

Editorial

Dear Colleague:

Welcome to volume 28(5) of the Intelligent Data Analysis (IDA) Journal.

Dear reader, welcome to the fifth issue of IDA's 28th year. For this issue, we have prepared a selection of papers covering different theoretical and applied topics in the field of Intelligent Data Analysis.

The first part of the issue includes theoretical and methodological contributions, the second part presents application papers in different areas, and we have included a third part of the issue to include three papers that address support methods, interactive/exploratory visualization and other techniques.

We open the issue with an interesting review by Malthora and Cherukuri, in which the authors explore the impact of hyperparameter tuning techniques on software quality prediction models, highlighting the importance of optimizing hyperparameters for improved model performance. The study identifies that tuning the parameters of classification algorithms can significantly enhance the predictive capability of software quality models, with some algorithms showing high sensitivity to tuning while others do not require extensive parameter adjustments.

The following four papers complete the theoretical and methodological contributions that are part of this issue. First, in Hu et al., the authors introduce a novel nonlinear classification network model called PEDCC that utilizes evenly-distributed class centroids to enhance classification performance. By maximizing inter-class distance and minimizing intra-class distance, PEDCC achieves effective nonlinearity while providing an analytical solution for network weights. The model incorporates techniques like PCA, CReLU activation function, and a multi-stage network structure based on latent feature norm to improve classification performance, showcasing advantages in training speed and recognition accuracy especially on small datasets. The second paper, by Yu, et al., introduces a novel double granularity filtering method to improve process discovery from business process logs by detecting and filtering noises at both event and trace levels. By analyzing directly-following and parallel event relations to identify infrequent behaviors, redundant events, and missing events, this approach enhances the understandability of discovered models. Experimental results on synthetic and real-life datasets show that the proposed method surpasses existing techniques, enhancing the performance of noise removal and model extraction from process logs. The third of these theoretical contributions is Shi et al.'s paper that proposes a novel TextCNN-based Two Ways Active Learning model (TCTWAL) that combines global and local features to enhance active learning effectiveness in natural language processing tasks. By utilizing the maximum normalized log-probability (MNLP) for query strategy and leveraging TextCNN with minimal hyper-parameter tuning, the proposed model outperforms manual-designed instance query strategies in terms of accuracy, precision, recall, and F1 score on text classification tasks, showing significant improvements after 39 iterations on the AG's News corpus. We finalize the theoretical contributions with Ji Zhang et al. who present a novel efficient ViT model called Dual-Granularity Former (DGFormer) designed to address limitations in lightweight Vision Transformers (ViTs) for resource-constrained devices. By incorporating Dual-Granularity Attention and Efficient Feed-Forward Network modules, DGFormer outperforms existing models on tasks like ImageNet image recognition, COCO object detection, and ADE20K semantic segmentation, showcasing improvements in accuracy and performance across various benchmarks.

The second part of this issue is devoted to applied techniques and we have a first block of papers addressing modeling and forecasting problems in the domain of road traffic. The first one is Lin et al.'s paper addressing the importance of traffic object detection, highlighting the need for specialized traffic datasets to improve target detection for Chinese road situations. The research proposes a cross-augmentation method for image datasets, utilizing YOLOX for target detection, achieving significant improvement in mAP compared to other algorithms and resulting in a specialized detector suitable for Chinese traffic conditions. The second paper of this block is presented by Zhou, et al. and it introduces a novel graph neural network architecture called Attention-based Spatial-Temporal Adaptive Integration Gated Network (AST-AIGN) to enhance traffic forecasting by capturing the topological structure of road networks and incorporating spatial-temporal features effectively. By combining Graph Attention Network (GAT) and Jumping Knowledge Net (JK-Net) within AST-AIGN, and utilizing spatial-temporal adaptive integration gates, the model outperforms existing baselines on real-world traffic datasets like PEMS04 and PEMS08, showcasing improved accuracy in traffic condition predictions. Finally, the third paper of this topic is signed by Shuo Zhang, et al. and introduces an unsupervised generative neural network, MSST-VAE (Multiple Streams Spatial Temporal-VAE), designed for traffic raster data imputation in Intelligent Transportation Systems despite missing data. By treating traffic raster data as a multiple channel image and leveraging a novel architecture combining VAEs with Sylvester Normalizing Flows and an ECB model, MSST-VAE outperforms traditional imputation methods, showcasing strong generalization capabilities and robust performance across varying missing rates in real traffic flow datasets.

To complete the series of applied papers, we include Zhong and Shao who present a novel approach for aspect-based sentiment analysis utilizing multimodal data, focusing on both aspect term extraction and aspect sentiment classification tasks to enhance practical applications. By combining text and visual information through a cross-model hierarchical interactive fusion network, the proposed model achieves higher effectiveness in sentiment analysis compared to existing methods, as demonstrated in experiments with publicly available multimodal fine-granular emotion datasets.

The final part of the issue includes a miscellanea of papers that can be described as support techniques for general Intelligent Data Analysis problems. First, the paper authored by Singh et al. discusses the challenges of implementing blockchain technology in Internet-of-Things (IoT) applications due to scalability issues and computational expenses. The paper presents a lightweight blockchain approach tailored for IoT requirements, emphasizing end-to-end security and decentralization achieved through a network of high-resource devices collaborating to maintain the blockchain efficiently. Additionally, the proposal includes a distributed execution time-based consensus algorithm and a randomized node-selection algorithm to reduce mining overhead, prevent double-spending, and mitigate 51% attacks, showcasing promising results in addressing these challenges in IoT blockchain implementation. A second contribution of this types is the study proposed by Al-Jumaili et al. that presents a fast-density peak clustering technique combining Canopy and K-means algorithms within Apache Mahout's distributed machine learning environment to analyze parallel power load anomalies efficiently. By leveraging Apache Hadoop's tools for data storage and processing, the study showcases how the Canopy clustering approach as a preprocessing step significantly reduces computational effort in managing and analyzing vast quantities of parallel power load abnormalities, ultimately providing a scalable and effective solution for power system monitoring. The next paper in this blocy is Boullé's paper, an interesting contribution that discusses the use of histograms as non-parametric density estimators in exploratory analysis, highlighting the challenges in parameter inference for real-world datasets. The focus is on the G-Enum histogram method, utilizing the Minimum Description Length principle to build histograms without user-defined parameters and extending it for improved performance with outliers and heavy-tailed distributions.

We conclude this last part and the issue with Subha & Bharathi's paper that tackles the challenges faced by content creators on platforms like YouTube in understanding viewers' sentiments within the vast amount of data, leading to the development of spark-based machine learning algorithms like the Improved Novel Ensemble Method (INEM) to predict viewer sentiments based on comments. By utilizing such algorithms, content creators can gain valuable insights to refine their strategies, optimize revenue, and enhance channel performance, ultimately allowing them to better connect with their audience and improve content quality.

With our best wishes,

Dr. A. Famili *Dr. J.M. Pena*
Founder *Editor-in-Chief*