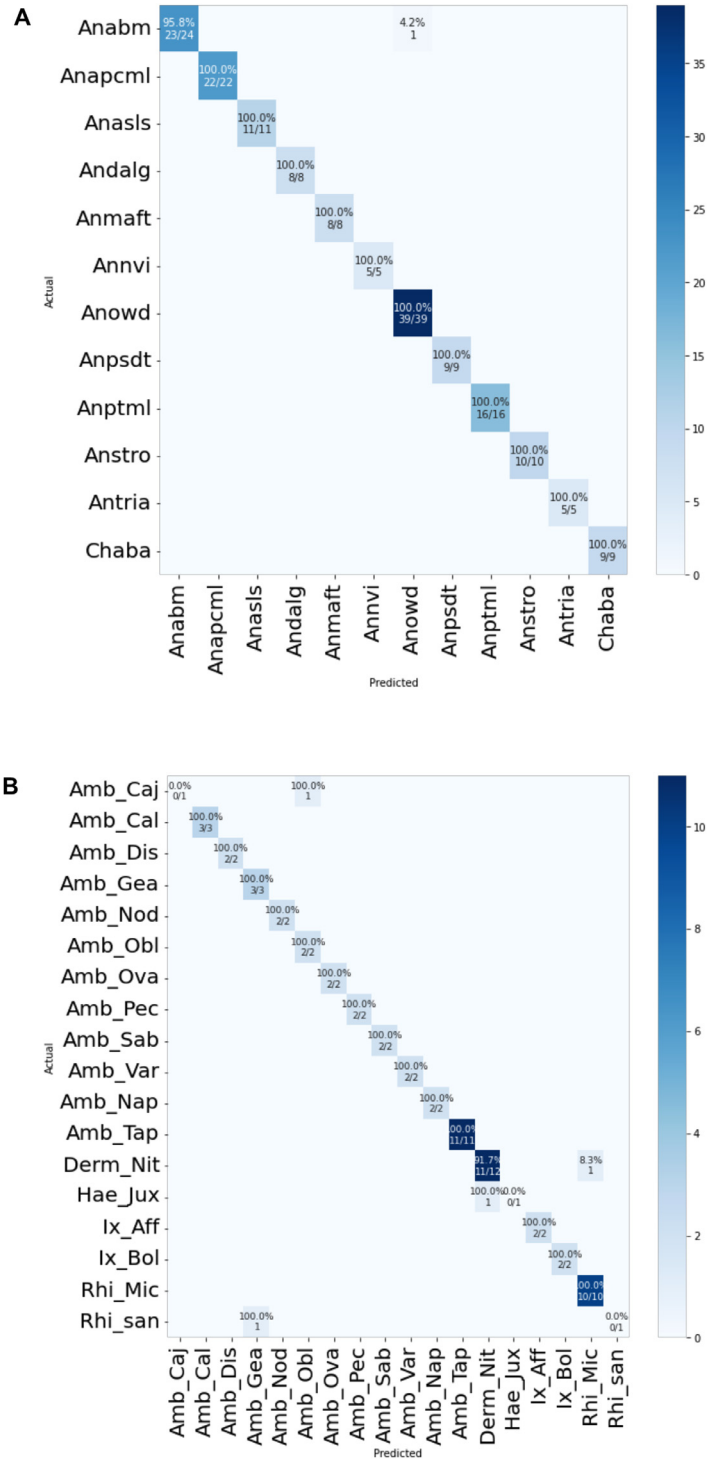


Fig. 9. Confusion matrices of the *TNN* model for the mosquito (A) and ticks (B) data sets.

to 99%), and it actually reduced the success rate for ticks (*i.e.*, from 94.2% to 93%), the improvements in lack of manual prep, speed and computing availability more than justify its implementation.

Our results show that both SNNs and TNNs were capable of recognizing and classifying MALDI spectra from roughly 50% of all hard tick taxa and 70% of all anopheline taxa (e.g., both ecologically dominant and rare species) reported for Panama. These results are equivalent and sometimes even superior to those from Loaiza et al. [2] and Gittens et al. [7]. Still, the *pacmap* visualization of the embedding confirmed that these algorithms are not perfect. Future research will need to look into what caused some samples in *Ana_bm, Amb_Caj, Derm_Nit, Hae_Jux*, and *Rhi_San* species to be misclassified. We can only speculate that some of these misclassifications may be related to poor quality spectra or degraded samples, and not necessarily to biologically-relevant inferences.

There are some additional limitations to this work that need to be discussed. The data sets used in this work come from biological samples that were gathered from the field as opposed to being reared in the laboratory. Besides foreign contaminants that are hard to control in field-collected specimens (*i.e.*, host blood, sap, pathogens), which already present an important challenge to the classification, the data sets also allowed us to test the reliability of the algorithms regarding sample storage. Fresh mosquito specimens had been stored dry-frozen as opposed to ticks that had been stored in ethanol and frozen for several years prior to protein extraction and sample processing with the MALDI-TOF. Ethanol is a well-known protein denaturing agent and we know

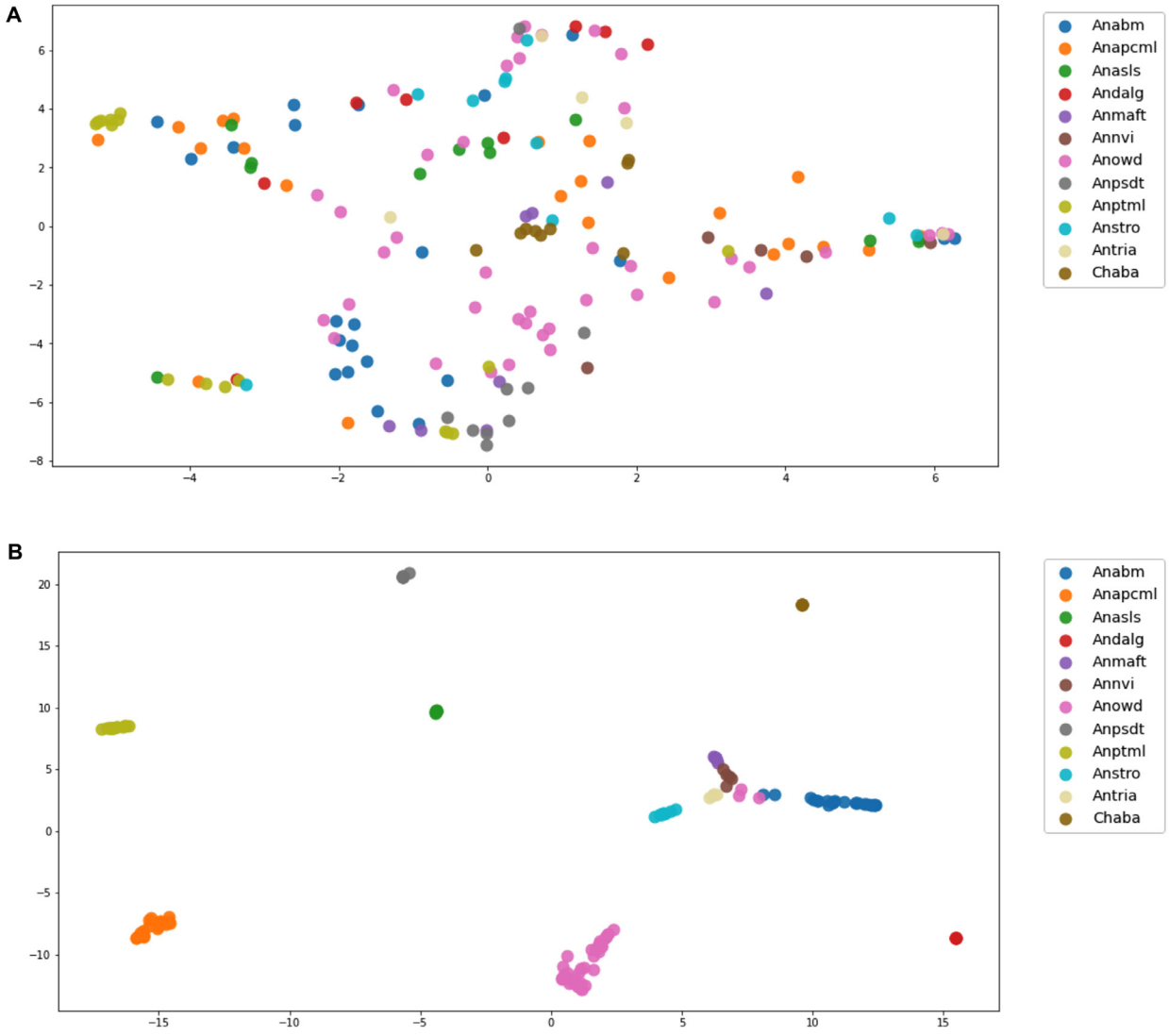


Fig. 10. Visualization with *pacmap* of the *embedding* produced by the *TNN* model before (A) and after (B) training for the mosquitos data set.

now that protein mass spectra quality of arthropod specimens stored in 70% ethanol is considerably reduced [7,48,49].

Sample degradation, lower quality spectra and lower number of specimens/spectra per species in the tick data set forced us to adjust the training and testing data split for the TNN model. Although the initial experimental plan was to work with the same ratio of 80%/20% training/testing sets for mosquitoes and ticks, we had to adjust the tick data split to 70%/30% to allow the TNN model to find the appropriate triplets for validation and avoid computational errors. Still, 70%/30% or 80%/20% are both typical data splits considered "adequate" for machine learning algorithms, especially when using cross-validation strategies [50–52].

According to Liu et al. [53], the classification of Deep Learning spectrometry signals presents significant benefits, because the CNN training process acts as a parallel mechanism for pre-processing (*e.g.*, correction of the signal base, curve smoothing) and feature extraction (*e.g.*, peak detection, non-linear features) from the raw spectra. Therefore, a unified architecture would eliminate the need to apply the pre-processing scheme required with the implementation of classical methods like PCA and LDA [2,7]. Additionally, the effectiveness of one-dimensional CNN with a Deep Metric Learning-based schema was demonstrated to minimize the difficulties in training a model with a reduced or unbalanced data set [54]. That is, a model with variable availability of samples for multiple classes, a case that is common in applications where there are several classes of elements to be analyzed and there is not enough information available to model each class uniformly. This situation generates in most of the implemented models either *overfitting* or *underfitting*, causing erroneous predictions and decreased performance [55]. An example would be the case of the mining of precious minerals [53] or the collection of tick specimens [7] where some species (*i.e.*, "classes") are very rare or difficult to obtain and therefore few samples are available for the data analysis.

Artificial neural networks (ANN) coupled to MALDI-TOF mass spectrometry have been recently used to predict drivers of malaria transmission using lab-reared *Anopheles stephensi* mosquitoes. ANN accurately recognized spectral patterns associated with mosquito age (0–10 days, 11–20 days, and 21–28 days), blood feeding, and *Plamodium berghei* infection, with the best prediction accuracy of 73%, 89%, and 78%, respectively [56]. Moreover, MALDI-TOF, coupled with CNN was used to detect *Wolbachia* bacteria in laboratory reared and field samples of the dengue vector, *Aedes aegypti*. Findings showed that CNN recognized *Aedes aegypti* spectral patterns associated with *Wolbachia*-infection with extremely high sensitivity (i.e., 93%), specificity (i.e., 99%), and accuracy (97%), being as efficient as qPCR, but surpassing the efficiency of traditional detection methods such as loop-mediated isothermal amplification (LAMP) [57]. Together, these studies illustrate the potential that

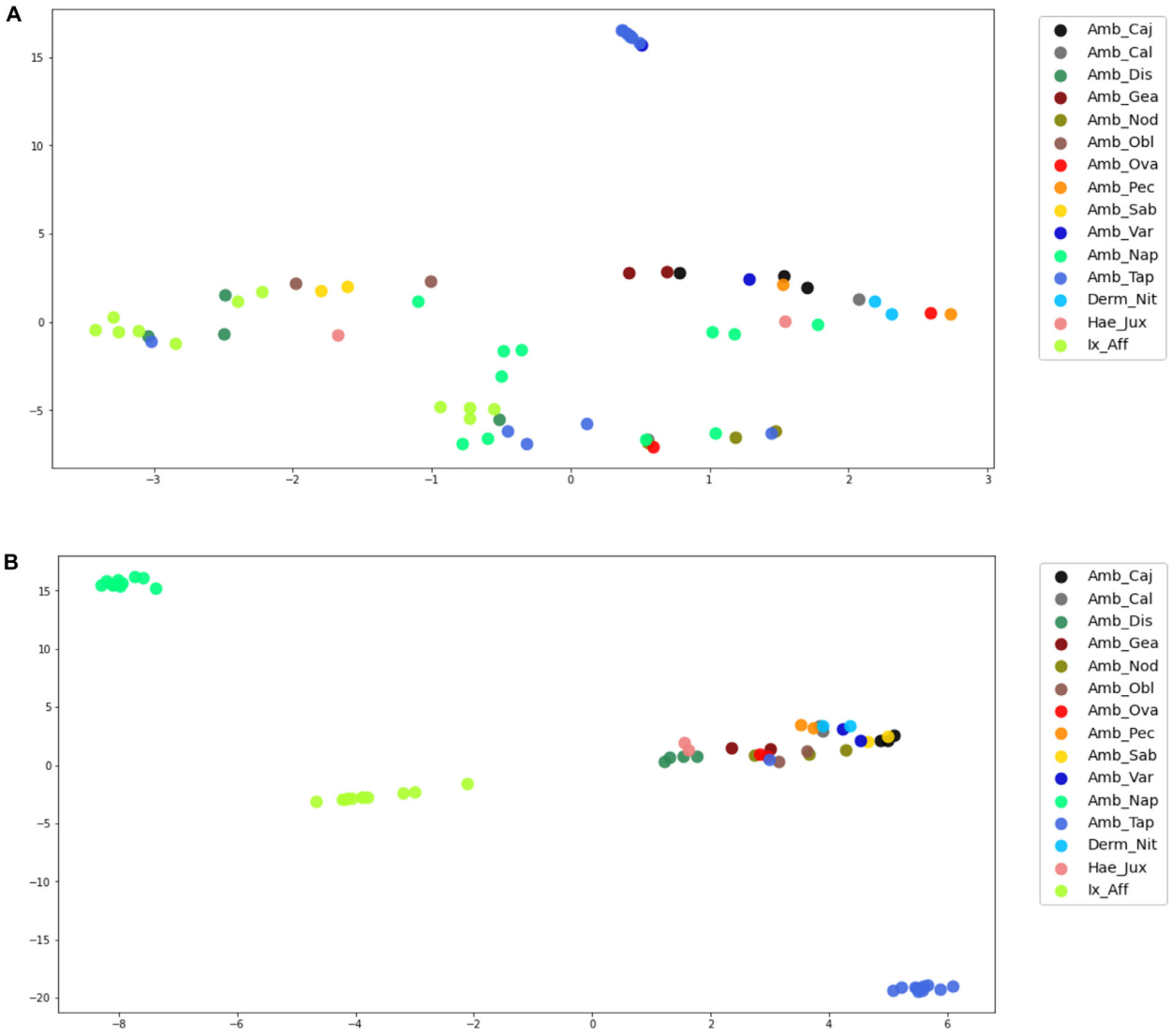


Fig. 11. Visualization with *pacmap* of the *embedding* produced by the *TNN* model before (A) and after (B) training for the ticks data set.

MALDI-TOF MS coupled with artificial intelligence has to predict entomological drivers of disease transmission, providing potential new tools for vector control.

To the best of our knowledge, this is the first attempt to classify MALDI-TOF data from arthropods from the Neotropics using Deep Metric Learning techniques that employ Metric Learning techniques with neural networks as a feature extraction method, including models that combine multiple neural networks with identical architectures commonly known as Siamese and Triplet Neural Networks. Our results indicate that Deep Metric Learning is yet another practical machine learning tool for quickly and precisely classifying MALDI-TOF-generated mass spectra of public-health-relevant arthropod species.

5. Conclusion

Rapid identification of arthropod species using mass spectra fingerprinting continues to offer a promising tool for entomologist and public-health specialist, specially in biodiverse regions such as the Neotropics. One of the long term goals of our group is to develop an online, crowd-sourced and open-source database, based on novel machine learning algorithms that can continuously grow and learn across arthropod families, and can improve the efficiency and accuracy of species identification regardless of sample preparation and analysis parameters. Our results here presented, where we successfully applied Deep Metric Learning techniques to two relevant and distinct data sets from the Culicidae and Ixodidae arthropod families, bring us one step closer to that goal.

We developed an analytical method for the evaluation of MALDI-TOF spectra utilizing SNN and TNN networks to quickly identify mosquito and tick species based on their protein mass spectra without the use of spectral pre-processing techniques or classification algorithms. Our results show that both SNN and TNN networks were capable of accurately classifying spectra from up to 30 different field-collected arthropod species, some of which are important disease vectors in Panama and more broadly in the Neotropics. Future studies should look into how well this approach can differentiate between other important Neotropical arthropod assemblages, including the triatomine bugs and Phlebotominae sand flies that transmit Leishmaniasis and Chagas disease. Simply put, our approach is adaptable and can be applied more broadly in other tropical regions of the world with high species diversity of arthropods and cryptic species complexes.

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

Data will be made available on request.

Acknowledgments

Financial support for this work was provided by SENACYT through the research grant GRID15-002 to JRL, JSG, and RAG. INDICASAT-AIP, UTP and STRI provided additional economic and logistic support. The SNI supports research activities by JRL, JSG, FM and RAG. RAG is also supported by SENACYT grants FID14-066, ITE15-016. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

[1] Weaver SC, Reisen WK. Present and future arboviral threats. Antiviral Res 2010;85(2):328–45.
[2] Loaiza JR, Almanza A, Rojas JC, Mejía L, Cervantes ND, Sanchez-Galan JE, et al. Application of matrix-assisted laser desorption/ionization mass spectrometry to identify species of Neotropical anopheles vectors of malaria. Malar J 2019;18(1):1–12.
[3] Lainhart W, Dutari LC, Rovira JR, Sucupira IM, Póvoa MM, Conn JE, et al. Epidemic and non-epidemic hot spots of malaria transmission occur in indigenous comarcas of Panama. PLoS Negl Trop Dis 2016;10(5):e0004718.
[4] Miller MJ, Esser HJ, Loaiza JR, Herre EA, Aguilar C, Quintero D, et al. Molecular ecological insights into Neotropical bird–tick interactions. PLoS ONE 2016;11(5):e0155989.
[5] Loaiza JR, Dutari LC, Rovira JR, Sanjur OI, Laporta GZ, Pecor J, et al. Disturbance and mosquito diversity in the lowland tropical rainforest of central Panama. Sci Rep 2017;7(1):1–13.
[6] Loaiza JR, Rovira JR, Sanjur OI, Zepeda JA, Pecor JE, Foley DH, et al. Forest disturbance and vector transmitted diseases in the lowland tropical rainforest of central Panama. Trop Med Int Health 2019;24(7):849–61.
[7] Gittens RA, Almanza A, Bennett KL, Mejía LC, Sanchez-Galan JE, Merchan F, et al. Proteomic fingerprinting of Neotropical hard tick species (Acari: Ixodidae) using a self-curated mass spectra reference library. PLoS Negl Trop Dis 2020;14(10):e0008849.
[8] Jiménez MG, Babayan SA, Khazaeli P, Doyle M, Walton F, Reedy E, et al. Prediction of mosquito species and population age structure using mid-infrared spectroscopy and supervised machine learning. Wellcome Open Res 2019;4.
[9] Ye S, Lu S, Bai X, Gu J. ResNet-Locust-BN network-based automatic identification of east asian migratory locust species and instars from RGB images. Insects 2020;11(8):458.
[10] Guglielmone AA, Robbins RG, Apanaskevich DA, Petney TN, Estrada-Peña A, Horak IG. The hard ticks of the world. Dordrecht: Springer; 2014. doi:101007/978-94-007-7497-1.
[11] Murugaiyan J, Roesler U. MALDI-TOF MS profiling-advances in species identification of pests, parasites, and vectors. Front Cell Infect Microbiol 2017;7:184.
[12] Perez-Lao E, Sanchez-Galan J, Gittens R. Assessing the performance of different sample targets for a MALDI-TOF mass spectrometer. In: 2014 IEEE Central America and Panama Convention (CONCAPAN XXXIV). IEEE; 2014. p. 1–4.
[13] Mansilla E.C., Moreno R.C., García M.O., Sánchez B.R., Pérez J.d. D. C., Bellido J.L.M.. Aplicaciones de la espectrometría de masas maldi-tof en microbiología clínica.
[14] Levasseur M, Hebra T, Elie N, Guérineau V, Touboul D, Eparvier V. Classification of environmental strains from order to genus levels using lipid and protein MALDI-ToF fingerprintings and chemotaxonomic network analysis. Microorganisms 2022;10(4):831.
[15] Yssouf A, Almeras L, Raoult D, Parola P. Emerging tools for identification of arthropod vectors. Future Microbiol 2016;11(4):549–66.
[16] Bennett KL, Gómez Martínez C, Almanza A, Rovira JR, McMillan WO, Enriquez V, Barraza E, Diaz M, Sanchez-Galan JE, Whiteman A, et al. High infestation of invasive Aedes mosquitoes in used tires along the local transport network of Panama. Parasites Vectors 2019;12(1):1–10.
[17] Dieme C, Yssouf A, Vega-Rúa A, Berenger J-M, Failloux A-B, Raoult D, Parola P, Almeras L. Accurate identification of culicidae at aquatic developmental stages by MALDI-TOF MS profiling. Parasites Vectors 2014;7(1):1–14.
[18] Sevestre J, Lemrabott MAO, Bérenger J-M, Zan Diarra A, Ould Mohamed Salem Boukhary A, Parola P. Detection of arthropod-borne bacteria and assessment of MALDI-TOF MS for the identification of field-collected immature bed bugs from mauritania. Insects 2023;14(1):69.
[19] Bittremieux W. Spectrum_utils: a Python package for mass spectrometry data processing and visualization. Anal Chem 2019;92(1):659–61.
[20] Ràfols P, Vilalta D, Brezmes J, Cañellas N, Del Castillo E, Yanes O, et al. Signal preprocessing, multivariate analysis and software tools for MA (LDI)-TOF mass spectrometry imaging for biological applications. Mass Spectrom Rev 2018;37(3):281–306.
[21] Yssouf A, Parola P, Lindström A, Lilja T, LAmbert G, Bondesson U, et al. Identification of european mosquito species by MALDI-TOF MS. Parasitol Res 2014;113(6):2375–8.
[22] Turk M, Pentland A. Eigenfaces for recognition. J Cogn Neurosci 1991;3(1):71–86.
[23] Belhumeur P, Hespanha J, Kriegman D. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. IEEE Trans Pattern Anal Mach Intell 1997;19(7):711–20. doi:10.1109/34.598228.
[24] López-Fernández H, Santos HM, Capelo JL, Fdez-Riverola F, Glez-Peña D, Reboiro–Jato M. Mass-up: an all-in-one open software application for MALDI-TOF mass spectrometry knowledge discovery. BMC Bioinformatics 2015;16(1):1–12.
[25] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–44.
[26] Kaya M, Bilge HŞ. Deep metric learning: a survey. Symmetry (Basel) 2019;11(9):1066.
[27] Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory 1967;13(1):21–7.
[28] Rakotonirina A, Caruzzo C, Ballan V, Kainiu M, Marin M, Colot J, et al. Wolbachia detection in Aedes aegypti using MALDI-TOF MS coupled to artificial intelligence. Sci Rep 2021;11(1):1–11.
[29] Khalighifar A, Komp E, Ramsey JM, Gurgel-Gonçalves R, Peterson AT. Deep learning algorithms improve automated identification of chagas disease vectors. J Med Entomol 2019;56(5):1404–10.
[30] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–44.
[31] Liu K, Li S, Wang L, Ye Y, Tang H. Full-spectrum prediction of peptides tandem mass spectra using deep neural network. Anal Chem 2020;92(6):4275–83.
[32] LeCun Y, et al. LeNet-5, convolutional neural networks. URL: http://yann lecun com/exdb/lenet 2015;20(5):14.
[33] Dubey S.R., Singh S.K., Chaudhuri B.B.. A comprehensive survey and performance analysis of activation functions in deep learning. arXiv preprint arXiv:2109145452021;.
[34] Lewkowycz A, Gur-Ari G. On the training dynamics of deep networks with l_2 regularization. Adv Neural Inf Process Syst 2020;33:4790–9.
[35] Zhang D, Li Y, Zhang Z. Deep metric learning with spherical embedding. Adv Neural Inf Process Syst 2020;33:18772–83.
[36] Byrd J, Lipton Z. What is the effect of importance weighting in deep learning?. In: Chaudhuri K, Salakhutdinov R, editors. Proceedings of the 36th international conference on machine learning. Proceedings of Machine Learning Research, vol. 97. PMLR; 2019. p. 872–81. https://proceedings.mlr.press/v97/byrd19a.html
[37] Santurkar S, Tsipras D, Ilyas A, Madry A. How does batch normalization help optimization? Adv Neural Inf Process Syst 2018;31.
[38] Kingma D.P., Ba J.. Adam: a method for stochastic optimization. arXiv preprint arXiv:14126980 2014;.
[39] Hadsell R, Chopra S, LeCun Y. Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer society conference on computer vision and pattern recognition (CVPR'06), Vol. 2. IEEE; 2006. p. 1735–42.
[40] Buja A, Stuetzle W, Shen Y. Loss functions for binary class probability estimation and classification: structure and applications. In: Working draft, Vol. 3; 2005. p. 13. November
[41] Schroff F, Kalenichenko D, Philbin J. FaceNet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 815–23.
[42] Shan G. Monte carlo cross-validation for a study with binary outcome and limited sample size. BMC Med Inform Decis Mak 2022;22(1):1–15.
[43] Abadi M., Agarwal A., Barham P., Brevdo E., Chen Z., Citro C., Corrado G.S., Davis A., Dean J., Devin M., Ghemawat S., Goodfellow I., Harp A., Irving G., Isard M., Jia Y., Jozefowicz R., Kaiser L., Kudlur M., Levenberg J., Mané D., Monga R., Moore S., Murray D., Olah C., Schuster M., Shlens J., Steiner B., Sutskever I., Talwar K., Tucker P., Vanhoucke V., Vasudevan V., Viégas F., Vinyals O., Warden P., Wattenberg M., Wicke M., Yu Y., Zheng X.. TensorFlow: large-scale machine learning on heterogeneous systems. 2015. Software available from URL: https://www.tensorflow.org/.
[44] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020;17:261–72. doi:10.1038/s41592-019-0686-2.
[45] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825–30.
[46] Bisong E. Building machine learning and deep learning models on Google cloud platform: a comprehensive guide for beginners. Apress; 2019.
[47] Wang Y, Huang H, Rudin C, Shaposhnik Y. Understanding how dimension reduction tools work: an empirical approach to deciphering t-SNE, UMAP, TriMAP, and PaCMAP for data visualization. J Mach Learn Res 2021;22(201):1–73. http://jmlr.org/papers/v22/20-1061.html
[48] Rakotonirina A, Pol M, Kainiu M, Barsac E, Tutagata J, Kilama S, Oconnor O, Tarantola A, Colot J, Dupont-Rouzeyrol M, et al. MALDI-TOF MS: optimization for future uses in entomological surveillance and identification of mosquitoes from new caledonia. Parasites Vectors 2020;13:1–12.
[49] Yssouf A, Socolovschi C, Leulmi H, Kernif T, Bitam I, Audoly G, et al. Identification of flea species using MALDI-TOF/MS. Comp Immunol Microbiol Infect Dis 2014;37(3):153–7.
[50] Liu L, Özsu MT. Encyclopedia of database systems, Vol 6. Springer; 2009.
[51] Chicco D. Ten quick tips for machine learning in computational biology. BioData Min 2017;10(1):35.
[52] An C, Park YW, Ahn SS, Han K, Kim H, Lee S-K. Radiomics machine learning study with a small sample size: single random training-test set split may lead to unreliable results. PLoS ONE 2021;16(8):e0256152.
[53] Liu J, Gibson SJ, Mills J, Osadchy M. Dynamic spectrum matching with one-shot learning. Chemom Intell Lab Syst 2019;184:175–81.
[54] Liu J, Osadchy M, Ashton L, Foster M, Solomon CJ, Gibson SJ. Deep convolutional neural networks for raman spectrum recognition: a unified solution. Analyst 2017;142(21):4067–74.

[55] Kotsiantis S, Kanellopoulos D, Pintelas P, et al. Handling imbalanced datasets: a review. GESTS Int Trans ComputSci Eng 2006;30(1):25–36.
[56] Nabet C, Chaline A, Franetich J-F, Brossas J-Y, Shahmirian N, Silvie O, et al. Prediction of malaria transmission drivers in anopheles mosquitoes using artificial intelligence coupled to MALDI-TOF mass spectrometry. Sci Rep 2020;10(1):1–13.
[57] Rakotonirina A, Caruzzo C, Ballan V, Kainiu M, Marin M, Colot J, et al. Wolbachia detection in Aedes aegypti using MALDI-TOF MS coupled to artificial intelligence. Sci Rep 2021;11(1):1–11.