DC-NAS: Divide-and-Conquer Neural Architecture Search for Multi-Modal Classification

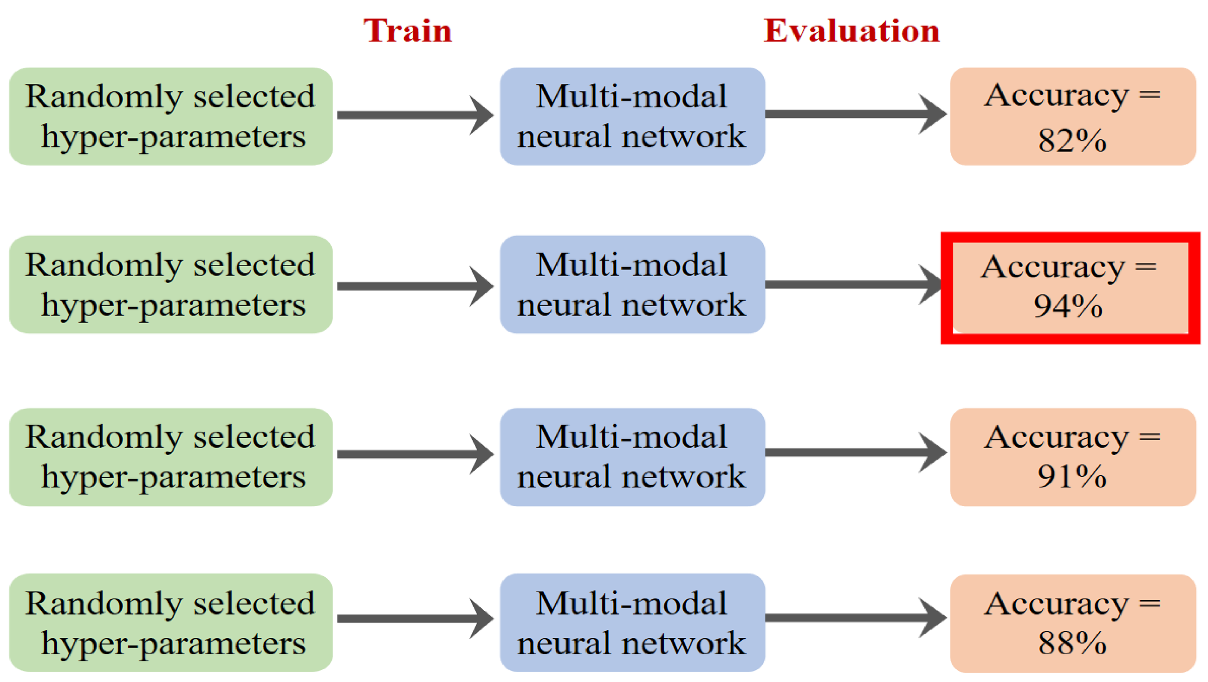
Xinyan Ling[1], Pinhan Fu[1], Qian Guo[2], Keyin Zheng[1], Yuhua Qian[1,*]

[1] Institute of Big Data Science and Industry, Shanxi University, China,

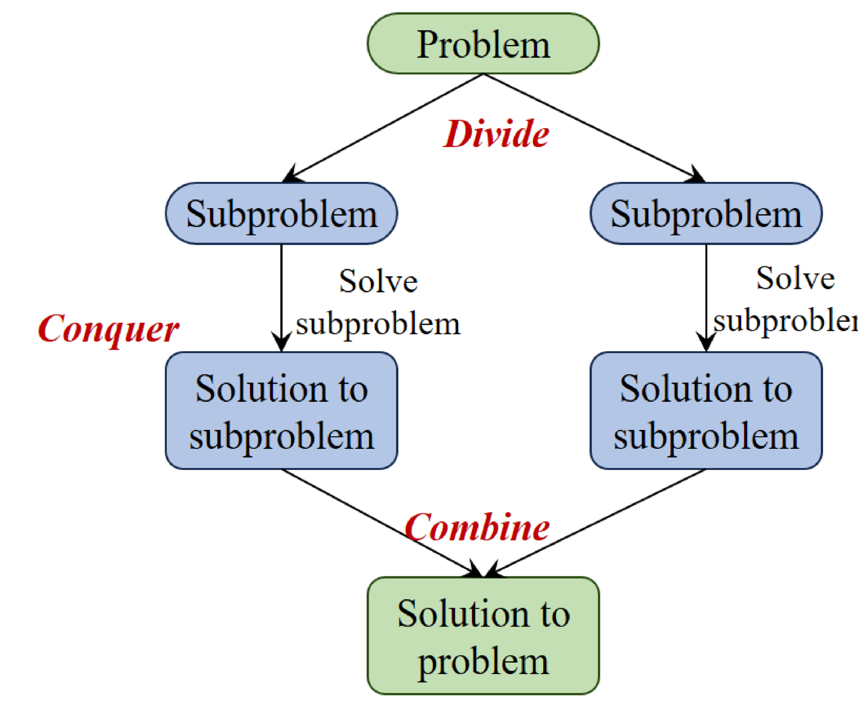
[2] School of Computer Science and Technology, Taiyuan University of Science and Technology, China. Email: liangxinyan48@163.com

Introduction

Challenge: Despite the encouraging results achieved by existing multi-modal NAS methods in various multi-modal tasks, for example, MMTM[1], MFAS[2], and BM-NAS[3], most methods require to train ***a large number of multi-modal neural networks*** in each update step, often consuming more time than non-NAS methods.

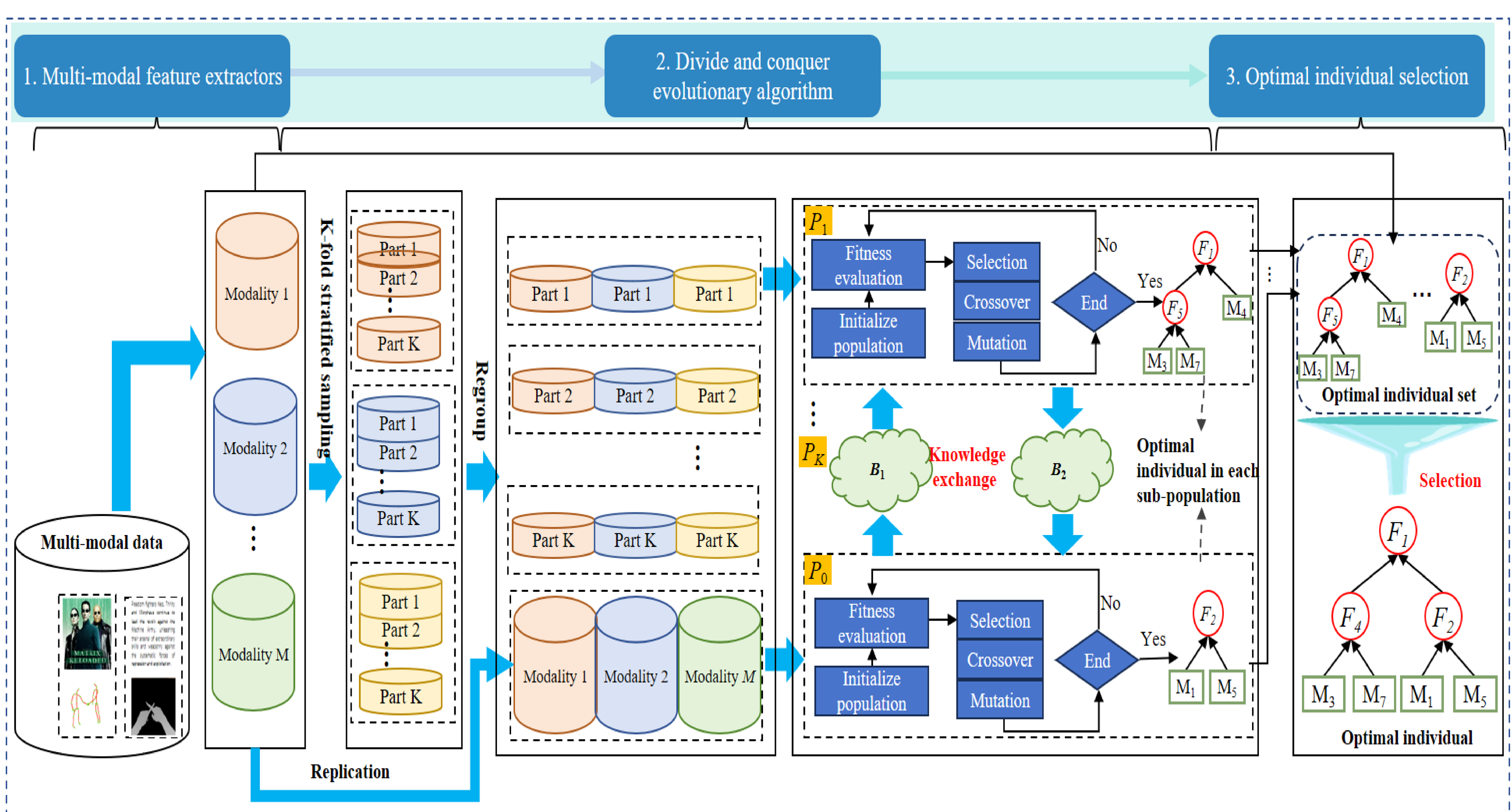


Solution: Propose a population-based multi-modal NAS method called Divide-and-Conquer Neural Architecture Search (DC-NAS), which exhibits high computational efficiency and scalability to large search spaces. DC-NAS can effectively adapt to various multi-modal feature fusion strategies and learn DNN architectures to handle different multi-modal classification tasks.



Advantages: ADC-NAS where most individuals evolve with the partial data, only few individuals evolve with the entire data, and knowledge is allowed to exchange between them achieves the comparable performance with one where all individuals evolve with the entire data. This design theoretically and empirically reduces the computation time.

Proposed Method



Uinmodal feature extraciton: $X = \{(X_1(s_i), X_2(s_i), X_v(s_i), y_i)\}_{i=1}^{n}$. Here, $X_j(s_i)$ represents the j-th feature representation extracted from the multi-modal dataset.

Primitive operations: Feature fusion operator set $\boldsymbol{F}$ including five basic fusion operators: *concatenation, addition, multiplication, max, average*.

Main steps of the DC-NAS framework include *population initialization, fitness evaluation, offspring generation,* and *selection*.

- **Population initialization**: A population P with M individuals is randomly generated, and then divide it into $K+1$ sub-populations.
- **Fitness evaluation:** Each individual is decoded into a multi-modal classification model, which is then trained and evaluated using the corresponding sub-dataset.

Experiments

➢Performance evaluation

Method	Modality	F1-W(%)
Unimodal methods		
Maxout MLP (ICML13)	Text	57.54
VGG Transfer (ICLR15)	Image	49.21
Multi-modal methods		
Two-stream (NIPS14)	Image + Text	60.81
GMU (ICLR17)	Image + Text	61.70
CentralNet (ECCV18)	Image + Text	62.23
MFAS (CVPR19)	Image + Text	62.50
BM-NAS (AAAI22)	Image + Text	62.92±0.03
DC-NAS (ours)	Image + Text	**63.70**±0.11

Table1: Multi-label genre classification results on MM- IMDB dataset

Method	Modality	Acc (%)
Unimodal methods		
Inflated ResNet-50 (CVPR18)	Video	83.91
Co-occurence (IJCAI18)	Pose	85.24
Multi-modal methods		
Two-stream (NIPS14)	Video + Pose	88.60
GMU (ICLR17)	Video + Pose	85.80
MMTM (CVPR20)	Video + Pose	88.92
CentralNet (ECCV18)	Video + Pose	89.36
MFAS (CVPR19)	Video + Pose	89.50±0.60
BM-NAS (AAAI22)	Video + Pose	90.48±0.24
DC-NAS (ours)	Video + Pose	**90.85±0.05**

Table 2: Action recognition results on NTU RGB-D dataset

Method	Modality	Acc (%)
VGG-16 + LSTM (CVPR17)	RGB + Depth	81.40
C3D + LSTM + RSTTM	RGB + Depth	92.20
I3D (CVPR17)	RGB + Depth	92.78
MMTM (CVPR20)	RGB + Depth	93.51
MTUT (3DV19)	RGB + Depth	93.87
3D-CDC-NAS2 (TIP21)	RGB + Depth	94.38
BM-NAS (AAAI22)	RGB + Depth	94.96±0.07
DC-NAS (ours)	RGB + Depth	**95.22±0.05**

Table 3: Gesture recognition results on EgoGesture dataset

Method	Dataset	Parameters	Time	CP (%)
MMTM	NTU	8.61M	-	88.92
MFAS	NTU	2.16M	603.64	89.50
BM-NAS	NTU	0.98M	53.68	90.48
DC-NAS(ours)	NTU	**0.26M**	**13.63**	**90.85**
BM-NAS	Ego	0.61M	20.67	94.96
DC-NAS(ours)	Ego	**0.19M**	**4.57**	**95.22**
BM-NAS	MM-IMDB	0.65M	1.24	62.94
DC-NAS(ours)	MM-IMDB	**0.42M**	**1.19**	**63.70**

Table 4: Comparison of model size, search cost, and performance (CP)

- DC-NAS outperforms state-of-the-art multi-modal methods comprehensively in terms of parameters, efficiency, and performance.

➢ Ablation Study

Feature selection strategies	ACC (%)
Random	88.81±0.11
Late fusion	89.47±0.07
Searched (MFAS)	89.50±0.60
Searched (BM-NAS)	90.48±0.24
Searched (DC-NAS)	**90.85±0.05**

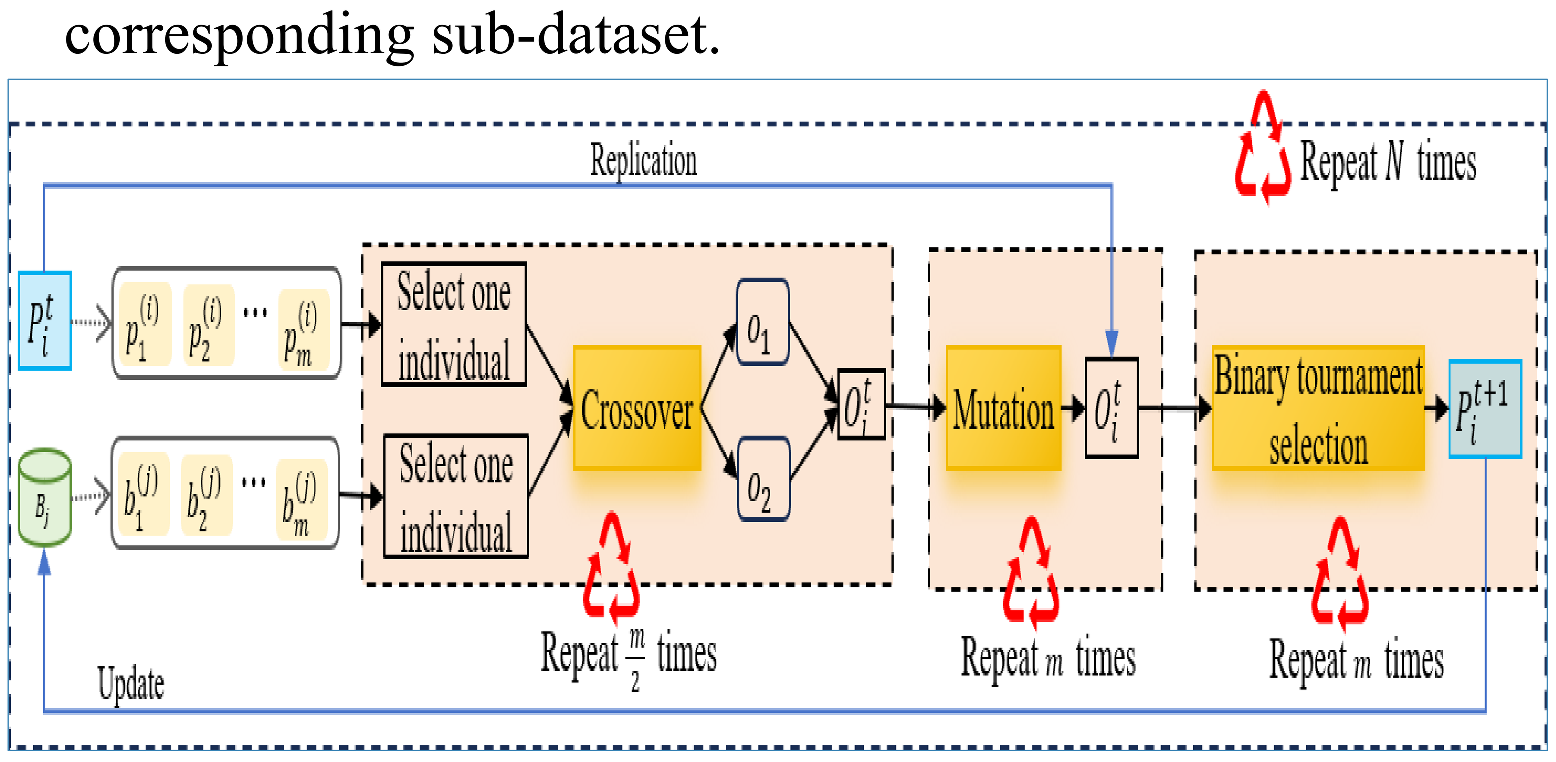
Table 7: Impact Analysis on Fusion Strategy

Version	DCE	KT	Time	ACC (%)
$DC\text{-}NAS_1$	False	False	20.67	90.86±0.03[8.0e-01]
$DC\text{-}NAS_2$	True	False	11.10	90.52±0.06[9.5e-06]
DC-NAS	True	True	13.63	90.85±0.05

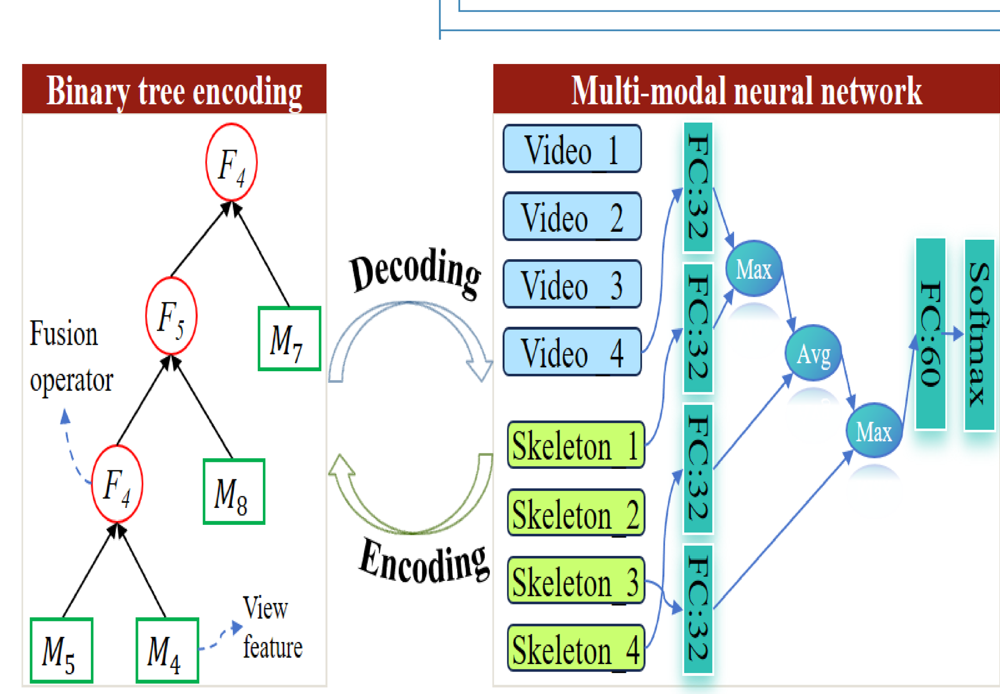
Table 5: Impact Analysis of Each Component of DC-NAS

Add	Mul	Cat	Max	Avg	DC-NAS
89.54	88.71	89.20	88.84	88.07	**90.85**

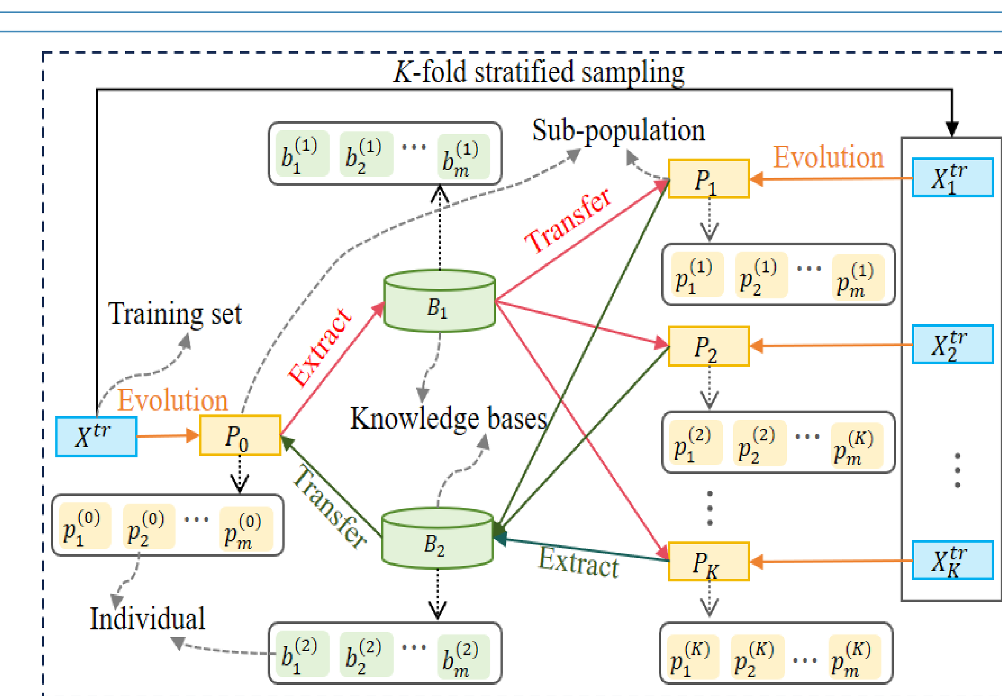
Table 6: Impact Analysis on Feature Selection Strategies



Crossover, mutation and selection



Model encoding and decoding



Relationship among variables

References

[1] Vaezi Joze, H. R.; Shaban, A.; Iuzzolino, M. L.; and Koishida, K. 2020. MMTM: Multimodal Transfer Module for CNN Fusion. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition,13286–13296.

[2] Perez Rua, J.; Vielzeuf, V.; Pateux, S.; Baccouche, M.; and Jurie, F. 2019. MFAS: Multimodal Fusion Architecture Search. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6959–6968.

[3] Yin, Y.; Huang, S.; Zhang, X.; and Dou, D. 2022. BM-NAS: Bilevel Multimodal Neural Architecture Search. In Association for the Advancement of Artificial Intelligence, 8901–8909.