Systematic review of using machine learning in imputing missing values

Alabadla, Mustafa[a]; Sidi, Fatimah[a]; Ishak, Iskandar[a]; Ibrahim, Hamidah[a]; Affendey, Lilly Suriani[a]; Che Ani, Zafienas[a]; Jabar, Marzanah A.[b]; Bukar, Umar Ali[b, c]; Devaraj, Navin Kumar[d]; Muda, Ahmad Sobri[e]; Tharek, Anas[e]; Omar, Noritah[f]

[a] Universiti Putra Malaysia (UPM), Faculty of Computer Science and Information Technology, Department of Computer Science, Selangor, Serdang, 43400, Malaysia

[b] Universiti Putra Malaysia (UPM), Faculty of Computer Science and Information Technology, Department of Software Engineering and Information System, Selangor, Serdang, 43400, Malaysia

[c] Taraba State University, Computer Science Unit, Department of Mathematical Sciences, Jalingo, 00234, Nigeria

[d] Universiti Putra Malaysia (UPM), Faculty of Medicine and Health Sciences, Department of Family Medicine, Selangor, Serdang, 43400, Malaysia

[e] Universiti Putra Malaysia (UPM), Faculty of Medicine and Health Sciences, Department of Radiology, Selangor, Serdang, 43400, Malaysia

[f] Universiti Putra Malaysia (UPM), Faculty of Modern Languages and Communication, Department of English, Selangor, Serdang, 43400, Malaysia

[g] Universiti Malaysia Pahang (UMP), Faculty of Computing, Department of Software Engineering, Pahang, Pekan, 26600, Malaysia

ABSTRACT

Missing data are a universal data quality problem in many domains, leading to misleading analysis and inaccurate decisions. Much research has been done to investigate the different mechanisms of missing data and the proper techniques in handling various data types. In the last decade, machine learning has been utilized to replace conventional methods to address the problem of missing values more efficiently. By studying and analyzing recently proposed methods using machine learning approaches, vital adoptions in accuracy, performance, and time consumed can be highlighted. This study aimed to help data analysts and researchers address the limitations of machine learning imputation methods by conducting a systematic literature review to provide a comprehensive overview of using such methods to impute missing values. Novel proposed machine learning approaches used for data imputation are analyzed and summarized to assist researchers in selecting a proper machine learning method based on several factors and settings. The review was performed on research studies published between 2016 and 2021 on adopting machine learning to impute missing values, focusing on their strengths and limitations. A total of 684 research articles from various scientific databases were analyzed using search engines, and 94 of them were selected as primary studies. Finally, several recommendations were given to guide future researchers in applying machine learning to impute missing values.

KEYWORDS

Data imputation; Data mining; Data preprocessing; Data quality; Missingness; Systematic literature review

ACKNOWLEDGEMENT
All opinions, _ndings, conclusions, and recommendations in this article are those of the authors and do not necessarily reflect the views of the funding agencies. The authors would like to thank the anonymous reviewers for their comments.