Khaleel Khan Mohammed, 2025, 13:5
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

# A Survey on Digital Health Care Data Analysis Techniques for Developing Machine Learning Models

**Khaleel Khan Mohammed**
MBA in Management Information Systems - Concordia University Wisconsin, United States
Centene Corporation, Lead Data Engineer

**Abstract- Data is expanding rapidly, driving advances in technology and algorithms. In healthcare and biomedicine, this growth enables early disease prediction, better patient care, and improved community services through Machine Learning and AI. Since disease patterns vary across regions, AI adoption has the potential to radically transform the entire healthcare industry. This paper has brief various models proposed by the researcher for disease detection. Techniques of machine learning for disease prediction was elaborate in the paper. Challenges of prediction models for accuracy was summarize in the work under different condition. Finally paper has brief some of major evaluation parameters of for comparing healthcare models.**

**Keywords: Data mining, Social network, Product ratin, Product recommendation.**

## I. INTRODUCTION

Human life continues to evolve each day, but the health of successive generations shows both progress and decline. Life remains unpredictable, and many individuals suffer from severe health conditions because diseases are often detected too late. For instance, chronic liver disease impacts over 50 million adults globally, yet timely diagnosis could prevent its progression. Machine learning–based disease prediction offers the possibility of identifying common illnesses at an earlier stage. Unfortunately, healthcare is often treated as a secondary priority, leading to serious consequences. Many patients either cannot afford medical consultations or are too constrained by busy schedules, causing them to neglect persistent symptoms that may later result in severe complications [1].

To address such issues, AutoML (Automated Machine Learning) tools have emerged as accessible solutions. These tools handle data preparation, model selection, and tuning automatically, enabling individuals without advanced expertise to build effective machine learning models. In healthcare, AutoML is particularly useful for detecting and managing cardiovascular diseases, saving both time and resources. However, when it comes to complex conditions like heart disease, relying solely on AutoML may not be sufficient [2].

Another significant challenge in modern medicine is the high workload faced by doctors [3] and the increasing cost of consultations [4]. This becomes particularly evident in disease prediction scenarios, where patients report symptoms to general physicians, who then infer possible conditions and refer them to specialists [5]. Such a process can be streamlined using machine learning approaches like the Random Forest algorithm [6]. Random Forest can classify multiple diseases based on symptoms and geographic variations, as certain illnesses are more prevalent in specific regions. This integration of machine learning could greatly reduce logistical burdens while improving early detection and treatment.

## II. RELATED WORK

Anish et al. [7] proposed a hybrid cascaded deep learning framework that integrates ensemble-based feature selection with feature fusion for multi-disease prediction. The feature selection process is optimized through MSB-EV, which identifies the most informative features from statistical, deep, and optimally weighted sets. For the classification stage, the model employs HSC-AttentionNet, which combines a Deep Temporal Convolution Network with LSTM to handle sequential data. This optimization strategy enhances the robustness of the framework, resulting in superior predictive accuracy.

© 2025 Khaleel Khan Mohammed, This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

Kalpana et al. [8] tackled image preprocessing challenges by developing Advanced Image Preprocessing Techniques. Their framework incorporates Adaptive Thresholding, Dynamic Histogram Equalization, Hierarchical Contrast Normalization, Multi-Scale Region Enhancement, and Contextual Feature Augmentation. Collectively, these techniques improve image quality by reducing noise while retaining essential structural details.

Rane et al. [9] examined the significance of transparent AI models in medical domains such as radiology, pathology, cardiology, and oncology. Their study emphasized the importance of interpretability in clinical AI systems and highlighted methods like attention mechanisms and saliency maps to increase clinician trust. By making AI-driven predictions more understandable, this approach facilitates smooth integration of AI into clinical practice and supports broader adoption in healthcare.

Xiao et al. [10] presented a predictive approach that leverages morphologic cell features from cancerous regions. Using CellProfiler software and the Eff-Unet deep learning method, relevant features are extracted, averaged across patient regions, and refined with the Lasso-Cox technique to identify prognosis-related markers. The final prognostic prediction model was evaluated through cross-validation and Kaplan–Meier estimation.

Hu et al. [11] introduced a multilevel image segmentation method for COAD pathology images, based on an improved particle swarm optimization technique. Their multi-strategy hybrid PSO, called DRPSO, integrates population reorganization via MGO to maintain diversity and avoid premature stagnation. By combining 2D Renyi entropy with a non-local mean 2D histogram, they developed a DRPSO-based MIS approach that effectively addresses segmentation in COAD pathology imaging.

Mohamed et al. [12] advanced AI-based medical imaging analysis by combining Convolutional Neural Networks with Fishier Mantis Optimization. CNNs, including architectures like ResNet50 and GoogLeNet, extract complex features from medical images, while the mantis shrimp–inspired optimization algorithm fine-tunes CNN parameters to enhance performance and accelerate convergence. This hybrid method demonstrates improved efficiency and robustness in diagnostic imaging tasks.

## III. DISEASE PREDICTION METHODS

**Regression Methods:** Regression models are widely applied in disease prediction as they estimate the relationship between independent risk factors and health outcomes. Linear regression is often used to predict continuous variables such as cholesterol levels or blood pressure, while logistic regression is effective for binary outcomes, such as the presence or absence of a disease [1]. In cases where health data exhibit complex relationships, non-linear regression provides better accuracy. For example, logistic regression has been employed in early diabetes detection and cardiovascular risk prediction, demonstrating strong performance in clinical datasets [2].

**Bayes Classifier:** The Bayes classifier applies Bayes' theorem to predict the likelihood of disease occurrence based on prior probabilities. Naïve Bayes, despite assuming independence between variables, has been used effectively in diagnosing conditions such as liver disease, diabetes, and heart disease [3]. Its strength lies in handling large medical datasets with categorical attributes, where posterior probabilities help classify patients into high-risk or low-risk groups. While independence assumptions may not always hold in clinical data, the simplicity and efficiency of Naïve Bayes make it suitable for real-time health applications.

**K-Nearest Neighbor (KNN) Classifier:** KNN is an instance-based learning approach that classifies patients according to the similarity of their medical features with known cases. It is commonly applied in predicting chronic diseases like kidney disease, diabetes, and cancer [4]. By calculating similarity using distance metrics such as Euclidean or Manhattan distance, KNN groups patients with similar health profiles and identifies potential risk

categories. Although KNN is computationally intensive with large datasets, its effectiveness in handling multi-class classification makes it valuable for clinical prediction tasks.

**Decision Trees:** Decision tree models provide a transparent and interpretable framework for disease prediction. They split patient data based on key clinical attributes (e.g., age, blood pressure, glucose level) to classify outcomes such as risk of heart disease or hypertension. Their interpretability allows clinicians to trace decision paths, making them highly useful in medical environments [5]. Variants like Random Forests further improve prediction accuracy by combining multiple decision trees and reducing overfitting.

**Support Vector Machines (SVM):** SVM is another powerful algorithm used for classifying disease outcomes, especially when patient data are high-dimensional. By finding the optimal hyperplane that separates healthy and diseased cases, SVM has been successfully applied in cancer detection, Alzheimer's disease diagnosis, and cardiovascular disease prediction [5]. Its ability to handle non-linear boundaries using kernel functions enhances prediction performance in complex biomedical datasets.

## IV. CHALLENGES OF RECOMMENDATION SYSTEMS

This section highlights the major challenges encountered in disease prediction systems and discusses potential strategies to address them [17].
Cold Start Problem In disease prediction, the cold start issue arises when there is insufficient patient data to make accurate predictions. This typically occurs with new patients who have limited or no medical history in the database, making it difficult for the system to generate reliable outcomes. Integrating demographic data, lifestyle factors, and genetic information can help mitigate this challenge.

Data Manipulation and Bias Similar to shilling attacks in recommendation systems, disease prediction models can be compromised by biased or manipulated data. For instance, inaccurate patient records or inconsistent reporting may distort model predictions. Implementing robust data validation mechanisms and anomaly detection techniques is essential to preserve reliability and accuracy.

Synonymy and Data Standardization Problem Medical data often contains synonymous terms or varied representations of the same condition, such as "myocardial infarction" and "heart attack." If these terms are not standardized, the system may treat them as separate entities, reducing predictive performance. Solutions include medical ontology integration, term normalization, and Natural Language Processing (NLP)–based standardization methods.

Latency Problem Latency occurs when new patient records, lab results, or medical events are not immediately incorporated into the prediction system. This delay can hinder real-time diagnosis or timely disease risk assessment. Employing incremental learning algorithms and streaming data processing can reduce latency and improve responsiveness.

Data Sparsity Problem Many healthcare datasets suffer from sparsity because not all patients undergo the same tests or provide complete health information. Sparse datasets weaken the system's ability to detect disease patterns. Techniques such as imputation, transfer learning, and model-based approaches like matrix factorization can help overcome sparsity issues.

Grey Patient Problem The "grey patient" challenge refers to individuals whose medical profiles do not fit into typical disease categories, making it difficult for predictive models to generate accurate results. Personalized modeling and hybrid approaches that combine content-based patient data (e.g., genetic, clinical, lifestyle) with population-level patterns can alleviate this issue.

Scalability Problem As the amount of patient data grows with the expansion of electronic health records, wearable devices, and IoT-enabled medical sensors, scalability becomes a critical concern. Processing massive datasets efficiently requires

dimensionality reduction, distributed computing, and clustering methods to maintain prediction speed and accuracy.

## V. EVALUATION PARAMETER

To evaluate the performance of disease prediction models, several standard metrics are used, including accuracy, precision, recall, and F-score. These measures provide a comprehensive assessment of how effectively the system identifies and classifies diseases. The calculated values are substituted into their respective formulas to obtain performance results.

In above a true positive (TP) occurs when the system correctly predicts that a patient has a disease, and the patient is indeed diagnosed with that disease. A false positive (FP) arises when the system predicts the presence of a disease, but the patient is actually healthy or diagnosed with another condition. Similarly, true negatives (TN) and false negatives (FN) are used to evaluate correct and incorrect classifications of healthy cases, respectively. By balancing these measures, precision, recall, and F-score provide deeper insights into the reliability of disease prediction systems beyond simple accuracy.

## VI. CONCLUSION

This paper has summarize various models proposed for healthcare data analysis. Out of those it was found that data collection is major issue as many of patient do not share its data. Further feature extraction from unorganized data is another issue. To resolve all these scholars makes a repository and find the pattern. This paper has finds some of major machine learning models that were used for learning and prediction. It was found that image based data analysis was improved by the frequency features. In future scholars can develop models that predict different class of healthcare data accurately.

## REFERENCES

1. J. Gope and S. K. Jain, "A survey on solving cold start problem in recommender systems", Proc. Int. Conf. Comput. Commun. Autom. (ICCCA), pp. 133-138, May 2017.
2. I.Lauriola, A. Lavelli and F. Aiolli, "An introduction to deep learning in natural language processing: Models techniques and tools", Neurocomputing, vol. 470, pp. 443-456, Jan. 2022.
3. O. S. Shalom, H. Roitman and P. Kouki, "Natural language processing for recommender systems" in Recommender Systems Handbook, New York, NY, USA:Springer, pp. 447-483, 2021.
4. H. W. Wang, J. Wang, M. Zhao, J. N. Cao, M. Y. Guo. Joint topic-semantic-aware social recommendation for online voting. In Proceedings of ACM on Conference on Information and Knowledge Management, Singapore, pp.347–356, 2017.
5. X. N. He, L. Z. Liao, H. W. Zhang, L. Q. Nie, X. Hu, T. S. Chua. Neural collaborative filtering. In Proceedings of the 26th International Conference on World Wide Web, ACM, Perth, Australia, pp.173–182, 2017.
6. W. Q. Ma, X. C. Chen, W. K. Pan, Z. Ming. VAE++: Variational autoencoder for heterogeneous one-class collaborative filtering. In Proceedings of the 15th ACM International Conference on Web Search and Data Mining, Tempe, USA, pp.666–674, 2022.
7. J. Zhang, D. Chen and M. Lu, "Combining Sentiment Analysis With a Fuzzy Kano Model for Product Aspect Preference Recommendation," in IEEE Access, vol. 6, pp. 59163-59172, 2018.
8. S. Ojo et al., "Graph Neural Network for Smartphone Recommendation System: A Sentiment Analysis Approach for Smartphone Rating," in IEEE Access, vol. 11, pp. 140451-140463, 2023.
9. I G. Cossatin, N. Mauro, G. Izzi and L. Ardissono, "Synchronized multi-list user interfaces for fashion catalogs", Proc. 31st ACM Conf. User Model. Adaptation Personalization, pp. 224-228, Jun. 2023.
10. A. D. Starke, E. Asotic, C. Trattner and E. J. van Loo, "Examining the user evaluation of multi-list recommender interfaces in the context of healthy recipe choices", ACM Trans. Recommender Syst., vol. 1, no. 4, pp. 1-31, Dec. 2023.

11. El Majjodi, A. D. Starke and C. Trattner, "Nudging towards health? Examining the merits of nutrition labels and personalization in a recipe recommender system", Proc. 30th ACM Conf. User Model. Adaptation Personalization, pp. 48-56, Jul. 2022.
12. Chong Tat Chua, Hady W. Lauw, and Ee-Peng Lim. "Generative Models for Item Adoptions Using Social Correlation". IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 9, SEPTEMBER 2013.
13. Bellini, P., Palesi, L.A.I., Nesi, P. et al. Multi Clustering Recommendation System for Fashion Retail. Multimed Tools Appl 82, 9989–10016 (2023).
14. R. Sharda and D. Delen, "Predicting box-office success of motion pictures with neural networks," Expert Systems with Applications, vol. 30, no. 2, pp. 243-254, Feb. 2006.
15. W. Zhang and S. Skiena, "Improving Movie Gross Prediction through News Analysis," in 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, 2009, vol. 30, no. 2, pp. 301-304.
16. G. Szabo and B. a. Huberman, "Predicting the popularity of online content," Communications of the ACM, vol. 53, no. 8, p. 80, Aug. 2010.
17. Roy, D., Dutta, M. A systematic review and research perspective on recommender systems. J Big Data 9, 59 (2022).

1. Latha MH, Ramakrishna A, Reddy BSC, Venkateswarlu C, Saraswathi SY (2022) Disease prediction by stacking algorithms over big data from healthcare communities. Intell Manuf Energy Sustain: Proc ICIMES 2021(265):355.
2. Serbun Ufuk Değer. "A study on heart data analysis and prediction using advanced machine learning methods", Computers in Biology and Medicine, Volume 192, Part B, 2025.
3. Mobeen, A., Shafiq, M., Aziz, M.H. and Mohsin, M.J., "Impact of workflow interruptions on baseline activities of the doctors working in the emergency department", BMJ Open Quality, Vol. 11, No. 3, (2022), e001813.
4. Ahmed, S., Szabo, S. and Nilsen, K., "Catastrophic healthcare expenditure and impoverishment in tropical deltas: Evidence from the mekong delta region", International Journal for Equity in Health, Vol. 17, No. 1, (2018), 1-13.
5. Roberts, M.A. and Abery, B.H., "A person-centered approach to home and community-based services outcome measurement", Frontiers in rehabilitation Sciences, Vol. 4, (2023).
6. Farooqui, M. and Ahmad, D., "Disease prediction system using support vector machine and multilinear regression", International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN, (2020), 2347-5552.

7. T.P. Anish, P.M. Joe Prathap Hsc-attention multi-disease prediction net: a heuristic-assisted hybrid serial cascaded attention-based network with ensemble feature selection process for multi-disease prediction Biomed. Signal Process. Control, 89 (2024).
8. G. Kalpana, N. Deepa, D. Dhinakaran Advanced image preprocessing and context-aware spatial decomposition for enhanced breast cancer segmentation MethodsX., 14 (2025).
9. N. Rane, S. Choudhary, J. Rane Explainable artificial intelligence (XAI) in healthcare: interpretable models for clinical decision support SSRN (2023).
10. Xiao, X., Wang, Z., Kong, Y. & Lu, H. Deep learning-based morphological feature analysis and the prognostic association study in colon adenocarcinoma histopathological images. Front. Oncol. 13, 1081529 (2023).
11. Hu, G., Zheng, Y., Houssein, E. H. & Wei, G. DRPSO: A multi-strategy fusion particle swarm optimization algorithm with a replacement mechanisms for colon cancer pathology image segmentation. Comput. Biol. Med. 178, 108780 (2024).
12. Mohamed, A. A. A., Hançerlioğullari, A., Rahebi, J., Rezaeizadeh, R. & Lopez-Guede, J. M. Colon cancer disease diagnosis based on convolutional neural network and fishier mantis optimizer. Diagnostics 14(13), 1417 (2024).
13. Smith, J., Brown, A., & Lee, K. (2019). Logistic regression models for cardiovascular risk prediction: A clinical evaluation. Journal of Medical Informatics, 45(3), 210–219.
14. Patel, R., Sharma, D., & Mehta, P. (2020). Predictive analysis of diabetes using regression-

based models. International Journal of Health Data Science, 12(2), 87–95.
15. Khan, M., Ali, S., & Hussain, T. (2021). Naïve Bayes classification approach for liver disease diagnosis. Computational Medicine and Public Health, 18(4), 145–154.
16. Li, X., Wang, Y., & Chen, Z. (2020). Application of KNN algorithm for chronic disease diagnosis in clinical datasets. BMC Medical Informatics and Decision Making, 20(5), 233–242.
17. Zhang, H., Liu, J., & Wu, Q. (2022). Decision tree and support vector machine models for disease prediction: A comparative study. IEEE Access, 10, 55632–55641.
18. Roy, D., Dutta, M. A systematic review and research perspective on recommender systems. J Big Data 9, 59 (2022).