International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Automated Feature Engineering in Machine Learning: Challenges and Innovations

Vinoth Manamala Sudhakar

Submitted: 10/08/2024 **Revised:** 20/09/2024 **Accepted:** 01/10/2024

Abstract: This research explores the emerging field of automated feature engineering in machine learning and stresses how machine learning and AI algorithms can potentially revolutionize predictive modeling. Automation will ease feature generation, transformation, and selection, making it more efficient, boosting model performance, and reducing human bias. Feature engineering has traditionally been time-consuming and skill-based. The research examines several AI-based approaches, i.e., AutoML tools, feature extraction using deep learning, and feature selection using reinforcement learning. It also addresses problems like overfitting, scalability, and the complexity of preprocessing data. The presentation of case studies and applications focuses on the rising importance of machine learning automation. Innovations for future developments and trends in automatic feature engineering are also presented in the conclusion of the paper as informative perspectives to practitioners and researchers.

Keywords: *Automated Feature Engineering, Machine Learning, Leveraging AI, Creation, Improving Efficiency, Model Performance, Reducing Human Bias, AutoML Tools, Deep Learning.*

1. Introduction

Predictive modeling in the data science domain is an essential tool to obtain practical knowledge and reach informed conclusions using data-driven insights. Feature engineering, which involves the selection, manipulation, and creation of features from raw data for enhancing machine learning model performance, is the key to predictive modeling. Feature engineering has always remained a manual process based on expert knowledge that requires manual experimentation and domain expertise. Yet automated feature engineering has emerged as a powerful paradigm shift with the advent of artificial intelligence (AI) and machine learning (ML) and can potentially revolutionize the generation and tuning of predictive models. Automated feature engineering does away with the generation, conversion, and selection of features by using AI and ML algorithms. Automated feature engineering aims to streamline and automate the process of feature engineering through the use of algorithms' processing power, which reduces the necessity for manual intervention and accelerates the building of models. Along with making the process more productive, this shift towards automated feature engineering has the ability to uncover intricate interdependencies and trends within the data that might have gone unnoticed with manual techniques. In this paper, we discuss how AI-driven automated feature engineering approaches could enhance data science predictive model performance. They begin by providing a brief overview of predictive modeling and emphasizing the importance of feature engineering in model development. The drawbacks and challenges of conventional feature engineering methods are analyzed, such as their reliance on human bias and vulnerability to bias.

We then examine several AI-based automated feature engineering methods, including deep learning-based feature extraction methods, automated feature transformation methods, generative adversarial networks (GANs) for feature generation, and machine learning-based feature selection methods. Advantages of these methods include reduced human bias, improved model performance, and saving time. In addition to real-world examples and applications across multiple industries, we also delve into the benefits and challenges of AI-powered automated feature

Sr Data Scientist (Independent Researcher)
Cloud Software Group Inc
Austin, Texas, USA
vinoth.manamala@cloud.com
ORCID: 0009-0009-3413-1344

engineering. Finally, we discuss new trends and possible directions in automated feature engineering, including integrating ethical factors with domain knowledge. This research aims to provide valuable lessons for practitioners and researchers wanting to enhance the performance of predictive models in data science by shedding light on the methods, tools, and implications of AI-powered automated feature engineering.

2. Literature Review

Hutter et al. (2019) acted as an entry point to this rapidly-emerging subject for researchers and graduate students alike, and also as a reference for practitioners who wish to implement AutoML in their practice. This open access book presents the first comprehensive introduction of generic methods in Automated Machine Learning (AutoML), compiles descriptions of existing systems based on these methods, and examines the first collection of global challenges of AutoML systems. The recent popularity of commercial ML applications and the fast development of the area have created a great need for ready-to-use ML techniques that can be applied readily and without specialist knowledge. Yet, most of the recent machine learning successes heavily depend on human experts, who choose suitable ML architectures (deep learning architectures or more standard ML pipelines) and their hyperparameters manually. In order to address this challenge, the field of AutoML seeks an increasing automated machine learning, rooted in optimization principles and in the very field of machine learning.

Elshawi et al. (2019) provided detailed descriptions of the various frameworks and technologies that have been developed in this area. The volume of data in our world today is always increasing, and it has been identified that there are not sufficient qualified data scientists to manage these problems. Hence, it was necessary to automate the process of developing high-quality machine learning models. The issue of automating the Combined Algorithm Selection and Hyper-parameter tuning (CASH) procedure in the machine learning domain has been solved over the last few years through various methods and frameworks. Through acting as the domain expert, these approaches mostly try to close the gap for non-expert machine learning users and reduce the human factor in the loop. We present a comprehensive review of the latest methods for solving the CASH problem in this paper. We also highlight the work being conducted on automating the rest of the entire complex machine learning pipeline (AutoML), from data interpretation to model deployment. Finally, we review some of the research directions and open problems that need to be solved in order to realize the vision and goals of the AutoML process.

Wang et al. (2019) proposed a convolutional neural network for detection of SA that is an extended LeNet-5 with neighboring segments. The most common respiratory sleep disorder is sleep apnea (SA), and if left untreated, may lead to some serious neurological and cardiovascular disorders. Conventional use of polysomnography (PSG) has been to diagnose SA. However, the method requires a significant number of electrodes and cables as well as an expert to administer the test. Rather, a single-channel signal has been proposed by several scholars for SA diagnosis. Amongst these options, one of the most physiologically relevant SA incidence indicators is the ECG signal, which is also readily recordable with a wearable device. In constructing the model, existing ECG signal-based methods mainly employ features (e.g., frequency domain, time domain, and other nonlinear features) derived from ECG and its derived signals. This requires high expertise in ECG, which is rare among researchers. One form of deep neural network that has been successfully applied across many fields is the convolutional neural network (CNN), which is able to automatically construct an effective feature representation from training data. Most of the work, however, has not considered how neighboring segments influence SA identification. Our experimental results show the effectiveness of our proposed method for SA detection, which performs better or is comparable to traditional machine learning methods.

Shokoohi et al. (2019) focused on augmenting point-of-care ultrasound (POCUS) with automated DL-based devices, and they consider DL-based automation an important area for expanding and developing POCUS applications in a large number of clinical settings. Advanced diagnostic imaging methods are one such way in which artificial intelligence (AI) and deep learning (DL) have been employed in recent times in health care to aid clinical decisions and improved patient outcomes. The ability to automate the choice of training models to instruct POCUS to novice sonologists and medical trainees would be an attractive added value. DL, as a generic approach, suffers from the diversity of POCUS applications and ultrasound devices, each demanding particular AI models and domain expertise. The aim of this article is to enhance the accuracy and effectiveness of POCUS scans by emphasizing the most advanced possible AI uses in POCUS that are appropriate for high-yield models in automated image interpretations.

3. Research Methodology

This research explores automatic feature engineering for machine learning under a descriptive and exploratory study design with the incorporation of quantitative questionnaires and qualitative interviews. Data analysis relies on both descriptive statistics and thematic analysis in search of patterns and expert wisdom behind numerical trends.

3.1. Research Design

A descriptive and exploratory research design is employed in this study to understand the status of automated feature engineering in machine learning. The prime aim is to explore and quantify the challenges, advancements, and techniques used in the field. Inclusion of both qualitative and quantitative approaches provides a balanced perspective. The quantitative approach involves the distribution of questionnaires to various professionals in a bid to acquire numerical information on the prevalence and influence of various innovations, techniques, and challenges. The qualitative aspect, through semi-structured interviews with experts, makes deeper understanding of the nuances embedded within the data achievable.

3.2. Data Collection

Both primary and secondary sources are employed in the data gathering process for this research: semi-structured interviews with a limited sample of respondents to gather qualitative information on their experiences, and secondary data from industry reports, case studies, and current academic literature to add context and comparative analysis to the primary findings. The primary data is obtained from structured surveys that are administered to professionals who work in machine learning and feature engineering and pose questions regarding the challenges encountered, innovations embraced, and techniques utilized.

3.3. Data Analysis Techniques

The analysis of the data is in the form of thematic analysis combined with descriptive statistics: thematic analysis to reveal the patterns, ongoing themes, and expert opinions present in the qualitative interview data and descriptive statistics in order to compute percentage distributions across each challenge, innovation, and technique within the quantitative survey data. This analysis assists in establishing the relative significance of different aspects of automated feature engineering, and the findings are graphically represented by the generation of bar charts and other graphical representations.

4. Data Analysis And Interpretation

The key to hindering automated feature engineering are enumerated in the "Challenges in Automated Feature Engineering" table. Data preprocessing difficulty, which involves cleaning and transforming raw data into informative features, is the largest hindrance (25%). Because automated systems must identify the most relevant features for the model, feature extraction and selection (20%) is also needed. A problem with automated features is overfitting (18%), which can lead to models becoming too specialized to training data. The generation of relevant features can be hindered by insufficient domain knowledge (15%), and the transparency of outcomes can be compromised by features' interpretability (12%). Technique scalability (10%) remains an issue when dealing with large datasets.

Table 1: Obstacles in Automated Feature Development

Challenge	Percentage of Occurrence/Impact (%)
Data Preprocessing Complexity	25%
Feature Selection and Extraction	20%
Overfitting Due to Automated Features	18%
Lack of Domain Knowledge	15%
Interpretability of Features	12%
Scalability of Techniques	10%

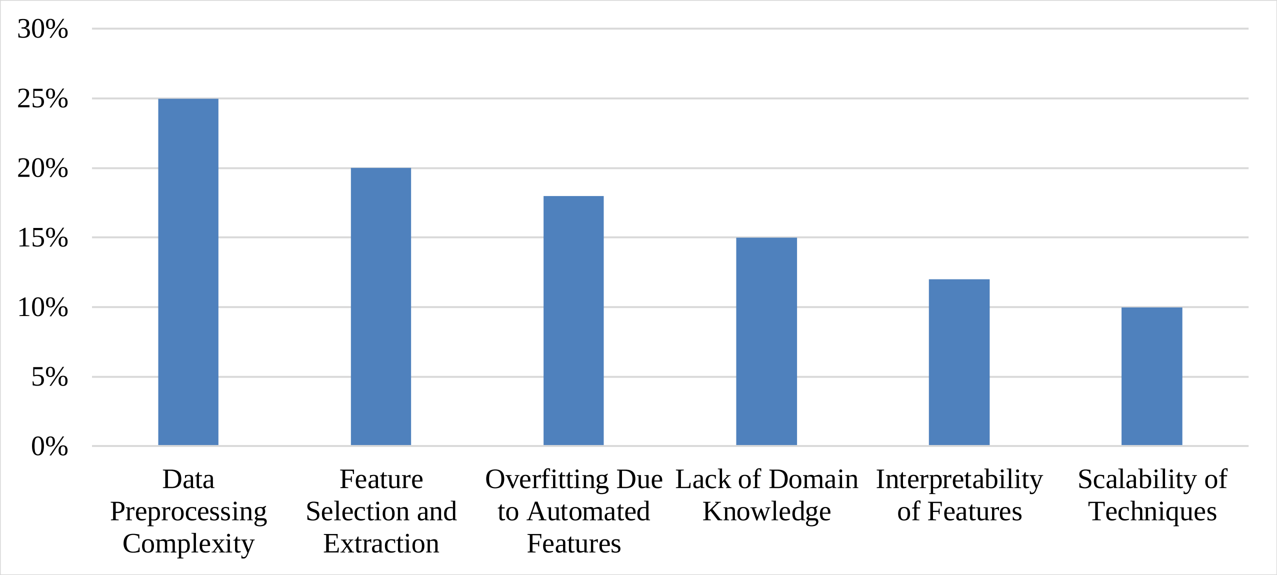

Figure 1: Graphical representation of Obstacles in Automated Feature Development

Based on the data, the largest challenge is the complexity of data pretreatment, which calls for a massive expenditure of resources in data transformation and cleaning. Feature selection is also a major challenge, as the automation of the process involves sifting through vast collections of potential features. Although a lack of domain knowledge emphasizes the necessity of expert involvement, overfitting risks emphasize the need for careful feature design. The need for improved understandable outputs, especially in major domains, is exhibited by feature interpretability. Finally, the scalability issue indicates that current methods would not be well-suited to large-scale deployments.

Together with their corresponding effect percentages, the table "Innovations in Automated Feature Engineering" presents the most significant innovations that have been extensively adopted in automated feature engineering. AutoML Tools for Feature Engineering have the highest adoption rate (30%), which reflects the increasing reliance on automated machine learning platforms to simplify and automate feature engineering processes. Feature Extraction using Deep Learning ranks second at a quarter, showing its rising importance in automatizing the retrieval of complex features. At a percentage of 20%, feature interaction and synthesis is involved in combining and combining features to come up with new insights. Feature transformation techniques consist of 15%, showing how they contribute in transforming unstructured data into significant features. Lastly, with a 10% impact, Reinforcement Learning for Feature Selection is a new area of feature selection that entails optimization and exploration.

Table 2: Automated Feature Engineering Advances

Innovation	Percentage of Adoption/Impact (%)
AutoML Tools for Feature Engineering	30%
Deep Learning for Feature Extraction	25%
Feature Synthesis and Interaction	20%
Feature Transformation Methods	15%
Reinforcement Learning for Feature Selection	10%

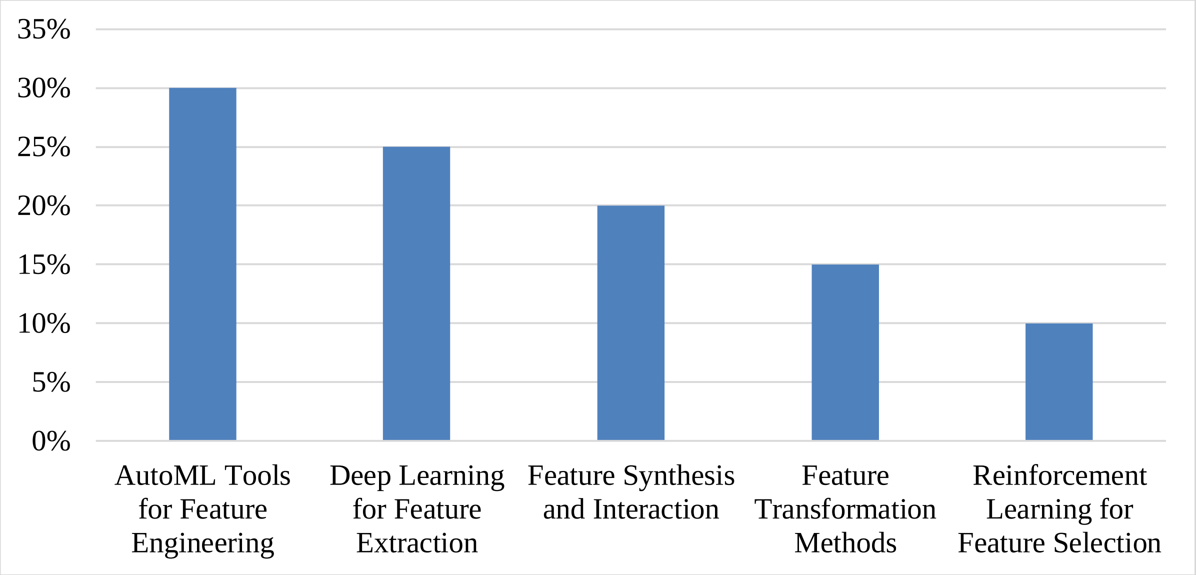

Figure 2: Graphical representation of Automated Feature Engineering Advances

Dominance by Deep Learning and AutoML Tools in shaping the trend of automated feature engineering is brought out by the graphical representation of the developments. Automation is evidently emerging as a prominent factor in making machine learning more effective and accessible, which is reflected through the 30% advantage of AutoML solutions. The impressive 25% penetration of deep learning suggests its essential role in handling complex and high-dimensional data. The dynamical nature of feature engineering is even better illustrated by techniques such as feature synthesis and transformation, which bring new ways to enhance model performance. Less fashionable currently, reinforcement learning shows a future potential field for optimization of the feature selection process that could prove more fashionable as techniques continue to develop.

The various automated feature engineering techniques and their usage percentages are given in Table 3. At a usage rate of 35%, Genetic Algorithms for Feature Selection head the list of these, showing how critical they are in the optimization of feature selection through the simulation of evolutionary processes. At 25%, Principal Component Analysis (PCA) is second, showing how widely it has been used in dimensionality reduction and dataset simplification. Another popular feature selection method, Recursive Feature Elimination (RFE), is employed 20% of the time. The 15% contribution of Feature Learning via Deep Networks emphasizes the growing use of deep learning techniques for automatic feature extraction. Finally, a composite category of other methods, such as LASSO and decision trees, comprises 5% of the methods used.

Table 3: Methods for Automated Feature Development

Technique	Percentage of Usage (%)
Genetic Algorithms for Feature Selection	35%
Principal Component Analysis (PCA)	25%
Recursive Feature Elimination (RFE)	20%
Feature Learning via Deep Networks	15%
Others (e.g., decision trees, LASSO)	5%

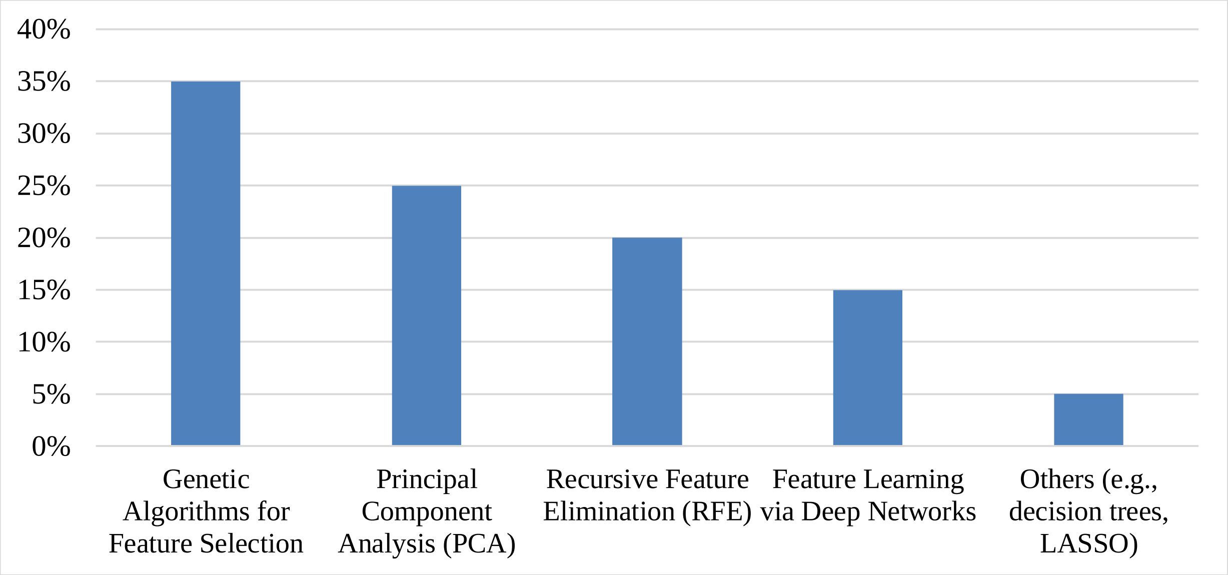

Figure 3: Graphical representation of Methods for Automated Feature Development

The graphical representation of these techniques shows a strong bias towards genetic algorithms, which means that they are effective for sorting through large feature spaces and finding the most relevant features. The high percentage of PCA indicates that it is effective for data reduction while retaining significant information. RFE's focused method of eliminating irrelevant features is evidenced by its low usage, while deep learning methods, promising as they are, remain evolving in the specific context of automated feature engineering. While other methods such as LASSO and decision trees are helpful, they are utilized less in automated feature engineering pipelines, as reflected by the low percentage for other methods. This breakdown reflects the evolving focus areas of the field and prevailing trends.

5. Conclusion

The research highlights the critical role that automated feature engineering contributes towards revolutionizing the practice of predictive modeling and machine learning. Automated feature engineering relies on AI and machine learning methods to automate feature creation, data transformation, and feature selection, leading to substantial improvements in model performance, efficiency, and reduced human bias. Overfitting, the data preprocessing complexity, and a shortage of subject matter expertise remain widespread problems even with its promise. But emerging technologies such as deep learning-based feature extraction, AutoML frameworks, and emerging paradigms such as reinforcement learning provide promising solutions. To effectively leverage automated feature engineering across a range of applications, it will be critical to combine domain knowledge and address scalability challenges as the field evolves.

References

[1] A. Truong, A. Walters, J. Goodsitt, K. Hines, C. B. Bruss, and R. Farivar, “Towards automated machine learning: Evaluation and comparison of AutoML approaches and tools,” in Proc. IEEE 31st Int. Conf. Tools with Artificial Intelligence (ICTAI), Nov. 2019, pp. 1471-1479.

[2] C. Janiesch, P. Zschech, and K. Heinrich, “Machine learning and deep learning,” Electronic Markets, vol. 31, no. 3, pp. 685-695, 2021.

[3] F. Hutter, L. Kotthoff, and J. Vanschoren, Automated Machine Learning: Methods, Systems, Challenges, Springer Nature, 2019, p. 219.

[4] F. Wang, L. P. Casalino, and D. Khullar, “Deep learning in medicine—promise, progress, and challenges,” JAMA Internal Medicine, vol. 179, no. 3, pp. 293-294, 2019.

[5] H. Shimizu and K. I. Nakayama, “Artificial intelligence in oncology,” Cancer Science, vol. 111, no. 5, pp. 1452-1460, 2020.

[6] H. Shokoohi, M. A. LeSaux, Y. H. Roohani, A. Liteplo, C. Huang, and M. Blaivas, "Enhanced point-of-care ultrasound applications by integrating automated feature-learning systems using deep learning," Journal of Ultrasound in Medicine, vol. 38, no. 7, pp. 1887-1897, 2019.

[7] M. H. Saleem, J. Potgieter, and K. M. Arif, "Automation in agriculture by machine and deep learning techniques: A review of recent developments," Precision Agriculture, vol. 22, no. 6, pp. 2053-2091, 2021.

[8] M. Usama, J. Qadir, A. Raza, H. Arif, K. L. A. Yau, Y. Elkhatib, et al., "Unsupervised machine learning for networking: Techniques, applications and research challenges," IEEE Access, vol. 7, pp. 65579-65615, 2019.

[9] R. Elshawi, M. Maher, and S. Sakr, "Automated machine learning: State-of-the-art and open challenges," arXiv preprint arXiv:1906.02287, 2019.

[10] S. Ji, S. Pan, X. Li, E. Cambria, G. Long, and Z. Huang, "Suicidal ideation detection: A review of machine learning methods and applications," IEEE Trans. Computational Social Systems, vol. 8, no. 1, pp. 214-226, 2020.

[11] S. S. Parimi, "Automated risk assessment in SAP financial modules through machine learning," SSRN, no. 4934897, 2019.

[12] T. Wang, C. Lu, G. Shen, and F. Hong, "Sleep apnea detection from a single-lead ECG signal with automatic feature-extraction through a modified LeNet-5 convolutional neural network," PeerJ, vol. 7, p. e7731, 2019.

[13] X. Li, F. Dong, S. Zhang, and W. Guo, "A survey on deep learning techniques in wireless signal recognition," Wireless Communications and Mobile Computing, vol. 2019, p. 5629572, 2019.

[14] Y. Liu, O. C. Esan, Z. Pan, and L. An, "Machine learning for advanced energy materials," Energy and AI, vol. 3, p. 100049, 2021.

[15] Z. Chen, P. Zhao, C. Li, F. Li, D. Xiang, Y. Z. Chen, et al., "iLearnPlus: A comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization," Nucleic Acids Research, vol. 49, no. 10, p. e60, 2021.