Feature engineering for classification: A bibliometric literature review

Frank M. A. Klingert

frank.klingert@informatik.hs-fulda.de,
Department of Applied Computer Science,
Fulda University of Applied Sciences,
Leipziger Str. 123, 36037 Fulda, Germany

Abstract. With the growing interest in machine learning, feature engineering is increasingly applied to improve model accuracy, robustness and understandability. As feature engineering is often related to domain- and model-specific aspects, existing literature reviews are usually domain-specific. However, there are also more general and domain-independent aspects, which could benefit from a more domain-independent approach with a separated research community. This paper extends the literature by performing a domain-independent citation and co-citation analysis on 999 papers from the feature engineering literature in the context of classification problems. The paper concludes that feature engineering is not yet a separate research area. Existing research often focuses on feature selection and extraction while feature construction is rarely discussed. Most papers are strongly embedded in domain and/or model-specific research. The co-citation analysis leads to 8 different clusters. While the cluster "dimensionality reduction (in graphs)" is mostly related to a feature extraction aspect, the other clusters are related to specific application areas, i.e., heartbeat classification (ECGs), malware detection, remote sensing, sleep apnea (ECGs), time series analysis, human activity recognition. With a growing community, more workshops, conferences, papers and journals, feature engineering could become a research area in its own right.

Keywords: Feature Engineering, Data Preparation, Classification, Literature Review, Citation-Analysis, Co-Citation Analysis, Bibliometrics

1 Introduction

Feature Engineering "is a process of preparing transforming, constructing, and filtering features with the goal of optimizing the performance of a data analysis task." [1]. It is a crucial step in the data science process, which is often underestimated [2]. Adding domain knowledge from stakeholders to the data can significantly improve predictive results [e.g., 3, 4]. While existing research applying feature engineering often focuses on performance improvement, it is important to improve the robustness and understandability of models as well [5].

There are already literature reviews on domain-specific feature engineering papers. For example, Wang et al. [6] provide a literature review on feature engineering for energy prediction applications. Wang et al. [6] demand generalizability, e.g. the comparison of "feature engineering methods on a more general data platform".

However, Wang et al. [6] only focus on the application area of energy prediction and results might differ with a domain-independent literature search. Existing papers and books demonstrate that a domain-independent approach is generally possible [e.g., 2, 5, 7]. Nevertheless, a manual literature review that covers a broader field might not be able to include all existing papers.

Moreover, traditional literature reviews involving manual classification leave some questions open as they discuss different aspects compared to semi-automated bibliometric approaches. Is feature engineering already a research area on its own or is it only a minor dependent part of machine learning research? Are researchers linking their results to domain-specific questions or do they also embed it in existing research tackling specific feature engineering topics? Which research topics are already discussed often, and which research topics do not have a wide research body? Those questions are especially relevant to researchers new to the field of feature engineering.

Analysis of citations and co-citations allows the identification of important clusters bottom-up as the decision about citing publications is made by each author and not by a central organization. The number of citations is often shown on scientific search engines like Google Scholar or Web of Science. Restricting the analysis to citations in papers about feature engineering reveals the emergent structure of the research area by considering the decentralized decisions of authors from the field.

This research paper extends the existing feature engineering literature by conducting a citation and co-citation analysis. It identifies important papers and relevant research clusters based on 999 papers and more than 50,000 citations. Thereby, it can disclose the intellectual structure of the research area. Feature engineering research is still at an early stage of development. Although there are generic aspects, feature engineering research is often done domain-specific and closely linked to machine learning research. No paper in the top10 of the most-often cited papers and only one cluster in the co-citations analysis are about feature engineering aspects. Usually, the papers are strongly embedded in machine learning (ML) literature and/or linked to application domains.

The research paper is structured as follows: based on existing literature, the existing knowledge about feature engineering literature is summarized and research questions are derived. Second, the dataset as well as citation and co-citation analysis are introduced. Third, the paper identifies important publications and relevant research clusters in a citation and co-citation analysis. Finally, it discusses the results and concludes.

2 Prior research and research gap

The definition from the introduction implies that feature engineering includes the steps of feature construction, feature extraction and feature selection as depicted in figure 1 [for further details see 6]. It is important to note that deviating definitions exist, e.g. not

mentioning one of these steps [e.g., 5]. Additionally, different papers assign feature engineering to different data science steps. For instance, Amershi et al. [8] contextualize feature engineering as a separate model-oriented step between data labeling and model training. Feature engineering indeed is often model-specific as its effect differs depending on the algorithms used. However, it is also possible to define certain mathematical relationships and categorize new features independent of the model [9].

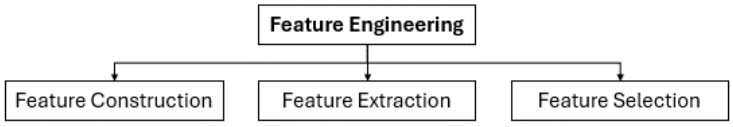

Fig. 1. Aspects of Feature Engineering [6]

This paper contextualizes feature engineering as the last step of the data preparation phase in CRISP-DM process [10] following data preprocessing and followed by the modeling phase. While data preprocessing for example tackles data quality issues and changes the row-oriented content of variables, feature engineering is often column-oriented. It selects, extracts or combines variables into new meaningful features. Still, there are similar tasks in both steps as mapping values to a certain distribution might solve the quality issue of extreme values, but also might be applied during feature engineering to support model training.

A manual literature review on feature engineering applied to energy prediction problems is offered by Wang et al. [6]. They took 172 energy prediction articles from the Web of Science database into account and found an increase in the number of papers which is even accelerated in 2020 and 2021. Consistently, with the above given definition, they see feature engineering as the last step of the data preparation phase and include the tasks feature construction, feature selection and feature extraction. Feature selection is by far the most common method in the reviewed papers. This observation is also made in two co-citation analyses about a machine learning journal [11] and big data [12]. Wang et al. [6] observe that feature engineering methods are often verified, but not systematically compared to other methods. Therefore, the "study's findings are limited to their specific cases due to the insufficient generalizability of the case study" [6].

Other literature reviews, e.g., Croft et al. [13], focus on data preparation and explicitly exclude feature engineering as they categorize this step as model-oriented. Several articles and books give a comprehensive overview of the topic [e.g., 2, 5, 7], however, a domain-independent literature review is still missing.

A more general literature review would support researchers to learn from results in other domains – especially when literature about feature engineering is rare in the own application field. There are domain-specific and domain-independent aspects of feature engineering, but in both cases, it might make sense to identify general methods and their effects (e.g. division, ration division [9]). A more general example is a publication in PLOS Computational Biology about "eleven quick tips for data cleaning and feature engineering" [2] which mostly are relevant in other domains as well.

Young scholars and researchers new to the field face difficulties finding relevant papers. This is often easier, if the research field already has conferences, journals and a profound body of research. Currently, it is an open question if feature engineering already has begun to develop into a research area on its own. On the one hand, there are already a special issue and books, but on the other side, most papers are domain-specific and a special issue in a journal of machine learning [14] again demonstrates the strong embedding in machine learning research. Therefore, it is worth asking research question 1 (RQ1): *"Is feature engineering already a research area in its own right or is it only a minor dependent part of machine learning research?"*

Feature engineering can be seen as a domain- and algorithm-specific task that is strongly linked to the model [8]. Contrary, feature construction, selection and extraction share domain-independent methods like principal component analysis (PCA) and even domain-specific feature construction can be categorized mathematically, e.g. counts, differences, ratio, root distance and others [9]. Papers showing the effect of feature engineering in comparison to the effect of model selection and tuning also demonstrate the high importance of feature engineering [3, 4] which could lead to the assumption that this is reflected in the structure of scientific literature as well. Thus, this paper asks research question 2 (RQ2): *"Do researchers link their results only to domain-specific questions or do they also embed it in existing research about feature engineering methods?"*

Researchers in feature engineering who would like to embed their papers into existing literature can profit from the identification of relevant research topics. Wang et al. [6] identify feature selection as the most often applied aspect of feature engineering in prior research, but it is still questionable if a separate research community already has developed. Consequently, one could additionally ask research question 3 (RQ3): *"Which research topics are already discussed often, and which research topics do not have a wide research body?"*

3 Method and dataset

Both methods, citation and co-citation analysis, have been applied to quantitatively measure different aspects of research. The citation analysis aims to show the influence of certain publications within the literature about feature engineering. In contrast, the co-citation analysis focuses on the emergent structure of research communities. In the past, citation and co-citation analyses have been applied to several research fields including machine learning, big data, prediction markets and open source innovation [e.g., 11, 12, 15, 16].

The search for citing papers in the Web of Science database uses the following search terms. Web of Science (www.webofknowledge.com) is a citation database by Thomson Reuters and was also used in one of the past literature reviews about feature engineering [6]. In this paper, the following search terms are used: "feature engineering" AND "classification" AND "data". While "feature engineering" is the main term, the other terms exclude different meanings like feature engineering in software engineering by focusing on classification tasks and data. Other terms like "feature extraction" or

"feature selection" are intentionally not used to reduce the papers to a manageable amount and to show which aspects are most relevant, when the generic term "feature engineering" is used.

This leads to a data set with 999 papers and 50,153 citations, which have been collected in May 2024. Thus, every paper has about 50 citations on average which is rather high compared to other co-citation analyses [e.g., 15, 17]. Similar to Klingert [15], the papers are mainly journal papers as those are in focus of Web of Science database. Consequently, conference proceedings, webpages, bachelor and master theses as well as other work are systematically underrepresented in this selection. Nevertheless, journal publications contain references to other publication types as well. Thus, these publications are included in the data set if they are cited within journal publications.

The papers in the data set represent a diverse range of application areas. Feature engineering is a new term and research area because no publication of the data set was published prior 2008. The topics include disciplines like tunnel exploration, earth observation, ECG heartbeat classification, bank customer behavior and more domain-independent methodological papers about dimensionality reduction, classification algorithms and automated feature engineering. Therewith, the citation and co-citation analysis is independent of certain application areas. As papers might be cited because of minor-relevant aspects, a threshold of citations is applied to the cited publications in the following step.

After the identification of the citing papers and the collection of their citations, a unique ID is assigned to the cited publications. This data preparation task is done in two steps: (1) The DOI as a unique identifier is used to match unique papers to a unique ID. (2) If the DOI is not contained or does not match, author and publication year are used. Here, shortened first names are matched as well, e.g. "Breiman L 2001" is matched with "Breiman Leo 2001". Besides one completely unrelated paper from the field of economics, which was present only in Web of Science database and not in the PDFs of the citing papers, no further data cleaning followed.

The citation analysis focuses on the most important publications by including publications only in the top10. Thus, publications must be cited at least 45 times in 999 papers.

For the co-citation analysis, two inclusion criteria are applied like prior co-citation research. First, the publications must be cited at least five times. This ensures that only relevant publications are contained and restricts the size of the network. Second, a co-citation value is derived for each link between the two publications. If there are two cited publications A and B, the co-citations value of AB is calculated as follows [18]:

$$CoCitationValue_{AB} = \frac{(CoCitation_{AB})^2}{min\ (Citation_A; Citation_B) * mean\ (Citation_A; Citation_B)}$$

$CoCitation_{AB}$ represents the number of conjoint citations of publications A and B, while $Citation_A$ represents the number of citations of publication A and $Citation_B$ the number of citations of publication B. Thus, the co-citation value is not only based on the co-citations but relates them to the number of citations as well. Publications that are often cited, are also more likely to be cited together with other publications. Only if the conjoint citations cover a significant number of their total citations, the link has a high

co-citation value and is contained in the analysis. This ensures, that papers which are often cited in general are not included, while papers cited often together with other papers about a certain sub-aspect are shown in the graph. The links are included if the co-citation value is above a certain threshold. A threshold of 0.3 is applied in line with prior research [15, 16] to focus on relevant co-citations within the co-citation network. To increase interpretability, minor clusters with only two publications are eliminated from the network as those clusters are considered less important.

4 Results

The results are presented in two subsections. Within the first sub-section, the most important publications are identified. Within the second sub-section, research clusters are identified, and relevant research topics are derived.

4.1 Important publications

Table 1 shows the most often cited publications, i.e., the top10 publications. The rank is derived based on the number of citations in column "Cit.". The publication is specified by year, first author and the title which has been shortened in some instances. The column "Research line" categorizes the papers.

Table 1. Top10 most cited publications

Rank	Cit.	Year	First author	Title	Research line
1	104	2001	Breiman L	Random forests	ML algorithm
2	80	2015	He K	Deep residual learning for image recognition	ML application
3	71	2016	Chen TQ	XGBoost: a scalable tree boosting system	ML algorithm
4	71	2017	Krizhevsky A	ImageNet classification with deep…	ML application
5	59	2002	Chawla NV	SMOTE: synthetic minority over-sampling…	Data preparation
6	58	1997	Hochreiter S	Long short-term memory	ML algorithm
7	56	2015	Lecun Y	Deep learning	ML algorithm
8	53	1995	Cortes C	Support-vector networks	ML algorithm
9	51	2016	Goodfellow I	Deep learning	ML algorithm
10	45	2014	Kingma DP	Adam: a method for stochastic optimization	Optimization

The most often cited publications are not from the field of feature engineering but from machine learning. Consistently, with 104 in 999 citations the most influential publication is the paper about "Random forests" from Breiman L [19]. Further descriptions or applications of machine learning algorithms follow. Only two papers are not in one of those two categories. Most papers are published in journals, followed by conference proceedings and books.

An indication for the dominance of "feature selection" gives the following Table 2 with the same structure, but filtering papers with a focus on feature engineering. The most-often cited paper is the introduction of a special issue, where Guyon and Elisseeff [14] give an overview of the topic of feature selection but also discuss some feature

construction aspects. They provide a checklist to solve feature selection problems and give an overview of methods. In the end, they recommend starting with a linear method and selecting variables with a ranking method and a nested subset selection method. Among the top10, five papers are tackling feature selection, four of them are in the top5. For example, Tibshirani [20] introduces the lasso method for regression which can be used for feature selection. Chandrashekar and Sahin [21] provide a survey on feature selection methods and apply them to standard datasets. Feature extraction is also important and covered by four papers. For example, Hinton and Salakhutdinov [22] reduce the dimensionality of data with neural networks. Feature construction is much less important and only represented by one paper ranked on the 10th place.

Table 2. Top10 most cited publications with a focus on feature engineering

Rank	Cit.	Year	First author	Title	Research line
18	31	2003	Guyon I	An introduction to variable and feature selection	Feature selection
36	18	2006	Hinton GE	Reducing the dimensionality of data with…	Feature extraction
43	16	1996	Tibshirani R	Regression shrinkage and selection via the Lasso	Feature selection
46	15	2014	Chandrashekar G	A survey on feature selection methods	Feature selection
54	14	2005	Peng HC	Feature selection based on mutual…	Feature selection
57	14	1987	Wold S	Principal component analysis	Feature extraction
61	14	2018	Li JD	Feature selection: A data perspective	Feature selection
67	13	2020	McInnes L	Umap: Uniform manifold approximation...	Feature extraction
70	12	2017	Hamilton WL	Representation learning on graphs: Methods…	Feature extraction
76	12	1973	Haralick RM	Textural features for image classification	Feature construction

Feature engineering seems to be at an early stage of development as no paper from the top10 is related to feature engineering. While the top10 papers in Table 1 range from 1995 to 2017, the papers about feature engineering in Table 2 range from 1973 to 2020. Considering that only 40 (about 4 %) of the 999 citing papers were published in 2017 or earlier, the term “feature engineering” is quite new. Nevertheless, papers about “feature selection” and “feature extraction” exist for a longer time and there would be enough potential to cite them. Instead, machine learning is still dominating the top10. This gives a first indication, that feature engineering still has not developed into an independent research area in its own right (RQ1).

All in all, feature engineering can be considered a young discipline. There is no big variety of “classic” papers and papers are still related to application- or method-domains. Feature selection seems to be the most developed branch of feature engineering at first glance.

4.2 Relevant research areas

The co-citation analysis leads to 8 different clusters in figure 2. The nodes of the network represent the publications and the edges represent the co-citations. The node size reflects the number of citations in the dataset and the edge size corresponds with the co-citation value. No publication from Table 1 or Table 2 is contained in the co-citation

network as the top-papers usually are cited often in papers about feature engineering in general, but not specifically in a context of a certain aspect. Therefore, the co-citation analysis can reveal the aspects relevant in a research area. As small clusters are removed, clusters range from three to seven nodes including three major clusters with six or seven nodes. The co-cited publications begin in 2000 (with one exception from 1985), but most of the publications are from 2015 or newer. This fact indicates that feature engineering is still at an early stage of developing its own body of research (RQ1).

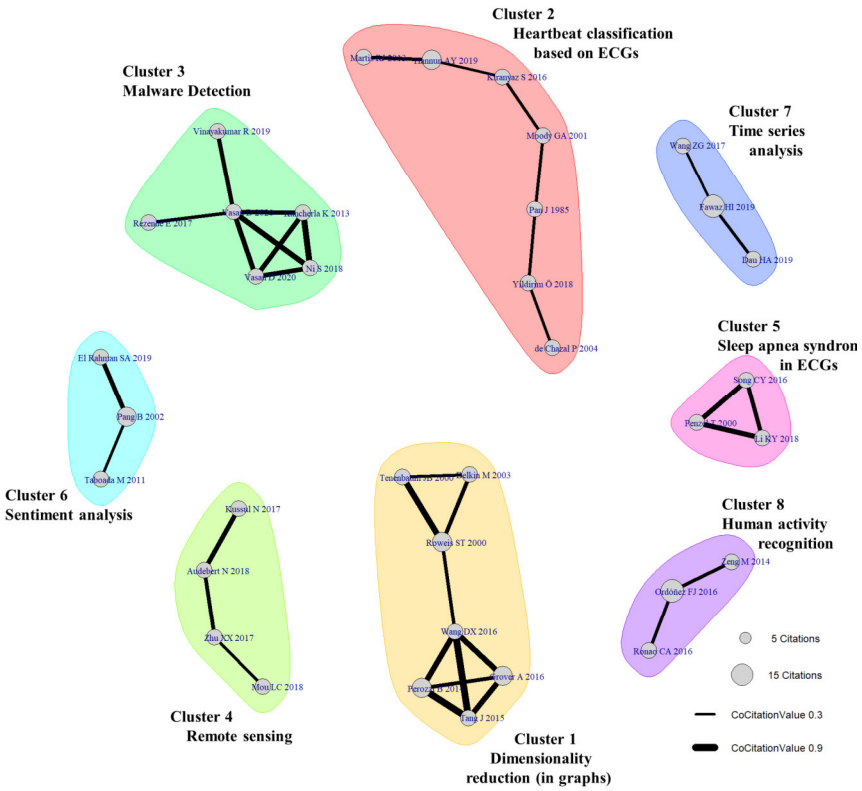

Fig. 2. Co-citation network of citations (nodes: number of citations >= 5 and edges: co-citation value >= 0.3)

All clusters are linked to a certain ML application as shown in Table 3. Seven out of eight clusters mainly share a certain application which includes heartbeat classification based on ECGs, malware detection, remote sensing, sleep apnea syndrome in ECGs, sentiment analysis, time series analysis and human activity recognition. At least one cluster is mainly dominated by the topic of dimensionality reduction which belongs to

the area of feature extraction. However, most of the papers are applying dimensionality reduction to the application area of graph analysis. Taking this result into account, feature engineering has not yet been able to generate a research field on its own where researchers from different application fields are strongly connected (RQ1). Instead, researchers link their results to domain-specific problems (RQ2). Interestingly, the presence of a feature extraction problem contradicts the results of prior research that feature selection is the research branch most developed.

Table 3. Classification of Cluster topics

Cluster	Cluster topic	Category	Nodes
1	Dimensionality reduct. (in graphs)	Feature Extract. (/ML Appl.)	7
2	Heartbeat classification based on ECGs	ML Application	7
3	Malware Detection	ML Application	6
4	Remote sensing	ML Application	4
5	Sleep apnea syndrome in ECGs	ML Application	3
6	Sentiment analysis	ML Application	3
7	Time series analysis	ML Applications	3
8	Human activity recognition	ML Application	3

The first major cluster is the cluster with the label “Dimensionality reduction (in graphs)”. All papers are tackling the task of dimensionality reduction and three of them explicitly mention it in their title. While all major clusters are rather weakly connected, this cluster has 10 edges and, therewith, is the most densely connected cluster among them. The paper with four edges from Wang et al. [23] proposes a structural deep network embedding method to capture the network structure and is related to the application field “graphs”. This recent paper is connected to the oldest paper in the cluster which has three nodes [24]. Roweis and Saul [24] introduce locally linear embedding and, therewith, map inputs (not only graphs) into a coordinate system of lower dimensionality. All in all, the papers in this cluster are related to the area of feature extraction and some link it to graphs.

The second major cluster is called “Heartbeat classification based on ECGs”. While 4 of 7 papers have “ECG” in their title, all papers are tackling the classification of ECG results. For instance, Moody and Mark [25] reduce false detections by an automated algorithm to be able to increase detection sensitivity. Pan and Tomkins [26] have the same goal and tackle QRS complexes in ECGs to lower thresholds and finally increase detection sensitivity. Therefore, this cluster is related to this medical ML application.

The third major cluster is again clearly related to an ML application and tackles the problem of “malware detection”. From the six papers in this cluster, the two papers by Danish Vasan and co-authors are strongly connected. Vasan D et al. [27] are tackling image-based malware detection and applying convolutional neural networks. They claim that their method is resilient to common hacker techniques. In the second paper [28], they use ensemble convolutional neural networks to detect malware. All papers in the cluster are related to this computer science application field.

The other clusters are smaller and investigate other ML application fields. The papers in the remote sensing cluster are trying to detect certain characteristics from area data gathered from a high distance [e.g., 29]. The papers on sleep apnea syndrome build a separate cluster although using ECG data as well [e.g., 30]. Sentiment analysis is done by the papers in the respective clusters, e.g. on twitter data [e.g., 31]. Time series analysis tries to facilitate the sequential information, e.g. based on neural networks [e.g., 32]. Human activity recognition uses smartphone data to detect human activities by ML algorithms [e.g., 33]. These clusters all reflect the broad variety of ML application fields (RQ3). However, they show that feature engineering is often applied domain-specific and is still not a research area on its own (RQ1).

Summarizing, three main and four minor research clusters are identified. Among them, there is only one cluster with a feature engineering specific topic, cluster 1 about dimensionality reduction which belongs to the field of feature extraction. Taking the result of co-citation into account, domain-independent problems from a more generic perspective are not in focus of past research.

5 Discussion and Conclusion

This paper identifies important feature engineering papers and research clusters. The citation analysis shows the top10 papers, mostly about ML algorithms. Additionally, the top10 papers about feature engineering aspects are presented. The co-citation analysis identifies important research clusters, mostly from different application domains. Based on these findings, the following conclusions are derived.

Feature engineering is still not a research area on its own (RQ1). The top10 is dominated by ML papers showing that papers applying feature engineering are embedded in ML research and, therewith, are often model-specific. Additionally, apart from one research cluster, all research clusters are application-oriented and, thus, feature engineering approaches are typically applied and described in an application context. Domain-independent papers on certain aspects of feature engineering are less often cited.

Feature engineering research is usually embedded in domain-specific research and often weakly integrated into specific feature engineering literature (RQ2). Only one research cluster on a feature engineering topic emerged, the one about “Dimensionality reduction (in graphs)”. However, this is again focused on the application in graphs. In the top10 no paper is about feature engineering. Following, an often-cited paper from 2003, “An introduction to variable and feature selection”, gave a domain-independent and model-independent overview of the feature engineering aspect “feature selection”. More domain-independent publications have been present for decades but are less often cited. Therefore, feature engineering is still strongly connected with ML research and has not developed as a separate research field. Maybe the increasing interest in machine learning research in general and feature engineering in specific will change this in the future. Recent books like “Feature engineering and selection: A practical approach for predictive models” [5] and “Feature engineering for machine learning models: Principles and techniques for data scientists” [34] are examples for a domain-independent approach.

Feature selection and feature extraction seem to be more often discussed than feature construction (RQ3). In line with an existing literature review [6], feature selection is an important aspect as reflected by the most often cited feature engineering papers and in total five papers in the top10 feature engineering papers. Feature extraction is important as well with four papers and the only cluster related to a feature engineering topic about "Dimensionality reduction (in graphs)". Again, this cluster contains several papers which focus on the graph aspect. With only one paper in the top10 feature engineering papers and no cluster, one might ask if feature construction is less important. As feature construction can have a major effect on a model result [4, 9], feature construction should not be neglected. Feature construction is highly relevant in practice and domain-independent aspects of it demand more discussion.

These results can support young scholars linking their research to existing publications. The top10 feature engineering papers might be a good starting point into the field, starting with the introductory papers. Beyond, the clusters show which aspects are already discussed in a separated research sub-stream and which are not. Research about those aspects might profit from some of the papers.

As usual, this research faces several limitations. For instance, the Web of Science database does not reflect all papers and it might be meaningful to repeat the analysis with a different data set, e.g., including conference papers as citing papers. Nevertheless, one of the biggest disadvantages of co-citation analysis is the time lag. Feature engineering is a new term, and it needs time to reflect major changes in a co-citation analysis. There is also a time lag in traditional literature reviews as very new publications might not be included. However, it even needs more time until those papers are cited by other papers. Therefore, the results are only a snapshot reflecting the situation of research a few years ago and an update of the citation and co-citation might already be valuable soon.

Finally, feature engineering research might profit from domain-independent and more data-centric papers, workshops, conferences and journals. This research area will always be connected to ML research and a lot of model- and domain-specific research will follow. But domain-independent advice on feature engineering will become more important, when ML methods are broadly applied, even in smaller companies with fewer resources for specialized teams. Therefore, a mixture of domain- and model-specific and domain- and model-independent approaches might foster the development of the field. Eventually, new research clusters will emerge, and feature engineering might develop into a research area on its own.

Acknowledgments. The author thanks Marlene Knapp for supporting the data preparation and visual graph generation and the anonymous referees for valuable feedback.

References

1. Sun, Y., Haghighat, F., Fung, B.C.M.: A review of the-state-of-the-art in data-driven approaches for building energy prediction. Energy and Buildings. 221, 110022 (2020).
2. Chicco, D., Oneto, L., Tavazzi, E.: Eleven quick tips for data cleaning and feature engineering. PLOS Computational Biology. 18, e1010718 (2022).

3. Baesens, B., Höppner, S., Verdonck, T.: Data engineering for fraud detection. Decision Support Systems. 150, 113492 (2021).
4. Bocca, F.F., Rodrigues, L.H.A.: The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modelling. Computers and Electronics in Agriculture. 128, 67–76 (2016).
5. Kuhn, M., CA Johnson, K.: Feature Engineering and Selection: A Practical Approach for Predictive Models. Taylor & Francis Ltd, Boca Raton London New York (2019).
6. Wang, Z., Xia, L., Yuan, H., Srinivasan, R.S., Song, X.: Principles, research status, and prospects of feature engineering for data-driven building energy prediction: A comprehensive review. Journal of Building Engineering. 58, 105028 (2022).
7. Verdonck, T., Baesens, B., Óskarsdóttir, M., vanden Broucke, S.: Special issue on feature engineering editorial. Mach Learn. (2021).
8. Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., Zimmermann, T.: Software Engineering for Machine Learning: A Case Study. In: 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP). pp. 291–300 (2019).
9. Heaton, J.: An empirical analysis of feature engineering for predictive modeling. In: SoutheastCon 2016. pp. 1–6 (2016).
10. Wirth, R., Hipp, J.: CRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining. (2000).
11. Xu, Z., Yu, D., Wang, X.: A bibliometric overview of International Journal of Machine Learning and Cybernetics between 2010 and 2017. Int. J. Mach. Learn. & Cyber. 10, 2375–2387 (2019).
12. López-Robles, J.R., Otegi-Olaso, J.R., Porto Gomez, I., Gamboa-Rosales, N.K., Gamboa-Rosales, H., Robles-Berumen, H.: Bibliometric Network Analysis to Identify the Intellectual Structure and Evolution of the Big Data Research Field. In: Yin, H., Camacho, D., Novais, P., and Tallón-Ballesteros, A.J. (eds.) Intelligent Data Engineering and Automated Learning – IDEAL 2018. pp. 113–120. Springer International Publishing, Cham (2018).
13. Croft, R., Xie, Y., Babar, M.A.: Data Preparation for Software Vulnerability Prediction: A Systematic Literature Review. IEEE Trans. Softw. Eng. 49, 1044–1063 (2023).
14. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182 (2003).
15. Klingert, F.M.A.: The Structure of Prediction Market Research: Important Publications and Research Clusters. The Journal of Prediction Markets. 11, 51–65 (2017).
16. Raasch, C., Lee, V., Spaeth, S., Herstatt, C.: The rise and fall of interdisciplinary research: The case of open source innovation. Research Policy. 42, 1138–1151 (2013).
17. Zaggl, M.A.: Eleven mechanisms for the evolution of cooperation. Journal of Institutional Economics. 10, 197–230 (2014).

18. Gmür, M.: Co-citation analysis and the search for invisible colleges: A methodological evaluation. Scientometrics. 57, 27–57 (2003).
19. Breiman, L.: Random Forests. Machine Learning. 45, 5–32 (2001).
20. Tibshirani, R.: Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological). 58, 267–288 (1996).
21. Chandrashekar, G., Sahin, F.: A survey on feature selection methods. Computers & Electrical Engineering. 40, 16–28 (2014).
22. Hinton, G.E., Salakhutdinov, R.R.: Reducing the Dimensionality of Data with Neural Networks. Science. 313, 504–507 (2006).
23. Wang, D., Cui, P., Zhu, W.: Structural Deep Network Embedding. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1225–1234. Association for Computing Machinery, New York, NY, USA (2016).
24. Roweis, S.T., Saul, L.K.: Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science. 290, 2323–2326 (2000).
25. Moody, G.B., Mark, R.G.: The impact of the MIT-BIH Arrhythmia Database. IEEE Engineering in Medicine and Biology Magazine. 20, 45–50 (2001).
26. Pan, J., Tompkins, W.J.: A Real-Time QRS Detection Algorithm. IEEE Transactions on Biomedical Engineering. BME-32, 230–236 (1985).
27. Vasan, D., Alazab, M., Wassan, S., Naeem, H., Safaei, B., Zheng, Q.: IMCFN: Image-based malware classification using fine-tuned convolutional neural network architecture. Computer Networks. 171, 107138 (2020).
28. Vasan, D., Alazab, M., Wassan, S., Safaei, B., Zheng, Q.: Image-Based malware classification using ensemble of CNN architectures (IMCEC). Computers & Security. 92, 101748 (2020).
29. Audebert, N., Le Saux, B., Lefèvre, S.: Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. ISPRS Journal of Photogrammetry and Remote Sensing. 140, 20–32 (2018).
30. Penzel, T., Moody, G.B., Mark, R.G., Goldberger, A.L., Peter, J.H.: The apnea-ECG database. In: Computers in Cardiology 2000. Vol.27. pp. 255–258 (2000).
31. El Rahman, S.A., AlOtaibi, F.A., AlShehri, W.A.: Sentiment Analysis of Twitter Data. In: 2019 International Conference on Computer and Information Sciences (ICCIS). pp. 1–4 (2019).
32. Fawaz, H.I., Forestier, G., Weber, J., Idoumghar, L., Muller, P.-A.: Deep learning for time series classification: a review. Data Min Knowl Disc. 33, 917–963 (2019).
33. Ordóñez, F.J., Roggen, D.: Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. Sensors. 16, 115 (2016).
34. Zheng, A., Casari, A.: Feature Engineering for Machine Learning Models: Principles and Techniques for Data Scientists. O'Reilly Media, Beijing : Boston (2018).