Present Condition and Future Outlook on the Application of Machine Learning and Big Data Analytics in Multimorbidity Research

Ardihan Adjani[1]
[1] University in Falun, SWEDEN

ABSTRACT

Multimorbidity, the coexistence of two or more chronic conditions in an individual, presents complex challenges for healthcare management and research. Traditional epidemiological and statistical methods often fail to capture the intricate interactions among diseases, functional impairments, and patient outcomes. Recent advances in machine learning (ML) and big data (BD) analytics offer promising strategies to address these challenges by enabling patient phenotyping, risk stratification, and prediction of disease progression using multimodal datasets, including electronic health records. Techniques such as clustering, deep learning, reinforcement learning, and temporal modeling facilitate the identification of disease patterns, trajectories, and treatment optimization for individuals with multimorbidity. Despite these advances, practical implementation in clinical settings is limited by data heterogeneity, methodological variability, and insufficient integration of domain expertise. Collaboration among clinicians, data scientists, and AI specialists is essential to develop standardized, validated research protocols and to translate ML/BD insights into actionable clinical strategies. Ultimately, the integration of these technologies promises to improve personalized care, population health management, and the creation of multidisciplinary guidelines for the management of multimorbidity.

Keywords: Multimorbidity; Machine Learning; Big Data Analytics; Electronic Health Records; Patient Stratification; Chronic Disease Patterns; Reinforcement Learning; Deep Learning; Clinical Decision Support

Introduction

Innovative Strategies in Multimorbidity Research Related to Patterns and Clusters

The ML/BD methodologies in multimorbidity research were initially employed in epidemiological surveys to diminish the extensive array of disease combinations, derived from enumerating disease dyads and triads, and to attempt to standardise disease patterns across studies by analysing the natural clustering of diseases [1]. Various methodologies have been used, including logistic regression analysis, hierarchical clustering techniques like latent class analysis (LCA), and exploratory factor analysis, while the data sources were electronic health records (eHRs), national registries, and surveys completed by primary care doctors [2]. In summary, recent comparative method analysis has validated that multimorbidity patterns, while exhibiting some overlap due to the prevalence of common diseases, differ based on the disease spectrum, the studied population, and the analytical method employed, underscoring the necessity for standardisation in study design [3]. This research suggests that factorisation approaches are superior for delineating comorbidity associations, whereas clustering methods are more advantageous for exploratory analyses during in-depth investigations. ARM is often used to examine illness correlations and analyse prevalent patterns [4]. The tree-based methodology yields findings that facilitate the identification of certain combinations of chronic illnesses or syndromes.

Our research group employed non-hierarchical (k-means) and hierarchical (LCA) cluster analyses in the pilot exploratory study to gain insights into the mechanisms underlying the clustering of physical frailty and cognitive impairment, the two principal states of ageing entropy [5]. Initially, we focused on these functional impairments and then delineated the found clusters by evaluating variations in chronic illness diagnoses and other clinical and socio-demographic factors, via the phenotyping of diverse individuals at risk for these outcomes. Conversely, the expert author group from Italy conducted illness-based clustering and then evaluated the disparities in clinical and functional status among people within the clusters, therefore offering insights into the processes driving disease clustering [5]. The authors used the fuzzy c-means algorithm as a clustering technique, often referred to as "the soft clustering method" [6]. This approach is more scalable than traditional hard clustering techniques, since it uses probability distributions instead of similarity levels to give membership to clusters.

In recent years, there has been a growing trend in employing deep learning techniques and electronic health record data for tasks such as patient phenotyping, disease feature detection or classification, and predicting clinical outcomes based on longitudinal event sequences, which are anticipated to enhance the management of multimorbidity. Numerous advancements in deep learning are used to map dependable ideas in raw or slightly processed data inside electronic health records (embedding approaches), including textual medical notes, which are then utilised for temporal sequencing and outcome prediction. For instance, Utilised data on therapeutic regimens for cancer patients from an extensive insurance claims database to delineate treatment paths. The authors developed an algorithm to extract patient-specific lines of treatment and consolidated this data using clustering and visualisation techniques to identify temporal phenotypes and facilitate predictions of illness progression. A modified non-negative tensor-factorization method, utilised for uncovering latent object variables in image analysis, using electronic health records data to discern phenotypic subgroups in individuals predisposed to cardiovascular disease. By integrating ARM with the projected risk of each subtype for cardiovascular disease development, these scientists were able to uncover many previously unrecognised phenotypes [100]. Nguyen et al. devised an enhanced convolutional neural network (CNN) model to predict the likelihood of hospital readmission, using medical history data represented as a series of ideas [1]. An enhanced recurrent neural network (RNN) model to forecast diagnoses and medicine prescriptions for future visits [2]. Despite these case studies targeting the intricacies of chronic illnesses, they still inadequately address the complexities of multimorbidity.

Certain methodologies from the reinforcement learning paradigm are used to provide doctors with data-driven decision assistance for therapy alternatives that are likely to optimum avert illness progression, predicated on forecasting future health states using longitudinal sequences from electronic health records [3].

A contemporary trend in the evaluation of multimorbidity is the transition from a disease-centric approach to a multi-modal representation of phenotypes, incorporating data on medications, laboratory results, and functional health status alongside disease classifications to enhance the comprehension of disease pathways. In response to this problem, new algorithms and matrices have been created, exhibiting enhanced capacities to manage

extensive and multi-modal datasets, as well as to uncover concealed information, as detailed in the recent review by [3]. The primary advances include a transition from static to probabilistic applications of fundamental machine learning techniques, facilitating a move from qualitative descriptive to quantitative testing methodologies, with advancements in "deep phenotyping." The latter phrase denotes endeavours to build characteristics of patient subgroups that are sufficiently comprehensive to be stable throughout the intricate layers of data structures and during the course of the phenotypic or the emergence of new illness associations. The depiction of temporal dynamics associated with multiple interactions inherent to multimorbidity poses a significant challenge for data scientists, necessitating advanced algorithms and multi-step analytics that surpass mere case-control classification and linear regression analysis to illustrate its progression over time. The research approaches for investigating the temporal dynamics of this complexity remain inadequately developed and are confined to the paradigm of unsupervised learning (where outcomes are unknown) and disease-disease interactions. While data scientists possess definitive expertise in developing advanced data analytics, the absence of domain knowledge in formulating research tasks and assessing the interpretability of the analytics will render the practical applicability of these innovations questionable. Addressing issues related to multimorbidity may include making inferences amid ambiguity and insufficient or contradictory information. Alternative methods to conventional machine learning, such as argumentation theory, may exhibit superior performance in learning certain tasks. For these novel models, domain expertise

Strategies for Enhancing the Application of Machine Learning and Big Data Techniques in Multimorbidity Research

The following examples demonstrate critical capabilities of ML/BD analytics in healthcare, including the identification of novel concepts from routinely collected data to enhance diagnosis and improve disease classification, as well as the utilisation of unstructured data, such as plain text or images, which would otherwise be unattainable. Automation, a depiction of features without the need for manual efforts and impromptu contributions from an expert provide an alternative option. Ultimately, ML/BD analytics may be achieved by integrating electronic health records from various platforms and healthcare environments.

A clinical decision support system that used a deep learning model (a modified deep belief network) to extract general ideas from medical notes (an unsupervised job) sourced from outpatient clinic and hospital electronic health record. The initial datasets included unstructured data (narrative symptoms and signs in plain English) and structured data (laboratory and socio-demographic information). Once the deep learning model's network was trained to extract features from the original data, the model's parameters (weights) were adjusted in a supervised way using support vector machines like a traditional classification model. The retrieved characteristics, encoded by the hidden layers of the deep learning model, were further modelled to align with the intended outcome measurements. These were diagnoses of many prevalent illnesses categorised according to the international categorisation system.

Numerous analogous initiatives are documented in recent literature, whereby writers used either illness-specific registries or electronic health records (eHRs), using either time-sequential or cross-sectional patterns to model patient characteristics for the purpose of supporting or predicting disease activity diagnosis. Norgeot et al. used a longitudinal deep learning model to analyse structured data (including drugs, laboratory results, patient demographics, and disease activity) from two extensive hospital registries of rheumatoid arthritis patients, aiming to predict disease activity at subsequent visits [10]. Comparing the models' performances across the two environments allows for the assessment of care quality and the evaluation of model interoperability. The consolidated electronic health records of about 700,000 patient entries from the Mount Sinai data repository were used to execute unsupervised patient representation models. The models were assessed by evaluating the likelihood of individuals having certain illnesses [11].

The provided examples illustrate the significant potential of the ML/BD research methodology in enhancing the diagnosis and prognosis of individual chronic illnesses, which would otherwise involve complex and laborious processes. Nonetheless, the design of these research fails to include the whole complexity of multimorbidity, since individual illnesses are used as target variables and there is an absence of patient stratification into subgroups that may more accurately represent variations in developmental disease phases.

The research design by Peng et al. aligns more closely with this approach [13]. The authors utilised the Taiwanese national health insurance research database to create a cumulative deficit frailty index using a machine learning random forest method, and compared it to the traditional index, which relies on expert opinion for feature selection, representing a hypothesis-driven approach [14]. This study is noteworthy as the authors stratified patients into various groups based on the risk of significant adverse outcomes, including all-cause mortality, hospitalisations, and admissions to intensive care units, employing survival analyses (specifically, Kaplan-Meier survival curves and Cox models), which are traditional statistical methods. The design of this research aligns with the integrated theory of ageing, which posits that ageing results in a loss in the functional capacities of older persons owing to disruptions in physiological connections and the dispersion of phenotypes [3].

In line with the premise of our work, our research group posits that, while examining multimorbidity issues, the multimodal data characterising patients across several dimensions should serve as the input for classification or prediction models. Health status measurements reflecting functional deterioration, rather than illness classifications, should be used as outcome metrics [93,94]. Consistent with the paradigm of complex thinking, the authors of this research used a combination of established approaches, emphasising the problem-solving task rather than the evaluation of novel ways.

Administering treatment suggestions for individuals with multimorbidity is a challenging issue owing to intricate illness and drug interdependencies. Zhang et al. introduced an algorithm that disaggregates treatment recommendations into a sequential decision-making process, simultaneously ascertaining the optimal amount of drugs using reinforcement learning (RL) [15]. It eliminates 99.8% of detrimental medication interactions within the prescribed treatment regimens. Zheng et al. developed a reinforcement learning system that

prescribes a treatment regimen aimed at optimising patients' cumulative health outcome use their own attributes and medical background. The use of reinforcement learning in healthcare mostly centers on the therapy suggestion issue. In actuality, it is somewhat impractical to ascertain pertinent consequences for every step undertaken throughout the medical procedure.

Additionally, in elderly individuals with multimorbidity, anticipating medical risks is a significant challenge. This subject has often been addressed by seasoned physicians, who possess extensive patient experience and knowledge of clinical recommendations, or by linear prediction models using well stated risk variables. Both techniques are more appropriate for a single condition than for multimorbidity [16]. Pham et al. Introduced a sophisticated temporal architecture using a deep dynamic neural network approach for forecasting illness development, suggesting therapies, and anticipating hospital admissions, grounded on medical history data [16]. The authors addressed significant challenges in forecasting future risks utilising temporal sequences derived from electronic health records (eHRs), including long- and short-term dependencies in health status changes, irregular timing and episodic documentation of care episodes, and the interplay between interventions and disease progression. This method is inadequate for addressing the intricacies of multimorbidity, since it relies only on coded information from electronic health records (diagnoses, treatments, and drugs) and focuses on the course of individual illnesses. The research conducted by Hassaine et al. demonstrated, via the use of matrix factorisation, the temporal progression of illness clusters, resulting in the formation of multimorbidity networks. The interpretation of the data is dubious, since the actual therapeutic value is not readily apparent [17].

The study team from Catalonia has focused on delineating multimorbidity patterns in a substantial senior population (≥65 years) within primary care, using data from the Information System for study in Primary Care (SIDIAP) [117–119]. The primary care setting, where elderly people with chronic medical illnesses often visit and where data on their health history and other elements of treatment are systematically recorded in electronic health records, serves as a distinctive locus for longitudinal investigation of multimorbidity patterns. In a recent cross-sectional study (2019), fuzzy cluster analysis was employed to identify multimorbidity clusters, enabling individuals to be associated with multiple clusters simultaneously, which aligns more closely with clinical experience than alternative methodologies. This yielded distinct clusters compared to an earlier study (2012), which utilised a combination of multiple correspondence analysis for categorical variables and k-means clustering for numerical variables to identify clusters [18]. In this earlier work, the authors monitored the identified clusters for six years and demonstrated that the originally formed clusters exhibit relative stability over time, as seen by the number and percentage of patients remained in the cluster at the conclusion of the follow-up period.

In their latest study, this group of authors employed a cross-sectional design and fuzzy cluster analysis to identify patterns of multimorbidity. They subsequently modelled the longitudinal trajectories of these patterns using a Hidden Markov Model, which serves as an approximation for addressing complex machine learning and reinforcement learning challenges, particularly in modelling transitions between multimorbidity patterns and

mortality risk. Additional algorithms were utilised to connect the inter-cluster transition probabilities with the initial cluster probabilities [19]. The authors evaluated five-year survival rates for multimorbidity patterns using Cox regression models. Supplementary variables, including socio-demographics and the quantity of drugs and visits, were used to examine the pathophysiological basis of illness clustering and the temporal progression of the clusters. The trajectories of multimorbidity were typically consistent across time, providing a foundation for timely targeting of particular multimorbidity patterns with preventative strategies.

The findings reported by the Catalonian research group are more clinically robust than most others and provide guidance for the treatment of multimorbidity in the community. A primary concern is the exclusion of functional limitations, such as frailty, from the study, despite evidence indicating that frailty status may substantially alter disease presentation and mortality risk [120]. Furthermore, data on laboratory examinations and body shape measurements as shown in our research, measures may provide further insights for stratifying disease severity owing to the progressive characteristics of chronic illnesses [9].

The quantity and nature of chronic illnesses affect physical and overall functioning throughout time and with ageing, as shown in the research by Stenholm et al. and Vetrano et al. [12]. Modelling the temporal changes of disease clusters with the reduction in functional skills of older individuals with multimorbidity is an additional difficulty in addressing the complexity of multimorbidity.

The aforementioned investigations have uncovered perplexing circumstances that hinder current research on multimorbidity. The issue encompasses both a clinical science approach and a data science perspective. Regarding the clinical aspect of the issue, a major worry is the absence of agreement on the definition of multimorbidity and the diverse range of illnesses (or other disorders) used to form the clusters [12]. The diversity of cluster identification techniques further enhances the unpredictability of multimorbidity patterns seen across different research [13]. A limited number of research on the temporal development of multimorbidity patterns are inadequate for informing future study designs.

The challenge in data science stems from the fast progression of machine learning and big data algorithms, which continually enhance modelling efficacy, complicating the comparison of studies and hindering the application of study findings. Furthermore, discrepancies in datasets and the multitude of ML/BD techniques accessible for analysing identical tasks may lead to result differences, prompting researchers to concurrently use many approaches to evaluate their efficacy [15]. Although regularly obtained data from electronic health records (eHRs) undeniably benefit research, they have mostly been developed for administrative and clinical trial objectives, rendering them possibly insufficient for the intended study. Electronic health records (eHRs) exhibit deficiencies like data incompleteness, irregular sampling, and data imbalance, necessitating different methodologies for the assessment of machine learning algorithms in medical categorisation and diagnostic tests [18]. Data scientists are devising diverse pre-processing and dimensionality reduction techniques to address these deficiencies [129]. To address difficult jobs, data scientists use a mix of methodologies or develop intricate protocols, which vary

in efficiency from one another. Despite the inclination towards complete automation of the analytical process to minimise human labour, the absence of domain expert involvement at all stages of data analysis may provide outcomes that are too complex, perplexing, and ultimately impractical.

To address these issues, consensus committees of medical experts, data scientists, and AI specialists should be established, with each contributing to progress in their respective domains. Collaborating in multidisciplinary teams, these expert groups might enhance the validation and standardisation of methodologies and the formulation of unified research procedures, akin to the development of clinical recommendations. The formulation of a comprehensive list of precise research questions or desired outcomes by medical professionals is essential for identifying patient risk categories, analysing data from electronic health records, and developing optimal treatment plans. Population health management would be significantly enhanced if an effective and streamlined approach for evaluating risk in older, multimorbid patients within primary care were developed and perpetually refined. To engage more effectively in this process, physicians, particularly general practitioners who first encounter multimorbid patients and bear the primary responsibility for data collection in electronic health records, should receive enhanced training in the functionalities of machine learning and big data techniques, as well as quantitative data analysis methods.

Conclusions

Population-centric health management for individuals aged 60 years and older is insufficient. The provision of personalised treatment and preventative interventions is primarily constrained by the deficiencies of conventional study designs and data processing techniques. This substantial demographic is marked by an increasing intricacy of chronic conditions.

Diseases and multimorbidity, defined as the coexistence of two or more chronic conditions in an individual. An alternative study methodology using ML/BD techniques has emerged as a viable solution with significant promise to tackle the challenges related to multimorbidity. These issues include, for instance, the phenotyping of patients and risk classification predicated on the modelling of many connected features that intersect among people. Additionally, recognising patterns of chronic health issues among the community and monitoring the deterioration of health conditions over time in persons with various ailments. The challenges to successfully integrate this research approach into routine clinical practice primarily involve the necessity for enhanced coordination among medical professionals, data scientists, AI researchers, and IT specialists to establish standardised and validated research protocols tested in real-world conditions and to develop a genuine interdisciplinary knowledge base. This expanding knowledge base, including case studies addressing diverse issues in clinical practice, will enable the future development of new multidisciplinary guidelines and recommendations for managing multimorbidity.

References

[1] Sethuraman, P., & Chennareddy, R. K. (2022a). Intelligent Vehicular Traffic Flow Prediction Using Learning-Based Spatio-Temporal Models for Data-Driven Wireless Transportation and Urban Analytics Systems. *International Journal of Emerging Trends in Computer Science and Information Technology, 3*(2), 111-121.

[2] Gudepu, B. K., & Jaladi, D. S. (2022a). Data Discovery and Security: Protecting Sensitive Information. *International Journal of Acta Informatica, 1*(1), 176-187.

[3] Gudepu, B. K., & Jaladi, D. S. (2022b). Why Real-Time Data Discovery is a Game Changer for Enterprises. *International Journal of Acta Informatica, 1*(1), 164-175.

[4] Jaladi, D. S., & Vutla, S. (2022a). Artificial Intelligence's Influence on Design: A New Era of Creative Collaboration. *International Journal of Acta Informatica, 1*(1), 188-198.

[5] Jaladi, D. S., & Vutla, S. (2022b). Medical Decision-Making with the Help of Quantum Computing and Machine Learning: An In-Depth Analysis. *International Journal of Acta Informatica, 1*(1), 199-215.

[6] Sethuraman, P., & Chennareddy, R. K. (2022b). Machine Learning Assisted Design of Wireless Access Systems for Reliable and Low-Latency Financial and Smart Commerce Services. *International Journal of Artificial Intelligence, Data Science, and Machine Learning, 3*(4), 133-142.

[7] Chennareddy, R. K. (2021). Designing Data and Analytics Ecosystems for High Volume Transaction Processing Applications. *International Journal of AI, BigData, Computational and Management Studies, 2*(2), 95-106.

[8] Gudepu, B. K., & Jaladi, D. S. (2021). GDPR Compliance Challenges and How to Overcome Them. *International Journal of Modern Computing, 4*(1), 61-71.

[9] Jaladi, D. S., & Vutla, S. (2021a). Exploring the Current Landscape and Applications of Artificial Intelligence in Healthcare. *The Computertech*, 28-38.

[10] Gudepu, B. K., & Jaladi, D. S. (2018a). The Role of Data Profiling in Improving Data Quality. *The Computertech*, 21-26.

[11] Goriparthi, R.G. (2022) Interpretable Machine Learning Models for Healthcare Diagnostics: Addressing the Black-Box Problem. *Revista de Inteligencia Artificial en Medicina. 13*(1): 508-534.

[12] Jaladi, D. S., & Vutla, S. (2021b). Quantum AI: Accomplishments and Obstacles in the Convergence of Quantum Computing and Artificial Intelligence. *International Journal of Modern Computing, 4*(1), 86-95.

[13] Jaladi, D. S., & Vutla, S. (2017). Harnessing the Potential of Artificial Intelligence and Big Data in Healthcare. *The Computertech*, 31-39.

[14] Gudepu, B. K., & Jaladi, D. S. (2018b). The Role of Data Quality Scorecards in Measuring Business Success. *The Computertech*, 29-36.

[15] Jaladi, D. S., & Vutla, S. (2018a). An Analysis of Big Data Analytics in Relation to Artificial Intelligence and Business Intelligence. *The Computertech*, 37-46.

[16] Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain?. *arXiv preprint arXiv:1712.09923*..

[17] Jaladi, D. S., & Vutla, S. (2018b). The Use of AI and Big Data in Health Care. *The Computertech*, 45-53.

[18] Wang, F., & Preininger, A. (2019). AI in health: state of the art, challenges, and future directions. *Yearbook of medical informatics*, *28*(01), 016-026.

[19] Sethuraman, P. (2022). Latency-Aware Scheduling and Resource Control Algorithms for Emergency and Public Safety Wireless Networks. *International Journal of Emerging Research in Engineering and Technology, 3*(4), 133-140.

[20] Jaladi, D. S., & Vutla, S. (2019a). Deploying Breiman's Random Forest Algorithm in Machine Learning. *The Computertech*, 45-57.

[21] Chennareddy, R. K. (2020). Engineering Intelligence Systems Using Big Data and Cloud Architectures for Modern Data Intensive Applications. *International Journal of AI, BigData, Computational and Management Studies, 1*(2), 41-50.

[22] Jaladi, D. S., & Vutla, S. (2019b). Revolutionizing Healthcare Through Quantum Computing: Insights and Future Directions. *International Journal of Modern Computing, 2*(1), 60-83.

[23] Goriparthi, R.G. (2022) Interpretable Machine Learning Models for Healthcare Diagnostics: Addressing the Black-Box Problem. *Revista de Inteligencia Artificial en Medicina. 13*(1): 508-534.

[24] Jaladi, D. S., & Vutla, S. (2020a). Leveraging Data Mining to Innovate Agricultural Applications. *International Journal of Modern Computing, 3*(1), 34-46.

[25] Kim, D. W., Jang, H. Y., Kim, K. W., Shin, Y., & Park, S. H. (2019). Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. *Korean journal of radiology*, *20*(3), 405-410.

[26] Jaladi, D. S., & Vutla, S. (2020b). Machine Learning Demystified: Concepts, Algorithms, and Use Cases. *The Computertech*, 1-12.