AGRICULTURE AND
NATURAL RESOURCES

Journal homepage: http://anres.kasetsart.org

Research article

Classifying DNA barcode sequences of four insects belonging to Orthoptera order using tensor network

Pradeep Bhadola[a,†], **Yash Munnalal Gupta**[b,†,*]

[a] *Centre for Theoretical Physics & Natural Philosophy, Nakhonsawan Studiorum for Advanced Studies, Mahidol University, Nakhon Sawan Campus, Phayuha Khiri, Nakhon Sawan 60130, Thailand*

[b] *Department of Biology, Faculty of Science, Naresuan University, Phitsanulok 65000, Thailand*

Article Info

Article history:
Received 8 October 2021
Revised 6 February 2022
Accepted 11 May 2022
Available online 26 August 2022

Keywords:
Cytochrome c oxidase subunit I (COI),
DNA barcode,
Machine learning,
Matrix product states (MPS),
Tensor network

Abstract

Importance of the work: Orthoptera species are one of the most rapidly increasing groups of insects being used as food and feed. However, identifying edible insects can be difficult due to their small size and the similar morphological features in closely related species. Therefore, classification of insects is often conducted by amplifying their DNA barcode sequence and comparing it with databases containing reference sequences. However, the absence of reference DNA sequences (such as cytochrome c oxidase subunit I (COI)) may confound predictions of the taxonomic community of interest and make it difficult to characterize biodiversity from DNA samples. **Objective**: To develop a quantum-inspired tensor network-based machine-learning model to categorize COI sequences for four insects belonging to the Orthoptera order. **Materials & Methods**: For alignment-free classification, each DNA barcode was represented as a tensor product of *k*-mers encoded in a *D*-dimensional space, which acts as the feature map and input for a tensor network layer for the classification. The developed model was tested with two different numbers of tensor units as well as different *k*-mer sizes. **Results**: The presented model was effective for making accurate predictions for unseen DNA barcodes and can be generalized for any DNA/RNA sequence categorization. The tensor network classifier could assign COI sequences of varying lengths to four different classes with an accuracy greater than 99% and with fewer hyper-parameters. **Main finding**: The developed model is free and publicly available through GitHub: https://github.com/yashmgupta/DNA-barcode-sequence-classification-

† Equal contribution.

* Corresponding author.
E-mail address: yashmunnalalg@nu.ac.th (Y.M. Gupta)

https://doi.org/10.34044/j.anres.2022.56.4.05

Introduction

Insect species represent the largest class within the kingdom Animalia (Grimaldi et al., 2005). They have a tremendous capacity to contribute to food and feed demand as the high percentage protein content in their bodies provides the opportunity to boost food security (Elorduy et al., 1998; Rumpold and Schlüter, 2013). However, their small body size and the similar morphology of related species pose a challenge in the taxonomy of the diverse insect species (Floyd et al., 2009). Furthermore, entomophagy is becoming very popular around the world, where insects and their body parts are used in processed food (Van Huis, 2020). Hence, it would be impossible to identify whole insects and their body parts based on morphological features in the processed food. Evidently, scientific research on DNA barcode (Hebert et al., 2003), metabarcoding (Alberdi et al., 2018), and minibarcoding (Hajibabaei et al., 2006) is resolving the problem of classifying insect species. Metabarcoding has been applied in studies to identify species used for food and for metagenomic analysis (Cuvelier et al., 2010; Staats et al., 2016; Dobrovolny et al., 2019; Xing et al., 2019). Several Orthoptera order species have been utilized as food and feed (Van Huis, 2020). Consequently, the goal of the current study was to create a machine-learning model that could classify four related insects from the Orthoptera order. In addition, it was expected to be trained using the DNA/RNA sequences of different lengths for developing a custom sequence classifier.

Cytochrome c oxidase I (COI) has been intensively used for species classification; however, species recognition can be problematic if variation emerges within primer binding sites (Elbrecht et al., 2016). Even with a wider acceptance of these strategies, a fundamental problem continues to exist, which is reliable and reproducible biodiversity characterization from mass meta-barcoding yet also influenced by research protocols (Braukmann et al., 2019). DNA barcoding is a powerful technique in food forensics for verifying organisms in food items (Paracchini et al., 2019; Xiong et al., 2019). The COI region of mitochondrial DNA is widely used for species classification. A DNA barcode based on COI region has been developed for cricket identification (Gupta et al., 2019). However, these investigations heavily rely on species identification using the DNA barcode but this does not always guarantee precision due to amplification of nonspecific sequences (Zinger et al., 2019).

Shotgun sequencing of DNA from food products can solve primer biding problems, but it also creates considerable sequencing noise (Bhargava et al., 2014). The elimination of sequencing error is a vital prerequisite for correct taxonomic analysis (Oulas et al., 2015). To overcome the shortcomings associated with current approaches, classification using machine-learning algorithms and accumulated genomic data may well be the key in this next generation of metagenomic methodologies. Recent advances in deep machine learning have resulted in an increased interest in nucleotide sequence analysis. Traditionally, nucleotide sequence classification has relied on the realignment of sequences against a pre-defined collection of target sequences through similarity algorithms (Johnson et al., 2008). These collections are massive and are skewed towards unique sequences that are not found in the reference datasets. Consequently, pre-trained models have been developed and adopted for rapid classification, such as the profile Hidden Markov Model (profile-HMM) which is also used by Pfam to group homologous sequences (Punta et al., 2012). Recently, investigators have examined the use of deep machine-learning algorithms for sequence classification using accessible barcode sequences (Nugent and Adamowicz, 2020) and also to identify viral sequences from metagenomic data (Ren et al., 2017). There is no record in the literature of a quantum machine-learning algorithm being used to classify barcode sequences. Quantum machine learning can be a powerful tool to overcome the limitations of classical machine learning (Schuld et al., 2015) and may provide a better understanding of genomic data.

Tensor network

The last few years have witnessed a huge interest in tensor-based methods in machine learning. For example, physicists are using tensor-based methods, especially tensor networks, to analyze quantum many-body systems to approximate high dimensional information in relatively low dimensions (Oseledets, 2011). Tensor networks are the decomposition of large high-dimensional tensors into networks of smaller low-rank tensors (Oseledets, 2011). Tensor networks have been widely applied in quantum computation, chemistry, machine learning, applied mathematics, physics and many other fields (Sharma and Alavi, 2015; Stoudenmire and Schwab, 2016; Han et al., 2018; Zhang et al., 2019). Tensor networks are best suited for nonlinear, kernel-based methods that require transforming the low dimensional inputs into a feature map in high dimensional space for pattern recognition (Stoudenmire and Schwab, 2016). More recently, tensor networks have been used for supervised learning, particularly for image classification (Stoudenmire and Schwab, 2016; Selvan and Dam, 2020), as well as for generative modelling (Han et al., 2018) and language modelling (Zhang et al., 2019). For supervised image classification, each pixel of the image is encoded into a high dimensional vector space and the classification is done in this high dimensional space (Stoudenmire and Schwab, 2016; Efthymiou et al., 2019). Tensor network notation makes it easier to follow the tensor operation (Penrose, 1971) (Fig. S1).

Matrix product or tensor contraction is the most common tensor operation that has been inspired by the Einstein summation convention for tensor contractions. The general rule for contracting tensors is that the two edges connected with each other imply a contraction (or summation) over the connected indices (Fig. S2). The current study used a quantum-inspired tensor network classifier based on matrix product states (MPS) for supervised learning. MPS) are the best-understood tensor networks and are also known as a tensor train, being a family of tensor networks where a large tensor with N dimensions (N edge index) is factorized as a linear chain of small rank tensors (Oseledets, 2011; Orús, 2014). The advantage of replacing a big tensor with an MPS is the reduction of the number of parameters required to specify the tensor (Orús, 2014). Apart from the reduction in the number of parameters, a tensor network based algorithms also provide linear scaling cost with training data compared to the quadratic cost for most kernel-based methods (Stoudenmire and Schwab, 2016). Standard machine-learning models are considered as black boxes in the sense that they give predictions without providing any insights on the underlying logic. On the other hand, a tensor network not only provides useful insights about data by capturing high-order correlations between features, but also offers a large number of efficient and tractable algorithms for a better understanding of the training process (Oseledets, 2011; Orús, 2014). Tensor networks offer a better interpretation of trained models due to their linear structure. Furthermore, one can estimate the entanglement entropy which characterizes the information flow in tensor networks. The entanglement also characterizes the importance of features and can be used to improve the performance of the algorithm as well as gives a better understanding of the correlations in data (Tangpanitanon et al., 2022). In this work, MPS are used for the classification of the COI region. A slight modification was made to the MPS by adding an extra classification leg (dimension) to the central tensor of the MPS. The dimension of the output edge is the classification dimension which is equivalent to the number of classes to categorize the data.

The core objective of the present research was to create a barcode sequence classification tool based on a tensor network MPS for four insects of the Orthoptera order. Therefore, the model was trained using cytochrome c oxidase subunit I (COI) sequences collected from BOLD or the barcode of life data system (Ratnasingham and Hebert, 2007). The *k*-mer features were applied and tested on input sequences for the model process. The performance of the model was tested with different *k*-mer values.

Materials and Methods

The Barcode of life data system (BOLD) was evaluated to compile all publicly accessible DNA barcode sequences (Ratnasingham and Hebert, 2007). Cytochrome oxidase subunit I (COI) sequences of the Orthoptera orders of four insects (cave cricket, grasshoppers, katydids and true crickets) were used as labels for four classes to train the tensor network. Statistical data of raw nucleotide sequences are shown in Table 1. For the classification task, the barcode sequences were limited to the range 500–700 base pairs (bp) for all classes before splitting each sequence into optimal *k*-mers. The dataset is highly imbalanced, which means that some classes have a high number of sequences whereas other classes have very few sequences. Most machine-learning algorithms perform best when the classes are balanced with nearly equal samples in each class. Imbalanced data often lead to a biased classifier, resulting in overtraining (or under training) of the model for some classes. A re-sampling technique can be utilized to correct the data imbalance. The current dataset used oversampling from the class with the lower number sequences and under-sampling from the class with the higher number of sequences. For each class, 5,500 sequences were taken as samples with lengths in the range 500–700 base pairs from the raw barcodes data. Different *k*-mer values were tested for training. All the sequences after converting to *k*-mer were padded to a length of 700 base pairs by adding zeros at the end.

Table 1 Number of Cytochrome oxidase subunit I barcode sequences obtained from the BOLD database for each order used for classifier development

Order	Raw barcodes	Minimum length	Maximum length
Cave crickets	1463	205	1731
Grasshoppers	11944	208	1743
Katydidas	7065	156	1737
True crickets	3503	219	1711

Model

Consider a COI sequence with N bases converted into a *k*-mer sequence with n *k*-mers (where *n* is the number of *k*-mers). The first step is to encode each *k*-mer into a vector in a d-dimensional space. This vector act as a feature for the classification task. Therefore, an embedding layer was used in the model to convert each *k*-mer into a d-dimensional vector. The embeddings layers were trained along with the supervised learning task. Define ψ^{xi} as the vector representation of *k*-mer at the i^{th} position. Each ψ^{xi} has *d* components, where *d* is the embedding dimension and a hyper parameter of the model. The set of the *k*-mer vectors (ψ) act as the features for the classification task. These local features ψ^{xi} are combined using the tensor product to define the feature map, as shown in Equation 1.

$$\psi^{x1,x2,\ldots,xn} = \psi^{x1} \otimes \psi^{x2} \otimes \ldots \otimes \psi^{xn} \quad (1)$$

The indices x_i run from 1 to *d*, where *d* is the embedding dimension. $\psi^{x1,x2,\ldots,xn}$ (in short denoted as Ψ(x)) is defined as the tensor product of local feature map. Ψ(x) can be viewed as a vector in a d^n-dimensional space that is an n-order tensor. The tensor network representation of Ψ(x) is shown in Fig. 1.

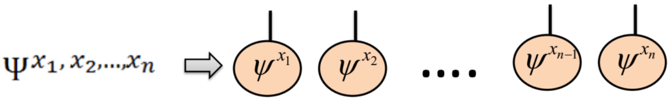

Fig. 1 Input sequence S with n *k*-mers mapped to a normalized tensor Ψ(x) of order *n* with rank 1 product structure

In the model, a COI region is represented by the tensor product of the *k*-mer vectors. The tensor Ψ(x) contains information about the interactions among the local feature vectors. Tensor networks are able to model the interaction among different *k*-mers. The tensor Ψ(x) maps each COI sequence as a vector in the d^N-dimensional feature space. For a given input COI sequence S, converted to high dimensional feature map Ψ(x), a decision function can be defined for classification, as given in Equation 2:

$$f^C(S) = W^C \Psi(x) \quad (2)$$

The input is classified into the class *c* for which $|f^C(S)|$ is largest. Here $c = 0, 1, \cdots C - 1$ are the *C* classes and W^C is an $N + 1$ (N input legs and one output leg)-order tensor with the output dimension having *c* components, each component represents a different class. The same feature map Ψ(x) is applied to all COI *k*-mer vectors; therefore, only the weight tensor W^C depends on the label. Equation 2 in the tensor notation is shown in Fig. 2. The weight tensor W^C has $N + 1$ legs out of which *N* legs take the input feature map Ψ(x) and one output leg (marked with *c*) that maps the input feature map to classes. The input legs of the weight tensor W^C are labelled with x_1, x_2 $\cdots x_N$ in Fig. 2. The weight tensor W^c has $d^N \times C$ components, where *C* is the number of classes in the classification task. This tensor grows exponentially with the size of the COI region. For efficiency, the weight tensor W^c is converted to matrix product states as shown in Fig. 3.

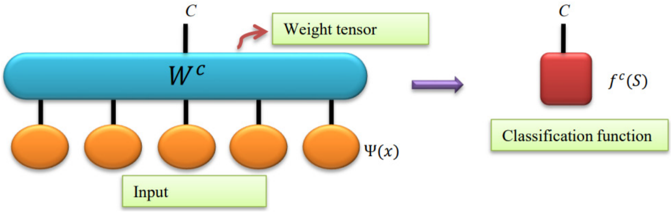

Fig. 2 Weight tensor and decision function

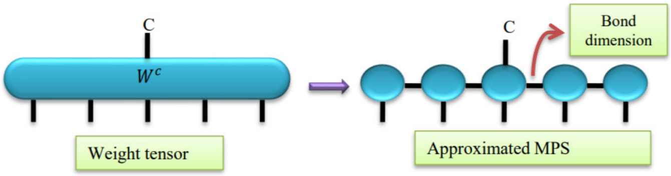

Fig. 3 Weight tensor to matrix product states (MPS)

The bond dimension of the MPS is a hyper parameter of the model. The classification model can now be represented as a tensor network. The elements of the weight MPS are not constrained and can take any real value. The classification function $f^C(S)$ is given in Equation 3:

$$f^C(S) = \langle \Psi(x) \rangle \quad (3)$$

This is simply a number for the binary classification and a vector for multiclass classification. The elements of $f^c(S)$ do not represent the probability distribution and can take any real value. The full model consists of two layers, with the first being the embedding layer to embed the COI *k*-mers into the embedding vectors and the second layer is a tensor network layer for the classification. The model is implemented using Python with TensorFlow version 2.4.1 (Abadi et al., 2016) and Google tensor network (Roberts et al., 2019) libraries.

Model training and optimization

Different optimization methods can be used to train the model. Some of the previous implementations of the tensor networks for machine learning used the density matrix renormalization group algorithm to update the MPS by minimizing the cost function (Stoudenmire and Schwab, 2016).

Herein, the gradient descent method is used to update the entries of MPS with a defined loss function. For implementation, the sparse categorical cross-entropy is used as the loss function. The training sequences are fed in batches of 50 sequences. For optimization, the Adam optimizer is used, which is an adaptive algorithm for the stochastic gradient descent method to minimize sparse categorical entropy loss. In TensorFlow version 2.4.1 (Abadi et al., 2016), the optimizer does the automatic differentiation during the back-propagation step to optimize MPS. The entire tensor network is trained for 20 epochs.

The schematic of the evaluation of the model is shown in Fig. 4. The input layer contains the *k*-mer embedding vectors. The MPS consist of random weights that are initialized for the classification task. The number for units N in the MPS and the bond dimension m connecting the MPS units are the hyper parameters for the model. The output leg of the MPS has the dimension *c*, which is equal to the number of classes in the dataset. The complete evaluation of the model contains two steps. The first step is the contraction of the input *k*-mer embedded vectors with the MPS, while the second step is the full contraction of the model to give the classifier vector of dimension *c*.

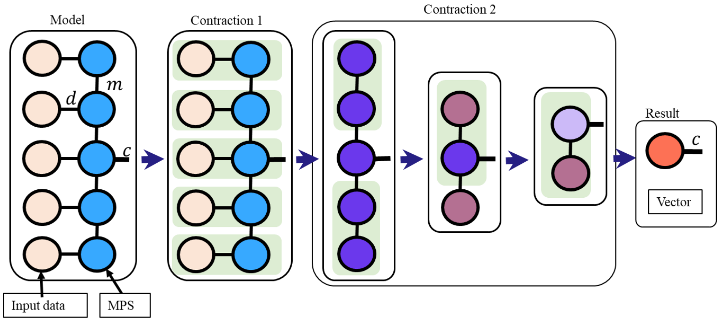

Fig. 4 Model in tensor notation and series of contraction to obtain classification function

Results and Discussion

Other reported research has shown that DNA sequences can be reliably classified using the machine-learning approach, (Weitschek et al., 2014), which prompted the current study in which a quantum machine-learning model was created that can be generalized to other biological sequences. The quantum-inspired tensor network is implemented to the supervised learning task of classifying the COI sequences of four insects from the Orthoptera order. Recently researchers have encountered difficulties in amplifying the COI region for DNA barcoding from crickets (*Tarbinskiellus* spp.; Pradit et al., 2021). A recent separation event in a species will be difficult to distinguish using the barcoding methods that rely on BLAST and a reference DNA barcode database. Researchers have emphasized that DNA barcoding should not be limited by reference sequences because only a fraction of all existing species have been identified (Austerlitz et al., 2009). In contrast, the alignment-free methods for biological sequence classification have proven to be efficient and accurate. In terms of memory utilization, a machine-learning model for sequence classification can be more efficient than an alignment-based method (Chu et al., 2020). Recently, researchers have used a machine-learning method for biological sequence classification but their method was limited for kingdom level classification (Nugent and Adamowicz, 2020). Herein, a quantum-inspired tensor network has been implemented for barcode sequence classification for four insect groups belonging to the Orthoptera order, which uses MPS for embedding the *k*-mer into embedded vectors and then performing the classification task. The model was evaluated with each sequence padded to the length of 700 base pairs. The model was trained with different *k*-mer sizes (2–5). An embedding dimension of 2 was chosen to embed the *k*-mer into embedding vectors for the classification task. After comparing the results of training and validation of MPS model with different *k*-mers and number of MPS units, a final optimum value of *k*-mer and the number of MPS units were finalized to select the final classification model. The validation accuracy, loss and validation loss from each training event, using two different numbers of units (2 and 3) and *k*-mer values in the range 2–5 are given in Table 2. The similarity scores were checked between the sequences used for training, with most of the similarity scores in the range 70–90% across different classes, whereas within the class, the similarity was sometimes as high as 99.9%. Even with the high similarity between the training sequences, the tensor network-based classifier could group the sequences with an accuracy greater than 99%. MPS model accuracy, validation accuracy, loss and validation loss with MPS units of 2 and k-mer of size 5 for 20 epochs are shown in Fig. 5 (For plots using different *k*-mer and MPS units, see Figs. S3 and S4). Researchers have used 4 *k*-mer for DNA sequence datasets for training the deep neural network and have achieved validation accuracy of 97.6 % (Nugent and Adamowicz, 2020).

Table 2 Training and validation accuracy and loss of tensor network model for 2 and 3 units, with different *k*-mers (2–5)

Number of units	*k*-mer Size	Accuracy (%)	Val accuracy (%)	Loss	Val Loss
2	1	99.4	97.3	0.018	0.100
	2	99.3	97.5	0.014	0.117
	3	99.6	97.7	0.010	0.113
	4	99.8	99.0	0.006	0.047
	5	99.9	99.4	0.002	0.025
3	1	99.7	97.7	0.008	0.120
	2	99.8	98.5	0.009	0.076
	3	99.9	98.6	0.005	0.074
	4	99.4	98.8	0.012	0.053
	5	99.8	99.0	0.010	0.043

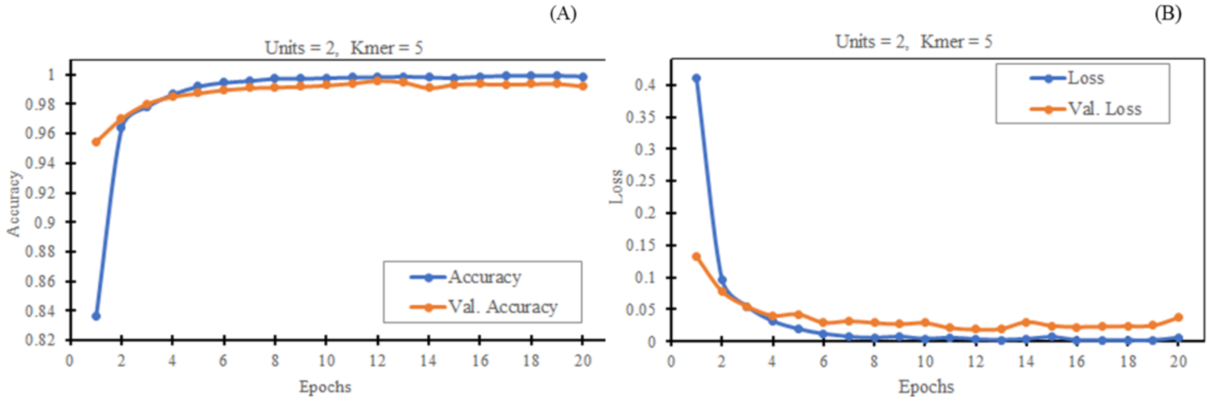

Fig. 5 (A) Training and validation accuracy; (B) loss and validation loss of tensor network model for unit = 2 and *k*-mer = 5

The test data (untrained sequences, with 200 sequences from each class in the length range 600–730, padded to the same length as the input in the model) was used to create the confusion matrix and to check the error in prediction between the classes, as shown in Fig. 6. For training, the sequence length ranges were from 500 to 700; however, the prediction for the unseen barcode can be made with variable-length sequences padded to the same length, as input to the model. The prediction accuracy of the unknown barcode sequences varied with the deviation from the training length range (500–700). It was found that the tensor network model (units 2 and 5 *k*-mer) out of the four classes could determine three, namely cave cricket, katydidas and grasshoppers, with an accuracy of more than 99%. For the class of true cricket, the model sometimes became confused with cave cricket and grasshopper but the accuracy was higher than 95%. The confusion may be attributed to the low number of sequences of cave crickets compared to other classes (see Fig. S5, for the normalized confusion matrix for prediction on the test set for TN for units 2 and 3).

The model with 3 MPS (tensor) units performed well for smaller *k*-mers (*k*-mers =1, 2, 3) and larger *k*-mers (*k*-mers = 4, 5) performance was better for 2 MPS (tensor) units. Among all *k*-mers and tensors units, the best performance in terms of training and validation accuracy was achieved with 2 MPS units and 5 *k*-mer sizes. Therefore, a 5 *k*-mer size and 2 MPS units were selected to train the final classification model (validation accuracy = 99.4%, validation loss = 2.5%). The performance metrics (micro averages) for the multiclass confusion matrix (Markoulidakis et al., 2021) for the tensor network showed that the model made the best prediction for unit size 2 with *k-mer*

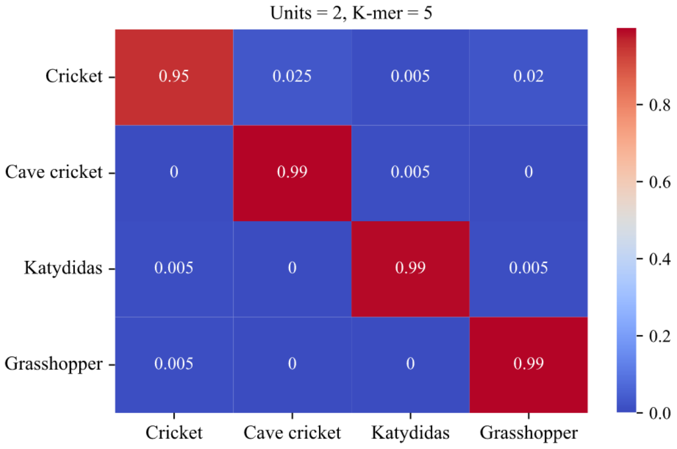

Fig. 6 Normalized confusion matrix for prediction on test set for tensor network for units = 2 and *k*-mer = 5

value 5. However, variation in performance of the model was minor with a change in hyper parameters (units and *k*-mers). The performance metrics based on the confusion matrix for all units (2 and 3) and *k*-mers (2–5) are compared in Table 3.

In summary, a quantum-inspired tensor network-based classifier created and tested for an alignment-free classification. The model takes the *k*-mer as features and maps each *k*-mer in a high dimensional feature space. The tensor network model could classify COI DNA barcode fragments at the order level with an overall accuracy greater than 99 %. The testing of the model was done using variable-length barcodes different from the training set, which ensured that the model performed to variable-length input sequences. The model was adaptive to both full-length barcodes and short sequence segments, making it an appropriate classifier for both barcode and meta-barcoding studies on edible insects. In addition, the presented model can be customized and modified for DNA or RNA fragments and supplemented with an additional default methodology for tasks, such as separating mitochondrial sequences and characterizing biodiversity from large sequencing datasets. Finally, the DNA barcode classifier presented in the paper can be modified to perform diverse sequence classification tasks. The pipeline also converts DNA sequences to appropriate length *k*-mers and provides embedding and padding for machine-learning applications.

Conflict of Interest

The authors declare that there are no conflicts of interest.

Acknowledgements

Our sincere gratitude goes to Satwika Nindya Kirana from Silpakorn University for proofreading the manuscript. YG wants to thank Department of Biology, Faculty of Science, Naresuan University, Thailand for research support “R2564C033”. PB wants to thank Mahidol University for research support “MRCMGR 04/2565”.

References

Abadi, M., Barham, P., Chen, J., et al. 2016. TensorFlow: A system for large-scale machine learning. In: Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation. Savannah, GA, USA, pp. 265–283.

Alberdi, A., Aizpurua, O., Gilbert, M.T.P., Bohmann, K. 2018. Scrutinizing key steps for reliable metabarcoding of environmental samples. Methods Ecol. Evol. 9: 134–147. doi.org/10.1111/2041-210X.12849

Austerlitz, F., David, O., Schaeffer, B., Bleakley, K., Olteanu, M., Leblois, R., Veuille, M., Laredo, C. 2009. DNA barcode analysis: A comparison of phylogenetic and statistical classification methods. BMC Bioinformatics 10: S10. doi.org/10.1186/1471-2105-10-S14-S10

Bhargava, V., Head, S.R., Ordoukhanian, P., Mercola, M., Subramaniam, S. 2014. Technical variations in low-input RNA-seq methodologies. Sci. Rep. 4: 3678. doi.org/10.1038/srep03678

Braukmann, T.W., Ivanova, N.V., Prosser, S.W., et al. 2019. Metabarcoding a diverse arthropod mock community. Mol. Ecol. Resour. 19: 711–727. doi: 10.1111/1755-0998.13008

Chu, J., Mohamadi, H., Erhan, E., Tse, J., Chiu, R., Yeo, S., Birol, I. 2020. Mismatch-tolerant, alignment-free sequence classification using multiple spaced seeds and multiindex Bloom filters. Proc. Natl. Acad. Sci. USA 117: 16961–16968. doi.org/10.1073/pnas.1903436117

Cuvelier, M.L., Allen, A.E., Monier, A., et al. 2010. Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton. Proc. Natl. Acad. Sci. USA 107: 14679–14684. doi.org/10.1073/pnas.1001665107

Dobrovolny, S., Blaschitz, M., Weinmaier, T., Pechatschek, J., Cichna-Markl, M., Indra, A., Hufnagl, P., Hochegger, R. 2019. Development of a DNA metabarcoding method for the identification of fifteen mammalian and six poultry species in food. Food Chem. 272: 354–361. doi.org/10.1016/j.foodchem.2018.08.032

Efthymiou, S., Hidary, J., Leichenauer, S. 2019. Tensornetwork for machine learning. arXiv preprint arXiv: 1906.06329. doi.org/10.48550/arXiv.1906.06329 https://github.com/google/TensorNetwork

Elbrecht, V., Taberlet, P., Dejean, T., et al. 2016. Testing the potential of a ribosomal 16S marker for DNA metabarcoding of insects. Peer J 4: e1966. doi.org/10.7717/peerj.1966

Elorduy, J.R., Pino, J.M., Correa, S.C. 1998. Edible insects of the State of Mexico and determination of their nutritional value. Anales del Instituto de Biología. Serie Zoología. 69: 65–104. [in Spanish]

Floyd, R.M., Wilson, J.J., Hebert, P.D.N. 2009. DNA barcodes and insect biodiversity. In: Foottit, R., Adlder, P. (Eds.). Insect Biodiversity: Science and Society. Blackwell Publishing. New Jersey, NJ, USA, pp. 417–431.

Table 3 Performance metrics (micro averages) for multiclass confusion matrix for tensor network with given number of units (U) and *k*-mers (K)

	U=2 K=2	U=2 K=3	U=2 K=4	U=2 K=5	U=3 K=2	U=3 K=3	U=3 K=4	U=3 K=5
Accuracy (%)	98.30	97.90	98.40	99.10	98.40	98.50	98.70	99.00
Precision (%)	96.50	95.80	96.80	98.30	96.80	97.00	97.40	98.00
Recall (%)	96.50	95.80	96.80	98.30	96.80	97.00	97.40	98.00
Specificity (%)	98.80	98.60	98.90	99.40	98.90	99.00	99.10	99.30
False negative rate (%)	3.50	4.30	3.20	1.80	3.30	3.00	2.60	2.00
False positive rate (%)	1.20	1.40	1.10	0.60	1.10	1.00	0.90	0.70

Grimaldi, D., Engel, M.S., Engel, M.S., Engel, M.S. 2005. Evolution of the Insects. Cambridge University Press. Cambridge, UK.

Gupta, Y.M., Buddhachat, K., Peyachoknagul, S., Homchan, S. 2019. Novel DNA barcode sequence discovery from transcriptome of *Acheta domesticus*: A partial mitochondrial DNA. Materials Science Forum 967: 59–64. doi.org/10.4028/www.scientific.net/MSF.967.59

Hajibabaei, M., Smith, M.A., Janzen, D.H., Rodriguez, J.J., Whitfield, J.B., Hebert, P.D. 2006. A minimalist barcode can identify a specimen whose DNA is degraded. Mol. Ecol. 6: 959–964. doi.org/10.1111/j.1471-8286.2006.01470.x

Han, Z.-Y., Wang, J., Fan, H., Wang, L., Zhang, P. 2018. Unsupervised generative modeling using matrix product states. Phys. Rev. X 8: 031012. doi.org/10.1103/PhysRevX.8.031012

Hebert, P.D., Cywinska, A., Ball, S.L., DeWaard, J.R. 2003. Biological identifications through DNA barcodes. P. Roy. Soc. B-Biol. Sci. 270: 313–321. doi.org/10.1098/rspb.2002.2218

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., Madden, T.L. 2008. NCBI BLAST: A better web interface. Nucleic Acids Res. 36: W5–W9. doi.org/10.1093/nar/gkn201

Markoulidakis, I., Rallis, I., Georgoulas, I., Kopsiaftis, G., Doulamis, A., Doulamis, N. 2021. Multiclass confusion matrix reduction method and its application on Net Promoter Score classification problem. Technologies 9: 81. doi.org/10.3390/technologies9040081

Nugent, C.M., Adamowicz, S.J. 2020. Alignment-free classification of COI DNA barcode data with the Python package Alfie. Metabarcoding and Metagenomics 4: e55815. doi: 10.3897/mbmg.4.55815

Orús, R. 2014. A practical introduction to tensor networks: Matrix product states and projected entangled pair states. Ann. Phys. 349: 117–158. doi.org/10.1016/j.aop.2014.06.013

Oseledets, I.V. 2011. Tensor-train decomposition. Siam J. Sci. Comp. 33: 2295–2317. doi.org/10.1137/090752286

Oulas, A., Pavloudi, C., Polymenakou, P., Pavlopoulos, G.A., Papanikolaou, N., Kotoulas, G., Arvanitidis, C., Iliopoulos, L. 2015. Metagenomics: Tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. Bioinform. Biol. Insights 9: 75–88. doi.org/10.4137/BBI.S12462

Paracchini, V., Petrillo, M., Lievens, A., Kagkli, D.-M., Angers-Loustau, A. 2019. Nuclear DNA barcodes for cod identification in mildly-treated and processed food products. Food Additi. Contam. Part A 36: 1–14. doi.org/10.1080/19440049.2018.1556402

Penrose, R. 1971. Applications of negative dimensional tensors. In: Welsh, D.J.A. (Ed.). Combinatorial Mathematics and Its Applications. Academic Press. London, UK, pp. 221–244.

Pradit, N., Saijuntha, W., Pilap, W., Suksavate, W., Agatsuma, T., Jongsomchai, K., Kongbuntad, W., Tantrawatpan, C. 2021. Genetic variation of *Tarbinskiellus portentosus* (Lichtenstein 1796) (Orthoptera: Gryllidae) in mainland Southeast Asia examined by mitochondrial DNA sequences. Int. J. Trop. Insect Sci. 42: 955–964. doi.org/10.1007/s42690-021-00622-4

Punta, M., Coggill, P.C., Eberhardt, R.Y., et al. 2012. The Pfam protein families database. Nucleic Acids Res. 40: D290–D301. doi.org/10.1093/nar/gkr1065

Ratnasingham, S., Hebert, P.D. 2007. Bold: The barcode of life data system (http://www. barcodinglife.org). Mol. Ecol. Notes 7: 355–364. doi: 10.1111/j.1471-8286.2007.01678.x

Ren, J., Ahlgren, N.A., Lu, Y.Y., Fuhrman, J.A., Sun, F. 2017. VirFinder: A novel *k*-mer based tool for identifying viral sequences from assembled metagenomic data. Microbiome 5: 69. doi.org/10.1186/s40168-017-0283-5

Roberts, C., Milsted, A., Ganahl, M. et al. 2019. Tensornetwork: A library for physics and machine learning. arXiv preprint arXiv: 1905.01330. https://github.com/google/TensorNetwork

Rumpold, B.A., Schlüter, O.K. 2013. Nutritional composition and safety aspects of edible insects. Mol. Nutr. Food Res. 57: 802–823. doi.org/10.1002/mnfr.201200735

Schuld, M., Sinayskiy, I., Petruccione, F. 2015. An introduction to quantum machine learning. Contemporary Physics 56: 172–185. doi.org/10.1080/00107514.2014.964942

Selvan, R., Dam, E.B. 2020. Tensor networks for medical image classification. In: Proceedings of Machine Learning Research. 121: 721–731.

Sharma, S., Alavi, A. 2015. Multireference linearized coupled cluster theory for strongly correlated systems using matrix product states. J. Chem. Phys. 143: 102815. doi.org/10.1063/1.4928643

Staats, M., Arulandhu, A.J., Gravendeel, B., Holst-Jensen, A., Scholtens, I., Peelen, T., Prins, T. W., Kok, E. 2016. Advances in DNA metabarcoding for food and wildlife forensic species identification. Anal. Bioanal. Chem. 408: 4615–4630. doi.org/10.1007/s00216-016-9595-8

Stoudenmire, E.M., Schwab, D.J. 2016. Supervised learning with tensor networks. In: Lee, d., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R. (Eds.). Advances in Neural Information Processing Systems 29. Neural Information Processing Systems Foundation, Inc. Barcelona, Spain, pp. 1–9.

Tangpanitanon, J., Mangkang, C., Bhadola, P., et al. 2022. Explainable natural language processing with matrix product states. New J. Phys. 24: 053032. doi.org/10.1088/1367-2630/ac6232

Van Huis, A. 2020. Insects as food and feed, a new emerging agricultural sector: A review. Journal of Insects as Food Feed 6: 27–44. doi.org/10.3920/JIFF2019.0017

Weitschek, E., Fiscon, G., Felici, G. 2014. Supervised DNA barcodes species classification: Analysis, comparisons and results. BioData Mining 7: 4. doi.org/10.1186/1756-0381-7-4

Xing, R.-R., Wang, N., Hu, R.-R., Zhang, J.-K., Han, J.-X., Chen, Y. 2019. Application of next generation sequencing for species identification in meat and poultry products: A DNA metabarcoding approach. Food Control 101: 173–179. doi.org/10.1016/j.foodcont.2019.02.034

Xiong, X., Yuan, F., Huang, M., Lu, L., Xiong, X., Wen, J. 2019. DNA barcoding revealed mislabeling and potential health concerns with roasted fish products sold across China. J. Food Prot. 82: 1200–1209. doi.org/10.4315/0362-028X.JFP-18-514

Zhang, L., Zhang, P., Ma, X., Gu, S., Su, Z., Song, D. 2019. A generalized language model in tensor space. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 7450–7458.

Zinger, L., Bonin, A., Alsos, I.G., et al. 2019. DNA metabarcoding-Need for robust experimental designs to draw sound ecological conclusions. Mol. Ecol. 28: 1857–1862. doi.org/10.1111/mec.15060