applied sciences

Editorial

Machine Learning-Based Feature Extraction and Selection: 2nd Edition

David Ruano-Ordás

SING Research Group, Galicia Sur Health Research Institute (IIS Galicia Sur), SERGAS-UVIGO, 36213 Vigo, Spain; drordas@uvigo.gal; Tel.: +34-988-387-016; Fax: +34-988-387-001

1. Introduction

Since the release of the initial version of this Special Issue in 2024, there has been growing importance in using pre-trained models to extract features. Many NLP and CV tasks now use features that have been learned on large datasets, reducing the need to create bespoke features for a specific problem. However, not all data will be useful. Features can often be redundant or unimportant, and their effect can be hard to interpret. Therefore, deciding what information to keep is still an important challenge in ML, despite recent advances in automatic feature extraction.

An example of this can be seen in natural language processing. Using transformer-based models, words, sentences and documents can be embedded into contextual representations. Recently, instruction tuning and self-supervised learning have been used to improve the performance of these models for information retrieval tasks where there is little or no labelled data available [1]. This has led to new ways to extract useful information from scientific texts, social media and other text. However, there is a difference between learning a semantic representation and learning to find useful information. The quality of the extracted features will depend on different factors, including characteristics of the data, the learning objective and how well the representation retains the discriminative information.

The field of representation learning has seen significant developments in computer vision. New models such as SigLIP 2 [2] and DINOv3 [3] have shown an increasing interest in creating generic visual representations. SigLIP 2 looks at multilingual vision-language representations, while DINOv3 looks at large-scale self-supervised learning to create visual features without requiring many manually annotated examples. Similar techniques appear in other data types, such as TabPFN [4], which exploits prior learning to solve tabular prediction problems. This is an important step in improving the reusability of knowledge acquired in pre-training. However, these models raise questions about the complexity of these representations and the ease of understanding them. In addition, they may not be suitable for certain specific tasks. While having a powerful pre-trained model is useful, we still need to determine if the information being used is relevant.

At the same time, feature selection remains an active research area. Recently, the goal of feature selection has become more than simply reducing dimensionality and improving classification accuracy. When the original data contains heterogeneous variables, complex dependencies, and redundancy, it is challenging to select a concise set of features. To address these issues, various studies have been conducted, including feature-selection stability [5], improvements in rough set-based feature reduction [6], and causal information incorporation [7]. In this work, different aspects of the issue have

Received: 15 November 2026
Revised: 1 January 1970
Accepted: 27 November 2026
Published: 29 November 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

been considered. Attributes may be correlated with the target variable, but they may not be reliable or useful. Furthermore, different subsets may be obtained by applying feature selection on different samples from the same dataset. As such, when evaluating the quality of a feature subset, it is also important to consider the redundancy, consistency and relevance of the subset.

This problem becomes even more critical in scenarios where data is changing over time. ML models are often built from data which has been obtained under certain conditions and implicitly assume these features will remain statistically representative when the model is deployed. However, changes to how the data was obtained, how the population differs, or differences in the relationships between variables can all cause issues with previously selected features. These changes can happen without affecting the names/types of attributes and can be hard to detect using traditional data validation techniques. Therefore, selecting features once is not always sufficient. It may also be important to monitor the behaviour of candidate attributes and identify potentially unstable features.

The developments described above show that feature extraction and selection have evolved beyond simple variable reduction. Instead, we now seek to understand how to generate informative features, identify attributes relevant to the task without destroying critical relationships, and ensure that the extracted features remain appropriate for the conditions. This second edition of the Special Issue "Machine Learning-Based Feature Extraction and Selection" has been developed within this scientific context. This edition contains ten papers exploring three main aspects of feature representation: (i) extraction, (ii) reduction and (iii) stability. The papers employ a wide range of traditional ML techniques along with deep learning, semantic modelling, optimisation algorithms and uncertainty-aware techniques. The papers consider a broad scope of applications, including software engineering, natural language processing, medical imaging, educational assessment, and urban data. Together, the papers consider multiple perspectives for improving how meaningful information can be represented and exploited by ML systems.

2. An Overview of Published Articles

With the increasing use of semantic representations, new opportunities have arisen to enhance predictive tasks that are traditionally based on manually defined features. Yalçıner et al. (Contribution 1) propose a method that aims to predict the effort required to complete software development tasks in Agile environments. This is achieved by using an SBERT model together with gradient-boosted tree models to convert text descriptions of user stories into useful numerical features. The model was evaluated in an experiment with 31,960 issues from 26 open-source Agile projects, showing that it outperformed traditional approaches and those previously proposed. The model demonstrated a mean absolute error of 2.15 and a median absolute error of 1.85. This demonstrates the potential of using context-aware feature extraction techniques to solve problems where the important information is primarily described in natural language.

Another interesting application of semantic feature representation is presented by Kılıçaslan et al. (Contribution 2), who investigate the automatic estimation of suitable age values for children's literature. The authors propose an AI-assisted methodology that combines conventional feature engineering, transformer-based language representations, and large language models (LLMs). Concretely, this study uses a dataset of 300 Turkish children's books to compare different regression methods, including random forests, support vector regression and hybrid models using BERT embeddings and structured features. Additionally, zero-shot and few-shot prediction strategies based on LLMs are explored to provide insight into how much semantic information is contributing to

prediction. Results show that traditional feature-based models remain competitive. However, combining random forest predictions with LLM-generated predictions improves performance. This study demonstrates the complementary nature of traditional feature engineering and new language-model-based methods, especially in low-sample-size scenarios.

Extracting relevant textual characteristics also plays an essential role in understanding public opinion and identifying emerging behavioural patterns. In this context, the work proposed by Özkara et al. (Contribution 3) combines sentiment analysis, text mining, and topic modelling to investigate social media conversations about electric vehicles. The authors analyse messages written in English and Turkish to identify consumer perceptions and the principal factors influencing electric vehicle adoption. A long short-term memory (LSTM) classification model achieves accuracy values of 92.1% and 96.7% for English and Turkish messages, respectively. Furthermore, the extracted information reveals relevant differences between the analysed linguistic contexts, particularly regarding environmental benefits, economic considerations, charging infrastructure, and battery performance. The results show how ML-based textual analysis can transform unstructured user-generated information into meaningful indicators for understanding public attitudes toward emerging technologies.

Another important issue is detecting the semantic structure of text data. This can be challenging when dealing with documents that cover multiple topics and contain ambiguous phrases. In this regard, Parlak et al. (Contribution 4) investigate different topic modelling approaches for analysing Turkish news articles. The proposed study compares conventional probabilistic methodologies with embedding-based and fuzzy models, including Latent Dirichlet Allocation, BERTopic, Top2Vec, and Fuzzy Latent Semantic Analysis. Additionally, knowledge graphs are used alongside entropy measurements to assess the relationships between the extracted topics. It was found that Top2Vec performed well in terms of topic coherence and interpretability. The fuzzy methods were beneficial for representing topics with overlapping semantics. This highlights the importance of considering semantic uncertainty and interpretability when extracting useful characteristics.

Identifying relevant linguistic features becomes especially important when the objective is to extract structured relationships from scientific documents. In this context, Liu et al. (Contribution 5) introduce a Feature Words Adversarial Scheme (FWAS) for improving relational quintuple extraction from the unstructured scientific literature. The proposed methodology combines transformer-based representations with a dual-pointer mechanism that distinguishes informative feature words from less relevant expressions. Through an adversarial learning strategy, the system dynamically reinforces the contribution of relevant contextual information while reducing the influence of potentially interfering words. Experimental evaluations on three public scientific information extraction datasets show improvements over competitive baselines, with F1-score increases of up to 4.9%. This work contributes to developing more effective mechanisms for semantic feature identification and illustrates their potential to improve scientific knowledge extraction and knowledge graph construction.

Although NLP applications constitute an important research direction, the benefits of feature extraction and selection extend to other forms of complex information. Medical imaging is a particularly relevant example, since identifying discriminative visual characteristics is essential for developing reliable computer-aided diagnosis systems. Mjahad and Rosado-Muñoz (Contribution 6) propose a hybrid framework for detecting pneumonia from chest X-ray images by combining deep learning-based feature extraction with fuzzy classification methodologies. The proposed system evaluates several

convolutional neural network architectures, including VGG16, EfficientNetV2, MobileNetV2, and ResNet50, together with different fuzzy classifiers. In particular, combining MobileNetV2 features with an Adaptive Neuro-Fuzzy Inference System (ANFIS) with Gaussian membership functions achieves an overall classification accuracy of 98.52% on the evaluated dataset. These findings demonstrate the potential of integrating discriminative deep representations with fuzzy reasoning mechanisms to address the uncertainty inherent in medical image classification.

Learning meaningful representations has many useful applications, including exploring high-dimensional data without predefined labelling. This work by Nadworska et al. (Contribution 7) explores the use of autoencoder networks for reducing the dimensionality of crime data. The authors convert the variable nature of crime-related information into numerical data. Different autoencoder types are explored, including standard, convolutional, denoising and variational autoencoders. After creating these representations, a variety of clustering techniques were applied, such as k-means, Gaussian mixtures, and spectral clustering. The experiments were performed on real crime data for Chicago, showing that using standard autoencoder representations (with either k-means or Gaussian mixtures) gave excellent results when using mid-level dimensions in the latent space. This work shows how learning representations can be very useful for finding hidden structures in high-dimensional and variable data.

While deep learning-based methods aim to learn alternative representations of the data, other studies have focused on selecting the most relevant features within a set of existing features. Gules et al. (Contribution 8) present a novel binary feature selection algorithm using the White Shark Optimiser (WSO). This new method, entitled bWSO-log, uses a logarithmic transfer function to transform the original continuous WSO into a binary version. The proposed method differs from traditional transfer functions that use S-shaped or V-shaped functions to transform continuous values into binary values. The experiments were carried out on nineteen benchmark datasets and compared to eight other metaheuristic optimisation algorithms. This demonstrated that the proposed method either competes with or outperforms the other methods in terms of classification accuracy and number of features selected. This demonstrates the importance of developing effective search and transformation strategies for feature selection problems with large search spaces.

While Gules et al. (Contribution 8) focus on the search strategy used for selecting features, Zhang et al. (Contribution 9) present another attribute-reduction contribution, proposing an Information Flow-based neighbourhood rough set methodology for processing high-dimensional heterogeneous information. The proposed IF-EMD-SPA approach integrates Information Flow theory, Earth Mover's Distance, and Set Pair Analysis to establish a common representation of continuous and discrete attributes. Based on this framework, the authors develop a three-stage greedy reduction mechanism incorporating dependency preservation, structural consistency, and redundancy elimination. The proposed methodology is evaluated on five benchmark datasets and two high-dimensional gene expression datasets, achieving an average classification accuracy of 93.9% using support vector machines. These findings highlight the potential of combining information-theoretic representations with rough set-based reduction strategies to preserve relevant relationships among heterogeneous attributes while identifying compact feature subsets.

Finally, feature selection reliability becomes important when dealing with regular updates of the datasets or when the information source varies. In this context, Piryankov et al. (Contribution 10) investigate feature stability assessment in high-dimensional Extract–Transform–Load (ETL) pipelines. The authors extend a previously developed

Canberra distance-based monitoring framework by introducing statistical hypothesis testing to identify potentially unstable attributes. The method combines univariate and bivariate statistical deviations with hypothesis test results. The resulting *p*-values are aggregated using a harmonic mean *p*-value formulation and incorporated into a feature stability score. This gives a measure of stability for each feature. This approach differs from traditional feature selection approaches because it aims to determine whether a feature is consistent enough. This investigation provides an alternative way to assess features and shows the importance of identifying unreliable features.

3. Conclusions

Taken together, the contributions collected in this Special Issue illustrate how feature extraction and selection methodologies are evolving in response to increasingly complex learning requirements. The first group of investigations demonstrates the potential of semantic embeddings, deep representations, and hybrid architectures to transform unstructured information into meaningful characteristics. Other contributions focus on identifying compact and informative attribute subsets through optimisation and uncertainty-aware reduction mechanisms. Finally, the investigation of feature stability highlights the importance of evaluating whether previously useful information remains reliable under changing data conditions. These complementary perspectives suggest that improving ML systems requires attention not only to the characteristics initially extracted or selected, but also to their interpretability, relevance, and consistency throughout the analytical process.

Despite the advances described above, several challenges remain unresolved. The transferability of learned representations between different domains, the computational cost of processing high-dimensional features, and the preservation of meaningful information during dimensionality reduction continue to motivate new methodological developments. Furthermore, the increasing integration of foundation models with conventional learning algorithms raises important questions regarding explainability, feature redundancy, and the contribution of individual attributes to predictive decisions. Future investigations should therefore focus on developing efficient, interpretable methodologies that combine heterogeneous feature representations, maintain stable attribute subsets, and adapt to evolving data distributions. Particular attention should also be paid to experimental reproducibility and validation across independent datasets, especially in application domains where predictive reliability is a fundamental requirement.

The variety of works included in this second edition clearly indicates that feature extraction and selection remain essential steps in developing effective and reliable ML methods. We hope the works included will help readers understand recent advancements and identify new opportunities for scientific research. We thank all authors for contributing to this Special Issue and all reviewers for their useful comments and efforts to improve the quality of the papers.

Conflicts of Interest: The author declares no conflicts of interest regarding this publication.

List of Contributions

1. Yalçıner, B.; Dinçer, K.; Karaçor, A.G.; Efe, M.Ö. Enhancing Agile Story Point Estimation: Integrating Deep Learning, Machine Learning, and Natural Language Processing with SBERT and Gradient Boosted Trees. *Appl. Sci.* **2024**, *14*, 7305, doi:10.3390/app14167305.
2. Kılıçaslan, F.N.; Genç, B.; Saglam, F.; Altun, A. Artificial Intelligence-Assisted Determination of Suitable Age Values for Children's Books. *Appl. Sci.* **2025**, *15*, 11438, doi:10.3390/app152111438.

3. Özkara, Y.; Bilişli, Y.; Yildirim, F.S.; Kayan, F.; Başdeğirmen, A.; Kayakuş, M.; Yiğit Açıkgöz, F. Analysing Social Media Discourse on Electric Vehicles with Machine Learning. *Appl. Sci.* **2025**, *15*, 4395, doi:10.3390/app15084395.
4. Parlak, I.B.; Şahin, M.Ş.; Acarman, T.; Adel, M.; Bourennane, S. Quantification of Overlapping and Network Complexity in News: Assessment of Top2Vec and Fuzzy Topic Models. *Appl. Sci.* **2025**, *15*, 9627, doi:10.3390/app15179627.
5. Liu, Y.; Fu, L.; Xia, X. Quintuple Extraction Method for Scientific Papers Based on Feature Words Adversarial Scheme. *Appl. Sci.* **2026**, *16*, 2187, doi:10.3390/app16052187.
6. Mjahad, A.; Rosado-Muñoz, A. Combining Deep Learning Architectures with Fuzzy Logic for Robust Pneumonia Detection in Chest X-Rays. *Appl. Sci.* **2025**, *15*, 10321, doi:10.3390/app151910321.
7. Nadworska, W.; Piłat-Rożek, M.; Łazuka, E. Clustering of Crimes Using Latent Representations Obtained via Autoencoders. *Appl. Sci.* **2026**, *16*, 7351, doi:10.3390/app16147351.
8. Gules, S.; Kılıç, A.; Kiran, M.S.; Gunduz, M. A Logarithmic Transfer Function for Binary Swarm Intelligence Algorithms: Enhanced Feature Selection with White Shark Optimizer. *Appl. Sci.* **2025**, *15*, 8710, doi:10.3390/app15158710.
9. Zhang, C.; Chen, C.; Yang, G.; Lan, S.; Zhang, Q. IF-EMD-SPA: An Information Flow-Based Neighborhood Rough Set Approach for Attribute Reduction. *Appl. Sci.* **2026**, *16*, 2789, doi:10.3390/app16062789.
10. Piryankov, K.; Grigorova, I.; Karamfilov, A.; Efremov, A. Hypothesis-Informed Feature Stability Scoring for High-Dimensional ETL Pipelines. *Appl. Sci.* **2026**, *16*, 6445, doi:10.3390/app16136445.

References

1. Zeng, Q.; Qiu, Z.; Hwang, D.Y.; He, X.; Campbell, W.M. Unsupervised Text Representation Learning via Instruction-Tuning for Zero-Shot Dense Retrieval. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*; Association for Computational Linguistics: Miami, FL, USA, 2024; pp. 269–279. [Google Scholar]
2. Tschannen, M.; Gritsenko, A.; Wang, X.; Naeem, M.F.; Alabdulmohsin, I.; Parthasarathy, N.; Evans, T.; Beyer, L.; Xia, Y.; Mustafa, B.; et al. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. *arXiv* **2025**, arXiv:2502.14786. [Google Scholar]
3. Siméoni, O.; Vo, H.V.; Seitzer, M.; Baldassarre, F.; Oquab, M.; Jose, C.; Khalidov, V.; Szafraniec, M.; Yi, S.; Ramamonjisoa, M.; et al. DINOv3. *arXiv* **2025**, arXiv:2508.10104. [Google Scholar]
4. Hollmann, N.; Müller, S.; Purucker, L.; Krishnakumar, A.; Körfer, M.; Hoo, S.B.; Schirrmeister, R.T.; Hutter, F. Accurate Predictions on Small Data with a Tabular Foundation Model. *Nature* **2025**, *637*, 319–326. [Google Scholar] [CrossRef] [PubMed]
5. Barbieri, M.C.; Grisci, B.I.; Dorn, M. Analysis and Comparison of Feature Selection Methods towards Performance and Stability. *Expert Syst. Appl.* **2024**, *249*, 123667. [Google Scholar] [CrossRef]
6. Liu, K.; Yang, X.; Ding, W.; Ju, H.; Li, T.; Wang, J.; Yin, T. A Survey on Rough Feature Selection: Recent Advances and Challenges. *IEEECAA J. Autom. Sin.* **2026**, *13*, 521–542. [Google Scholar] [CrossRef]
7. Lamsaf, A.; Carrilho, R.; Neves, J.C.; Proença, H. Causality, Machine Learning, and Feature Selection: A Survey. *Sensors* **2025**, *25*, 2373. [Google Scholar] [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.