Chemometrics and Intelligent Laboratory Systems 265 (2025) 105491

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

Enhancing one-class classification performance through variable selection: A review based on advanced literature search approaches

Mateus P. Schneider [a], Cristina Malegori [b,*], Adriano de A. Gomes [a], Paolo Oliveri [b]

[a] Universidade Federal do Rio Grande do Sul – UFRGS, PPGQ, Brazil
[b] Università degli Studi di Genova, Dipartimento di Farmacia – UniGe, DIFAR, Italy

ARTICLE INFO

Keywords:
One-class classification
Variable selection
Chemometrics
Rigorous modeling
Bibliometric analysis
AI-Assisted review

ABSTRACT

Variable selection is a key step in improving One-Class Classification (OCC), especially when applied to high-dimensional datasets common in chemometrics and anomaly detection tasks. This systematic literature review explores how different strategies—filter, wrapper, embedded, and hybrid methods—have been employed to enhance OCC models' accuracy, interpretability, and robustness. A comprehensive search was conducted using Scopus, complemented by AI-powered tools such as Elicit and Litmaps, and visual analytics platforms including VOSviewer and Bibliometrix. The review highlights methodological trends across both chemometric and machine learning domains, revealing a predominance of embedded approaches and a growing interest in hybrid strategies. Embedded methods, particularly LASSO, Elastic Net, and autoencoder-based architectures, were favored for their scalability and model integration. Approximately 69 % of the reviewed studies adopted a rigorous OCC approach—relying solely on target class data—demonstrating a preference for bias-resistant modeling. Additionally, bibliometric analysis revealed a disciplinary division, with chemometric studies emphasizing analytical applications and model interpretability, while computer science-driven studies prioritized automation and scalability. The findings emphasize the need for flexible, domain-adapted variable selection pipelines capable of handling class imbalance and high dimensionality. This work also introduces a reproducible framework combining traditional and AI-assisted literature review tools to support future systematic analyses. The review concludes by identifying emerging trends and suggesting future research directions in OCC and variable selection, with a focus on hybrid modeling, domain adaptability, and performance benchmarking across application fields.

1. Introduction

Pattern recognition in chemistry, particularly in the context of analytical chemistry, is closely tied to qualitative and screening methodologies. The chemometric approaches used for this purpose can be broadly categorized into two main groups. The first one involves differentiating between samples of type A and type B (or more than two types, although a binary example is used here for simplicity). In this approach, after the training stage, the model can classify an unknown sample as either A or B—a process known as discrimination. The second group focuses on training a model to recognize samples of a single target class (e.g., type A) exclusively. Once trained, this model can identify samples of type A, while also flagging those that do not belong to type A. When there is more than one class of interest, the procedure can be repeated, individually, for each class. This strategy is referred to as class modeling [1].

One-Class Classification (OCC) is a class modeling approach tailored for scenarios where there is only one target class of interest, and well-represented in the training data. It focuses on identifying deviations from the target class without relying on alternative class data, which is particularly suitable for tasks like anomaly detection, novelty identification, and authentication in fields such as food fraud detection, medical diagnostics, and quality control. Consequently, modeling the target class considering only representative features plays a pivotal role in enhancing the performance of OCC [2–9].

OCC methodologies can be categorized by training data composition and algorithmic modeling approaches. The availability of training data include, among others, learning with target samples only, and with target samples and some non-target ones [10,11]. These paradigms are often referred to as Rigorous and Compliant OCC, respectively [12].

* Corresponding author.
E-mail address: cristina.malegori@unige.it (C. Malegori).

https://doi.org/10.1016/j.chemolab.2025.105491
Received 4 May 2025; Received in revised form 20 July 2025; Accepted 20 July 2025
Available online 21 July 2025

Algorithmically, OCC includes One-Class Support Vector Machines (OSVM)-based methods, which rely on kernel techniques to define decision boundaries, neural networks, ensemble methods, and statistical models, offering great flexibility across diverse domains [11].

The underlying mechanisms of OCC include boundary-, density-, and reconstruction-based approaches. Boundary-based methods define decision boundaries around the target class, meanwhile density-based methods estimate the target class probability distribution. Reconstruction-based approaches, finally, assess anomalies based on deviations between input and reconstructed data [5].

In such scenarios, selecting a small subset of informative variables can significantly enhance multivariate models, improving both their figures of merit and chemical interpretability. Especially in the case of rigorous OCC, where a model is trained to learn in the absence of counterexamples, removing redundant and uninformative variables might play a pivotal role in enhancing the performance of OCC. Variable selection can be part of OCC, with methods broadly classified as Filter, Wrapper, Embedded, or Hybrid approaches. Filter methods independently evaluate features based on statistical metrics, while Wrapper methods iteratively refine feature subsets by optimizing model performance. Embedded approaches integrate variable selection within model training. Hybrid methods combine elements of these strategies to balance computational efficiency and model robustness. By refining feature spaces, variable selection enhances model accuracy, interpretability, and efficiency, particularly in high-dimensional datasets with class rarity [1,13–15]. A first aim of the present review study is to systematically present and discuss established methods and current trends employed to address the issue of variable reduction/selection in the OCC context.

In the last years, significant advancements in variable selection for OCC have been achieved, but several challenges remain open in adapting these methods to high-dimensional chemical datasets. The present review systematically employs diverse tools, including bibliometric platforms, and AI-driven literature analysis, to identify trends, compare methods, and highlight future directions [16,17]. By bridging traditional chemometric approaches with AI-driven innovations, it underscores the critical role of variable selection in advancing OCC and addresses enduring challenges in anomaly detection and classification, fostering interdisciplinary progress [18–20].

An additional aim of the present study is, therefore, to provide the reader with a comprehensive and critical guide on advanced AI-based tools for bibliographic search and analysis, to support efficient examinations of the scientific literature.

2. Methodology

2.1. Description of tools

To emphasize both comprehensiveness and efficiency in analyzing variable selection approaches for multidisciplinary OCC methodologies, the present study integrates queries from traditional databases with AI-powered tools in a structured and innovative literature review. Importantly, to avoid the risk of generating nonexistent references, a known limitation of generative language models such as ChatGPT, this study relied exclusively on platforms that query structured bibliographic databases. In more detail, insights from Scopus, Elicit, and Litmaps, which allow free access to publication metadata but often require institutional or individual subscriptions for full-text access, are complemented by VOSviewer (a freely downloadable bibliometric mapping software) and Bibliometrix (a free R package for bibliometric analysis). Together, these tools support a diverse and integrated review strategy based on verifiable literature sources [16,21].

2.1.1. Scopus

Scopus is a leading multidisciplinary database providing access to peer-reviewed literature across diverse fields, with strengths in comprehensive metadata, including citation relationships, author collaborations, and keyword trends. Scopus search page is available to registered users on its website through www.scopus.com/search. Its rigorous curation by the Content Selection and Advisory Board (CSAB) ensures high-quality data [22]. Scopus has emerged as a key resource for bibliometric studies, challenging Web of Science's dominance and gaining significant adoption in emerging economies like China, Brazil, and Iran, alongside developed countries [23].

Features like "Analyze Data" enable users to explore publication trends, contributors, and collaboration networks, while the advanced Boolean search helps tailoring results. This last feature is critical especially while managing broad interdisciplinary searches through Scopus' queries. Despite its "Limit to" and "Exclude" functionalities, unrelated fields occasionally appear in search results, introducing unwanted sources and requiring additional manual curation. Furthermore, Scopus' document cap of 20,000 items per search demands a careful prioritization of exclusion criteria, potentially complicating efforts to explore expansive datasets comprehensively.

2.1.2. VosViewer

VOSviewer is a bibliometric visualization software designed to map and analyze relationships within scientific literature. The software is found and can be downloaded on its latest version by checking on vosviewer.com website. Vosviewer employs the "Visualization of Similarities" (VOS) technique to create intuitive visualizations of co-authorship networks, citation relationships, and keyword co-occurrences [24]. Known for its ability to handle large datasets, VOSviewer offers network, overlay, and density visualizations, making it ideal for identifying trends, gaps, and connections in research. Its integration with databases like Scopus and Web of Science, along with support for thesaurus files to harmonize metadata, ensures precise and meaningful analyses [25]. A well-curated thesaurus introduces challenges while using VosViewer, as inconsistent terminology or synonymous terms can fragment data, reducing the clarity of visualized networks. Consequently, this drawback underscores the need for researchers to meticulously standardize terms.

2.1.3. Bibliometrix

Bibliometrix, an R-based bibliometric analysis package, provides comprehensive tools for quantitative scientific mapping and visualization [26]. The downloadable package is found at www.bibliometrix.org. It supports analysis of co-citation networks, collaboration patterns, and thematic evolution, quantitatively complementing academic tools by offering a robust statistical and computational foundation. Bibliometrix allows researchers to extract bibliographic data from sources like Dimensions, Scopus, and Web of Science, enabling detailed exploration of citation relationships, author dominance, and institutional collaborations. In the present study, Bibliometrix was instrumental in uncovering thematic clusters and identifying research gaps in variable selection approaches for One-Class Classification (OCC), providing critical insights to contextualize trends and influence within the field [27].

2.1.4. Elicit

Elicit, developed by Ought, is an AI-powered literature discovery tool designed to streamline the literature review process, with its website address at www.elicit.com. Leveraging Semantic Scholar's vast corpus and natural language processing techniques, Elicit enables researchers to input specific questions and to receive semantically relevant results. Beyond traditional search engines, Elicit provides customizable features like summarizing abstracts, extracting metadata, brainstorming research questions, and asking detailed questions about articles. Its unique ability to synthesize information, to iteratively refine queries, and to highlight interdisciplinary connections makes it a valuable tool for identifying trends and insights. While it excels in efficiency, Elicit complements rather than replacing traditional search methods, requiring researchers to critically evaluate its outputs for academic rigor

[28].

2.1.5. Litmaps

Litmaps is a web-based platform that visually tracks the evolution of research ideas through interactive literature mapping. Using natural language processing and machine learning, it generates dynamic citation maps that illustrate connections between studies, identify influential articles, and trace citation trajectories over time. This tool helps researchers in placing their work within the broader research landscape, uncovering emerging trends, and identifying gaps in the literature. By suggesting related articles based on semantic context or citation networks, Litmaps allows users to refine their focus and to discover papers aligned with their specific interests. By concentrating on a curated library, the platform enhances decision-making through clear visualizations, enabling researchers to effectively navigate their field of study [29].

2.2. Search strategy

The present study followed a systematic process, beginning with defining the review focus to encompass broad variable selection methodologies in OCC while considering both chemometric and non-chemometric contexts. This approach ensured a diverse and comprehensive reviewing, by avoiding an exclusive concentration on foundational sources within the chemometrics field. The research was conducted between January and March 2024, with database results covering the last five years and AI tools providing insights limited to the past ten years, both constrained to sources available up to March 2024. The subsequent steps involved data collection, exploration, and refinement, followed by analysis and summarization. Finally, the results were structured and visualized to provide the reader with clarity and coherence [16].

Bibliographic data collection was conducted using **Scopus**. Boolean operators were employed to construct queries combining keywords such as "Feature*," "Dimensionality Reduction," "One-Class*," and "Outlier Detection." To explore chemometric and non-chemometric contexts, additional terms like "AND Chemometr*" and "AND NOT Chemometr*" were incorporated. Filters were applied to narrow results by subject areas (e.g., Chemistry, Computer Science, Mathematics) and document types (e.g., articles, reviews, and conference proceedings). The results were exported in RIS and BibTeX formats for further analysis.

Visualization and exploration of the collected data were conducted using bibliometric tools. **VosViewer** generated network and overlay maps to reveal keyword interconnections, research hotspots, and emerging trends. **Bibliometrix** complemented this with thematic maps, citation analyses, and keyword co-occurrence trends, providing quantitative validation of the findings.

AI-powered tools such as Elicit and Litmaps were employed to deepen and refine the review process. **Elicit** dynamically refined queries by responding to targeted questions, such as "What algorithms are used for variable selection in OCC?" It presented the most relevant studies aligned with the research focus, considering citation counts, contextual relevance, and user-defined filters. Features like "show more like these" enabled iterative exploration, helping uncover key methodologies, themes, and gaps in the literature. **Litmaps** complemented this by focusing on the researcher's curated library, visualizing relationships among studies based on semantic context and shared citations. Its dual-axis visualization of citation count and publication year highlighted both seminal works and recent advancements, allowing for an evolving and targeted exploration of the field.

Finally, integrating AI-driven insights with bibliometric tools facilitated a comprehensive synthesis of the literature. Selected papers were summarized, and targeted questions about OCC concepts, algorithms, and performance improvements were answered, streamlining the analysis process. The workflow integrates Scopus, VosViewer, Bibliometrix, Elicit, and Litmaps to ensure a robust and innovative exploration of variable selection approaches in OCC, connecting traditional and modern research tools, uncovering emerging insights and laying the foundation for future advancements.

3. Review findings using different tools

3.1. Insights from scopus

The Scopus database provided a structured yet complex landscape for analyzing trends and contributions in variable selection algorithms for OCC. Given its extensive coverage of well-established and emerging fields, Scopus was evaluated using Vosviewer visualizations, supported by Bibliometrix quantitative features [**22**]. The results are summarized in Fig. 1, which presents the relative frequencies (%) of the most cited journals, algorithms, and research fields found in 157 documents retrieved under the "AND Chemometrics" query and 13,027 documents retrieved under the "AND NOT Chemometrics" query from Scopus.

Each subplot on Fig. 1 brings percentages that reflect proportions relative to the total number of documents in each group. The figure is organized into six panels: **(a)–(b)** show the top 10 journals in each group, **(c)–(d)** present the most frequently cited algorithms, and **(e)–(f)** summarize the most common subject areas based on author-defined keywords. Each category is coded on the x-axis (e.g., A1–J1, A3–J3, A5–G5), and the full legend is provided in the figure caption.

The "**AND Chemometrics**" group emphasized chemometric methodologies, with a strong predominance of Partial Least Squares (PLS, A3) and Principal Component Analysis (PCA, B3) as key modeling strategies. Other frequent techniques included Support Vector Machines (SVM, C3), Variable Importance in Projection (VIP, E3), Soft Independent Modeling of Class Analogy (SIMCA, F3), and Genetic Algorithms (GA, G3), often applied for variable selection or classification enhancement. These studies were mainly published in journals such as *Chemometrics and Intelligent Laboratory Systems* (A1), *Food Chemistry* (B1), and *Spectrochimica Acta A* (C1), and predominantly associated with research fields such as Chemistry (A5), Computer Science (B5), and Mathematics (C5).

In contrast, the "**AND NOT Chemometrics**" group exhibited a clear computational orientation, highlighting Support Vector Machines (A4), Random Forest (B4), Artificial Neural Networks (D4), and Autoencoders (E4) as central tools. The algorithmic diversity reflected a trend toward deep learning, hybrid, and ensemble methods. Key publication venues included *Applied Sciences* (Switzerland) (A2), *Electronics* (Switzerland) (B2), and *Mathematics* (C2). Subject area frequencies indicate a strong dominance of Computer Science (B6), with smaller contributions from Chemistry (A6) and Mathematics (C6).

These patterns reveal a distinct disciplinary divergence. In the "**AND Chemometrics**" group, approximately 46 % of studies were indexed under Chemistry, followed by 27 % in Computer Science and 18 % in Mathematics. Conversely, in the "**AND NOT Chemometrics**" group, 76 % of documents were categorized under Computer Science, while Chemistry and Mathematics accounted for only 8 % and 13 %, respectively (Fig. 1e and f). This contrast illustrates two paradigms: one grounded in experimental and analytical chemistry, and the other in machine learning and computational modeling.

Notably, SVM was cited in both groups (C3 in "AND" and A4 in "AND NOT"), underscoring its robustness across methodologies. However, its use in chemometric studies was often secondary—employed for comparison with PLS models—whereas in machine learning literature, it was a primary modeling strategy, frequently integrated with dimensionality reduction or deep learning.

Application domains also diverged. Chemometric studies focused on *food authentication*, *environmental monitoring*, and *quality control*, while machine learning studies emphasized *medical diagnostics*, *cybersecurity*, and *intelligent systems*. Interdisciplinary applications such as *remote sensing* and *healthcare* remain underexplored in both contexts, suggesting future research opportunities.

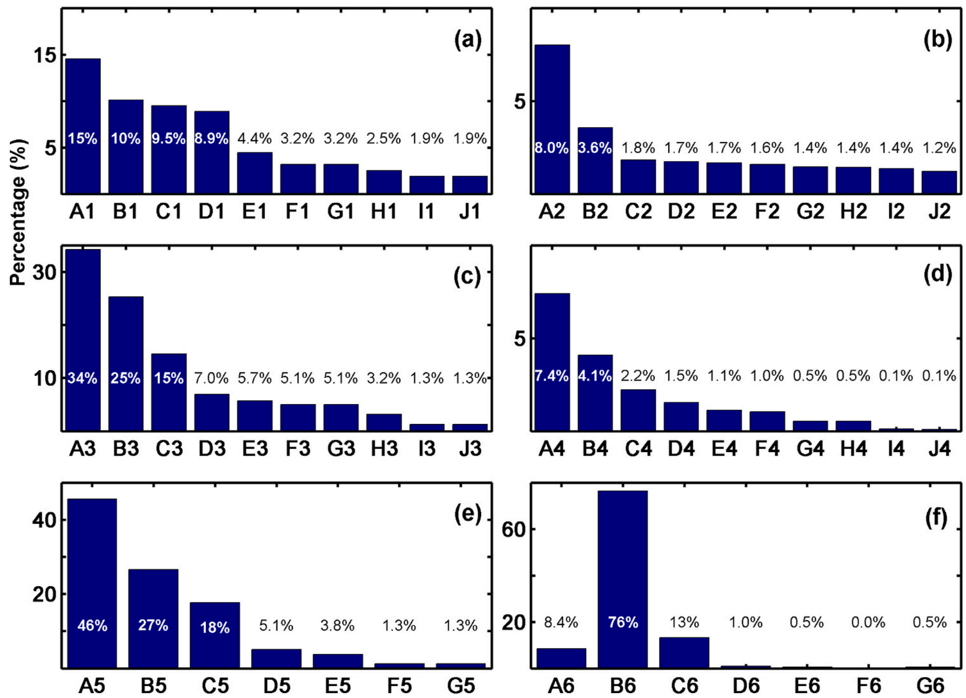

Fig. 1. Most frequently reported bibliometric elements in the literature screening.
The labels on the X-axis correspond to ranked categories, coded as follows:
Top Journals
(a) AND Chemometrics
A1 = Chemom. Intell. Lab. Syst.;
B1 = Food Chemistry;
C1 = Spectrochim. Acta A;
D1 = Molecules;
E1 = Analytica Chimica Acta;
F1 = Food Anal. Methods;
G1 = Talanta;
H1 = J. Agric. Food Chem.;
I1 = Comput. Electron. Agric.;
J1 = Sensors.
(b) AND NOT Chemometrics
A2 = Appl. Sci. (Switz.);
B2 = Electronics (Switz.);
C2 = Mathematics;
D2 = J. Image and Graphics;
E2 = IEEE Access;
F2 = Entropy;
G2 = Symmetry;
H2 = Sustainability (Switz.);
I2 = Int. J. Innov. Technol.;
J2 = Comput. Eng. (China).
Top Algorithms
(c) AND Chemometrics
A3 = PLS; B3 = PCA; C3 = SVM;
D3 = Random Forest; E3 = VIP; F3 = SIMCA;
G3 = GA; H3 = ANN; I3 = iPLS; J3 = LASSO.
(d) AND NOT Chemometrics
A4 = SVM; B4 = Random Forest; C4 = PCA;
D4 = ANN; E4 = Autoencoder; F4 = GA;
G4 = LASSO; H4 = PLS; I4 = VIP; J4 = Elastic Net.
Study Fields
(e) AND Chemometrics | (f) AND NOT Chemometrics
A5 | **A6** = Chemistry; **B5** |**B6** = Computer Science; **C5** |**C6** = Mathematics; **D5** |**D6** = Engineering;
E5 |**E6** = Statistics; **F5** |**F6** = Physics; **G5** |**G6** = Biology.

Together, the patterns shown in Fig. 1 illustrate how journal preferences, algorithmic choices, and research focus differ between chemometrics-driven and machine-learning-driven OCC research. These insights reinforce the need for cross-disciplinary dialogue, especially regarding variable selection techniques adapted to the specific challenges of one-class classification.

VOSviewer network visualizations reveal the thematic structures underlying the bibliometric analysis of the Scopus queries. For each query studied, after applying a thesaurus, setting a minimum occurrence threshold of 10 for binary term counts, and removing filler terms, the generated nodes (circles) represent recurrent terms from a publication abstract and title, graphically summarizing the gathered metadata. Each circle size indicates relevance based on occurrence and connections. General terms like “spectra” and “dataset” appear centrally, linking related research topics. In the “**AND Chemometrics**” query, three dominant clusters emerge, representing distinct aspects of chemometric research: chemical characterization, classification modeling, and quantitative analysis (Fig. 2).

The **red cluster** emphasizes chemometric approaches and authentication methods, with key terms such as “element”, “metabolite,” “component,” “antioxidant activity,” and “geographical origin”, highlighting the role of chemical characterization and authenticity verification. Analytical techniques like “HPLC”, and “GC-MS” are central to data acquisition and multivariate analysis for quality control and fingerprinting.

The **green cluster** focuses on classification and machine learning algorithms, featuring terms such as “LDA” (Linear Discriminant Analysis), “PLS-DA” (Partial Least Squares Discriminant Analysis), “SVM,” “machine learning,” and “random forest”, underscoring the importance of spectral data and cross-validation in building robust classification models.

The **blue cluster** centers on quantitative modeling and regression techniques, with key terms including “PLS”, “PCR” (Principal Component Regression), “SIMCA” (Soft Independent Modeling of Class Analogy), and “ANN” (Artificial Neural Network) reflecting the use of multivariate tools for “quantification”, along with “square error”, “prediction,” and “precision”, integrating quality validation applications.

Vosviewer overlay visualizations provide a temporal dimension, mapping the evolution of research priorities. In the “**AND Chemometrics**” overlay (Fig. 3), terms like “PLS-DA,” “SIMCA,” and “spectra” are associated with earlier publication years, indicating that these methodologies have been studied over the last years. Newer terms such as “VIP” (Variable Importance in Projection) and “random forest” suggest recent efforts to refine predictive “accuracy” and enhance chemometric models.

In contrast, the “**AND NOT Chemometrics**” query (Fig. 4) demonstrates a broader emphasis on machine learning and deep learning approaches, shifting the focus from traditional chemometric applications to emerging trends in *computational techniques*, *feature engineering*, and interdisciplinary domains like *IoT* and *security*.

The **Red Cluster** emphasizes the growing role of *computational techniques* and neural networks, with terms like “deep learning,” and “CNN” (Convolutional Neural Networks) showcasing the prominence of advanced algorithms. The **Blue Cluster** highlights the importance of *feature engineering* for machine learning performance, with terms such as “feature,” “combination,” “optimization,” “effectiveness,” “class,” and “robustness” underscoring the connection between feature selection and model robustness. Finally, the **Green Cluster** reflects the expanding applications of OCC in *technological domains* beyond chemistry, with terms like “environment,” “device,” and “integration” pointing to its relevance in “IoT” (Internet of Things) and security contexts.

Additionally, the “**AND NOT Chemometrics**” overlay (Fig. 5) highlights emerging applications in cybersecurity and environmental monitoring through the presence of terms like “IoT” and “privacy” in later years. Moreover, the plot reveals a more dynamic trend toward machine learning. Terms such as “deep learning,” “neural networks,” “feature extraction,” “effectiveness,” and “precision” have surged in recent years, demonstrating a shift in focus toward automated and scalable computational solutions.

To be noticed that, across both queries, there is limited mention of *variable selection algorithms* tailored specifically for OCC. However, some terms suggest interest in the approach, including “feature selection,” “dimensionality reduction,” and “prediction accuracy.” These terms appear more frequently in the “**AND Chemometrics**” query, where *regression-based methods* dominate, but their application in OCC-specific contexts remains sparse.

The fields of study differ significantly between queries. The “**AND**

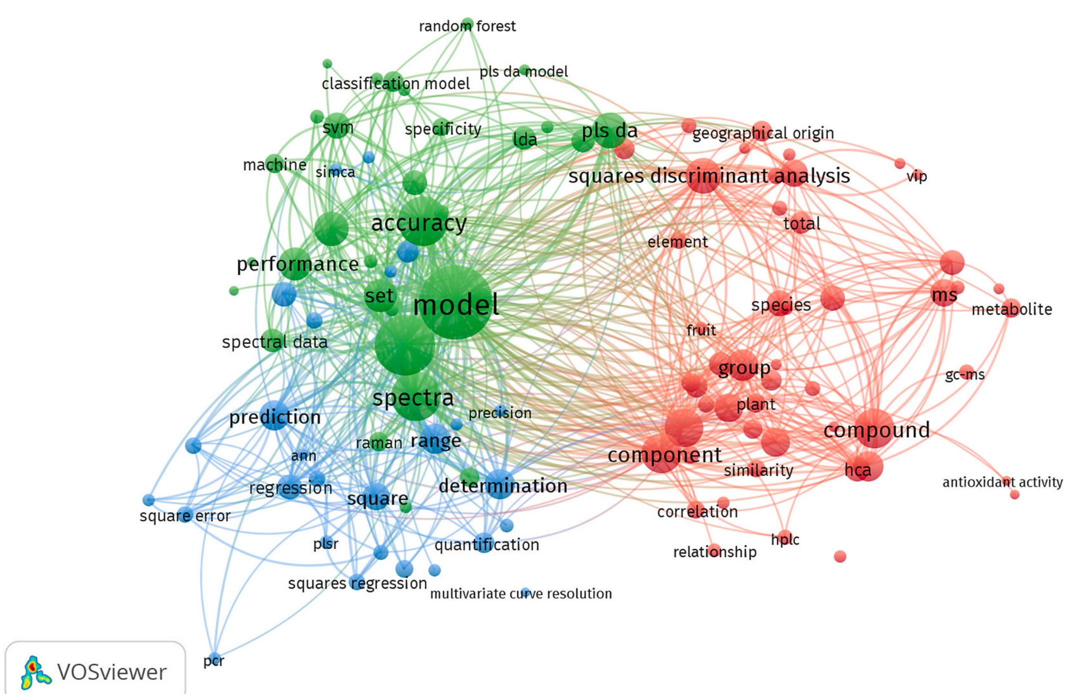

Fig. 2. Network Visualization of Terms for the “AND Chemometrics” Scopus’ query.
Thematic clusters derived from the bibliometric analysis of the “AND Chemometrics”. The green cluster key terms focused on chemometric methods. The red cluster emphasizes terms that reflect discriminant analysis and compositional studies. The blue cluster underscores terms that point to multivariate regression models. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

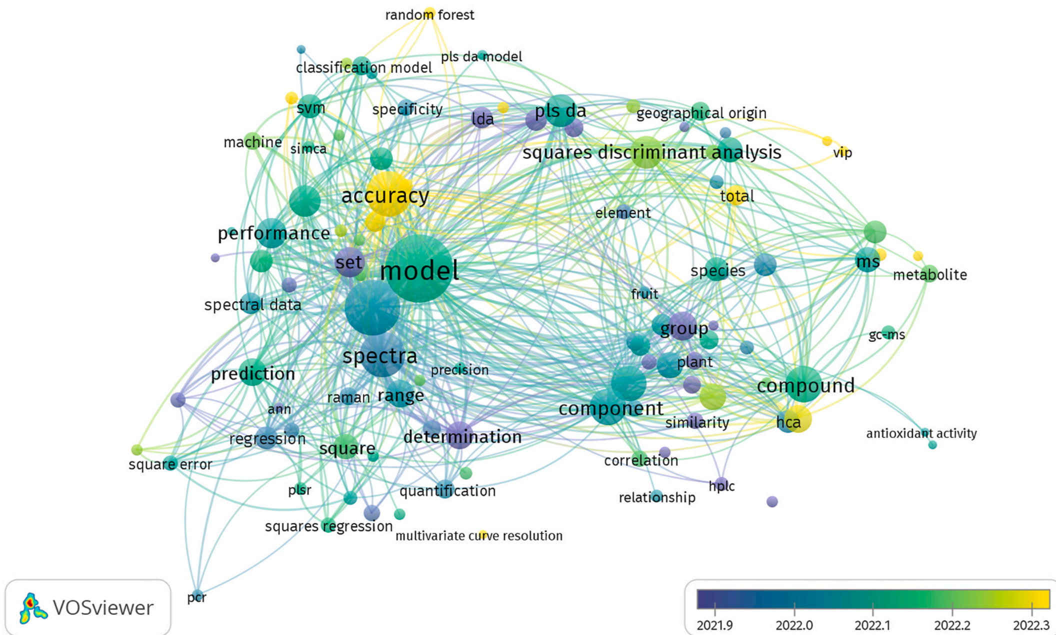

Fig. 3. Overlay Visualization of Terms for the AND Chemometrics Query.
The overlay visualization adds a temporal dimension to the network analysis of the AND Chemometrics query. Earlier research emphasizes foundational chemometric methods like “PLS-DA”, “PCR”, and “LDA,” also a focus on “specificity” as a figure of merit. Meanwhile, recent studies focus on "model performance" and variable selection approaches, like “VIP”. This shift demonstrates the evolution of chemometric methodologies toward improving predictive accuracy and model robustness.

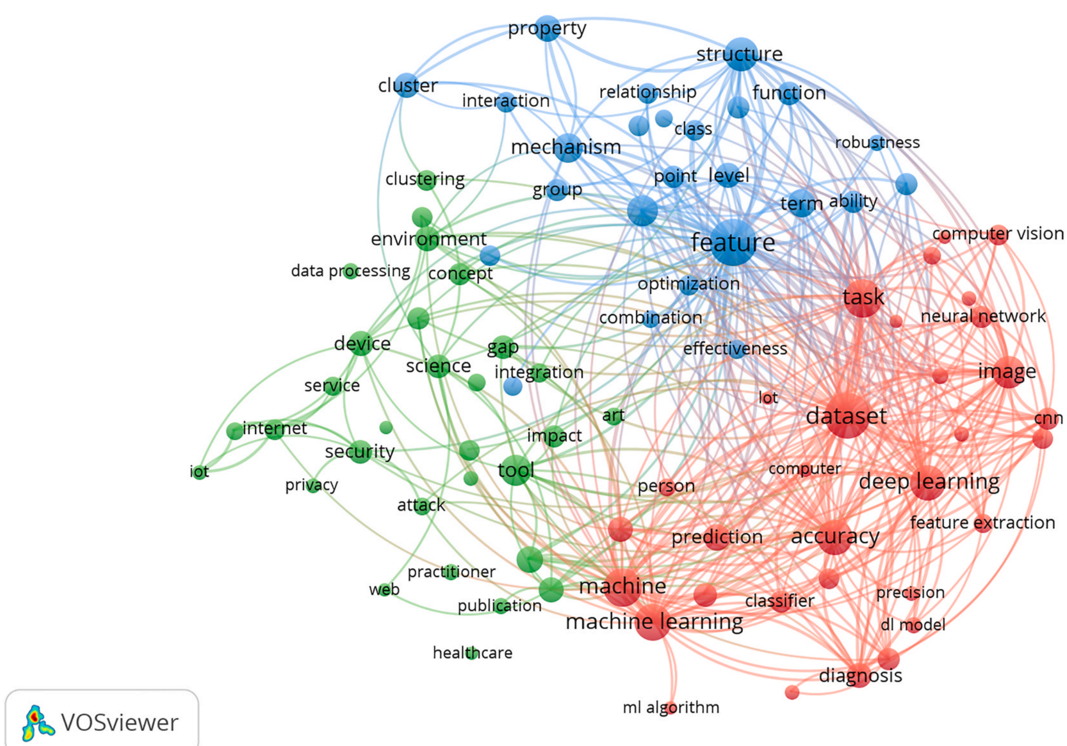

Fig. 4. Network Visualization of Terms for the AND NOT Chemometrics Query.
Thematic structure of research derived from the AND NOT Chemometrics query. The red cluster highlights terms connected to “feature extraction” like “deep learning,” “machine learning,” and “CNN,” reflecting advanced computational algorithms. The blue cluster emphasizes “feature,” “optimization”, “combination”, and “effectiveness,” tying feature engineering to dimension reduction purposes through machine learning models. The green cluster focuses on emerging applications with terms such as “environment,” “device,” and “integration.”. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

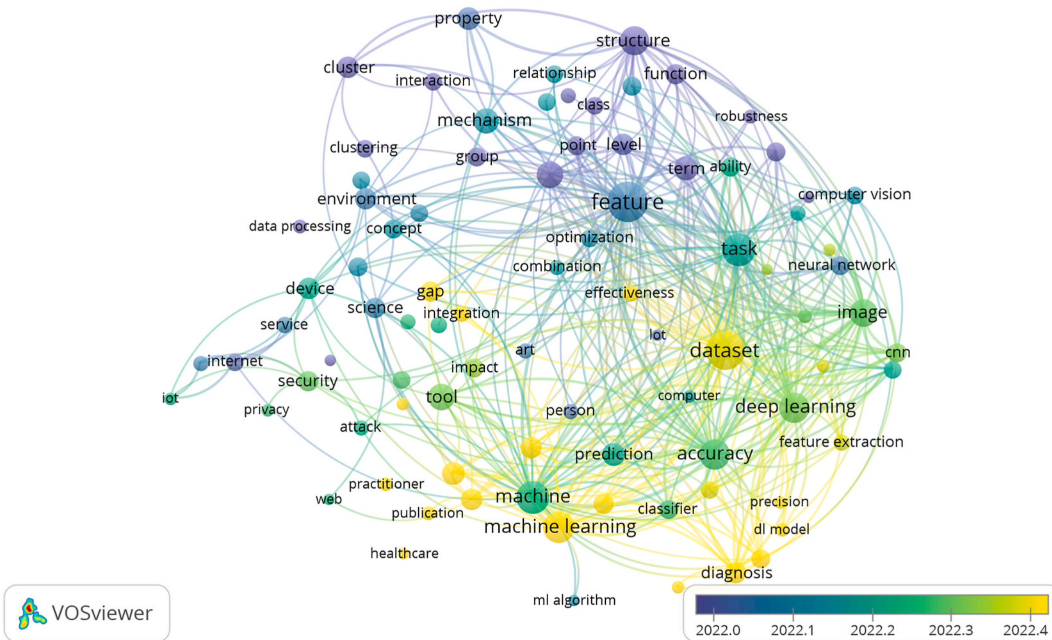

Fig. 5. Overlay Visualization of Terms for the AND NOT Chemometrics Query.
The overlay visualization for the AND NOT Chemometrics query shows temporal trends in research. Earlier studies focused on "machine learning" and "feature selection," while recent work emphasizes "deep learning," "neural networks," and applications in IoT and cybersecurity. This progression highlights the growing relevance of computational methods in OCC applications.

Chemometrics" results are concentrated in analytical chemistry and food sciences, with key journals such as *Food Chemistry* and *Microchemical Journal* leading the discourse. Conversely, the "**AND NOT Chemometrics**" results are dispersed across computer science, IoT, and applied machine learning, as evidenced by journals like *IEEE Access* and *Sensors*. This disparity suggests a gap in interdisciplinary research that bridges chemometric rigor with modern computational tools.

These thematic clusters reinforce the need for methodological

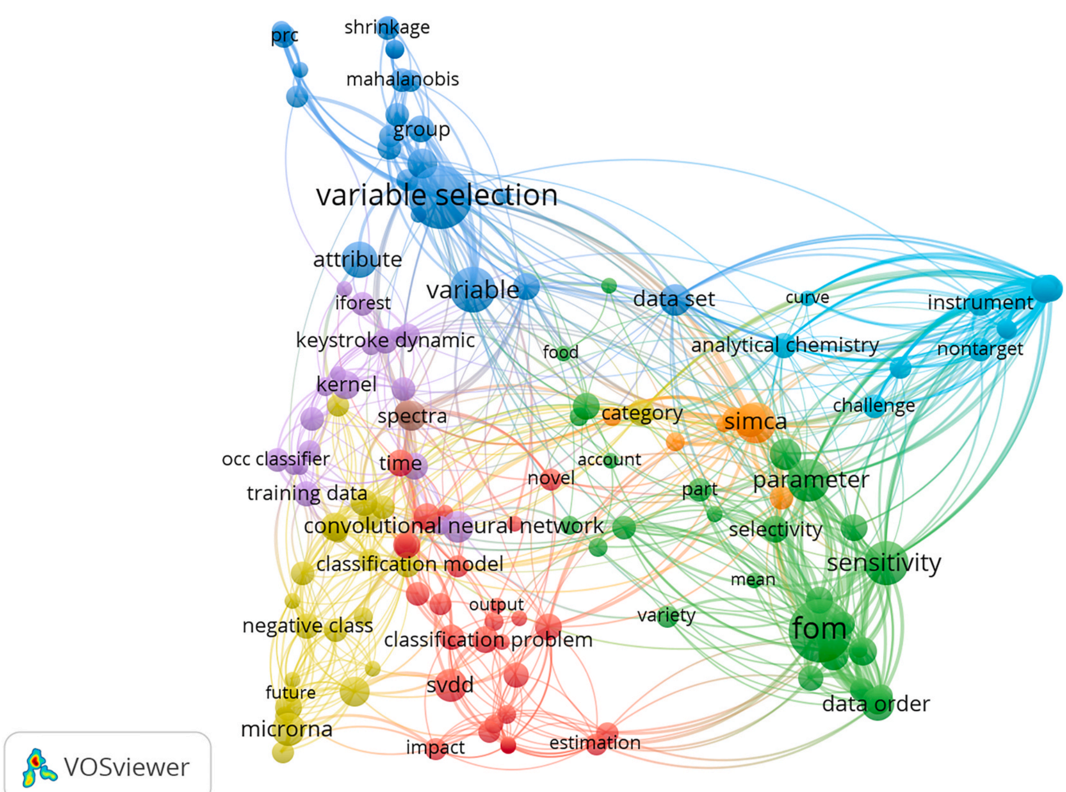

Fig. 6. Network Visualization of Elicit Data.
The network visualization of keyword terms for variable selection in OCC applications found through Elicit. Observed result was tailored by the application of a housemade thesaurus. Clusters are focused on figures of merit, variable selection, machine learning, and analytical chemical instrumentation.

innovation in variable selection applied to OCC methodologies. For instance, leveraging advanced computational techniques, as suggested in the "**AND NOT Chemometrics**" query, could complement the chemometric approaches emphasized in "**AND Chemometrics**". By integrating these findings, future research can address gaps in algorithms tailored for a rigorous approach of OCC, aligning analytical precision with computational scalability.

3.2. AI - powered literature discoveries

AI-powered tools like Elicit and Litmaps provided critical insights into variable selection for OCC, broadening the interdisciplinary scope beyond the traditional focus areas identified by Scopus. However, their features of filtering result in more limited suggestions, making them more accurate than Scopus. This accuracy ensures a focused approach aligned with the study objectives, further facilitating the extraction of specific literature and identifying pivotal connections [28,29].

For instance, Elicit's data, used in Vosviewer network visualizations (Fig. 6), identified clusters linking "variable selection" with "shrinkage" for dimensionality reduction, as well as figures of merit ("fom") like "sensitivity" tied to OCC classification parameters. Additionally, "OCC classifier" appeared in a cluster with "kernel" and "convolutional neural network", reinforcing its relevance to OCC methodologies, while "SIMCA" was linked to "classification problems" alongside "iforest," "keystroke dynamics," and Support Vector Data Descriptor ("SVDD").

Beyond identifying thematic clusters, Elicit and Litmaps facilitate the discovery of literature beyond conventional chemometric domains. As illustrated in Fig. 6, Elicit's recommendations often include multidisciplinary and applied studies surfacing Data Science, Biomedical Research, and Environmental Studies, indicating its semantic-driven capacity to surface relevant yet unexpected content. Complementarily, Litmaps' citation dynamic visualization (Fig. 7) reveals both foundational and emerging works, with emphasis on niche domains such as Behavioral Science and Industrial Applications. These tools not only support a broader exploration of literature but also reinforce the importance of the researcher's interpretive role: navigating, filtering, and strategically prioritizing findings based on domain expertise to build a comprehensive and meaningful synthesis [5,11,30,31].

In mapping the landscape of variable selection methodologies within OCC, we sought to identify relevant studies that provide a structured understanding of existing approaches. Given the broad scope of OCC applications, the implemented search strategy prioritized sources that could offer both precision and interdisciplinary perspectives. To achieve this, tools such as Elicit and Litmaps were leveraged enabling a more targeted exploration of key contributions in the field. These seed papers highlighted interdisciplinary tendencies within OCC research, with applications spanning *medical diagnostics*, *cybersecurity*, and *chemometrics*. This adaptability across fields underscores OCC scalability for high-dimensional data and its potential to address diverse classification challenges. In line with the goal of synthesizing methodological insights, Table 1 summarizes the key variable selection algorithms identified through the literature examination.

To optimize OCC models, the reviewed studies employed a range of variable selection strategies categorized as Filter, Wrapper, Embedded, and Hybrid approaches. Among these, embedded methods were the most frequently adopted (28 studies), followed by filter strategies (14 studies). Hybrid approaches—combining elements from filter, wrapper, and embedded methods—were applied in 6 studies, while wrapper-only strategies appeared in just 2. These proportions indicate a clear preference for techniques integrated into model training, particularly in high-dimensional or deep-learning-based OCC contexts.

Filter methods assess features independently of the model structure or use lightweight heuristics to evaluate variable relevance. Common techniques included Mutual Information, Fisher Scores, Interquartile Range, and Correlation Attribute Evaluation. These methods were often applied in density-based frameworks like Kernel Density Estimation (KDE) and remain valued for their simplicity and interpretability [32–37]. More advanced filtering techniques—such as Spectral Score, Information Score, and Rank Aggregation—further supported dimensionality reduction prior to modeling [38].

Wrapper methods, although less common, involved iterative selection of variable subsets based on model performance. Classical algorithms such as Genetic Algorithms (GA), Recursive Feature Elimination (RFE), Forward Selection, and Backward Elimination were used to optimize model boundaries, particularly in SVDD frameworks [32,36,39–41].

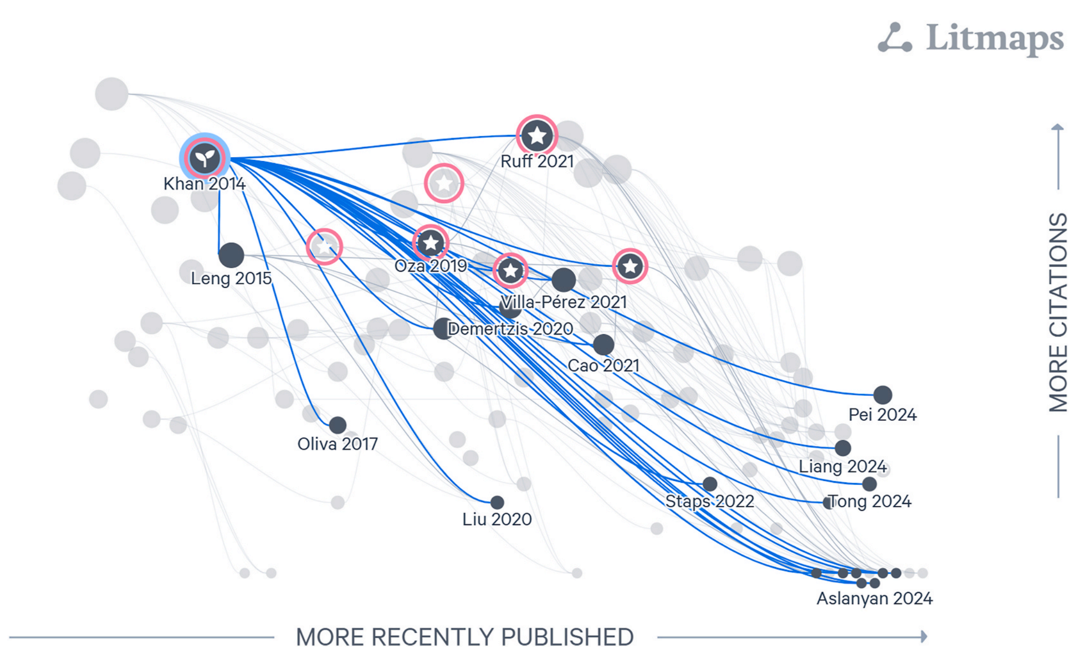

Fig. 7. Litmaps' library connections and suggestions
After Litmaps' library was constructed with Elicit results, some suggestions appeared, pointing to new papers regarding its content, which can configure a foundational paper for the literature review.

Table 1
Overview of variable selection strategies and one-class classification studies.

Study Title	OCC Approach	OCC Algorithm	Variable Selection Approach	Variable Selection Method	Ref.
A filter-augmented auto-encoder with learnable normalization for robust multivariate time series anomaly detection	Compliant	NormFAAE (Filter-Augmented Auto-Encoder with Learnable Normalization)	Embedded	Deep Hybrid Normalization (DZSN, DMMN)	[49]
A Joint Representation Learning and Feature Modeling Approach for One-class Recognition	Rigorous	Support Vector Data Description (SVDD) applied in latent space using an Adversarial Autoencoder with Augmented Latent Features	Embedded	Adversarial Autoencoder (AAE) for regularization, concatenated with reconstruction loss, followed by SVDD in the latent space	[50]
A New Feature Selection Method for One-Class Classification Problems	Rigorous	SVDD	Wrapper	SVDD-radius-RFE, SVDD-dual-objective-RFE	[41]
A Survey on Unbalanced Classification: How Can Evolutionary Computation Help?	Both	Autoencoder + PSO, Genetic Programming-based OCC	Embedded	Particle Swarm Optimization (PSO) for weight tuning, Genetic Programming (GP) using a distance-based fitness function, and implicit variable selection during model construction	[51]
Accuracy and diversity in classifier selection for one-class classification ensembles	Rigorous	SVDD + Random Subspace + ECOC	Wrapper	Multi-objective Memetic Algorithm (Accuracy + Diversity), Random Subspace	[39]
Active Authentication using an Autoencoder regularized CNN-based One-Class Classifier	Rigorous	OC-ACNN (Autoencoder-regularized CNN-based One-Class Classifier)	Embedded	Convolutional Neural Networks pre-trained on AlexNet, VGG16, and VGGFace, combined with an Autoencoder and a Gaussian Classifier	[30]
adVAE: A self-adversarial variational autoencoder with Gaussian anomaly prior knowledge for anomaly detection	Rigorous	adVAE (self-adversarial variational autoencoder)	Embedded	Encoder, Generator, and Transformer architecture with latent-space augmentation, and thresholding using Kernel Density Estimation (KDE)	[52]
An Evaluation of One-Class and Two-Class Classification Algorithms for Keystroke Dynamics Authentication on Mobile Devices	Rigorous	parzendd, knndd, mogdd (DdTools)	Filter	Gain Ratio Attribute Eval with Ranker (Weka), comparing feature sets consisting of 71, 17, and 3 variables	[33]
Class-imbalanced classifiers for high-dimensional data	Compliant	One-Class SVM and SVM-THR, with ensembles built using DLDA, Random Forest (RF), and Support Vector Machines (SVM)	Filter	Between–Within Variance Ratio, Mean Decrease in Accuracy, RFE	[53]
Comparison of Statistical and Machine Learning Techniques for Physical Layer Authentication	Rigorous	OCNN (11NN, 1KNN, J1NN, JKNN), OC-SVM	Filter	Extraction of real and imaginary components of subcarriers (Re + Im), followed by analysis of the impact of N variables	[54]
Constructing Kernels for One-Class Support Vector Machine	Rigorous	OCSVM com λ-RBF Kernel e Fisher Kernel	Embedded	Principal Component Analysis (PCA) for weighting in the λ-RBF Kernel, and Fisher Kernel constructed from a Gaussian generative model	[55]
Continuous Authentication Using One-Class Classifiers and Their Fusion	Rigorous	SV1C (One-Class SVM), LOF, Isolation Forest (IF), Elliptic Envelope (EE), and fusion of OCC models	Embedded	PCA (para EE), StandardScaler, Score Fusion (logistic, tanh, soft-sign)	[42]
Controlled graph neural networks with denoising diffusion for anomaly detection	Both	ConGNN (DDPM + GAT + hypersphere consistency)	Embedded	Denoising Diffusion Probabilistic Model (DDPM) for generating embeddings, Graph Attention Network (GAT) for local feature attention, and consistency regularization	[56]
Counterfactual Graph Learning for Anomaly Detection on Attributed Networks	Compliant	CFAD (Counterfactual Graph-based One-Class Classifier using contrastive loss and causal subgraphs)	Embedded	Granger causality for the extraction of relevant subgraphs, combined with Autoencoders (AE), Generative Adversarial Networks (GAN), and contrastive learning	[57]
Data-Driven Version of Multiway Soft Independent Modeling of Class Analogy (N-Way DD-SIMCA): Theory and Application	Both	DD-SIMCA, N-Way DD-SIMCA (Tucker 3), Multiblock DD-SIMCA	Hybrid (Filter + Embedded)	Selection of variable blocks based on efficiency (EFF), followed by Tucker3 decomposition	[46]
Deep Isolation Forest for Anomaly Detection	Rigorous	Deep Isolation Forest (DIF) enhanced with CERE and DEAS techniques	Embedded	Random neural networks (CERE) combined with DEAS for scoring, with implicit variable selection through latent representations	[58]
Distance-based one-class time-series classification approach using local cluster balance	Compliant	OCLCB (One-Class Local Cluster Balance)	Filter	K-means clustering applied to sliding windows, using the Local Cluster Balance vector	[59]
Dynamic classifier selection for one-class classification	Rigorous	OCDCS with a pool of: AENN, KMDD, MSTDD, MoGDD, NNDD, OCSVM, PDDD, PCADD, SOMDD, SVDD	Filter	implicit filtering performed through the predefined configurations of the classifiers in the pool	[34]
Dynamic Construction of Outlier Detector Ensembles With Bisecting K-Means Clustering	Rigorous	CBDS (an ensemble of Local Outlier Factor (LOF) models with feature selection using Bisecting K-Means)	Embedded	LOF with varying MinPts, variable selection based on correlation with simulated ground truth, and local clustering	[60]
Feature Encoding With Autoencoders for Weakly Supervised Anomaly Detection	Compliant	Autoencoder + MLP (inspired by DevNet/ DeepSAD)	Embedded	Autoencoder extracting latent factors (h, r, e), combined with an MLP trained using a weakly supervised loss function	[61]
Feature Selection and Ensemble Learning Techniques in One-Class Classifiers: An Empirical Study of Two-Class Imbalanced Datasets	Rigorous	LOF, OCSVM, IForest, homogeneous and heterogeneous ensembles	Hybrid (Filter + Wrapper + Embedded)	PCA, GA with K-means, C4.5	[40]

(continued on next page)

Table 1 (*continued*)

Study Title	OCC Approach	OCC Algorithm	Variable Selection Approach	Variable Selection Method	Ref.
Feature Selection for MicroRNA Target Prediction – Comparison of One-Class Feature Selection Methodologies	Rigorous	OC-Gaussian, OC-kMeans, OC-kNN	Filter	Zero-Norm Thresholding, Pearson Correlation	[35]
Feature Selection Has a Large Impact on One-Class Classification Accuracy for MicroRNAs in Plants	Rigorous	OCC via DDTools (e.g., OC-Gaussian)	Filter	PCF, ZNF, RFS, LIG, HIG, RFC, HIC, SFC	[62]
Filter Feature Selection for One-Class Classification	Rigorous	ν-SVM (One-Class SVM with kernels linear and RBF)	Filter	SPEC, IS, PC, ICD, IQR; aggregation using Borda count, Majority voting, and Mean rank	[38]
Fixing Bias in Reconstruction-Based Anomaly Detection with Lipschitz Discriminators	Rigorous	LAD (Lipschitz Anomaly Discriminator with corrupted data)	Embedded	Latent representation learning using a Lipschitz-constrained network, Wasserstein-1 score computation, and adversarial representation modeling	[63]
FOOR: Be Careful for Outlier-Score Outliers When Using Unsupervised Outlier Ensembles	Rigorous	FOOR, FOOR+, FOORM, and FOORM + applied to ensembles of: LOF, ODIN, MOD, kNN, CBLOF, HBOS, SOD, VAE, PCA, MCD, ABOD, AutoEncoder, COPOD, KNNAGG, etc.	Embedded	Implicit variable selection through the removal of noisy detectors (ODIN, LOF, MOD) based on scoring, combined with medoid filtering and score trimming.	[64]
Generative Adversarial Active Learning for Unsupervised Outlier Detection	Rigorous	SO-GAAL, MO-GAAL	Embedded	Adversarial learning using specialized sub-generators and deep neural networks (MLPs)	[65]
Gryphon: a semi-supervised anomaly detection system based on one-class evolving spiking neural network	Compliant	eSNN-OCC, Gryphon Framework	Embedded	ROPE encoding, 20 Gaussian Radial Functions per variable, OPAL framework, and Fuzzy c-means clustering	[66]
Interpretable One-Class Classification of Raman Spectra Using Prediction Bands Estimated by Wavelet Regression	Rigorous	Wavelet-based OCC, PC-based Band, DD-SIMCA	Hybrid (Filter + Embedded)	Wavelet thresholding, truncated PCA (wPC), and modified band depth (mBD)	[47]
Isolation-Based Anomaly Detection	Rigorous	Isolation Forest (iForest)	Filter	Kurtosis applied to high-dimensional subspace	[67]
KDE-OCSVM model using Kullback-Leibler divergence to detect anomalies in medical claims	Compliant	KDE-OCSVM	Filter	Variance Threshold, Pearson Correlation Coefficient	[68]
Learning Deep Features for One-Class Classification	Rigorous	Deep One-Class Classification (DOC)	Embedded	Compactness Loss, Descriptiveness Loss, CNN Fine-tuning	[69]
Mahalanobis-Based One-Class Classification	Rigorous	Kernel-based Mahalanobis One-Class Classifier (OCC) and Sparse Mahalanobis OCC	Embedded	Kernel PCA (KPCA) with whitening, eigenvalue-based projection, and support vector selection	[70]
Maximum Correntropy Criterion-Based Hierarchical One-Class Classification	Rigorous	MC-OCELM, HC-OCELM	Embedded	Stacked autoencoders trained using the correntropy criterion (MCC) and optimized via fixed-point iterations	[71]
Mutual information maximization for semi-supervised anomaly detection	Compliant	MIAD (Mutual Information-based Anomaly Detection)	Embedded	Mutual Information Maximization, Adaptive Filter	[72]
One-Class Classification with Extreme Learning Machine	Rigorous	Extreme Learning Machine (ELM) One-Class Classifier using random projections and kernel functions	Embedded	Nonlinear mapping using sigmoid or RBF functions, with selection of the number of neurons and the regularization parameter C	[73]
One-Class Convolutional Neural Network	Compliant	OC-CNN (Convolutional Neural Network using Gaussian noise as pseudo-negative class examples)	Embedded	Automatic feature extraction using AlexNet or VGG16, followed by fine-tuning with pseudo-negative samples	[31]
One-class machines based on the coherence criterion	Rigorous	Coherence-guided One-Class Least Squares (OC-LS) method	Filter	Coherence Criterion	[74]
One-class support vector classifiers: A survey	Both	ocSVM, SVDD, Ramp-OCSVM, etc.	Filter, Wrapper, Embedded	PCA, RFE, PCF, HIG/LIG, Target Supervised Subsetting	[36]
Prototype-based One-Class Classification Learning Using Local Representations	Rigorous	Support Vector Quantization One-Class Classifier (SVQ-OCC) with CSW, BSW, and CEW strategies	Embedded	Locally supervised prototyping using the function f (d, θ), with Gaussian, Student-t, and uniform clipping functions	[75]
Selecting locally specialized classifiers for one-class classification ensembles	Rigorous	OCClustE ensemble combining Weighted One-Class SVM (WOCSVM) and Kernel Fuzzy c-means clustering	Embedded	Soft clustering with membership weighting and partition evaluation using PE, I, FHV, and AIC criteria	[76]
Semi-supervised anomaly detection algorithms: A comparative summary and future research directions	Both	ocSVM, iForest, XGBOD, MO-GAAL, etc.	Filter	PCA	[77]
Sensory quality prediction of coffee assessed by physicochemical parameters and Multivariate model	Rigorous	PCA + One-Class Model	Filter	Correlation ($r > 0.3$), Self-scaling	[37]
Shrinkage methods for one-class classification	Rigorous	Sparse kernel-based One-Class Classifier (OCC) modeling a hypersphere in Reproducing Kernel Hilbert Space (RKHS)	Embedded	LARS, LASSO, and Elastic Net (LARSEN) with hypersphere center shrinkage	[43]
Sparse random projection isolation forest for outlier detection	Compliant	RP-IF, SP-IF, SSP-IF	Embedded	Sparse Random Projection, Soft Sparse Random Projection	[45]
Stable variable selection of class-imbalanced data with precision-recall criterion	Rigorous	SRLRS (Sparse Regularized Logistic Regression with Subsampling)	Embedded	LASSO, Elastic Net, Subsampling (LHO-LOO), Inclusion Frequency, AUPRC	[44]

(continued on next page)

Table 1 (*continued*)

Study Title	OCC Approach	OCC Algorithm	Variable Selection Approach	Variable Selection Method	Ref.
Statistical and Machine Learning-Based Decision Techniques for Physical Layer Authentication	Rigorous	OCNN (1KNN, JKNN, J1NN, 11NN)	Embedded	Features extracted from the channel (real and imaginary parts), evaluated using Euclidean distances and optimized through cross-validation	[78]
The impact of feature selection on one and two-class classification performance for plant microRNAs	Rigorous	OC-Gaussian (via DDTools), OCC com k-means adaptado	Filter	LIG, RFS, RFC, SFC, HIG, HIC, ZNF, PCF	[79]
PBC4occ: A novel contrast pattern-based classifier for one-class classification	Rigorous	PBC4occ (Pattern-Based Classifier for OCC)	Embedded	Contrast pattern filtering via decision tree rules	[80]

Embedded approaches, in contrast, integrate variable selection directly within the model's training process. Techniques such as LASSO and Elastic Net were widely employed to identify sparse and relevant feature subsets, while ensemble models like Random Forests and tree-based classifiers used internal importance metrics for selection [32, 42–45]. These methods offer a balance between accuracy and efficiency and have been successfully applied across OCC scenarios, from anomaly detection to continuous authentication.

Hybrid approaches combined the strengths of multiple strategies. Some studies merged filter and embedded methods [46,47], while others integrated all three (filter, wrapper, and embedded) into unified frameworks [40]. Hybrid models based on deep learning and generative adversarial networks illustrate the trend toward more flexible and adaptive OCC pipelines [32,48].

Together, these findings confirm that variable selection is not merely a preprocessing step, but a central component of OCC modeling. The predominance of embedded and hybrid strategies reflects a methodological shift toward model-aware, scalable, and interpretable feature selection—especially relevant in real-world applications where high dimensionality and class imbalance are prevalent.

Among the studies that applied variable selection in OCC (as shown in Table 1), 68.8 % adopted a rigorous modeling approach, relying solely on data from the target class. In contrast, 20.8 % employed a compliant strategy, and 10.4 % implemented both approaches in comparative designs.

Regarding application domains, 60 % of the studies were methodological, focusing on algorithm development or feature selection strategies. 26 % addressed anomaly detection tasks, including fault detection, industrial monitoring, and fraud prevention. Cybersecurity and biometric authentication represented 10 %, and 4 % were applied to biomedical or bioanalytical contexts, such as Raman spectroscopy or microRNA prediction.

4. Discussion

4.1. Tool effectiveness and research implications

The integration of Scopus, Bibliometrix, VOSviewer, Elicit, and Litmaps created a comprehensive framework for analyzing variable selection in OCC. Each tool contributed uniquely to this process. Scopus provided structured search results, particularly valuable for identifying OCC applications within chemometrics, while Bibliometrix offered deeper insights by analyzing metadata trends, revealing interdisciplinary connections between chemometrics, computer science, and mathematics. These tools also highlighted emerging research directions, such as clustering algorithms and deep learning models, that could enhance variable selection methodologies.

VOSviewer contributed by visually mapping relationships between algorithms, reinforcing the recurrent application of discriminant analysis, SIMCA, and machine learning-based classifiers, including Random Forest and k-Nearest Neighbors (k-NN). Meanwhile, AI-powered tools like Elicit and Litmaps broadened the scope of literature exploration, identifying novel hybrid approaches for outlier detection and methods for improving sensitivity and reducing false positives. However, these AI-driven tools exhibited limitations, particularly in the reliability of predefined filters, requiring validation against more structured bibliometric analyses.

The synergistic use of these tools led to a more dynamic and multidimensional literature review. Scopus and Bibliometrix provided the foundation with structured and quantitative data, while VOSviewer enhanced thematic clustering and visualized term relationships. Elicit and Litmaps extended these insights by uncovering interdisciplinary links and less-cited but potentially impactful studies, ensuring consideration of both established and emerging methodologies. By bridging traditional bibliometric approaches with AI-enhanced exploration, the integration of these tools facilitated a more systematic and forward-looking analysis of variable selection in OCC.

The present study underscores the evolving nature of research methodologies, where traditional literature review strategies are complemented by AI-assisted tools. The combination of structured database searches and AI-driven insights enhances the ability to identify gaps, emerging themes, and influential research directions. This shift supports a more data-driven and interdisciplinary approach, essential for advancing OCC methodologies and improving classification performance. By embracing these tools, researchers can modernize literature review strategies, fostering innovation in variable selection and ensuring that OCC studies remain adaptive and methodologically rigorous.

4.2. Variable selection in OCC

The findings of this literature review highlight the pivotal role of variable selection in enhancing the performance and interpretability of one-class classification (OCC) models. Across the selected studies, a variety of filter, wrapper, embedded, and hybrid strategies were employed to address challenges such as high dimensionality, class imbalance, and the need for generalizable models. Commonly used methods included Recursive Feature Elimination (RFE), Mutual Information (MI), and correlation-based filters, alongside less conventional approaches such as Conditional Mutual Information Distance (CMIDist) and Pearson Correlation-Based Filtering (PCF).

The integration of bibliometric and AI-enhanced tools revealed a growing reliance on hybrid and domain-adapted selection strategies. Rank aggregation methods—such as Borda count and majority voting—were frequently used to consolidate multi-criteria outputs, while evolutionary and swarm intelligence algorithms like Genetic Algorithms (GA) and Particle Swarm Optimization (PSO) featured prominently in optimization routines. Embedded approaches, including LASSO and Elastic Net, contributed to model sparsity and interpretability, particularly when used in combination with dimensionality reduction techniques like PCA. In more recent studies, deep learning-based frameworks, such as Convolutional Neural Networks (CNNs) and autoencoders, were applied for feature extraction and selection in high-complexity datasets.

These strategies were largely implemented within rigorous OCC frameworks, in which optimization procedures rely exclusively on data from the target class [12]. Indeed, among the studies that applied variable selection, 68.8 % followed a rigorous approach, while 20.8 % used a compliant strategy, incorporating information from non-target classes. A smaller portion (10.4 %) adopted both paradigms in comparative experiments. This distribution suggests a consistent preference for rigorous modeling in variable selection routines—likely due to concerns regarding model bias, overfitting, and the theoretical assumptions inherent to OCC.

Moreover, the domain distribution of studies reflects both methodological interest and practical relevance: 60 % were dedicated to methodological contributions, such as proposing new algorithms or improving variable selection workflows; 26 % addressed anomaly detection, including fault detection, industrial monitoring, and fraud prevention; 10 % focused on cybersecurity and biometric authentication, and the remaining 4 % applied OCC to bioanalytical contexts, including Raman spectroscopy and microRNA prediction. These findings reinforce the idea that rigorous variable selection plays a crucial role in advancing OCC research across diverse fields, particularly where data imbalance and the absence of negative class examples are central challenges.

5. Future research directions

Future research should emphasize the refinement of variable selection methodologies tailored specifically to the rigorous requirements of One-Class Classification (OCC). The findings from this review highlight key areas for development:

- **Advanced Hybrid Approaches**: The integration of hybrid methods which blends clustering techniques with evaluation metrics shows potential for capturing sub-concepts within high-dimensional data. Future work should explore extending hybrid models with advanced optimization techniques to enhance robustness across diverse datasets.
- **Adaptability Across Fields**: As studies increasingly apply OCC in fields such as environmental monitoring, cybersecurity, and biomedical research, variable selection algorithms should be designed with flexibility to adapt to field-specific challenges.
- **Scalability for High-Dimensional Data**: Addressing the scalability issues associated with high-dimensional data remains critical. Future studies should focus on dimensionality reduction strategies that minimize computational costs without sacrificing accuracy.
- **Multi-objective Optimization**: The reviewed studies often emphasize the need for multi-objective optimization to balance competing goals such as accuracy, generalization, and computational efficiency.
- **Algorithm Performance Evaluation**: The comparative performance of wrapper, filter, embedded, and hybrid methods remains an open question. Future research should focus on systematically benchmarking these approaches across various OCC scenarios, particularly emphasizing chemometric applications.

By pursuing these directions, researchers can advance the theoretical and practical frameworks of variable selection in OCC while addressing underrepresented applications and improving the scalability and adaptability of these methodologies. This progression will provide a robust foundation for deploying OCC in real-world, high-stakes contexts.

CRediT authorship contribution statement

Mateus P. Schneider: Writing – original draft, Visualization, Software, Investigation, Formal analysis, Data curation. **Cristina Malegori:** Writing – review & editing, Supervision, Conceptualization. **Adriano de A. Gomes:** Writing – review & editing, Supervision, Methodology. **Paolo Oliveri:** Project administration.

Permission for third-party material

Figs. 2–6 were generated by the authors using **VOSviewer**, and Fig. 7 using **Litmaps**. These tools were used in accordance with their respective terms of use, and proper attribution is given in the figure captions.

Funding

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil (CAPES) – Finance Code 001, through the CAPES/PrInt program (Process no. 88887.877611/2023-00).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil (CAPES) – Finance Code 001, for the doctoral scholarship awarded to Mateus P. Schneider under the PRINT program (Process nº 88887.877611/2023-00), which enabled the research stay at the Università degli Studi di Genova. This work was supported in part by CAPES under the International Cooperation Program CAPES/PRINT – Edital nº 41/2017.

Data availability

Data will be made available on request.

References

[1] A. de Araújo Gomes, S.M. Azcarate, P.H. Gonçalves Dias Diniz, D.D. de Sousa Fernandes, G. Veras, Variable selection in the chemometric treatment of food data: a tutorial review, Food Chem. 370 (2022) 131072, https://doi.org/10.1016/j.foodchem.2021.131072.

[2] L. Strani, M. Cocchi, D. Tanzilli, A. Biancolillo, F. Marini, R. Vitale, One class classification (class modelling): state of the art and perspectives, TrAC, Trends Anal. Chem. 183 (2025) 118117, https://doi.org/10.1016/j.trac.2024.118117.

[3] R. Vitale, M. Cocchi, A. Biancolillo, C. Ruckebusch, F. Marini, Class modelling by soft independent modelling of class analogy: why, when, how? A tutorial, Anal. Chim. Acta 1270 (2023) 341304, https://doi.org/10.1016/j.aca.2023.341304.

[4] A. Tugnolo, V. Giovenzana, C. Malegori, P. Oliveri, A. Casson, M. Curatitoli, R. Guidetti, R. Beghi, A reliable tool based on near-infrared spectroscopy for the monitoring of moisture content in roasted and ground coffee: a comparative study with thermogravimetric analysis, Food Control 130 (2021) 108312, https://doi.org/10.1016/j.foodcont.2021.108312.

[5] N. Seliya, A.A. Zadeh, T.M. Khoshgoftaar, A literature review on one-class classification and its potential applications in big data, J. Big Data 8 (1) (2021) 122, https://doi.org/10.1186/s40537-021-00514-x.

[6] L. Awhangbo, R. Bendoula, J.-M. Roger, F. Béline, Fault detection with moving window PCA using NIRS spectra for monitoring the anaerobic digestion process, Water Sci. Technol. 81 (2) (2020) 367–382, https://doi.org/10.2166/wst.2020.117.

[7] A. Biancolillo, F. Marini, C. Ruckebusch, R. Vitale, Chemometric strategies for spectroscopy-based food authentication, Appl. Sci. 10 (18) (2020) 6544, https://doi.org/10.3390/app10186544.

[8] P. Oliveri, C. Malegori, M. Casale, E. Tartacca, G. Salvatori, An innovative multivariate strategy for HSI-NIR images to automatically detect defects in green coffee, Talanta 199 (2019) 270–276, https://doi.org/10.1016/j.talanta.2019.02.049.

[9] O.Y. Rodionova, A.V. Titova, A.L. Pomerantsev, Discriminant analysis is an inappropriate method of authentication, TrAC, Trends Anal. Chem. 78 (2016) 17–22, https://doi.org/10.1016/j.trac.2016.01.010.

[10] Z. Małyjurek, R. Vitale, B. Walczak, Different strategies for class model optimization: a comparative study, Talanta 215 (2020) 120912, https://doi.org/10.1016/j.talanta.2020.120912.

[11] S.S. Khan, M.G. Madden, One-class classification: taxonomy of study and review of techniques, Knowl. Eng. Rev. 29 (3) (2014) 345–374, https://doi.org/10.1017/S026988891300043X.
[12] O.Y. Rodionova, P. Oliveri, A.L. Pomerantsev, Rigorous and compliant approaches to one-class classification, Chemometr. Intell. Lab. Syst. 159 (2016) 89–96, https://doi.org/10.1016/j.chemolab.2016.10.002.
[13] J. Buendía García, M. Lacoue-Nègre, J. Gornay, S. Mas-García, R. Bendoula, J.-M. Roger, Variable selection and data fusion for diesel cetane number prediction, Fuel 332 (2023) 126297, https://doi.org/10.1016/j.fuel.2022.126297.
[14] R. Vitale, J.-M. Roger, Editorial: variable selection in chemometrics, Front. Anal. Sci. 3 (2023) 1186952, https://doi.org/10.3389/frans.2023.1186952.
[15] G. Muller Beuren, M.J. Anzanello, Variable selection using statistical non-parametric tests for classifying production batches into multiple classes, Chemometr. Intell. Lab. Syst. 193 (2019) 103830, https://doi.org/10.1016/j.chemolab.2019.103830.
[16] J.S. Saltz, I. Krasteva, Current approaches for executing big data science projects—a systematic literature review, PeerJ Comput. Sci. 8 (2022) e862, https://doi.org/10.7717/peerj-cs.862.
[17] G. Wagner, R. Lukyanenko, G. Paré, Artificial intelligence and the conduct of literature reviews, J. Inf. Technol. 37 (2) (2022) 209–226, https://doi.org/10.1177/02683962211048201.
[18] P.B. Joshi, Navigating with chemometrics and machine learning in chemistry, Artif. Intell. Rev. 56 (2023) 9089–9114, https://doi.org/10.1007/s10462-023-10391-w.
[19] Z.J. Baum, X. Yu, P.Y. Ayala, Y. Zhao, S.P. Watkins, Q. Zhou, Artificial intelligence in chemistry: current trends and future directions, J. Chem. Inf. Model. 61 (7) (2021) 3197–3212, https://doi.org/10.1021/acs.jcim.1c00619.
[20] R. Houhou, T. Bocklitz, Trends in artificial intelligence, machine learning, and chemometrics applied to chemical data, Anal. Sci. Adv. 2 (3–4) (2021) 128–141, https://doi.org/10.1002/ansa.202000162.
[21] S. Robledo, A.M. Grisales Aguirre, M. Hughes, F. Eggers, "Hasta la vista, baby" – will machine learning terminate human literature reviews in entrepreneurship? J. Small Bus. Manag. 61 (3) (2023) 1314–1343, https://doi.org/10.1080/00472778.2021.1955125.
[22] J. Baas, M. Schotten, A. Plume, G. Côté, R. Karimi, Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies, Quant. Sci. Stud. 1 (1) (2020) 377–386, https://doi.org/10.1162/qss_a_00019.
[23] J. Zhu, W. Liu, A tale of two databases: the use of web of science and scopus in academic papers, Scientometrics 123 (2020) 321–335, https://doi.org/10.1007/s11192-020-03387-8.
[24] N.J. van Eck, L. Waltman, Software survey: vosviewer, a computer program for bibliometric mapping, Scientometrics 84 (2) (2010) 523–538, https://doi.org/10.1007/s11192-009-0146-3.
[25] J.T. McAllister, L. Lennertz, Z. Atencio Mojica, Mapping a discipline: a guide to using VOSviewer for bibliometric and visual analysis, Sci. Technol. Libr. 41 (3) (2022) 319–348, https://doi.org/10.1080/0194262X.2021.1991547.
[26] M. Aria, C. Cuccurullo, Bibliometrix: an R-tool for comprehensive science mapping analysis, J. Informetr. 11 (4) (2017) 959–975, https://doi.org/10.1016/j.joi.2017.08.007.
[27] D.O. Oyewola, E.G. Dada, Exploring machine learning: a scientometrics approach using bibliometrix and VOSviewer, SN Appl. Sci. 4 (2022) 143, https://doi.org/10.1007/s42452-022-05027-7.
[28] S. Whitfield, M.A. Hofmann, Elicit: AI literature review research assistant, Public Serv. Q. 19 (3) (2023) 201–207, https://doi.org/10.1080/15228959.2023.2224125.
[29] D. Sulisworo, Exploring research idea growth with litmap: visualizing literature review graphically, Bincang Sains Dan Teknol 2 (2) (2023) 48–54, https://doi.org/10.56741/bst.v2i02.323.
[30] P. Oza, V.M. Patel, Active authentication using an autoencoder regularized CNN-based one-class classifier, IEEE Int. Conf. Autom. Face Gesture Recognit. (2019) 1–8, https://doi.org/10.1109/FG.2019.8756525.
[31] P. Oza, V.M. Patel, One-class convolutional neural network, IEEE Signal Process. Lett. 26 (2) (2019) 277–281, https://doi.org/10.1109/LSP.2018.2889273.
[32] M.J. Anzanello, F.S. Fogliatto, A review of recent variable selection methods in industrial and chemometrics applications, Eur. J. Ind. Eng. 8 (5) (2014) 619–645, https://doi.org/10.1504/EJIE.2014.065731.
[33] M. Antal, L.Z. Szabó, An evaluation of one-class and two-class classification algorithms for keystroke dynamics authentication on Mobile devices, in: Proc. Int. Conf. Control Syst. Comput. Sci. (CSCS), IEEE, 2015, pp. 343–350, https://doi.org/10.1109/CSCS.2015.16.
[34] B. Krawczyk, M. Woźniak, Dynamic classifier selection for one-class classification, Knowl.-Based Syst. 107 (2016) 43–53, https://doi.org/10.1016/j.knosys.2016.05.054.
[35] M. Yousef, J. Allmer, W. Khalifa, Feature selection for microRNA target prediction: comparison of one-class feature selection methodologies, in: Proc. Int. Joint Conf. Biomed. Eng. Syst. Technol. (BIOSTEC) – Vol. 3: BIOINFORMATICS, SciTePress, 2016, pp. 216–225, https://doi.org/10.5220/0005701602160225.
[36] S. Alam, S.K. Sonbhadra, S. Agarwal, P. Nagabhushan, One-class support vector classifiers: a survey, Knowl.-Based Syst. 196 (2020) 105754, https://doi.org/10.1016/j.knosys.2020.105754.
[37] L.O.C. Domingues, A.O. Garcia, M.M.C. Ferreira, M.A. Morgano, Sensory quality prediction of coffee assessed by physicochemical parameters and multivariate model, Coffee Sci. 15 (2020) e151654, https://doi.org/10.25186/cs.v15i.1654.
[38] L.H.N. Lorena, A.C.P.L.F. Carvalho, A.C. Lorena, Filter feature selection for one-class classification, J. Intell. Rob. Syst. 80 (S1) (2015) S227–S243, https://doi.org/10.1007/s10846-014-0101-2.
[39] B. Krawczyk, M. Woźniak, Accuracy and diversity in classifier selection for one-class classification ensembles, Proc. IEEE Symp. Comput. Intell. Ensemble Learn. (2013) 46–51, https://doi.org/10.1109/CIEL.2013.6613139.
[40] Y. Zhang, J. Wang, Y. Li, X. Liu, A novel deep learning approach for network intrusion detection, IEEE Access 9 (2021) 123456–123467, https://doi.org/10.1109/ACCESS.2021.3051969.
[41] Y.-S. Jeong, I.-H. Kang, M.-K. Jeong, D. Kong, A new feature selection method for one-class classification problems, IEEE Trans. Syst. Man Cybern. C Appl. Rev. 42 (6) (2012) 1500–1509, https://doi.org/10.1109/TSMCC.2012.2196794.
[42] R. Kumar, P.P. Kundu, V.V. Phoha, Continuous authentication using one-class classifiers and their fusion, Proc. IEEE Int. Conf. Identity Secur. Behav. Anal. (2018) 1–8, https://doi.org/10.1109/ISBA.2018.8311467.
[43] P. Nader, P. Honeine, P. Beauseroy, Shrinkage methods for one-class classification, Eur. Signal Process. Conf. (EUSIPCO) 23 (2015) 135–139, https://doi.org/10.1109/EUSIPCO.2015.7362360.
[44] G.-H. Fu, F. Xu, B.-Y. Zhang, L.-Z. Yi, Stable variable selection of class-imbalanced data with precision-recall criterion, Chemometr. Intell. Lab. Syst. 171 (2017) 241–250, https://doi.org/10.1016/j.chemolab.2017.10.015.
[45] X. Tan, J. Yang, S. Rahardja, Sparse random projection isolation forest for outlier detection, Pattern Recognit. Lett. 163 (2022) 65–73, https://doi.org/10.1016/j.patrec.2022.09.015.
[46] A.P. Pagani, G. Camargo, G.A. Ibañez, A.C. Olivieri, A.L. Pomerantsev, O. Y. Rodionova, Data-driven version of multiway soft independent modeling of class analogy (N-way DD-SIMCA): theory and application, Anal. Chem. 96 (12) (2024) 4845–4853, https://doi.org/10.1021/acs.analchem.3c05096.
[47] T.H. Avohou, P.-Y. Sacré, P. Hubert, E. Ziemons, Interpretable one-class classification of Raman spectra using prediction bands estimated by wavelet regression, Anal. Chem. 94 (10) (2022) 4183–4191, https://doi.org/10.1021/acs.analchem.1c04098.
[48] G.M. Beuren, M.J. Anzanello, Variable selection using statistical non-parametric tests for classifying production batches into multiple classes, Chemometr. Intell. Lab. Syst. 193 (2019) 103830, https://doi.org/10.1016/j.chemolab.2019.103830.
[49] J. Yu, X. Gao, B. Li, F. Zhai, J. Lu, B. Xue, S. Fu, C. Xiao, A filter-augmented auto-encoder with learnable normalization for robust multivariate time series anomaly detection, Neural Netw. 170 (2024) 478–493, https://doi.org/10.1016/j.neunet.2023.11.047.
[50] P. Perera, V.M. Patel, A joint representation learning and feature modeling approach for one-class recognition, in: Proc. Int. Conf. Pattern Recognit. (ICPR), IEEE, 2021, pp. 10644–10651, https://doi.org/10.1109/ICPR48806.2021.9412390.
[51] W. Pei, B. Xue, M. Zhang, L. Shang, X. Yao, Q. Zhang, A survey on unbalanced classification: how can evolutionary computation help? IEEE Trans. Evol. Comput. 28 (2) (2024) 353–373, https://doi.org/10.1109/TEVC.2023.3257230.
[52] X. Wang, Y. Du, S. Lin, P. Cui, Y. Shen, Y. Yang, adVAE: a self-adversarial variational autoencoder with Gaussian anomaly prior knowledge for anomaly detection, Knowl.-Based Syst. 190 (2020) 105187, https://doi.org/10.1016/j.knosys.2019.105187.
[53] W.-J. Lin, J.J. Chen, Class-imbalanced classifiers for high-dimensional data, Briefings Bioinf. 14 (1) (2013) 13–26, https://doi.org/10.1093/bib/bbs006.
[54] L. Senigagliesi, M. Baldi, E. Gambi, Comparison of statistical and machine learning techniques for physical layer authentication, IEEE Trans. Inf. Forensics Secur. 16 (2021) 1506–1521, https://doi.org/10.1109/TIFS.2020.3033454.
[55] B. Zhang, J. Zhu, H. Tian, Constructing kernels for one-class support vector machine, in: *Proc.* 14th Int. Symp. Distrib. Comput. Appl. Bus. Eng. Sci., 2015, pp. 468–471, https://doi.org/10.1109/DCABES.2015.123.
[56] X. Li, C. Xiao, Z. Feng, S. Pang, W. Tai, F. Zhou, Controlled graph neural networks with denoising diffusion for anomaly detection, Expert Syst. Appl. 237 (2024) 121533, https://doi.org/10.1016/j.eswa.2023.121533.
[57] C. Xiao, X. Xu, Y. Lei, K. Zhang, S. Liu, F. Zhou, Counterfactual graph learning for anomaly detection on attributed networks, IEEE Trans. Knowl. Data Eng. 35 (10) (2023) 10540–10553, https://doi.org/10.1109/TKDE.2023.3250523.
[58] H. Xu, G. Pang, Y. Wang, Y. Wang, Deep isolation forest for anomaly detection, IEEE Trans. Knowl. Data Eng. 35 (12) (2024) 12591–12604, https://doi.org/10.1109/TKDE.2023.3270293.
[59] T. Hayashi, D. Cimr, F. Studnička, H. Fujita, D. Bušovský, R. Cimler, A. Selamat, Distance-based one-class time-series classification approach using local cluster balance, Expert Syst. Appl. 235 (2024) 121201, https://doi.org/10.1016/j.eswa.2023.121201.
[60] R.R.Z. Koko, I.A. Yassine, M.A. Wahed, J.K. Madete, M.A. Rushdi, Dynamic construction of outlier detector ensembles with bisecting K-means clustering, IEEE Access 11 (2023) 24431–24447, https://doi.org/10.1109/ACCESS.2023.3252004.
[61] Y. Zhou, X. Song, Y. Zhang, F. Liu, C. Zhu, L. Liu, Feature encoding with autoencoders for weakly supervised anomaly detection, IEEE Transact. Neural Networks Learn. Syst. 33 (6) (2022) 2454–2465, https://doi.org/10.1109/TNNLS.2021.3086137.
[62] M. Yousef, M.D. Saçar Demirci, W. Khalifa, J. Allmer, Feature selection has a large impact on one-class classification accuracy for microRNAs in plants, Adv. Bioinform. 2016 (2016) 5670851, https://doi.org/10.1155/2016/5670851.
[63] A. Tong, G. Wolf, S. Krishnaswamy, Fixing bias in reconstruction-based anomaly detection with Lipschitz discriminators, in: Proc. IEEE Int. Workshop Mach. Learn. Signal Process. (MLSP), IEEE, 2020, pp. 1–6, https://doi.org/10.1109/MLSP49062.2020.9231832.

[64] J. Yang, S. Rahardja, S. Rahardja, FOOR: be careful for outlier-score outliers when using unsupervised outlier ensembles, IEEE Trans. Comput. Soc. Syst. 11 (2) (2024) 2843–2852, https://doi.org/10.1109/TCSS.2023.3280593.

[65] Y. Liu, Z. Li, C. Zhou, Y. Jiang, J. Sun, M. Wang, X. He, Generative adversarial active learning for unsupervised outlier detection, IEEE Trans. Knowl. Data Eng. 32 (8) (2019) 1517–1528, https://doi.org/10.1109/TKDE.2019.2905606.

[66] K. Demertzis, L.S. Iliadis, I. Bougoudis, Gryphon: a semi-supervised anomaly detection system based on one-class evolving spiking neural network, Neural Comput. Appl. 32 (9) (2020) 4303–4314, https://doi.org/10.1007/s00521-019-04363-x.

[67] F.T. Liu, K.M. Ting, Z.-H. Zhou, Isolation-based anomaly detection, ACM Trans. Knowl. Discov. Data 6 (1) (2012) 39, https://doi.org/10.1145/2133360.2133363, 3:1–3.

[68] X. Wang, W. Hou, H. Zhang, J. Wang, M. Goh, Z. Tian, K. Shen, KDE-OCSVM model using Kullback-Leibler divergence to detect anomalies in medical claims, Expert Syst. Appl. 200 (2022) 117056, https://doi.org/10.1016/j.eswa.2022.117056.

[69] P. Perera, V.M. Patel, Learning deep features for one-class classification, IEEE Trans. Image Process. 28 (11) (2019) 5450–5463, https://doi.org/10.1109/TIP.2019.2917862.

[70] C. Nader, P. Honeine, C. Richard, Mahalanobis-based one-class classification, in: 2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP), IEEE, 2014, pp. 1–6, https://doi.org/10.1109/MLSP.2014.6958934.

[71] J. Cao, H. Dai, B. Lei, C. Yin, H. Zeng, A. Kummert, Maximum correntropy criterion-based hierarchical one-class classification, IEEE Transact. Neural Networks Learn. Syst. 32 (8) (2021) 3748–3754, https://doi.org/10.1109/TNNLS.2020.3015356.

[72] S. Liu, M. Tian, Mutual information maximization for semi-supervised anomaly detection, Knowl.-Based Syst. 284 (2024) 111196, https://doi.org/10.1016/j.knosys.2023.111196.

[73] Q. Leng, H. Qi, J. Miao, W. Zhu, G. Su, One-class classification with extreme learning machine, Math. Probl Eng. (2015) 1–11, https://doi.org/10.1155/2015/412957, 2015.

[74] Z. Noumir, P. Honeine, C. Richard, One-class machines based on the coherence criterion, IEEE Stat. Signal Process. Workshop (SSP) (2012) 600–603, https://doi.org/10.1109/SSP.2012.6319771.

[75] D. Staps, R. Schubert, M. Kaden, A. Lampe, W. Hermann, T. Villmann, Prototype-based one-class-classification learning using local representations, in: Int. Joint Conf. Neural Netw. (IJCNN), IEEE, 2022, pp. 1–8, https://doi.org/10.1109/IJCNN55064.2022.9892912.

[76] B. Krawczyk, B. Cyganek, Selecting locally specialised classifiers for one-class classification ensembles, Pattern Anal. Appl. 20 (2) (2017) 427–439, https://doi.org/10.1007/s10044-015-0505-z.

[77] M.E. Villa-Pérez, M. Álvarez-Carmona, O. Loyola-González, M.A. Medina-Pérez, J. C. Velazco-Rossell, K.-K.R. Choo, Semi-supervised anomaly detection algorithms: a comparative summary and future research directions, Knowl.-Based Syst. 218 (2021) 106878, https://doi.org/10.1016/j.knosys.2021.106878.

[78] L. Senigagliesi, M. Baldi, E. Gambi, Statistical and machine learning-based decision techniques for physical layer authentication, in: Proc. IEEE Glob. Commun. Conf. (GLOBECOM), IEEE, 2019, pp. 1–6, https://doi.org/10.1109/GLOBECOM38437.2019.9013609.

[79] W. Khalifa, M. Yousef, M.D. Saçar Demirci, J. Allmer, The impact of feature selection on one and two-class classification performance for plant microRNAs, PeerJ 4 (2016) e2135, https://doi.org/10.7717/peerj.2135.

[80] D.L. Aguilar, O. Loyola-González, M.A. Medina-Pérez, L. Cañete-Sifuentes, K.-K. R. Choo, PBC4occ: a novel contrast pattern-based classifier for one-class classification, Future Gener. Comput. Syst. 125 (2021) 71–90, https://doi.org/10.1016/j.future.2021.06.046.