International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

Impact of Data Hygiene on AI Performance: A Comprehensive Study in Healthcare Applications

Vinu V Pillai Sobhana

Abstract: ***Integrating clean data and Artificial Intelligence (AI) in health can completely transform medical diagnosis, therapy, and patient management. Clear data is very important for guaranteeing the validity and dependability of artificial intelligence models, therefore raising their capacity to give accurate medical observations. Still, difficulties including data quality, interoperability, bias, privacy concerns, scalability, and regulatory limits impede the seamless use of artificial intelligence in the medical field. Notwithstanding these hiccups, artificially intelligent healthcare solutions raise diagnostic accuracy, automate administrative works, and allow for personalized medicine. The vital need of artificial intelligence in healthcare data cleaning, its difficulties, and more general ramifications for the healthcare sector are addressed in this article. Maximizing AI's potential and guarantee ethical and secure deployment will depend in part on creative frameworks and rules handling these problems.***

Keywords: clean data, artificial intelligence, healthcare transformation, diagnostic accuracy, ethical deployment

1. Introduction

A fundamental part of data analytics, data cleaning handles such problems as missing, redundant, or imprecise values that could compromise analysis quality (Krishnan et al., 2015). In the confines of healthcare, "dirty data" means data that is wrong, missing, or inconsistent; this could cause improper diagnoses, useless therapy, and compromised patient safety. Especially difficult is big data environments, since effective management techniques are needed for vast quantities of structured and unstructured data (Jesmeen M. Z. H. et al. (2018). Employing machine learning algorithms to automate data cleaning procedures can enhance data quality (Z. H et al.2018; Singh and Dwivedi 2020). A current strategy, data quality mining uses data mining methods to find and fix data quality issues in extensive databases (Singh and Dwivedi, 2020). Data cleansing, though important, is sometimes left out of the data analytics curriculum, therefore leaving students unprepared for situations in reality for which most of an analyst's time is spend on data preparation (Snyder, 2019). SampleClean and other technologies have been designed to estimate query results when only a sample of data can be cleaned, hence dealing with problems in both database research and machine learning - based analytics (Krishnan et al., 2015).

Accurate and good data for analysis and decision - making depend greatly on thorough healthcare data cleaning. (Mavrogiorgos et al., 2022) suggest a multi - level strategy that uses validation, cleansing, checking, and logging methods to tackle several data quality problems. Rahul et al., For healthcare big data uses, highlight the difficulties of data cleaning, missing values, and outlier analysis, therefore recommending optimization - based models for dirty data identification and cleaning. (Mathew and Pillai, 2015) underline the need of using sophisticated analytics to enhance results as well as integrating disparate healthcare information from several sources. They chat about how big data systems could help improve scalability and conquer processing restrictions. The Terracina group of researchers. A logic programming - driven method for data cleaning is shown in application in the Italian National Healthcare System to study archives of disease caused by tumor.

(Thanikaivel and Ramakrishnan, 2019) argue that India's vast population, low literacy levels, and bad sanitation seriously hamper its healthcare system. With states having main control over healthcare provision, the system combines public and private spheres (Galhotra, 2020). Although public healthcare is organized according to primary, secondary, and tertiary levels, often poor is implementation and regulation (Thanikaivel and Ramakrishnan, 2019; Galhotra, 2020). This challenge can be reduced by the utilisation of AI.

Especially in the shape of predictive analysis, artificial intelligence can help medical professionals to predict patient results and spot people who might be at risk, therefore going a long way in solving these issues. Though still not enough, government healthcare spending is rising from states contributing around 80% of public funds (Galhotra, 2020; Ghosh, 2021). Rural and urban regions have a significant difference in healthcare facilities (Ghosh, 2021). Some health indicators—including life expectancy—improve notwithstanding obstacles (Mathai et al., 2014). (Mathai et al., 2014). recommend a balance between free market and legislation to rational healthcare distribution. Moreover, growing health insurance coverage among the public is a result of the soaring cost of healthcare (Ghosh, 2021).

Role of AI in Data Cleaning

Applied more and more to data cleaning activities, artificial intelligence (AI) provides great improvements in effectiveness and efficiency. Relational databases are subject to error checking, data correction, and imputation by means of artificial intelligence methods including machine learning and deep learning (Jingyu Zhu et al., 2024). Although they may forfeit interpretability contrasted with conventional methods, these techniques can manage difficult links and merge several cleaning signals. Systems for artificial intelligence planning have also been created to streamline data - cleaning operations. Still, the lack of transparency in data cleansing techniques might cause comparability problems in AI assessment, therefore requiring standard protocols (Tjoa et al., 2022). Recent developments show better accuracy and speed on data preparation activities as frameworks using natural language processing, supervised and unsupervised learning tackle several data quality issues by integrating those learning techniques (Thoutam, 2024).

These AI - powered methods let data experts concentrate on more valuable analytical projects while guaranteeing steady data quality.

Not only does artificial intelligence (AI) revolutionize healthcare, it also greatly helps patient results. Artificial intelligence is guaranteeing that patients get the finest possible care by increasing diagnostic accuracy, tailoring treatments, and simplifying administrative chores (Bhagat et al., 2024; Talati, 2023). Already very promising are artificial intelligence uses in medicine including analysis of medical images, predictive analytics, virtual health aids, and clinical decision support systems (Chatterjee et al., 2024). Early disease detection (Chatterjee et al., 2024) can be supported by machine learning algorithms that accurately analyze medical images including X - rays, MRIs, and CT scans. With artificial intelligence - driven predictive analytics, healthcare practitioners can forecast patient results and identify at - risk people (Chatterjee et al.2024). Natural language processing and computer vision technologies help with clinical decision - making and individualized treatment plans (Talati, 2023). Though there are very encouraging developments, issues around data privacy, ethical concerns, and regulatory frameworks are to be solved (Talati, 2023; Chatterjee et al., 2024). Fully maximizing the potential of artificial intelligence while guaranteeing ethical standards and patient safety depends on continuous research and regulatory oversight (Bhagat et al., 2024; Purohit and Kumar, 2024)

Improving several facets of patient care, diagnosis, and treatment, artificial intelligence is transforming healthcare (Chatterjee et al., 2024; Sharma, 2020). Medical imaging review, prognosis analytics, virtual health advisors, clinical decision support systems, and robotic surgery are among the AI applications used in healthcare (Chatterjee et al., 2024). By early detecting diseases, machine learning software can carefully analyze medical images. Permeable forecasts produced by artificial intelligence drive personalized treatment plans and early interventions (Chatterjee et al., 2024). Artificial intelligence speeds up the drug discovery process and autotarget discovery (Shaheen 2021). With help of artificial intelligence, clinical trials can manage great amounts of information and deliver reliable findings (Shaheen, 2021). Although there are many advantages, obstacles such as ethical issues, regulatory compliance, and data privacy persist (Sharma, 2020; Chatterjee et al., 2024). Ensuring responsible use of artificial intelligence in medicine while maximizing its full potential depends on cooperative projects among developers of AI, healthcare professionals, legislators, and regulators (Chatterjee et al., 2024).

AI in healthcare applications and benefits

Healthcare has great potential for enhancing patient outcomes and clinical decision - making through the use of artificial cleansed data and artificial intelligence integration. Artificial intelligence approaches such Natural Language Processing and Machine Learning can improve patient outcomes, treatment efficiency, and diagnosis precision (Divya, 2024). AI could enable patients, enhance relationships between patient and provider, and cut clinical visit expenses by merging electronic health records (EHRs) with patient - produced health information (PGHD) (Ye et al., 2024). Managing diverse datasets, discovering dynamic patterns, and using advanced algorithms for accurate suggestions depending on incorporated data all depend significantly on artificial intelligence (Ye et al., 2024). There are still obstacles, though, including data privacy, interoperability, and meaningful usage (Ye et al., 2024). Handling these problems calls for a single data cleaning approach including MLClean to preprocess data sets so accurate, fair, and robust models for sensitive applications can be assured (Tae et al., 2019). This method shows a bigger pattern of incorporating Big Data and artificial intelligence applications in the medical field.

Artificial Intelligence (AI) is revolutionizing healthcare, particularly in diagnosis and treatment. AI technologies, including machine learning and deep learning, have demonstrated superior performance in detecting and classifying various diseases, often surpassing human experts (Atiqur Rahman, 2021). In medical imaging, AI enhances diagnostic accuracy in radiology, pathology, and dermatology (Fardin Quazi et al., 2024). AI - powered systems can analyze large volumes of imaging data, patient information, and genetic data to provide personalized diagnostic recommendations and treatment plans (Kharibam Jilenkumari Devi et al., 2023). Applications of AI in healthcare extend to robotic surgery, virtual health assistants, and computer - aided diagnostics (Fardin Quazi et al., 2024). AI also shows promise in predicting mental health risks and optimizing disease management (Atiqur Rahman, 2021). Various automated systems and tools, such as artificial neural networks and fuzzy approaches, help minimize errors and control disease progression (Mishra Sg et al., 2017). Despite challenges, AI's integration into healthcare has the potential to improve patient outcomes and streamline healthcare delivery significantly.

Clean data in healthcare disease prediction

In healthcare, improving disease forecast accuracy depends much on data cleaning. Many studies have looked into this subject, emphasizing heart disease prediction and other long - term issues. Alkhafaji et al.2010 Data purging before forecasting resulted in excellent precision levels with decision trees, Bayesian classification, and neural networks, as shown in 2020. For their part, Lattar et al. For duplicate record removal, the 2020 suggested a deep - learning - based solution which enhanced the accuracy of heart disease prediction. Chen et al. 31 - 356. Using a latent factor model, a CNN - based multimodal disease risk prediction algorithm was developed and incomplete data problems were tackled, reaching 94.8% accuracy for cerebral infarction prediction. These research underline how critical data quality is in medical analysis. Sharma and colleagues. (2023) more thoroughly investigated how machine learning techniques could be applied on big healthcare community data from many sources, therefore emphasizing the promise of big data in early detection and disease forecast.

Particularly in early detection and disease forecast, artificial intelligence (AI) transforms healthcare. Using data from wearable devices, electronic health records, and medical images, artificial intelligence could spot possible illnesses and at - risk patients (Patil and Patil, 2023). By enabling doctors to make decisions and provide patient treatment, machine learning algorithms enable accurate forecasts to be pulled from vast databases (Diksha et al., 2023). Using methods such

support vector machines, neural networks, and natural language processing, AI applications in healthcare cover several fields—from cancer and neurology to cardiology. S. S. Ponde and others 2020). Integrating AI with other technologies such Internet of Things, cloud computing, and big data analytics improves health care system efficiency and patient results by means of advanced monitoring techniques and management. Deeba and S. A Patil 2021. On the whole, artificial intelligence in medical care offers the possibility of advanced early detection, proactive intervention, error correction, and improved decision - making, therefore lowering medical expenses and helping to improve patient care.

Integrating artificial intelligence and pristine data in health disease prognosis provides great promise for enhancing patient results and medical procedures. By examining large datasets to customize therapies based on specific genetic profiles, artificial intelligence approaches—machine learning and computational genomics—enable personalized medicine (Rane et al., 2023). Early disease identification and preventative health management are made easier by predictive modeling and machine learning techniques (Rane et al., 2023; Rao et al., 2022). More precise disease prediction results from wireless sensors and body sensors linked with neural networks can pick up physiological changes and behavioural signals (Rao et al., 2022; UmaMaheswaran et al., 2022). An AI - driven examination of medical imaging including pathology informatics and radiomics improves disease definition and understanding of progression (Rane et al., 2023). By enhancing data processing and analysis, the integration of AI with current healthcare systems produces more effective and accurate disease forecasting models (UmaMaheswaran et al., 2022). Still, ethical concerns and careful application of AI technologies in healthcare present major problems to solve (Rane et al., 2023).

2. Challenges of AI and Clean Data in Healthcare and Their Implications

1) **Data Quality and Accuracy**
   - Healthcare data is often incomplete, inconsistent, or erroneous.
   - AI models require high - quality, standardized data for accurate predictions.
   - Implication: Poor data quality can lead to misdiagnosis, incorrect treatment recommendations, and biased AI models.
2) **Data Integration and Interoperability**
   - Healthcare data comes from multiple sources (EHRs, wearables, medical imaging, etc.), often in incompatible formats.
   - Implication: Lack of standardization makes it difficult to aggregate and analyze data effectively, limiting AI's efficiency.
3) **Bias and Fairness**
   - AI models trained on biased or unrepresentative data can reinforce health disparities.
   - Implication: Certain populations may receive suboptimal care due to biased AI - driven recommendations.
4) **Privacy and Security Risks**
   - Healthcare data is sensitive, and AI applications require large datasets.
   - Implication: Risk of data breaches, ethical concerns over patient consent, and regulatory compliance challenges (e. g., HIPAA, GDPR).
5) **Scalability and Computational Costs**
   - AI algorithms require substantial computing power and storage, which may not be feasible for all healthcare institutions.
   - Implication: Small or underfunded healthcare providers may struggle to adopt AI - driven solutions.
6) **Regulatory and Ethical Challenges**
   - AI in healthcare faces strict regulations, requiring extensive validation and oversight.
   - Implication: Delays in AI deployment and legal risks for healthcare providers and AI developers.

Implications for Healthcare:

- **Enhanced Decision - Making:** Clean data enables AI to provide accurate diagnostics and treatment recommendations.
- **Patient Safety & Trust:** Addressing biases and ensuring data privacy strengthens patient confidence in AI - driven healthcare.
- **Operational Efficiency:** AI - powered automation reduces administrative burden and improves resource allocation.
- **Personalized Medicine:** High - quality data enhances AI's ability to tailor treatments based on individual patient profiles.

3. Conclusion

By allowing precise disease prediction, enhancing clinical decision - making, and optimizing patient care, artificial intelligence - driven healthcare solutions are turning the sector. Still, guaranteeing clean and top - quality data increases AI benefits. Building reliable and efficient AI - driven healthcare systems demands the resolution of several issues connected with AI acceptance including data inconsistency, bias, privacy risks, and legislative hurdles. For effective AI deployment, standardized data integration, strong ethical guidelines, and collaborative efforts among technology developers, medical professionals, and legislators will be absolutely key. Although artificial intelligence has shown encouraging results in reducing diagnostic delay and automating data cleaning, more study and rules are required to guarantee its ethical use and scalability across a wide range of medical settings.

4. Future Scope

Future of artificial intelligence and clean data in healthcare depends on creation of sophisticated data cleaning algorithms to guarantee greater accuracy and reliability in medicinal uses. Better interoperability requirements will enable natural data sharing between health systems, therefore increasing artificial intelligence effectiveness across many settings. Given the ethical AI setup, which will call for strong rules to tackle biases, data privacy worries, and transparency problems, this is still a critical field of attention. Furthermore, advancing will be AI - driven predictive analysis and customized medicine, which will permit early disease

diagnosis and individually matched therapies based on patient - specific information. The merger of artificial intelligence with developing technologies including blockchain, the Internet of Things (IoT), and cloud computing will improve data security, availability, and real - time decision - making possibilities. Furthermore, growing use of artificial intelligence in under - resourced medical institutions can help to close healthcare disparities and increase worldwide healthcare availability. Future studies need to investigate scalable AI - driven healthcare systems that adhere to regulatory standards yet maximize efficiency, precision, and patient - centered results

References

[1] Alkhafaji, M. J. A., Fadhil Aljuboori, A., & Ibrahim, A. A. (2020). Clean medical data and predict heart disease.*2020 International Congress on Human - Computer Interaction, Optimization and Robotic Applications (HORA)*, 1–7. https: //doi. org/10.1109/hora49412.2020.9152870

[2] Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease Prediction by Machine Learning Over Big Data From Healthcare Communities. *IEEE Access*, *5*, 8869–8879. https: //doi. org/10.1109/access.2017.2694446

[3] Chu, X., Ilyas, I., Krishnan, S., & Wang, J. (2016). Data Cleaning: Overview and Emerging Challenges. *Data Cleaning: Overview and Emerging Challenges*.

[4] Devi, K. J., Alghamdi, W., N, D., Alkhayyat, A., Sayyora, A., & Sathish, T. (2023). Artificial Intelligence in Healthcare: Diagnosis, Treatment, and Prediction. *E3S Web of Conferences*, *399*, 04043. https: //doi. org/10.1051/e3sconf/202339904043

[5] Dr. M. Thanikaivel, & Dr. R. Ramakrishnan. (2019). AN OVERVIEW ON INDIAN HEALTHCARE INDUSTRY. *International Journal of Technical Research & Science*, *04* (01), 25–26. https: //doi. org/10.30780/ijtrs. v04. i01.004

[6] Dr. N Sree Divya. (2024). AI ASSISTANCE FOR HEALTHCARE USING NLP. *Indian Scientific Journal of Research in Engineering and Management*, *08* (05), 1–5. https: //doi. org/10.55041/ijsrem34174

[7] Farha Deeba, & Patil, S. R. (2021). Implementation of Artificial Intelligence in Disease Prediction and Healthcare System - A Survey.*2021 Innovations in Power and Advanced Computing Technologies (I - PACT)*. https: //doi. org/10.1109/i - pact52855.2021.9696698

[8] Ghosh, S. (2021). An Overview of Healthcare Spending and Practice in India. *The Management Accountant Journal*, *56* (2), 45. https: //doi. org/10.33516/maj. v56i2.45 - 48p

[9] Indranil Chatterjee, Rajkumar Ghosh, Suchetan Sarkar, Krishna Das, & Monojit Kundu. (2024). Revolutionizing Innovations and Impact of Artificial Intelligence in Healthcare. *International Journal for Multidisciplinary Research*, *6* (3). https: //doi. org/10.36948/ijfmr.2024. v06i03.19333

[10] Isha, Miss., Komal, Miss., None Mr. Navoday Atul Kopawar, & None Prof. Dipali A. Sananse. (2024). Artificial Intelligence in Healthcare: A Review. *International Journal of Scientific Research in Science Engineering and Technology*, *11* (4), 133–138. https: //doi. org/10.32628/ijsrset24114107

[11] Krishnan, S., Wang, J., Franklin, M. J., Goldberg, K., Kraska, T., Milo, T., & Wu, E. (2015). SampleClean: Fast and Reliable Analytics on Dirty Data. *IEEE Data (Base) Engineering Bulletin*, *38*, 59–75.

[12] Lattar, H., Salem, A. B., & Ben Ghezala, H. H. (2020). Does data cleaning improve heart disease prediction? *Procedia Computer Science*, *176*, 1131–1140. https: //doi. org/10.1016/j. procs.2020.09.109

[13] M Srinivasa Rao, S. K. UmaMaheswaran, Naveen Chakravarthy Sattaru, Khairul Hafezad Abdullah, Umesh Kumar Pandey, & Laxmi Biban. (2022). A Critical Understanding of Integrated Artificial Intelligence Techniques for the Healthcare Prediction System.*2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*. https: //doi. org/10.1109/icacite53722.2022.9823678

[14] Mathai, K., Mathai, S., & Subba, G. (2014). Rationalizing health care in India: Challenges & strategies. *Journal of Marine Medical Society*, *16* (2), 98. https: //doi. org/10.4103/0975 - 3605.203387

[15] Mishra Sg, Takke Ak, Auti St, Suryavanshi Sv, & Oza Mj. (2017). Role of Artificial Intelligence in Health Care. *Biochemistry: An Indian Journal*, *11* (5).

[16] None Diksha, Verma, A. K., Shuchi Bhadula, & Sharma, S. (2023). *Artificial Intelligence Enabled Disease Prediction System in Healthcare Industry*.1–6. https: //doi. org/10.1109/icbds58040.2023.10346275

[17] Patil, N. P., & Patil, N. K. (2023). *A Review on Disease Prediction Using Artificial Intelligence*.1 (1), 1–10. https: //doi. org/10.59535/ece. v1i1.8

[18] Ponde, P., Swati, M., & Padghan, V. (2020). SMART HEALTH PREDICTION FOR AVOIDING FUTURE HEALTH RISK BY USING MACHINE LEARNING. *IJSDR2008049 International Journal of Scientific Development and Research*, *5*, 2455–2631. https: //www.ijsdr. org/papers/IJSDR2008049. pdf

[19] Quazi, F., Mohammed, PhD, A. S., & Gorrepati, N. (2024). Transforming Treatment and Diagnosis in Healthcare through AI. *International Journal of Global Innovations and Solutions (IJGIS)*. https: //doi. org/10.21428/e90189c8.072ffbe8

[20] Rahman, A. (2021). Artificial Intelligence in Healthcare: A Review of Diagnostic Applications and Impact on Clinical Practice. *Journal of Primeasia*, *2* (1), 1–5. https: //doi. org/10.25163/primeasia.219816

[21] Rane, N., Choudhary, S., & Rane, J. (2023). Towards Autonomous Healthcare: Integrating Artificial Intelligence (AI) for Personalized Medicine and Disease Prediction. *Social Science Research Network*. https: //doi. org/10.2139/ssrn.4637894

[22] Saraswati, P., & Kumar C. N., S. (2024). AI in Health Care: A Comprehensive Review. *A and v Pub Journal of Nursing and Medical Research*, 112–114. https: //doi. org/10.52711/jnmr.2024.26

[23] Shaheen, M. Y. (2021). Applications of Artificial Intelligence (AI) in healthcare: A review. *ScienceOpen Preprints*, *1* (1). https: //doi. org/10.14293/s2199 - 1006.1. sor - . ppvry8k. v1

[24] Sharma, K., Kaur, A., Ghosh, S., & Anshika Abrol.

(2023). Predicting Diseases Using Machine Learning on Big Data from Healthcare Communities.*2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. https: //doi. org/10.1109/icccnt56998.2023.10307964

[25] Sharma, R. (2020). Artificial Intelligence in Healthcare: A Review. *Türk Bilgisayar ve Matematik Eğitimi Dergisi*, *11* (1), 1663–1667. https: //doi. org/10.61841/turcomat. v11i1.14628

[26] Singh, S. K., & Dwivedi, Dr. R. K. (2020). Data Mining: Dirty Data and Data Cleaning. *SSRN Electronic Journal*. https: //doi. org/10.2139/ssrn.3610772

[27] Snyder, J. (2019). Data Cleansing: An Omission from Data Analytics Coursework. *Information Systems Education Journal*, *17* (6), 22–29.

[28] Tae, K. H., Roh, Y., Oh, Y. H., Kim, H., & Whang, S. E. (2019). Data Cleaning for Accurate, Fair, and Robust Models. *Proceedings of the 3rd International Workshop on Data Management for End - To - End Machine Learning - DEEM'19*. https: //doi. org/10.1145/3329486.3329493

[29] Talati, D. (2023). Artificial Intelligence (AI) In Mental Health Diagnosis and Treatment. *Journal of Knowledge Learning and Science Technology ISSN: 2959 - 6386 (Online)*, *2* (3), 251–253. https: //doi. org/10.60087/jklst. vol2. n3. p253

[30] Thoutam, P. (2024). Automated Data Preparation through Deep Learning: A Novel Framework for Intelligent Data Cleansing and Standardization. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *10* (6), 1867–1877. https: //doi. org/10.32628/cseit241061231

[31] Tjoa, S., Kieseberg, P., Marlies Temper, & Holzinger, A. (2022). *Comparing Apples and Oranges: Comparability problems caused by intransparent Data Cleansing for AI Applications*.5–13. https: //doi. org/10.1109/aeis59450.2022.00009

[32] Ye. (2024). The role of artificial intelligence for the application of integrating electronic health records and patient - generated data in clinical decision support. *AMIA Joint Summits on Translational Science Proceedings. AMIA Joint Summits on Translational Science*, *2024*. https: //pubmed. ncbi. nlm. nih. gov/38827061/

[33] Z. H, J. M., Hossen, J., Sayeed, S., Ho, C., K, T., Rahman, A., & Arif, E. M. H. (2018). A Survey on Cleaning Dirty Data Using Machine Learning Paradigm for Big Data Analytics. *Indonesian Journal of Electrical Engineering and Computer Science*, *10* (3), 1234. https: //doi. org/10.11591/ijeecs. v10. i3. pp1234 - 1243

[34] Zhu, J., Zhao, X., Sun, Y., Song, S., & Yuan, X. (2024). Relational Data Cleaning Meets Artificial Intelligence: A Survey. *Data Science and Engineering*. https: //doi. org/10.1007/s41019 - 024 - 00266 - 7