Big Data Integration with Machine Learning Towards Public Health Records and Precision Medicine

Dr. Maria Fernanda Castillo[1], and Dr. Ahmed Al-Mansouri[2]

[1] *Department of Medical Research, King Saud University, Saudi Arabia.*

[2] *Department of Medical Research, King Saud University, Saudi Arabia.*

Received: 24/October/2024; Revised: 21/November/2024; Accepted: 26/December/2024; Published: 30/January/2025

Abstract

Contemporary biological science studies associate the term 'precision' with health care and public health (PH) with concepts such as big data, data integration, and Machine Learning (ML). Technological breakthroughs enable the aggregation and integration of extensive heterogeneous information from many sources, including genomic sequences, social media postings, Public Health Records (PHR), and wearable devices. Moreover, sophisticated algorithms facilitated by high-performance computers enable the converting of big data into intelligence. Notwithstanding this progress, several obstacles persist in realizing Precision Medicine (PM) and PHR for the betterment of individuals and populations. A primary objective of PM is to incorporate and integrate big data from various PHRs regarding the cellular and ecological roots of disease into analytical backgrounds, facilitating the creation of personalized, context-specific diagnosis and treatment plans. In this context, artificial intelligence (AI) and ML methodologies may be used to develop analytical representations of intricate diseases to predict individualized health situations and outcomes. These frameworks must include the extensive variability among people regarding their genetic predispositions and contextual and social variables. Computational methodologies in medicine must effectively handle, display, and incorporate big data, including organized and unorganized forms. Effective big data integration and organization are essential for the effective use of ML methodologies in PM. Numerous obstacles emerge in developing effective medical big data analytics systems, given the present stringent performance requirements in PM, alongside limitations in time, computing resources, and bioethical considerations.

Keywords: Machine Learning, Big Data, Precision Medicine, Data Integration, Public Health Records.

1 INTRODUCTION

In the last ten years, interest in PM has grown. This domain investigates the advancement of personalized therapies for people, informed by biological, ecological, medical, and communal determinants. The National Institute of Health characterizes PM as an "approach for preventing and treating diseases that consider unique variations in environmental factors, genetics, and lifestyle for every person," aiming to precisely identify which therapies and preventative measures will be most effective for certain populations (Naqvi et al., 2020). PM involves a comprehensive knowledge of a person's wellness to provide more targeted treatments or preventative strategies for certain features and profiles within a population. Genetic data from people with a common ailment may be used to create novel pharmaceuticals for those affected by this condition. The Centre for Disease Control and

Prevention defines public wellness as the "science of safeguarding and enhancing the well-being of persons and their populations" (Wu et al., 2016). In PM, the person is the primary focus, while the healthcare system considers the population as the fundamental unit for treatments. This is accomplished by preventive measures and interventions within a population. Building upon the prior definition of PM, one can broaden the concept to encompass "precision public wellness": the examination of relationships among genes and biology alongside individual, ecological, and social elements of health, aimed at monitoring the prevalence of diseases within societies and directing successful measures at the population level. This would include categorizing populations based on certain qualities, actions, or genetic data to enhance therapy and intervention results.

Advances in PHR characterize precision PH as "the implementation and integration of novel and established technologies that more accurately delineate and assess people and their environments throughout their lifespan to customize preventive measures for at-risk populations and enhance the complete well-being of the community." Authors in (Thirunavukarasu et al., 2022) defined precision health care as "the potential to prevent illness, enhance health, and minimize health inequities in populations" via the use of developing approaches and technology. These novel methodologies include strategies for evaluating extensive volumes of heterogeneous data, facilitated by many advancements in health data acquisition. Among these advancements, we may mention intelligent technologies that gather patient-generated health data (PGHD), PHR, and genetic sequencing.

Big Data technology and methodologies assist academics in interpreting the vast and varied PH information now being produced. Data investigation is crucial since society has entered a stage when data creation exceeds the human ability to derive insights without computer assistance (Gupta & Kumar, 2023). To conduct these studies, robust data storage and computing methods and AI algorithms for interpreting and deriving insights from Big Data are required. Improvements in computational capacity, data acquisition methods, and storage capabilities allow researchers to analyze interactions within extensive databases, including genetics, environmental data, online communities, and many forms of medical and personal information. PM and healthcare will employ the fresh medical information created by modern technologies, with Big Data serving as the computer science domain that facilitates the identification of trends and findings within this data. AI and statistical analysis approaches are used to facilitate these studies.

Moreover, PM and public wellness are predominantly regarded as distinct study domains in modern healthcare. Both domains need substantial data: in PM, researchers often collect diverse data from a limited number of individuals, while in healthcare services, they get a limited number of data types from a large population. In reality, increased data gathered from a diverse population should provide superior outcomes in statistical studies. Genomic variation can only be observed with a sufficiently large population (Kraus et al., 2018).

In the pursuit of more effective medical care interventions grounded in data-optimized medical policy and procedure, AI or ML will undoubtedly play a pivotal role in transitioning from a traditional medicine approach—predominantly reliant on the expertise of well-trained clinicians—to one that incorporates comprehensive (often programmed) analyses of the intricate interactions among molecular dynamics, biological features, environmental factors, and social influences, thereby fulfilling the potential of customized healthcare (Hulsen et al., 2019). The formulation of this analytical methodology for customized medicine, often referred to as PM, encompasses several conceptual structures, including biological systems, mathematical biology, healthcare information systems, and digital medicine (Velmovitsky et al., 2021).

Healthcare and wellness are inherently multifaceted; thus, their research must include genetic, atomic, clinical, and demographic data. Wellness and health analytics must examine and address the complexity and inherent biases of the many health-related datasets, including atomic, medical, and statistical information. This underscores the need for tailored, flexible quantitative and statistical techniques for design identification and strategy formulation and testing. AI and ML are becoming fundamental to customized treatment.

2 ROLE OF BIG DATA INTEGRATION WITH ML TOWARDS PHR AND PM

The ongoing advancement of advanced and efficient AI or ML, coupled with the growth of big data sources in the medical sector, has heightened expectations concerning the numerous potential advantages arising from the combination of robust methodologies and quality data (Sahu et al., 2022). Nevertheless, for substantial data volumes to effectively contribute to developing robust AI/ML models, mere quantity is insufficient; this critical consideration is often neglected. Medical and biological data exists in many quantities, types, and forms; it is often intricate, diverse, inadequately documented, and usually unorganized. The efficient AI or ML model is now hampered by these problems: dimension, variation, layout, complexity, variation, poor annotation, and absence of structure, as given in part 1 of Figure 1.

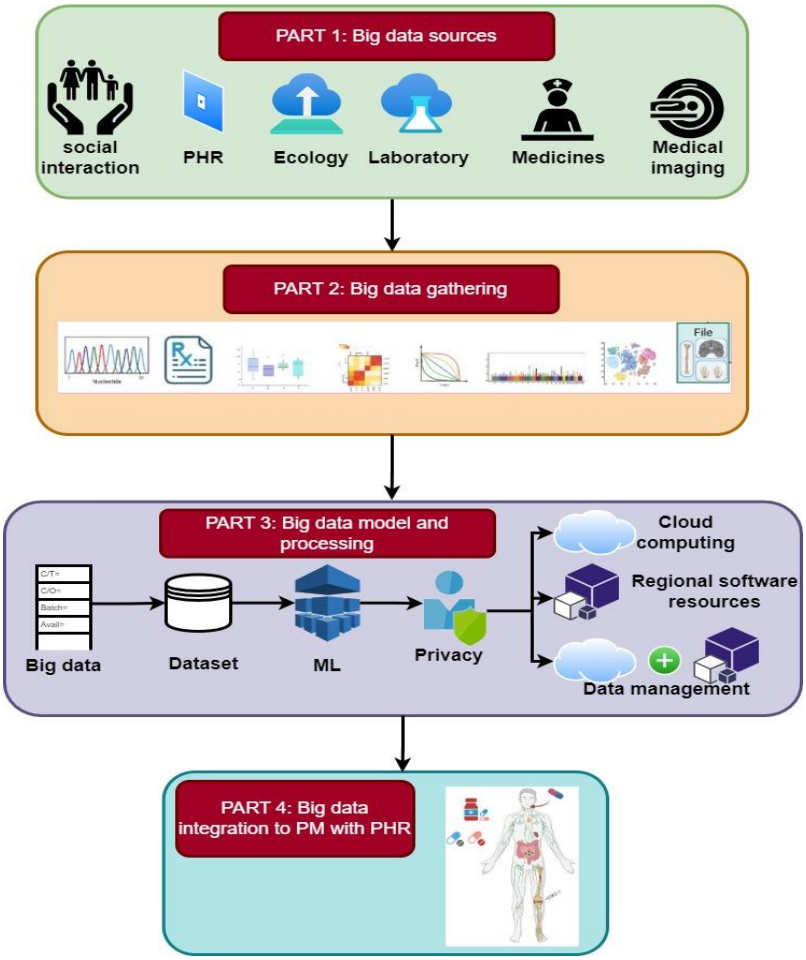

Figure 1: Big Data Integration with ML to PM with PHR

In terms of size, despite frequently engaging with big data—typically regarded as beneficial—it is prevalent for these sources of information to experience a calamity of dimensions (CoD). In this condition, the quantity of parameters or features significantly exceeds the number of test specimens or insights. CoD is especially pronounced in genomic and transcriptome investigations, where the quantity of genes or scripts often reaches numerous thousands. Still, the specimen size seldom exceeds a few hundred or several thousand. The situation becomes further intricate when assessing chemical alterations, such as DNA mutation; modern experimental techniques permit the simultaneous measurement of several thousand mutation probes. Elevated data size may lead to overestimation in AI or ML methodologies. Overestimating indicates that the algorithms exhibit great accuracy in training data but perform poorly in extrapolating or managing unknown information. Promising approaches may not succeed in practical implementations. One strategy to address the CoD is to reduce data size before training the ML algorithms. The main methods of data decreasing dimensionality include feature mining, which involves projecting data from an expansive space to a lower-dimensional one and selecting features, which decreases dimensions by choosing a relevant or useful sample of the originating variables.

Feature mining techniques, including principal component analysis (PCA), eigenvalue decomposition-based techniques, and t-diverged unpredictable neighbor integrating (t-UNI), facilitate

enhanced presentation of data, investigation, contraction, and hidden factor targeting. Conversely, the selection of feature approaches includes any combination of the following strategies: information filtering (IF), information encasing (IE), and information winding (IW). The objective of the former (IF) is to identify a subsection of pertinent characteristics in a model-independent manner, incorporating scientific methods including ANOVA, Pearson's correlation coefficient, data-theoretical strategies like entropy and information sharing, restricted regression analysis, and maximal Significance Limited Redundancy (mSLR) techniques (Martínez-García & Hernández-Lemus, 2022).

Data storage approaches seek the optimal mix of features a certain prediction model uses, including recursive feature removal (RFR), jackstraw, and Random Forests (RF). IW is an amalgamation of IF and IE that executes choosing features during the construction of a forecasting model; an important instance of the IE methodology is the minimal relative shrinkage and decision operator (MRSDO) along with its extensions.

The richness and diversity of data provide challenges for deploying AI and ML models in PM with PHR. Divergence arises from numerous circumstances, including significantly varied variables (or differing coding) across varied data sources (for instance, PHR from different medical facilities), incongruent variations or scaling, like distinct flexible ranges (for instance, integrated information on expression from DNA microarrays and RNA sequencing techniques), varied data procedures (consistent signals, numbers, periods, classes, routes, etc., obtained from atomic and visualization), and distinct formats as given in part 2 of Figure 1. Combining diverse data types may be performed simply by merging attributes from diverse data sources; however, this approach diminishes the efficacy of decision tree (DT) models, which are prone to overestimation. An alternate approach involves using graded regression methods, such as elastic nets, and various normalization algorithms; nevertheless, this may pose issues with the accessibility of findings. Enhanced outcomes may be achieved via the use of various kernel learning techniques.

Owing to the inherent complexity of biological and medical information, as well as challenges in subject or model procurement and information collecting (data generation/sampling equipment may malfunction), It is typical to encounter challenging situations, including absent data (from unmeasured or inaccurately measured cases), class inequalities (very disparate sample numbers across several characteristic groups), and scarcity (an extreme variant of class disparity). Numerous ML strategies exist to address missing information and class disparity, including list-wise elimination (i.e., entirely removing the error-prone samples from the learning) and attribution (i.e., estimating the missing value using anticipation methods based on sample or feature-wise accounts), employing techniques like k-NN substitution, reliant standards, random gradient promoted trees, and various ensemble regression structures.

AI or ML may assist medical professionals in being current with the latest scientific research in their disciplines, a challenge that has long plagued attending doctors. In summary, when physicians seek to

provide their patients with the most effective treatment alternatives, challenges emerge in determining what is now regarded as superior. The existing scholarly literature on a single medical specialty is already extensive. When managing multi-morbid patients, the circumstance deteriorates since medical recommendations and algorithms often focus on single-condition scenarios. By adopting the ML framework, medical professionals can utilize a novel array of tools that facilitate recommendations and assessments based on immediate patient data analysis, which encompass the intricacies of an individual's genetic profile, environmental factors, and relevant complications, in conjunction with accepted norms of care in the prevailing research as shown in part 3 of Figure 1.

In addition to traditional biological and medical records, big data analytics facilitates the integration of professional, interpersonal, biological, and behavioral data for particular patients, sourced from social networks, handheld devices, and other cloud-based materials, to augment clinical profiles. To get to this juncture, however, significant dilemmas must be resolved. Specifically, innovative computational and diagnostic frameworks must be developed to identify patients' comparisons and disparities, as well as to uncover patterns that elucidate their associations and differences, with the objective of computing customized illness risk characteristics, similar to inherited risk assessments, but from a broader perspective—incorporating all previously discussed data types—facilitating customized PM.

Therefore, by combining phenotypic and illness-history-based methodologies, big data analysis seeks to enhance individualized disease forecasting, optimize healthcare administration, and ultimately foster a beneficial effect on individual well-being, as illustrated in part 4 of Figure 1. ML methodologies contribute to a transition in medication from a disease-centric perspective to a patient-oriented training. AI or ML and big data analysis have already yielded significant advancements in PM. Nonetheless, an agreement has yet to be established regarding the integration of extensive electronic health record data, diverse databases containing cellular, genetic, and ecological information from substantial experimental, medical, and epidemiological studies, along with individual-specific data collected from various sources, including social networks and mobile devices, to formulate a PM approach using PHR.

Big data integration and standardization are becoming crucial in medical studies and customized clinical practice environments. Recent discussions have highlighted the necessity of medical study information sharing for the replication of outcomes, transparency of results, the enhancement of additional studies or sophisticated clinical trial phases, and the facilitation of digital comparisons of efficacy, which are significantly more efficient and cost-effective than traditional methods. Additionally, it accelerates result coverage, fosters continuous learning, and supports the development of startups or entrepreneurial ventures, among other considerations. Uniformity is essential to ensure the effective use of shared information.

Big data is not the only concern that requires evaluation and validation for the extensive use of AI and ML methodologies in PM and PHR. Researchers have lately advocated for the importance of computational governance for AI or ML skills in the medical domain. In this context, an algorithmic manager is an individual or group within a medical facility or medical organization tasked with responsibilities such as developing and maintaining a record of algorithms utilized in the organization, overseeing the continuing clinical application and performance of these computational resources, and assessing the security, effectiveness, and equity of the methods employed. Methods and information are the most prominent components of the medical analytics environment; however, big data integration is increasingly assuming a significant role in ML applications in PHR and PM, as it often provides insights for computerized labeling or categorization tasks, which can be further refined through AI and statistical learning techniques.

3 CONCLUSION

Recent research in the biological sciences links the term 'precision' to health care and PH, emphasizing concepts such as big data, data integration, and ML. Technological advancements facilitate aggregating and integrating diverse and extensive information from various sources, such as genomic sequences, social media posts, PHR, and wearable devices. Furthermore, advanced algorithms supported by high-performance computing allow for transforming large datasets into actionable insights. Despite this progress, various challenges remain in achieving PM and PHR to improve individuals and populations. The primary objective of PM is to integrate big data from various PHRs concerning the cellular and ecological origins of disease into analytical frameworks, thereby enabling the growth of personalized, context-specific diagnosis and treatment plans. AI and ML methodologies can create analytical models of complex diseases to predict personalized health conditions and outcomes in this context. These frameworks must account for the significant variability among individuals regarding their genetic predispositions and contextual and social factors. Computational methodologies in medicine must efficiently manage, present, and integrate big data, encompassing structured and unstructured formats. Efficient integration and organization of big data are crucial for the optimal application of machine learning methodologies in PM.

REFERENCES

[1] Naqvi, M. R., Jaffar, M. A., Aslam, M., Shahzad, S. K., Iqbal, M. W., & Farooq, A. (2020, June). Importance of big data in precision and personalized medicine. In *2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)* (pp. 1-6). IEEE. https://doi.org/10.1109/HORA49412.2020.9152842

[2] Wu, P. Y., Cheng, C. W., Kaddi, C. D., Venugopalan, J., Hoffman, R., & Wang, M. D. (2016). –Omic and electronic health record big data analytics for precision medicine. *IEEE Transactions on Biomedical Engineering*, *64*(2), 263-273. https://doi.org/10.1109/TBME.2016.2573285

[3] Thirunavukarasu, R., Gnanasambandan, R., Gopikrishnan, M., & Palanisamy, V. (2022). Towards computational solutions for precision medicine based big data healthcare system using deep learning models: A review. *Computers in Biology and Medicine*, *149*, 106020. https://doi.org/10.1016/j.compbiomed.2022.106020

[4] Gupta, N. S., & Kumar, P. (2023). Perspective of artificial intelligence in healthcare data management: A journey towards precision medicine. *Computers in Biology and Medicine*, *162*, 107051. https://doi.org/10.1016/j.compbiomed.2023.107051

[5] Kraus, J. M., Lausser, L., Kuhn, P., Jobst, F., Bock, M., Halanke, C., ... & Kestler, H. A. (2018). Big data and precision medicine: challenges and strategies with healthcare data. *International Journal of Data Science and Analytics*, *6*, 241-249. https://doi.org/10.1007/s41060-018-0095-0

[6] Hulsen, T., Jamuar, S. S., Moody, A. R., Karnes, J. H., Varga, O., Hedensted, S., ... & McKinney, E. F. (2019). From big data to precision medicine. *Frontiers in medicine*, *6*, 34.

[7] Velmovitsky, P. E., Bevilacqua, T., Alencar, P., Cowan, D., & Morita, P. P. (2021). Convergence of precision medicine and public health into precision public health: toward a big data perspective. *Frontiers in Public Health*, *9*, 561873. https://doi.org/10.3389/fpubh.2021.561873

[8] Sahu, M., Gupta, R., Ambasta, R. K., & Kumar, P. (2022). Artificial intelligence and machine learning in precision medicine: A paradigm shift in big data analysis. *Progress in molecular biology and translational science*, *190*(1), 57-100. https://doi.org/10.1016/bs.pmbts.2022.03.002

[9] Martínez-García, M., & Hernández-Lemus, E. (2022). Data integration challenges for machine learning in precision medicine. *Frontiers in medicine*, *8*, 784455. https://doi.org/10.3389/fmed.2021.784455