Inherently Interpretable Machine Learning

In recent years, machine learning (ML), especially deep neural networks (DNNs), have been intensively studied and applied to many scientific and industrial sectors where intelligent processing is required, leading to superior, sometimes unprecedented performance. Despite the extraordinary success, the interpretability of ML systems, especially the black-box nature of state-of-the-art (SOTA) ML architectures, has posed a big challenge, causing concerns about questionable performances and predictions in real applications. To address this black-box problem, interpretable ML (I-ML) has recently drawn considerable attention in the ML community. While most ML models experience certain level of black-box design, the consensus is that the contemporary neural network (NN)-based models, e.g., convolutional neural network (CNN), exhibit less interpretable characteristics, thus attracting more attention from both academia and industry. A plethora of publications on I-ML have been made available to the ML and intelligent processing communities, mostly focusing on using feed forward NN (FFNN)-based or DNN-based to explain the internal structure of a black- box. While acknowledging progress along this line of research, emphasis of this paper will be given to the class of models which are designed to be inherently interpretable from analytically inspired perspectives, especially those integrating statistics guided optimization (SGO) with NN architecture, coined as SGO-NN. The class of SGO-NN models features three distinct characters: a) Kolmogorov-Arnold (K-A) theorem as the foundation; b) the incorporation of certain neurobiological facts in architecture design; c) powerful optimization methods in the training process by solving SGO problems in the convolutional layers. Analytically, the recent progress in approximation theory has solidly verified K-A theorem under mild conditions. At the same time, numerous practical SGO-NN models with three or fewer layers have emerged and demonstrated their flexibility, effectiveness, and computational affordability. Amongst this branch of I-ML models, those based on canonical correlation analysis (CCA) stand out, demonstrating tremendous potential to address the interpretability challenge in ML research. To validate the power of the SGO-NN models, practical examples in text-image representation, facial analysis, and object recognition are presented. It is expected that the SGO-NN models and the inherently I-ML models in general would better inspire researchers and practitioners in the pursuit of powerful interpret ML models in their R&D endeavor.

Keywords: *machine learning, interpretability, Kolmogorov-Arnold (K-A) theorem, statistics guided optimization, neurobiological facts*

Introduction

Recently, machine learning (ML), especially deep neural networks (DNNs) and artificial intelligence (AI) in general, have been successfully utilized in a broad rand of applications, such as audio recognition, visual computing, video processing, image retrieval, amongst others [1-2]. Nonetheless, the interpretability of ML/AI becomes a persistent challenge. Specifically, the black-box nature of

contemporary ML architectures has posed a longstanding challenging problem, causing concerns about questionable performances and predictions in real applications. In order to address this black-box problem, interpretable ML (I-ML) methods have drawn considerable attention and interests [4]-[5]. As consensus suggests [6]-[7], the classical neural network (NN)-based models (e.g., neural network, convolutional neural network (CNN) and DNNs in general) eexhibit less interpretable characteristics, thus attracting more attention from both academic and industrial sectors, first attempting to explain the black-box and, more recently, designing new models that are inherently interpretable.

Although NN-based models stem from Kurt-Vladimir (K-V) Universal approximation (UA)theory [8]-[9], research into DNNs has dominated the landscape for the past 10 years in visual computing, natural language processing, video processing, and more [10]-[11].It is known that most deep learning (DL)-based models utilize the end-to-end architecture, which makes the DNN-based representations a black-box [12]-[13], implying that it is difficult to tell what the prediction relies on, and what features or representations play more important roles in a given task. As a result, the ultimate goal of studying interpretability is to construct the model architecture, which is inherently interpretable to avoid the black-box problem [14].

The core of this paper focuses on finding relationships either contained in the data or learned by the ML model. Several survey papers have been made available [15]-[19] to the ML and intelligent processing communities, mostly attempting to explaining the internal structure of a black-box. While this paper will touch recent advances along this line of research, emphasis will be given to another class of models which are designed to be inherently interpretable from analytically or mathematically inspired perspectives. Evaluation the I-ML models and comparisons with representative models pertinent to multi-modal image and multimedia analysis and recognition will be presented.

NN Based Methods

Recently, DNN-based methods have achieved great success and outperform humans in numerous difficult tasks, such as visual classification, natural language recognition, and video processing. However, the black-box nature of the contemporary methods presents a real challenge to understand mechanisms and behaviors of the networks. Essentially, for this class of methods, the word ‘interpretability’ refers to the ability to clarify and extract knowledge representations in different layers of NN-based methods as defined in [20]. In this section, the most studied methods in this class, feedforward neural network (FFNN) based and DL based, are surveyed.

FFNN based Methods

In the 1980’s, FFNN was already employed to interpret and design NNs and other networks [15]. Kuo *et al.* [21] designed an interpretable feedforward (FF)

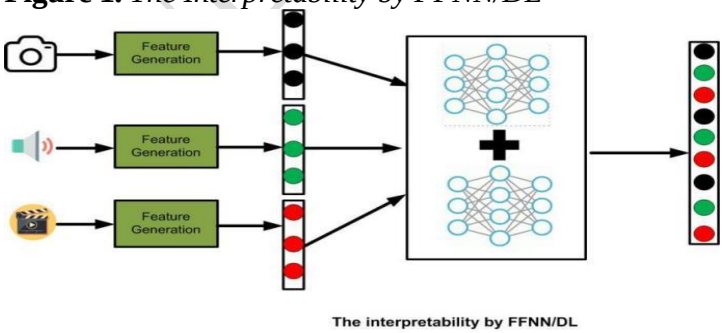
model by utilizing a data-centric strategy. As a result, the parameters in the current layer are able to be derived from the previous layer in a one-pass manner. Yosinski *et al.* [22] investigated the activation values of neurons in different layers according to different types of input data. Based on the experimental results, the live activation values are helpful for understanding inner mechanisms of networks, leading to an interpretable model.

DNN based Methods

Lately, a great number of works based on pure DNNs have been introduced, forming the mainstream tactics in identifying explainability of DNNs, especially for CNNs. Zhang *et al.* [20] introduced interpretable CNNs to clarify knowledge representations in high convolution layers, which aid in the understanding of intrinsic logic inside a CNN architecture. In [23], a CNN-INTE solution is presented and applied to explain deep CNNs. By employing global interpretation for any given samples in the hidden layers, the CNN-INTE is able to explain the inner mechanisms of CNN-based models. In [24], a prototype layer was proposed. With the extra prototype layer, the involved DL-based model is capable of generating several prototypes for different parts of the input samples, resulting in appropriate interpretation. In [25], the Locality Guided Neural Network (LGNN) method is presented and applied to explainable artificial intelligence (XAI). Since LGNN is able to preserve locality between neighbouring neurons within each layer of a deep network, it is able to alleviate the black-box nature of current AI methods.

The schematic diagram of the interpretability by FFNN/DNN is drawn in Figure 1. It is known the power of the reviewed methods on I-ML is confined by certain limitations such as the vanishing/exploding gradient problem and tuning of parameters manually. In order to address these limitations, some researchers and practitioners investigate the model interpretability from alternative angles, leading to inherently I-ML methods/approaches.

Figure 1. *The Interpretability by FFNN/DL*



Inherently I-ML Methods

In this section, coined as inherently interpretable models, this class of I-ML

obeys structural knowledge of different domains, e.g., monotonicity, causality, structural (generative) constraints, or physical constraints that come from domain knowledge and can, at least, be partially justified by theoretical analysis such as mathematical expressions and/or physics laws [26]. The representative members include physics-informed, model based, algorithm unrolling solutions and mathematics inspired methods. More detailed information is given as following subsections.

Physics-informed NN

Physics-informed NN based algorithms are mainly utilized to deal with the supervised learning tasks while respecting any given laws of physics is described by nonlinear partial differential equations [27]. In [28], a physics-informed NN model was employed to address two problems in ML: data-driven solution and data-driven discovery of partial differential equations, resulting in satisfying performance in computational science. In [29], a survey paper on physics-informed NN based models was published. The research emphasis was given on customizing this class of models through gradient optimization techniques, NN structures, and loss functions in ML.

Model based NN

Studies on interpretability of model-based NN mainly focus on the construction of models that readily provide insight into the relationships they have learned [30]. A model based NN framework was created for image reconstruction in [31]. Based on this framework, a systematic approach was introduced, producing an interpretable DNN model for different image applications. In [32], a model based NN was presented for optimized sampling and reconstruction. Benefiting from the combination of continuous optimization of the sampling pattern and the CNN parameters, it is able to improve image quality to some extent, generating certain levels of interpretability in ML.

Algorithm Unrolling

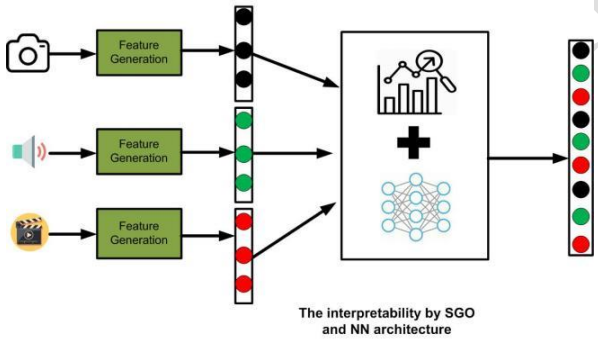
In I-ML studies, algorithm unrolling handles model interpretability by providing a concrete and systematic connection between iterative algorithms [33]. Koo *et al.* [34] proposed a Bayesian based unrolling algorithm for single-photon Lidar systems. Profiting from the integration statistical and learning based frameworks, it resulted in improved network interpretability. In [35], a graph unrolling network algorithm was presented with application to signal denoising, leading to an interpretation of the architecture design in ML.

Mathematics Inspired Methods

By integrating Statistics Guided Optimization (SGO) with NN architecture, this class of models, coined as SGO-NN, exhibits model agnostic properties and is ideal for global model interpretability. Essentially, the SGO-NN architecture is

designed according to three characters: a) Kolmogorov-Arnold (K-A) theorem [36] and K-V UA theory [8]-[9] as the foundation; b) biological justifications and scientific rules in architecture design; c) powerful optimization methods for a quality training process. Analytically, the recent progress in approximation theory solidly verified the K-A theorem/KV UA theory that three hidden layers are sufficient for a NN to approximate any nonlinear functions under mild conditions [37]. In addition, a great number of practical models with three or fewer layers have emerged and demonstrated their flexibility [38][69], effectiveness and computational affordability in I-ML. Examples consist of PCANet [39], DCTNet [40], CCANet [41], DDCCANet [42][43], ILMMHA [44], etc. Note that, several representative members of this model class, such as CCANet, DDCCANet and ILMMHA, are particularly prevalent to achieve the task of information processing by mimicking certain facts in neurobiological systems, handling multiple information streams coherently and simultaneously [45-46]. Apparently, such an architecture fits well with multimodal information processing, in which two or more different data sources are processed jointly. The schematic diagram of such a network architecture is given in Figure 2.

Figure 2. *The Interpretability by SGO and NN Architecture*



Exemplar Applications

In this section, the performances of different I-ML methods are evaluated on several applications, including cross-modal (text-image)-based and multi-view visual-based (face recognition and recognition of objects) examples. The involved algorithms/models are classified into three categories, a) without I-ML (WO-I-ML), b) with contemporary NN (C-NN), and c) SGO-NN.

Cross-modal (Text-image) Recognition

The Wiki Database

There are 2,866documents stored in text-image pairs and associated with supervised semantic labels of 10classes in the Wiki database. In our experiments, the total samples are divided into a training subset with 2173samples and a testing subset with 693 samples as practiced in past studies [47]-[50]. The SGO-NN model, ILMMHA, is applied to two different classical features (bag-of-visual

SIFT (BOV-SIFT) [49] and the Latent Dirichlet Allocation (LDA) [50]). Then, the accuracies by different types of methods are given in **Table 1.**

Table1. *Recognition accuracies on the Wiki database*

Methods	**Training Number**	**Accuracy**	**Type**
L21CCA[47]	2173	65.99%	WO-I-ML
MH-DCCM[48]	2173	67.10%	WO-I-ML
RE-DNN[49]	2173	63.95%	CNN
ILMMHA[44]	2173	74.28%	SGO-NN

Visual Examples

Face Recognition-The ORL Database

In this paper, we conduct experiments on the Olivetti Research Lab (ORL) database for face recognition. There are 40 people with 10different images for each subject, leading to 400 samples in total. In this experiment, all 400samples are used with 280images randomly selected as the training subset while the remaining samples utilized as the testing subset. The SGO-NN models are performed on two-view datasets (the original image and local binary patterns (LBP)-based image). The experimental results are shown in **Table 2**.

Table 2. *Recognition accuracies on the ORL database*

Methods	**Training Number**	**Accuracy**	**Type**
ESP[51]	280	96.00%	WO-I-ML
DL-SE[52]	280	96.08%	WO-I-ML
HMMFA[53]	280	94.17%	WO-I-ML
CNN[54]	280	95.92%	CNN
IKLDA+PNN[55]	280	96.35%	CNN
LiSSA[56]	280	97.51%	CNN
PCANet[39]	280	96.28%	SGO-NN
CCANet[40]	280	97.92%	SGO-NN
DDCCANet [42]	280	98.50%	SGO-NN

Recognition-The ETH-80 Database

The ETH-80 data set includes 3280 color RGB images. In this work, all images are normalized at a size of 64 × 64 pixels. During the evaluation, 1640 images are randomly chosen for training while the remaining samples are utilized in testing. Two raw data sources (R and G sub-channel images) are adopted as the two inputs for SGO-NN models. Then, recognition accuracies in three different

categories are tabulated in **Table 3**.

Table3. *Recognition accuracies on the ETH-80database*

Methods	**Training Number**	**Accuracy**	**Type**
SSL-TR[57]	1640	93.40%	WO-I-ML
SRC+DPC[58]	1640	94.00%	WO-I-ML
SML[59]	1640	94.02%	WO-I-ML
RMML[60]	1640	94.25%	WO-I-ML
TLRDA+PCA[61]	1640	92.00%	CNN
CMCM[62]	1640	92.50%	CNN
AlexNet [63]	1640	94.20%	CNN
CCANet[40]	1640	93.98%	SGO-NN
DDCCANet [42]	1640	94.40%	SGO-NN

Object Recognition-The Caltech256 Database

In the Caltech256 database, there are different images with a varying set of illumination, movements, backgrounds, etc. Totally, there are 256 classes and one background class. For fair comparison, the same settings used in other studies are adopted. Specifically, 60 images are chosen from each class as training samples. A relatively simple DNN architecture, VGG-19, is employed to extract DL-based features, which serve as the input to the SGO-NN models. The recognition rates are reported in **Table 4**.

Table4. *Recognition accuracies on the Caltech256database*

Methods	**Training Number**	**Accuracy**	**Type**
CMFA-SR[64]	15420	76.31%	WO-I-ML
LLKc[65]	15420	75.36%	WO-I-ML
Fine-tuning[66]	15420	83.80%	CNN
SMNN[67]	15420	84.70%	CNN
TransTailor[68]	15420	87.30%	CNN
CCANet[40]	15420	87.82%	SGO-NN
DDCCANet [42]	15420	88.34%	SGO-NN

According to the above aforementioned results, it clearly shows that both I-ML branches work well in the evaluated data sets, with SGO-NN having a slight edge in the three visual examples and being substantially better in cross-model recognition. The results evidently verify the necessity of integrating interpretability in ML, showing the effectiveness of inherently I-ML.

Note, the current version of the DDCCANet code has been uploaded to GitHub (https://github.com/09liukai08/GADDCCANet). Interested readers can download and test the code following the readme instructions. Feedbacks are very welcome!

Conclusions

This paper provides a survey on the inherently interpretable machine learning (I-ML). Performance and comparison on the collected exemplar applications indicate that interpretable I-ML methods had evidently led to performance gains. Moreover, methodology fusion of SGO principles and NN architecture (SGO-NN) moves studies on I-ML towards the next level, better satisfying human demands.

References

[1] M.I. Jordan, and T.M. Mitchell. "Machine learning: Trends, perspectives, and prospects." *Science*, vol. 349, no.6245, pp. 255–260, 2015.

[2] S. Das, N. Agarwal, D. Venugopal, F.T. Sheldon, and S.Shiva. "Taxonomy and Survey of Interpretable Machine Learning Method." *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 670–677, 2020.

[3] W. Zhu, X. Wang, and W. Gao. "Multimedia intelligence: When multimedia meets artificial intelligence." *IEEE Transactions on Multimedia*, vol. 22, no. 7, pp. 1823–1835, 2020.

[4] C. Molnar, G. Casalicchio, and B. Bischl. "Interpretable machine learning: a brief history, state-of-the-art and challenges." *European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 417–431, 2020.

[5] A. Adadi, and M. Berrada. "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)." *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[6] T. Adali, R.C. Guido, T.K. Ho, K. Mller, and S. Strother. "Interpretability, Reproducibility, and Replicability." *IEEE Signal Processing Magazine*, vol. 39, no. 4, pp. 5–7, 2022.

[7] Q. Teng, Z. Liu, Y. Song, K. Han, and Y. Lu. "A survey on the interpretability of deep learning in medical diagnosis." *Multimedia Systems*, vol. 28, pp. 2335-2355, 2022.

[8] K. Hornik, M. Stinchcombe, and H. White. "Multilayer feedforward networks are universal approximators." *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

[9]A. Kolmogorov. "On the representation of continuous functions of several variables by superpositions of continuous functions of a smaller number of variables." *Proceedings of the USSR Academy of Sciences*, vol. 108, pp. 179–182, 1956.

[10] A. Khan, A. Sohail, U. Zahoora, and A.S. Qureshi. "A survey of the recent architectures of deep convolutional neural networks." *Artificial Intelligence Review*, vol. 53, no. 8, pp. 5455–5516, 2020.

[11] Y. LeCun, Y. Bengio, and G. Hinton. "Deep learning." *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[12] V. Buhrmester, D. Munch, and M. Arens. "Analysis of explainers of black box deep neural networks for computer vision: A survey." *arXiv preprint arXiv:1911.12116*, 2019.

[13] T. Langner, R. Strand, H. Ahlstrom, and J. Kullberg. "Large-scale biometry with

interpretable neural network regression on uk-biobank body MRI." *Scientific reports*, vol. 10, no. 1, pp. 1–9, 2020.

[14] G. Xu, T.D. Duong, Q. Li, S. Liu, and X. Wang. "Causality learning: A new perspective for interpretable machine learning." *arXiv preprint arXiv:2006.16789*, 2020.

[15] X. Huang, D. Kroening, W. Ruan, J. Sharp, Y. Sun, E. Thamo, M. Wu, and X. Yi. "A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability." *Computer Science Review*, vol. 37, pp. 1–35, 2020.

[16] F.L. Fang, J. Xiong, M. Li, and G. Wang. "On interpretability of artificial neural networks: A survey." *IEEE Transactions on Radiation and Plasma Medical Sciences*, 2021 (Preprint).

[17] Q. Zhang, and S.C. Zhu. "Visual interpretability for deep learning: a survey." *arXiv preprint arXiv:1802.00614*, 2018.

[18] Y. Zhang, P. Tino, A. Leonardis, and K. Tang. "A survey on neural network interpretability." *arXiv preprint arXiv:2012.14261*, 2020.

[19] L.H. Gilpin, D. Bau, B.Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. "Explaining explanations: An overview of interpretability of machine learning." *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pp. 80–89, 2018.

[20] Q. Zhang, Y.N. Wu, and S.C. Zhu. "Interpretable convolutional neural networks." *2018 IEEE CVPR*, pp. 8827–8836, 2018.

[21] C-C.J. Kuo, M. Zhang, S. Li, J. Duan, and Y. Chen. "Interpretable convolutional neural networks via feedforward design." *Journal of Visual Communication and Image Representation*, vol. 60, pp. 346–359, 2019.

[22] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson. "Understanding neural networks through deep visualization." *arXiv preprint arXiv:1506.06579*, 2015.

[23] X. Liu, X. Wang, and S. Matwin. "Interpretable deep convolutional neural networks via meta-learning." *2018 IJCNN*, pp. 1–9, 2018.

[24] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su. "This looks like that: deep learning for interpretable image recognition." *2019 NIPS*, pp. 1–12, 2019.

[25] R. Tan, N. Khan, and L. Guan. "Locality guided neural networks for explainable artificial intelligence." *2020 IJCNN*, pp. 1–8, 2020.

[26] C. Rudin. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[27] G.E. Karniadakis, I.G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang. "Physics-informed machine learning." *Nature Reviews Physics*, vol. 3, no. 6 (2021): 422-440.

[28] M.Raissi, P. Perdikaris, and G.E. Karniadakis. "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations." *Journal of Computational physics*, vol. 378, pp. 686–707, 2019.

[29] S. Cuomo, V.S.D. Cola, F. Giampaolo, G. Rozza, M. Raissi, and F. Piccialli."Scientific machine learning through physics-informed neural networks: where we are and whats next." *Journal of Scientific Computing*, vol. 92, no. 3, pp. 1–62, 2022.

[30] W.J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu. "Definitions, methods, and applications in interpretable machine learning." *Proceedings of the National Academy of Sciences*, vol. 116, no. 44, pp. 22071–22080, 2019.

[31] H.K. Aggarwal, M.P. Mani, and M. Jacob. "MoDL: Model-based deep learning

architecture for inverse problems." *IEEE transactions on medical imaging*, vol. 38, no. 2, pp. 394–405, 2018.
[32] H.K. Aggarwal, and M. Jacob. "J-MoDL: Joint model-based deep learning for optimized sampling and reconstruction." *IEEE journal of selected topics in signal processing*, vol. 14, no. 6, pp. 1151–1162, 2020.
[33] V. Monga, Y. Li, and Y.C. Eldar. "Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing." *IEEE Signal Processing Magazine*, vol. 38, no. 2, pp. 18–44, 2021.
[34] J. Koo, A. Halimi, and S. McLaughlin. "A Bayesian based deep unrolling algorithm for single-photon Lidar systems." *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 4, pp. 762–774, 2022.
[35] S. Chen, Y.C. Eldar and L. Zhao. "Graph unrolling networks: Interpretable neural networks for graph signal denoising." *IEEE Transactions on Signal Processing*, vol. 69, pp. 3699–3713, 2021.
[36] V. Kurkova. "Kolmogorov's theorem is relevant." *Neural computation*, vol. 3, no. 4, pp. 617–622, 1991.
[37] Z. Shen, H. Yang, and S. Zhang. "Neural network approximation: Three hidden layers are enough." *Neural Networks*, vol. 141, pp. 160–173, 2021.
[38] Z. Qin, F. Yu, C. Liu, and X. Chen. "How convolutional neural networks see the world—A survey of convolutional neural network visualization methods." *Mathematical Foundations of Computing*, vol. 1, no. 2, pp. 149–180, 2018.
[39] T.H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng and Y. Ma. "PCANet: A simple deep learning baseline for image classification?" *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5017–5032, 2015.
[40] C.J. Ng, and A.B.J. Teoh. "DCTNet: A simple learning-free approach for face recognition." *2015 APSIPA*, pp. 761–768, 2015.
[41] X. Yang, W. Liu, D. Tao, and J. Cheng. "Canonical correlation analysis networks for two-view image recognition." Information Sciences, vol. 385, pp. 338–352, 2017.
[42] L. Gao, Z. Guo, L. Guan. "A Distinct Discriminant Canonical Correlation Analysis Network based Deep Information Quality Representation for Image Classification." *2020 ICPR*, pp. 7595–7600, 2020.
[43] L. Gao, Z. Guo, and L. Guan. "ODMTCNet: An Interpretable Multi-view Deep Neural Network Architecture for Image Feature Representation." *arXiv preprint arXiv:2110.14830*, 2021.
[44] L. Gao and L. Guan. "Interpretable learning-based multi-modal hashing analysis for Multi-view Feature Representation Learning." *IEEE MIPR*, pp. 1–6, 2022.
[45] B. Widrow, and J.C. Aragon. "Cognitive memory." *Neural Networks*, vol. 41, pp. 3–14, 2013.
[46] B.Widrow, A. Greenblatt, Y. Kim, and D. Park. "The no-prop algorithm: A new learning algorithm for multilayer neural networks." *Neural Networks*, vol. 37, pp. 182–188, 2013.
[47] M. Xu, Z. Zhu, X. Zhang, Y. Zhao, and X. Li. "Canonical correlation analysis with L2,1-Norm for multiview data representation." *IEEE transactions on cybernetics*, vol. 50, no. 11, pp. 4772–4782, 2020.
[48] L. Gao, and L. Guan. "A Discriminative Vectorial Framework for Multi-modal Feature Representation." *IEEE transactions on Multimedia*, vol. 24, pp. 1503–1514, 2022.
[49] K. Nguyen, D. Le, and D.A. Duong. "Efficient traffic sign detection using bag of visual words and multi-scales sift." *2013 NIPS*, pp. 433–441, 2013.
[50] D.M. Blei, A.Y. Ng, and M.I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research*, vol. 3, no. 1, pp. 993-1022, 2003.

[51] W. Wei, H. Dai, and W. Liang. "Exponential sparsity preserving projection with applications to image recognition." *Pattern Recognition*, vol. 104, pp. 1-11, 2020.
[52] Y. Zhang, W. Liu, H. Fan, Y. Zou, Z. Cui, and Q. Wang. "Dictionary learning and face recognition based on sample expansion." *Applied Intelligence*, vol. 52, no. 4, pp. 3766–3780, 2022.
[53] S. Zhao, W. Liu, S. Liu, J. Ge, and X. Liang. "A hybrid-supervision learning algorithm for real-time uncompleted face recognition." *Computers and Electrical Engineering*, vol. 101, pp. 1–15, 2022.
[54] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "ImageNet classification with deep convolutional neural network." *2012 NIPS*, pp. 10971105, 2012.
[55] A. Ouyang, Y. Liu, S. Pei, X. Peng, M. He, and Q. Wang. "A hybrid improved kernel LDA and PNN algorithm for efficient face recognition." *Neurocomputing*, vol. 393, pp. 214222, 2019.
[56] T. Wang, W.W. Ng, M. Pelillo and S. Kwong. "LiSSA: localized stochastic sensitive autoencoders." *IEEE Transactions on Cybernetics*, vol. 51, no. 5, pp. 2748–2760, 2019.
[57] W. Yan, Q. Sun, H. Sun, and Y. Li. "Semi-Supervised Learning Framework Based on Statistical Analysis for Image Set Classification." *Pattern Recognition*, vol. 107, pp. 1–13, 2020.
[58] K. Sharma, and R. Rameshan. "Image Set Classification Using a Distance-Based Kernel Over Affine Grassmann Manifold." *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 3, pp. 1082–1095, 2020.
[59] E. Vural, and C. Guillemot. "A study of the classification of low-dimensional data with supervised manifold learning." *Journal of Machine Learning Research*, vol. 18, no. 1, pp. 5741–5795, 2017.
[60] N. Sogi, L.S. Souza, B.B. Gatto, and K. Fukui. "Metric learning with a-based scalar product for image-set recognition." *IEEE/CVF CVPR Workshops*, pp. 850–851, 2020.
[61] J. Zhang, Z. Li, P. Jing, Y. Liu, and Y. Su. "Tensor-driven low-rank discriminant analysis for image set classification." *Multimedia Tools and Applications*, vol. 78, no. 4, pp. 4001–4020, 2019.
[62] N. Sogi, T. Nakayama, and K. Fukui. "A method based on convex cone model for image-set classification with cnn features." *2018 IJCNN*, pp. 1–8, 2018.
[63] X. Wu, Y. Wang, H. Tang, and R. Yan. "A structure–time parallel implementation of spike-based deep learning." *Neural Networks*, vol. 113, pp. 72–78, 2019.
[64] A. Puthenputhussery, Q. F. Liu, and C. J. Liu. "A sparse representation model using the complete marginal fisher analysis framework and its applications to visual recognition." *IEEE Transactions on Multimedia*, vol. 19, no. 8, pp. 1757–1770, 2017.
[65] Q. F. Liu, and C. Liu. "A novel locally linear KNN method with applications to visual recognition." *IEEE transactions on neural networks and learning systems*, vol. 28, no. 9, pp. 2010–2021, 2017.
[66] W. Ge, and Y. Yu. "Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning." *IEEE CVPR*, pp. 1086–1095, 2017.
[67] D. Wang, and K. Mao. "Learning Semantic Text Features for Web Text-Aided Image Classification." *IEEE Transactions on Multimedia*, vol. 21, no. 12, pp. 2985–2996, 2019.
[68] B. Liu, Y. Cai, Y. Guo, and X. Chen. "TransTailor: Pruning the pre-trained model for improved transfer learning." *2021 AAAI*, vol. 35, no. 10, pp. 8627–8634, 2021.
[69] L. Gao and L. Guan. "Interpretability of Machine Learning: Recent Advances and Future Prospects." *arXiv preprint arXiv:2305.00537*, 2023.