ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1.
Available from http://www.i6doc.com/en/.

Machine Learning and Information Theoretic Methods for Molecular Biology and Medicine

T. Villmann[1], J. Almeida[2], J.A. Lee[3], and S. Vinga[4]

1 - University of Applied Sciences Mittweida,
Saxon Institute for Comp. Intelligence and Machine Learning,
Mittweida - Germany

2 - National Institutes of Health, National Cancer Institute,
Division of Cancer Epidemiology and Genetics,
Maryland - USA

3 - UC Louvain, IREC & ICTEAM,
Brussels & Louvain-la-Neuve - Belgium

4 - Universidade de Lisboa,
Instituto Superior Técnico, Dept. Computer Science and Engineering /
Dept. Bioengineering
Lisboa - Portugal

Abstract. A short introduction to the application of information-theoretic and machine learning methods to biomolecular and medical data is provided as the motivating material that supports special session dedicated to this topic at ESANN 2022. In particular, we highlight current developments of foundation such as interpretability and model certainty. Further, we emphasize how theoretic models provide a natural framework to deal with heterogeneous and complex data structures as frequently occurring in biomedical research.

1 Introduction

Methods of artificial intelligence (AI) and machine learning models in particular increasingly contribute to critical advances in molecular medicine and biology, as well as in other branches of bio-medical research, which all have to deal with growing volume and variety of data to be processed and analyzed in an integrated manner. These approaches comprise approaches for data visualization, clustering, classification and (non-)linear regression as well as adaptive rule-based deduction systems like decision trees, as alternatives to traditional statistical methods.

As a matter of fact, data in medicine and biology often have complex structures , which raise a challenge to wrangling integrated representations. Moreover, (dis-)similarity of objects or individuals are frequently determined by external domain knowledge and respective heuristically motivated estimation procedures. The resulting similarity relations generally do not satisfy rigorous properties of (dis-)similarity measures defined in the mathematical or technical context [44]. For example, methods for alignment of protein sequences typically use substitution matrices, e.g. BLOSUM [29], with values for all possible exchanges of one amino acid with another to determine a respective alignment score evaluating

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

the similarity [45, 24]. In contrast, in epidemiology, psychology and psychotherapy, participant evaluations and classifications are usually achieved answering questionnaires [22, 41]. The contrast is even starker for digital pathology, , which plays a critical role in cancer research, where the data wrangling focus is set on analyzing and annotating images at scales ranging from cellular to whole tissues in order to associate morphological features with patient classifications [17].

This special session at ESANN 2022 will rely on specific examples to highlight methodological aspects that matter in the application of machine learning methods to molecular biology and medicine. Although the information theory discussion should be relevant to methods of machine learning in general, that aspect of the discussion will assume the use of neural networks.

2 Machine Learning in Context of Medical and Biological Applications

Currently, machine learning is largely dominated by deep learning [27] and, hence, also molecular biology and medicine applications of machine learning models are in the focus of deep learning methods [73, 72]. Those applications range from bioinformatics and computational biology [2, 59], post-genomic biology and personalized medicine [50], detection of bacterial colonies for the production of vaccines [11] up to medical imaging [7], to name just a few. The possibility of end-to-end learning of deep networks and the availability of pre-trained models for many application areas and in particular for image processing contribute to the big success of those neural networks also in biology and medicine. For a recent overview we refer to [4].

The main challenge of deep learning models is the absence or at least the difficulty of interpretation [54, 8]. Yet, in medicine and bio-medical molecular research in particular, the explanation of a trained model or the interpretability by model design remains an important component of successful applications [62, 6]. Similarly, assessing robustness and model confidence is an absolute requirement to achieve reliable and statistically consistent classifications [58, 31]. According to [38], explainable artificial intelligence (XAI) approaches comprise machine learning models which can be elaborated after training (post-hoc) by sophisticated tools, simpler surrogate models and experts in the field whereas interpretable artificial intelligence approaches comprise those models which are interpretable by design (ante-hoc) also for application domain experts. Following this work, the desidarata of explainable/interpretable models are

- explicitness and comprehensibility,
- faithfulness,
- stability,
- sparsity,
- transparency,
- model inspection,

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

which should be accompanied in the field of bio-medical research by

- data privacy

due to ethical and social reasons [16, 42]. Thereby, there is evidence that interpretable models should be preferred over explainable approaches [19, 52, 53, 75]. In particular, prototype-based machine learning models like supervised and unsupervised vector quantization belong to this class of algorithms [36, 12]. Supervised vector quantization for classification learning originates from a Bayesian perspective and has been coined as learning vector quantization (LVQ) [34]. The respective mathematical theory is provided in [56] and leads to the verification of robustness by means of implicit margin maximization during the classification learning process [20, 55, 69, 46]. The margin can be calculated for a given data also in the recall phase to evaluate the model decision confidence [32]. This proofed robustness is certainly useful also in the context of outlier detection and corresponding reject options for confident model predictions [9, 25, 28, 30, 61].

In medical applications, we frequently have to deal with imbalanced data or, even more complicated, with the so-called problem of one-class-classification (OCC), which is closely related to the reject problem [60, 47]. For example, this is the case when limited data is available for a subset of patients suffering from a particular disease. As a result, the representation of the target class for those patients is obscured. This can cause all non-target data to be rejected and/or counterfactuals have to be considered. Also for those problems, prototype-based methods seem to open a promising perspective [3, 57]. An ensemble method for outlier detection (and gene selection) for breast cancer data analysis was reported in [40].

Another important task in biomedical machine learning to be mentioned briefly here is the problem of heterogeneous and unstructured data. Respective data fusion processes have to be applied to wrangling data into such a format that machine learning algorithms can deal with them adequately [43]. For example, graph kernels and graph embedding can reflect relations between complex data objects while at the same time they can be handled efficiently in an appropriate framework [51, 15].

Last but not least, the special session will seek to explore a range of applications of machine learning models in bio-medical research that are equipped with evaluation measures and tools that assess the model certainty and confidence. This aspect should be considered as an important task to improve the acceptance of artificial intelligence by the experts in the field together with the above mentioned interpretability.

3 Information Theory and Data Representation in Context of Medical and Biological Applications

Information theory and information theoretic concepts contribute to many statistical and machine learning tasks also in medicine and molecular biology and provide a natural paradigm to process information [4, 5]. For example, the

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

information theoretic perspective on medical diagnostic inference is considered in [23] and its relevance for clinical diagnostic testing is investigated in [10], whereas their applications in computational biology and quantitative genetics are discussed in [18] and [26], respectively.

Considered broadly, one can distinguish foundational information theoretic concepts for data proximity evaluation (potentially leading to distance metrics) with statistical inference on the one hand and information theory based learning concepts in machine learning applied for bio-molecular and medical data analysis tasks on the other hand. The latter topic is explicitly addressed in [48, 49]. Yet, (deep) multi-layer perceptrons with cross-entropy loss for classification can also be seen as an information theoretic approach for consistent classification learning, For interpretable LVQ we may refer to the probabilistic variant proposed in [63].

The utilization of information theoretic concepts for the investigation of the origin of life was suggested in [74] and was successfully applied to distinguish RNA-virus signatures from structured artificial sequences by means of their information content estimated by the Kolmogorov complexity as reported in [64].

In molecular sequence analysis, several concepts of information theory inspired and innovative sequence coding schemes were proposed such that the transformed molecular sequence data can be compared and processed by machine learning tools. In this field, the molecular sequences can be seen as sequences over a finite alphabet of symbols (e.g. amino acids). The relevant approaches frequently belong to the family of alignment-free methods, which entail a lower computational cost than alignment-based methods [14]. Prominent examples are the representation of DNA-sequences by means of the chaos game approach, which was studied in [1, 68, 67, 66] constituting a bijective scale-free projection of sequences to the continuous numeric space. Specifically, it provides a sequence encoding procedure from which information profiles at multiple degrees (at any L-tuple length) can be more effectively processed. Such a unifying information domain comes with the promise of a comprehensive AI framework where, for example, digital pathology and biomolecular data can share machine learning methods.

The representation of sequences in protein analysis by so-called natural vectors using a sequence description in terms of the higher moments of the sequence alphabet elements with respect to their occurrences and co-occurrences provides another method based on information-theoretic statistics for molecular sequence encoding [71]. A generalization of this approach is the mutual information function [37], which can be used for efficient sequence comparison [13].

In this introduction to the special ESANN 2022 session, the representation of symbolic sequence data will be highlighted as a particular challenge, and also as an opportunity, for the use of AI on genomics data.

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

4 Conclusion and Outlook

In this short introductory and motivational paper we revisited and highlighted some aspects of machine learning and the application of information theoretic models and concepts in the context of medicine and molecular biology. In this special session of ESANN 2022, we will try to see beyond the methods that underpin the tremendous recent success of deep learning, in order to explore the promise of new approaches emerging in current research. Specifically, the combination of information theoretic concepts as intuitive models with the design of interpretable networks may offer new insights in machine learning and, frequently in the context of medical and molecular-biologic research, has to deal with complex data structures and incomplete information. That research context especially values methods and concepts that are robust, transparent, and information-preserving. Moreover, it is beneficial for the domain experts if knowledge acquired by the model for decision making is provided to them as users.

Finally, the expectation of interpretability/explainability is often compounded with the requirement that the computational platforms advancing machine learning in biology and medicine can facilitate collaborations across fields of expertise – preferably as experts in the loop [39] or/and by guided (visual) inspections of the results [21], where (data-) privacy is absolutely critical. This last aspect of the session is undergoing major advances with the development of distributed learning systems, e.g. multi-source learning algorithms or federated learning approaches [35, 65, 33, 70].

References

[1] J.S. Almeida, J.A. Carrico, A. Maretzek, P.A. Noble, and M. Fletcher. Analysis of genomic sequences by chaos game representation. *Bioinformatics*, 17(5):429–437, 2001.

[2] C. Angermueller, L. Parts T. Pärnamaa, and O. Stegle. Deep learning for computational biology. *Molecular Systems Biology*, 12(7):878:1–16, 2016.

[3] A. Artelt and B. Hammer. Efficient computation of counterfactual explanations and counterfactual metrics of prototype-based classifiers. *Neurocomputing*, 470:304–327, January 2022.

[4] D. Bacciu, G. Lisboa, and A. Vellido, editors. *Deep Learning in Biology and Medicine*. World Scientific, 2022.

[5] D. Bacciu, P. Lisboa, A. Sperduti, and T. Villmann. *Springer Handbook of Computational Intelligence*, chapter Probabilistic Modeling in Machine Learning, pages 545–575. Springer, Berlin-Heidelberg, 2015.

[6] A. Barragán-Montero, A. Bibal, M. Huet Dastarac, C. Draguet, G. Valdés, D. Nguyen, S. Willems, L. Vandewinckele, M. Holmström, F. Löfman, K. Souris, E. Sterpin, and J.A. Lee. Towards a safe and efficient clinical implementation of machine learning in radiation oncology by exploring model interpretability, explainability and data-model dependency. *Physics in Medicine and Biology*, 2022.

[7] A. Barragán-Montero, U. Javaid, G. Valdés, D. Nguyen, P. Desbordes, B. Macq, S. Willems, L. Vandewinckele, M. Holmström, F. Löfman, S. Michiels, K. Souris, E. Sterpin, and J.A. Lee. Artificial intelligence and machine learning for medical imaging: a technology review. *Physica Medica*, 83:242–256, 2021.

[8] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R., R. Chatila, and F. Herrera. Explainable artificial

intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.

[9] P. L. Bartlett and M.H. Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9:1823–1840, 2008.

[10] W.A. Benish. A review of the application of information theory to clinical diagnostic testing. *Entropy*, 22(e22010097):1–20, 2020.

[11] T. Beznik, P. Smyth, G. de Lannoy, and J.A. Lee. Deep learning to detect bacterial colonies for the production of vaccines. *Neurocomputing*, 470:427–431, 2022.

[12] M. Biehl, B. Hammer, and T. Villmann. Prototype-based models in machine learning. *Wiley Interdisciplinary Reviews: Cognitive Science*, 7(2):92–111, 2016.

[13] K.S. Bohnsack, M. Kaden, J. Abel, S. Saralajew, and T. Villmann. The resolved mutual information function as a structural fingerprint of biomolecular sequences for interpretable machine learning classifiers. *Entropy*, 23(10):1357–1378, 2021.

[14] K.S. Bohnsack, M. Kaden, J. Abel, and T. Villmann. Alignment-free sequence comparison: A systematic survey from a machine learning perspective. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, page accepted, 2022.

[15] K.S. Bohnsack, M. Kaden, J. Voigt, and T. Villmann. Efficient classification learning of biochemical structured data by means of relevance weighting for sensoric response features. In M. Verleysen, editor, *Proceedings of the 30th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2021), Bruges (Belgium)*, page in this volume, Louvain-La-Neuve, Belgium, 2022. i6doc.com.

[16] A. Boulemtafes, A. Derhab, and Y. Challal. Privacy-preserving deep learning for pervasive health monitoring: a study of environment requirements and existing solutions adequacy. *Haelth and Technology*, 12:285–304, 2022.

[17] G. Campanella, M.G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K.J. Busam, E. Brogiand V.E. Reuter, D.S. Klimstra, and T. Fuchs. Clinical grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, 2019.

[18] Pritam Chanda, Eduardo Costa, Jie Hu, Shravan Sukumar, John Van Hemert, and Rasna Walia. Information theory in computational biology: Where we stand today. *Entropy*, 22(6), 2020.

[19] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J.K. Su. This looks like that: Deep learning for interpretable image recognition. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 8930–8941. Curran Associates, Inc., 2019.

[20] K. Crammer, R. Gilad-Bachrach, A. Navot, and A.Tishby. Margin analysis of the LVQ algorithm. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing (Proc. NIPS 2002)*, volume 15, pages 462–469, Cambridge, MA, 2003. MIT Press.

[21] I. Diaz-Blanco, J.M. Enguita-Gonzalez, D. Garcia-Perez, A. Gonzalez-Muniz, A.A. Cuadrado-Vega, M.D. Chiara-Romero, and N. Valdes-Gallego. Interactive dual projections for gene expression analysis. In M. Verleysen, editor, *Proceedings of the 30th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2021), Bruges (Belgium)*, page in this volume, Louvain-La-Neuve, Belgium, 2022. i6doc.com.

[22] J. M. Digman. Personality structure: Emergence of the five–factor model. *Annual Review of Psychology*, 41:417–440, 1990.

[23] Nathaniel A. Eiseman, Matt T. Bianchi, and M. Brandon Westover. The information theoretic perspective on medical diagnostic inference. *Hospital Practice*, 42(2):125–138, 2014.

[24] Da-Fei Feng and Russell F. Doolittle. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of Molecular Evolution*, 25(4):351–360, August 1987.

[25] L. Fischer, B. Hammer, and H. Wersing. Efficient rejection strategies for prototype-based classification. *Neurocomputing*, 169:334–342, 2015.

[26] D.J. Galas, J. Kunert-Graf, L. Uechi, and N.A. Sakhanenko. Towards an information theory of quantitative genetics. *Journal of Computational Biology*, 28(6):527–559, 2021.

[27] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, Cambridge, MA,

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

2016.

[28] M.E. Hellman. The nearest neighbor classification rule with a reject option. *IEEE Transactions on Systems Science and Cybernetics*, 6:179–185, 1970.

[29] S. Henikoff and J.G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. of the National Academy of Science USA*, 89:10915–10919, 1992.

[30] R. Herbei and M.H. Wegkamp. Classification with reject option. *The Canadian Journal of Statistics*, 34(4):709–721, 2006.

[31] R. Hu, S.J. Delany, and B. MacNamee. Sampling with confidence: Using k-NN confidence measures in active learning. In *Proceedings of the UKDS Workshop at 8th International Conference on Case-based Reasoning (ICCBR 09)*, pages 181–192, 2009.

[32] M. Kaden, K.S. Bohnsack, M. Weber, M. Kudla, K. Gutowska, J. Blazewicz, and T. Villmann. Learning vector quantization as an interpretable classifier for the detection of SARS-CoV-2 types based on their RNA-sequences. *Neural Computing and Applications*, pages 1–26, 2021.

[33] G.A. Kaissis, M.R. Makowski, D. Rückert, and R.F. Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2:305–311, 2020.

[34] Teuvo Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, 1995. (Second Extended Edition 1997).

[35] J. Li, W. Wu, D. Xue, and P. Gao. Multi-source deep transfer neural network algorithm. *Sensors*, 19(18):3992–4008, 2019.

[36] O. Li, H. Liu, C. Chen, and C. Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 3531–3537. Association for the Advancement of Artificial Intelligence, AAAI Press, 2018.

[37] W. Li. Mutual information functions versus correlation function. *Journal of Statistical Physics*, 60(5/6):823–837, 1990.

[38] P. Lisboa, S. Saralajew, A. Vellido, and T. Villmann. The coming of age of interpretable and explainable machine learning models. In M. Verleysen, editor, *Proceedings of the 29th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2021), Bruges (Belgium)*, pages 547–556, Louvain-La-Neuve, Belgium, 2021. i6doc.com.

[39] M. Littmann, K. Selig, L. Cohen-Lavi, Y. Frank, P. Hönigschmid, E. Kataka, A Mösch, K. Qian, A. Ron, S. Schmid, A. Sorbie, L. Szlak, A. Dagan-Wiener, N. Ben-Tal, M.Y. Niv, D.Razansky, B.W. Schuller, D. Ankerst, T. Hertz, and B. Rost. Validity of machine learning in biology and medicine increased through collaborations across fields of expertise. *Nature Machine Intelligence*, 2:18–24, 2020.

[40] M.B. Lopes, A. Veríssimo, E. Carrasquinha, S. Casimiro, N. Beerenwinkel, and S. Vinga. Ensemble outlier detection and gene selection in triple-negative breast cancer data. *BMC Bioinformatics*, 19(168):1–15, 2018.

[41] L. Luborsky, J. Barber, and P. Schaffer. The assessment of the CCRT: comparison of tailor–made with standard category rating scales on a specimen case. In *Progress in Assessing Psychodynamic Functioning: A Comparison of Four Methods on a Single Case*. Society of Psychotherapy Research, Toronto, 1989.

[42] B. Murdoch. Privacy and artificial intelligence: challenges for protecting health information in a new era. *BMC Medical Ethics*, 22(122):1–5, 2021.

[43] E. Mwebaze, G. Bearda, M. Biehl, and D. Zühlke. Combining dissimilarity measures for prototype-based classification. In M. Verleysen, editor, *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2015)*, pages 31–36, Louvain-La-Neuve, Belgium, 2015. i6doc.com.

[44] D. Nebel, M. Kaden, A. Villmann, and T. Villmann. Types of (dis−)similarities and adaptive mixtures thereof for improved classification learning. *Neurocomputing*, 268:42–54, 2017.

[45] S.B. Needleman and C.D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970.

[46] B. Paaßen and T. Villmann. Prototype selection based on set covering and large margins.

Machine Learning Reports, 14(MLR-03-2021):35–42, 2021.

[47] P. Perera, P. Oza, and V.M. Patel. One-class classification: A survey. arXiv, 2021.

[48] J.C. Principe. *Information Theoretic Learning*. Springer, Heidelberg, 2010.

[49] M. Rao, S. Seth, J. Xu, Y. Chen, H. Tagare, and J.C. Principe. A test of independence based on a generalized correlation function. *Signal Processing*, 91:15–27, 2011.

[50] A. Ray. Machine learning in postgenomic biology and personalized medicine. *Wiley interdisciplinary reviews on data mining and knowledge discovery*, 12(e1451):1–26, 2022.

[51] K. Riesen and H. Bunke. *Graph Classification and Clustering Based on Vector Space Embedding*, volume 77 of *Machine Perception and Artificial Intelligence*. World Scuentific, Singapore, 2010.

[52] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[53] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges, 2021.

[54] W. Samek, G. Monatvon, A. Vedaldi, L.K. Hansen, and K.-R. Müller, editors. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, number 11700 in LNAI. Springer, 2019.

[55] S. Saralajew, L. Holdijk, M. Rees, E. Asan, and T. Villmann. Classification-by-components: Probabilistic modeling of reasoning over a set of components. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, pages 2788–2799. MIT Press, 2019.

[56] A. Sato and K. Yamada. Generalized learning vector quantization. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8. Proceedings of the 1995 Conference*, pages 423–9. MIT Press, Cambridge, MA, USA, 1996.

[57] D. Staps, R. Schubert, M. Kaden, A. Lampe, W. Hermann, and T. Villmann. Prototype-based one-class-classification learning using local representations. In *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN) - Padua*, page in press, Los Alamitos, 2022. IEEE Press.

[58] E. Sterpin, S.T. Rivas, F. Van den Heuvel, B. george, J.A. Lee, and K. Souris. Development of robustness evaluation strategies for enabling statistically consistent reporting. *Physics in Medicine and Biology*, 66(045002):1–17, 2021.

[59] B. Tang, Z. Pan, K. Yin, and A. Khateeb. Recent advances of deep learning in bioinformatics and computational biology. *Frontiers in Genetics*, 10(214):1–10, 2019.

[60] D. Tax. *One-Class-Classification*. PhD thesis, University Delft, 2001.

[61] A. Vailaya and A.K. Jain. Reject option for VQ-based Bayesian classification. In *International Conference on Pattern Recognition (ICPR)*, pages 2048–2051, 2000.

[62] A. Vellido. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Network and Applications*, 32:18069–18083, 2020.

[63] A. Villmann, M. Kaden, S. Saralajew, and T. Villmann. Probabilistic learning vector quantization with cross-entropy for probabilistic class assignments in classification learning. In L. Rutkowski, R. Scherer, M. Korytkowski, W. Pedrycz, R. Tadeusiewicz, and J.M. Zurada, editors, *Proceedings of the 17th International Conference on Artificial Intelligence and Soft Computing - ICAISC, Zakopane*, LNCS 10841, pages 736–749, Cham, 2018. Springer International Publishing, Switzerland.

[64] T. Villmann, M. Kaden, S. Wasik, M. Kudla, K. Gutowska, A. Villmann, and J. Blazewicz. Searching for the origins of life – Detecting RNA life signatures using learning vector quantization. In A. Vellido, K. Gibert, C. Angulo, and J.D.M. Guerrero, editors, *Advances in Self-Organizing Maps, Learning Vector Quantization, Clustering and Data Visualization – Proceedings of the 13th International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization, WSOM+2019, Barcelona*, volume 976 of *Advances in Intelligent Systems and Computing*, pages 324–333, Berlin-Heidelberg, 2019. Springer.

[65] T. Villmann, D. Staps, J. Ravinchandran, S. Saralajew, M. Biehl, and M. Kaden. A learning vector quantization architecture for transfer learning based classification by means of nullspace evaluation. In T. Bouadi, E. Fromont, and E. Hüllermeier, editors, *Advances in*

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

Intelligent Data Analysis XX – Proceedings of the 20th Symposium on Intelligent Data Analysis (IDA 2022), Lecture Notes in Computer Science, pages 354–364. Springer, 2022.

[66] S. Vinga. Information theory applictions for biological sequence analysis. *Bioinformatics*, 15(3):376–389, 2014.

[67] S. Vinga and J.S. Almeida. Rényi continuous entropy of DNA sequences. *Journal of Theoretical Biology*, 231:377–388, 2004.

[68] S. Vinga and J.S. Almeida. Local Rényi entropic profiles of dna sequences. *BMC Bioinformatics*, 8(393):1–19, 2007.

[69] V. Voráček and M. Hein. Provably adversarially robust nearest prototype classifiers. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceeding of Machine Learning Research (PMLR)*, pages 1–8, 2022.

[70] Q. Wang and Y. Zhou. FedSPL: federated self-paced learning for privacy-preserving disease diagnosis. *Briefings in Bioinformatics*, 23(1):1–12, 2022.

[71] Y. Wang, K. Tian, and S. Yau. Proteine sequence classification using natural vectors and the convex hull method. *Journal of Computational Biology*, 26(4):315–321, 2019.

[72] S. Webb. Deep learning for biology. *Nature*, 554:555–557, 2018.

[73] C. Xu and S.A. Jackson. Machine learning and complex biological data. *Genome Biology*, 20(76):1–4, 2019.

[74] H.P. Yockey. *Information Theory, Evolution, and the Origin of Life*. Cambridge University Press, 2011.

[75] J. Zeng, B. Ustun, and C. Rudin. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society, Series A*, 180(0964–1998/17/180000):1–34, 2017.

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.