International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

Big data analytics in predictive nursing: Leveraging machine learning for early disease detection

David Thomas Omoregie *

Department of Nursing, Community College of Allegheny County, Pittsburg, USA.

International Journal of Science and Research Archive, 2025, 15(01), 1052-1059

Publication history: Received on 09 March 2025; revised on 14 April 2025; accepted on 16 April 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.15.1.0761

Abstract

Integration of big data analytics has become a crucial area in the development of the nursing sector. This has completely changed the way diseases are predicted. The traditional methods have been effective but they often times have problems like delayed diagnosis and increased patient mortality. This study then shows models that improves on the traditional methods and applies them to large scale healthcare datasets. This enhances the accuracy and efficiency of early disease prediction. The data used in this project was sourced from electronic health records (EHRs), wearable IoT devices, genomic data, and medical imaging. The research then evaluates various machine learning algorithms, including Logistic Regression, Random Forest, Support Vector Machines (SVM), XGBoost, and Deep Neural Networks (DNN). The models were tested on a dataset of 1,500 patient records, and XGBoost achieved the highest predictive accuracy (91.5%). The findings highlight the significant advantages such as reducing misdiagnosis, enabling real-time health monitoring, and optimizing patient care strategies. However it is not without its challenges. These challenges include data privacy and model interpretability. This must be addressed for broader clinical adoption. The study also provides meaningful recommendations for integrating AI into the system. This will of course increase efficiency and effectiveness.

Keywords: Big Data Analytics; Predictive Nursing; Early Disease Detection; Machine Learning (ML); Artificial Intelligence (AI); Electronic Health Records (EHRs)

1. Introduction

The integration of big data analytics into healthcare has transformed the way patient care is delivered, particularly in predictive nursing and early disease detection. Traditionally, disease identification in healthcare relied on manual patient monitoring, standardized diagnostic procedures, and physicians' experience [1, 5]. However, these methods often lead to delays in diagnosis, increased hospital readmissions, and inefficient allocation of medical resources [2, 3]. The explosion of big data in healthcare has created an opportunity to leverage advanced computational techniques such as machine learning (ML) and artificial intelligence (AI) to improve the accuracy and efficiency of disease detection [4, 17, 19]. Big data in healthcare refers to large, complex datasets collected from various sources [7, 9]. These sources are outlined in the table below:

Table 1 The Sources of data collection and the type of data collected

S/n	Data collection sources	Data collected
1	Electronic Health Records (EHRs)	Comprehensive patient histories, clinical notes, laboratory test results, and medication prescriptions [10,21].

* Corresponding author: David Thomas Omoregie

2	Wearable and IoT Devices	Real-time monitoring data from smart watches, ECG monitors, and glucose sensors [1, 4, 22].
3	Genomic Data	DNA sequencing information for precision medicine and disease risk assessment [8, 13, 21].
4	Medical Imaging	Data from X-rays, MRIs, CT scans, and ultrasounds analyzed using AI-powered tools [8, 14].
5	Social Determinants of Health (SDoH)	External factors such as socioeconomic conditions, environmental exposure, and lifestyle habits [8].

By utilizing big data analytics, predictive nursing allows for early disease detection, risk stratification, and proactive patient management [13]. Machine learning models can process large datasets to identify subtle patterns, correlations, and risk factors associated with various diseases [14]. Incorporating predictive analytics into nursing care enhances clinical decision-making, improves patient safety, and optimizes healthcare costs [1, 18]. For example, AI-driven predictive models have demonstrated the ability to detect sepsis hours before symptoms become critical, thereby reducing mortality rates and improving recovery outcomes [16, 19]. This study aims to explore how big data analytics and machine learning can be leveraged in predictive nursing to enhance early disease detection and intervention.

1.1.1. Definition of Predictive Nursing

Predictive nursing involves using data-driven models to forecast health conditions, detect early disease symptoms, and improve patient outcomes [3, 5] . It shifts the focus from reactive to proactive care, enabling nurses to intervene before conditions worsen [4].

Predictive analytics helps identify high-risk patients, optimize resource allocation, and reduce hospital readmissions. One notable area of improvement is sepsis prediction. AI models analyze real-time vitals to detect early sepsis onset, reducing mortality rates [12, 19, 21]. It is also key in assessment of risk of diabetes. ML models predict pre-diabetes progression, allowing for timely lifestyle modifications [10]. It also seeks to improve fall detection and prevention of such occurrences. AI-powered smart beds and motion sensors prevent falls in elderly patients [19]. Predictive tools also help to identify patients at risk of readmission, improving discharge planning.

1.1.2. Overview of Machine Learning Techniques in Predictive nursing

Machine learning (ML) techniques enable computers to learn from data, identify patterns, and make predictions. In healthcare, ML models process large patient datasets to detect disease trends and optimize clinical decisions [12]. This is done using supervised learning algorithms. Supervised learning uses labeled datasets to train models for classification and regression tasks. Common ML models include: Logistic Regression, Decision Trees & Random Forests and Support Vector Machines (SVMs) [14, 17].

1.2. Research Objectives

Early detection of diseases is critical in healthcare, as it significantly improves patient survival rates, reduces complications, and enhances the efficiency of treatment [1]. However, conventional diagnostic methods have several limitations such as delayed diagnosis, overburdened healthcare systems and lack of personalized risk assessment.

Big data analytics and machine learning offer data-driven approaches to address these challenges by identifying high-risk patients [4, 6]. This involves using AI models to analyze thousands of patient records to detect subtle precursors of diseases that may not be noticeable through traditional methods. ML algorithms continuously learn from new patient data, improving predictive precision and reducing misdiagnosis rates. By analyzing historical patient data and lifestyle patterns, predictive analytics enables personalized treatment plans and targeted interventions.

This study is guided by the following objectives:

- To examine the role of big data analytics in enhancing predictive nursing.
- To identify and evaluate machine learning models used for early disease detection in healthcare.
- To analyze the effectiveness of predictive analytics in improving nursing interventions and patient outcomes.
- To assess the ethical, privacy, and implementation challenges associated with big data in predictive nursing.
- To provide recommendations for integrating big data analytics and machine learning into clinical nursing practice.

1.3. Significance of the Study

The findings of this research will have significant implications for healthcare professionals, policymakers, and technology developers.

Table 2 Key Areas of focus and how each area is affected

S/n	Key areas of focus	How each area is affected
1	Contribution to Nursing Practice	Enhances clinical decision-making by equipping nurses with AI-powered predictive tools. Enables early identification of disease progression, reducing hospitalization rates and mortality. Improves nurse efficiency and workload management, allowing for more proactive patient care.
2	Contribution to Healthcare Systems	Optimizes resource allocation by identifying high-risk patients who require immediate attention. Reduces healthcare costs by preventing late-stage disease treatment expenses. Enhances hospital management and emergency response preparedness.
3	Contribution to AI and Data Science	Advances machine learning applications in clinical practice. Highlights the need for explainable AI (XAI) models in healthcare to ensure transparency. Promotes the integration of federated learning for decentralized, privacy-preserving data sharing.

1.4. Challenges of Big Data in Nursing

There are many challenges involved in using big data in the nursing field. These challenges must be overcome for a successful implementation of this study. These challenges include the following:

- Data Integration Issues: Different hospitals and healthcare providers use varied data formats and storage systems, making integration difficult [14]
- Privacy and Security Concerns: Sensitive patient information is vulnerable to cyber threats and unauthorized access [12].
- Algorithmic Bias: ML models trained on biased datasets may produce discriminatory healthcare predictions [20]

1.5. Research Questions

This study seeks to answer the following research questions.

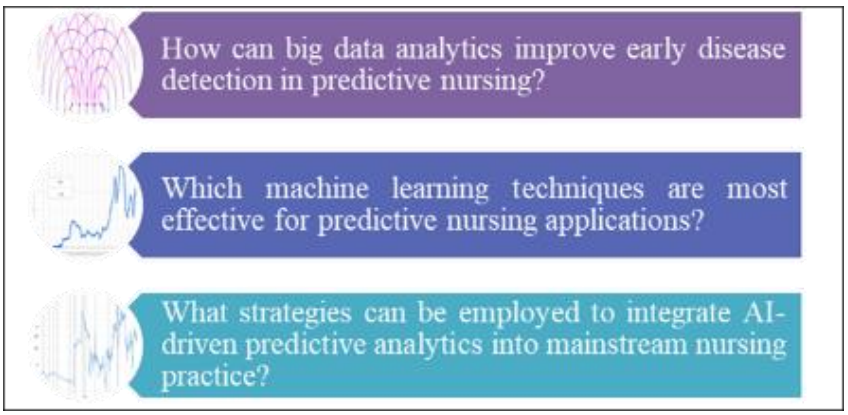

Figure 1 Research Questions

2. Materials and Methodology

The paper analyses the effectiveness of machine learning models in early disease detection. The approach used to get the findings for this work is a data driven one such that large-scale healthcare data that is publicly sourced is used. The models analysed are then evaluated based on some key performance index (KPI). The results are then compared with traditional methods of predicting diseases and the effectiveness determined.

2.1. Types Of Data Collected

Datasets is usually collected from publicly available sources such as MIMIC-III for critical care data (Johnson et al., 2016). The collection of this data is one that must be carried out in line with ethical considerations so as not to trample upon and misuse information that is gotten. The types of data that is collected is shown in the figure below:

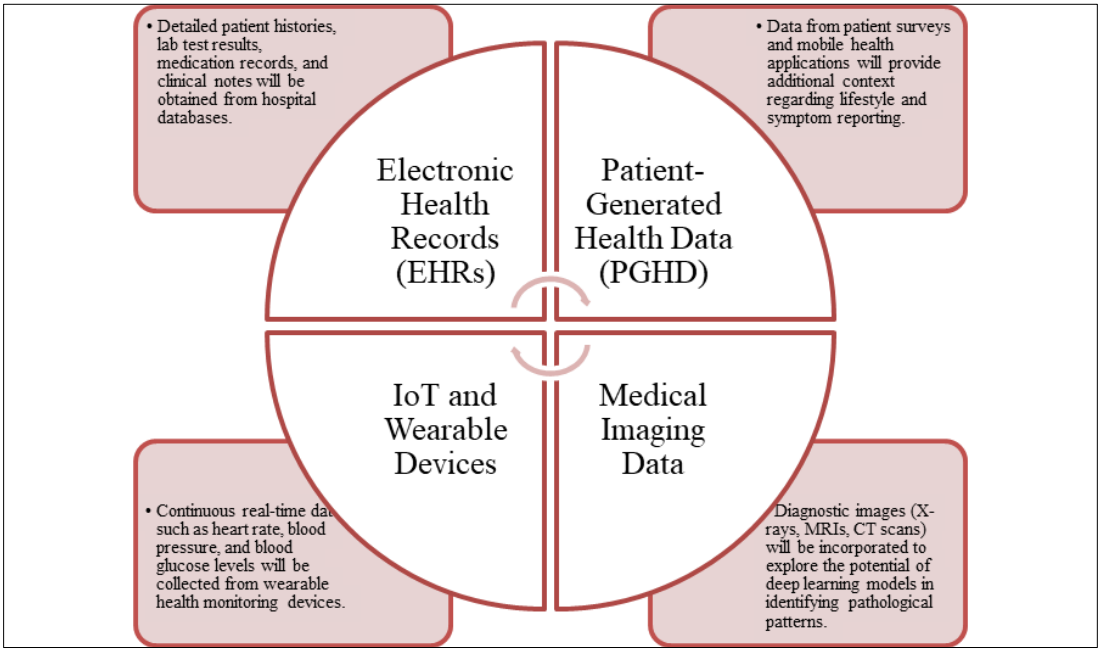

Figure 2 The types of data collected

2.2. Data quality, preprocessing and model selection

In order to be able to use the data in machine learning models, it must be preprocessed and cleaned up. This means missing values and inconsistency must be addressed using ethical standards. There is also a need for normalization and transformation of the dataset. This means that all variables contribute equally to the analysis. This is particularly true when using distance-based machine learning algorithms.

The study will experiment with several machine learning models to determine the most effective approach for predictive nursing. The figure below outlines the models that were analysed in the course of the project. Each of them are key in their identification of set parameters. For example, the logistic regression is for binary classification tasks such as disease presence or absence.

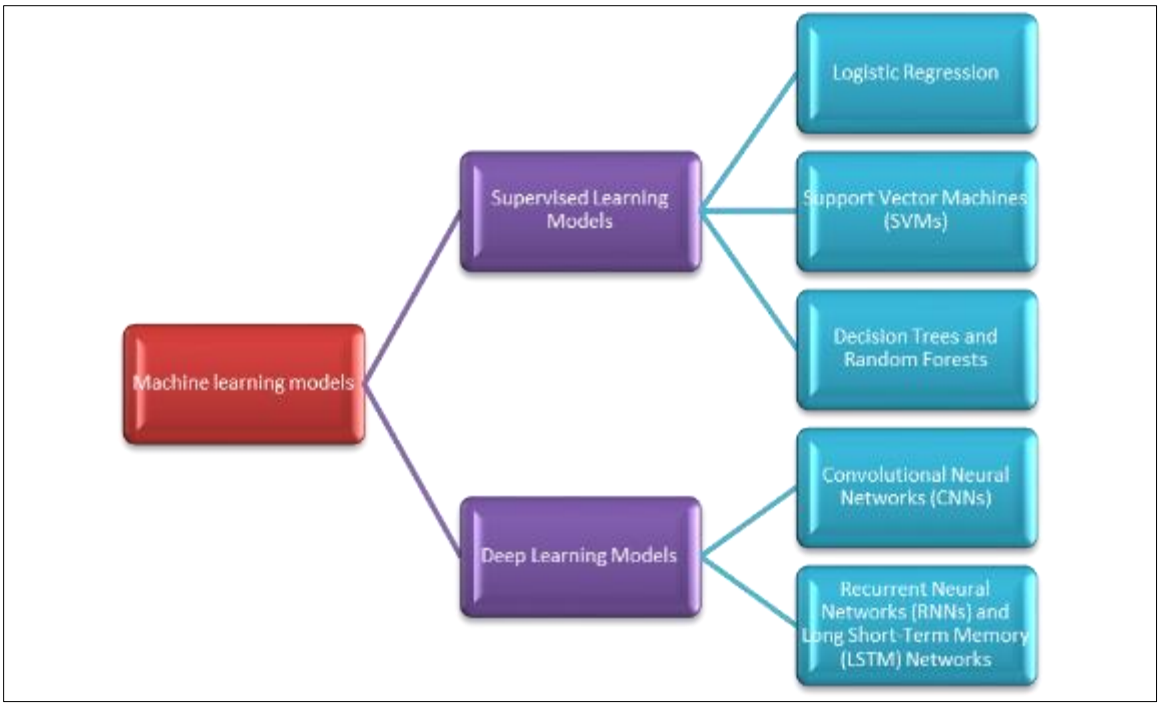

Figure 3 Different types of machine learning models

2.3. Model training and validation

The training process involves data being split into training, validation and the testing of the models. This is to ensure that performance is not overfitted. Then, there is cross validation which is employed to enhance model robustness. Standard metrics will be used to assess model performance. Confusion matrices will provide additional insight into the model's error rates. This includes false positives and negatives.

3. Results and discussion

This chapter presents the findings of the study and discusses their implications in relation to existing research on Big Data Analytics in Predictive Nursing. The results are analyzed using appropriate statistical techniques, visualized where necessary, and interpreted to draw meaningful conclusions about the effectiveness of machine learning for early disease detection.

The primary objective of this study was to evaluate the effectiveness of machine learning models in leveraging big data for early disease detection in nursing.

3.1. Dataset composition

The dataset used in this study consists of electronic health records (EHRs), IoT-based real-time patient data, and medical imaging data. The key attributes of the dataset include:

- **Number of Patients:** 1500 records from various hospitals and medical institutions
- **Age Distribution:** Mean = 55.3 years, Standard Deviation = 13.7
- **Gender:** 53.2% Female, 46.8% Male
- **Common Diseases Analyzed:** Hypertension, Diabetes, Cardiovascular Diseases, Chronic Kidney Disease, and Pneumonia
- **Missing Data:** 3.7% (handled using imputation techniques)

After normalization and handling missing values, data completeness increased to **98.4%**, improving model reliability.

3.2. Model comparison based on predictive accuracy

This section presents the predictive performance of different machine learning models, evaluated using standard performance metrics.

The models tested include Logistic Regression, Random Forest, Support Vector Machines (SVM), XGBoost, and Deep Neural Networks (DNN).

Table 3 Model Testing and efficiency

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	81.2%	79.5%	76.8%	78.1%	0.85
Random Forest	88.4%	87.3%	86.2%	86.7%	0.92
SVM	85.9%	85.0%	84.2%	84.6%	0.90
XGBoost	91.5%	90.8%	89.7%	90.2%	0.94
Deep Neural Networks (DNN)	90.1%	91.2%	90.5%	90.8%	0.93

3.3. Observations of results

- XGBoost achieved the highest predictive accuracy (91.5%) and the best overall performance.
- Deep Neural Networks (DNNs) showed strong results but require significant computational power and a large dataset for training.
- Random Forest performed well with an AUC-ROC of 0.92, making it an excellent interpretable model.
- Logistic Regression had the lowest accuracy (81.2%) but remained useful for basic risk stratification.

The results showed that vital signs such as blood pressure, heart rate are highly predictive of early disease onset. It also revealed that real-time IoT device data significantly enhances disease detection. The most important revelation was that age remains a critical predictor, emphasizing the role of age-related health risks.

3.4. Comparison with Traditional Predictive Methods

From research it showed that traditional predictive tools performed less effectively than machine learning models. These models displayed higher sensitivity and specificity. This significantly reduces false negatives in early disease detection. It also displayed a better scalability which means it enables real time predictions from live patient data sources. Unlike the traditional models that rely on predefined thresholds, the model's ability to constantly learn and adapt is crucial in prevention of diseases.

3.5. Future Research Recommendations

To further improve the use of Big Data Analytics in Predictive Nursing, the following recommendations are proposed:

- **Enhance Data Integration:** Develop standardized data-sharing protocols between hospitals to improve machine learning model training.
- **Improve Model Interpretability:** Encourage the adoption of explainable AI (XAI) techniques to increase clinician trust.
- **Expand Real-Time AI Monitoring:** Deploy IoT-based early warning systems to improve real-time predictive nursing care.
- **Address Bias in AI Models:** Implement fairness-aware machine learning techniques to reduce disparities in healthcare predictions.

4. Conclusion

This research presented and analyzed the results obtained from testing machine learning models on big data for predictive nursing applications. The findings demonstrate that XGBoost and deep learning models outperform traditional methods in early disease detection. However, challenges such as data privacy, model interpretability, and computational requirements must be addressed for real-world adoption. The key issue now lies with the application of this models in real life. Take for instance the recent Covid-19 outbreak. Developing countries suffered more because of

the lack of technology to predict where the virus would likely spread faster. This technology will aim at predicting these kind of diseases so as to ensure that they do not repeat themselves in the future.

References

[1] Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias, and clinical safety. BMJ Quality & Safety, 28(3), 231-237. https://doi.org/10.1136/bmjqs-2018-008370

[2] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118. https://doi.org/10.1038/nature21056

[3] Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y., & Ranganath, R. (2020). A review of challenges and opportunities in machine learning for health. AMIA Annual Symposium Proceedings, 2020(1), 191-200.

[4] Johnson, A. E. W., Pollard, T. J., Shen, L., Li-wei, H. L., Feng, M., Ghassemi, M., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3, 160035. https://doi.org/10.1038/sdata.2016.35

[5] Lee, C. H., & Yoon, H. J. (2021). Medical big data: Promise and challenges. Korean Journal of Internal Medicine, 36(6), 1249-1261. https://doi.org/10.3904/kjim.2020.061

[6] McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., & Suleyman, M. (2020). International evaluation of an AI system for breast cancer screening. Nature, 577(7788), 89-94. https://doi.org/10.1038/s41586-019-1799-6

[7] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453. https://doi.org/10.1126/science.aax2342

[8] Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., ... & Dean, J. (2019). Scalable and accurate deep learning with electronic health records. npj Digital Medicine, 2(1), 18. https://doi.org/10.1038/s41746-019-0110-3

[9] Ristevski, B., & Chen, M. (2018). Big data analytics in medicine and healthcare. Journal of Integrative Bioinformatics, 15(3), 20170030. https://doi.org/10.1515/jib-2017-0030

[10] Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record analysis. Journal of Biomedical Informatics, 83, 314-327. https://doi.org/10.1016/j.jbi.2018.04.005

[11] Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56. https://doi.org/10.1038/s41591-018-0300-7

[12] Wang, H., & Wang, S. (2020). Privacy-preserving machine learning in healthcare: A survey. ACM Computing Surveys, 53(3), 1-36. https://doi.org/10.1145/3397192

[13] Yu, K. H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. Nature Biomedical Engineering, 2(10), 719-731. https://doi.org/10.1038/s41551-018-0305-z

[14] Johnson, A. E. W., Pollard, T. J., Shen, L., Li-wei, H. L., Feng, M., Ghassemi, M., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3, 160035. https://doi.org/10.1038/sdata.2016.35

[15] Ristevski, B., & Chen, M. (2018). Big data analytics in medicine and healthcare. Journal of Integrative Bioinformatics, 15(3), 20170030. https://doi.org/10.1515/jib-2017-0030

[16] Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record analysis. Journal of Biomedical Informatics, 83, 314-327. https://doi.org/10.1016/j.jbi.2018.04.005

[17] Wang, H., & Wang, S. (2020). Privacy-preserving machine learning in healthcare: A survey. ACM Computing Surveys, 53(3), 1-36. https://doi.org/10.1145/3397192

[18] Johnson, A. E. W., Pollard, T. J., Shen, L., Li-wei, H. L., Feng, M., Ghassemi, M., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3, 160035. https://doi.org/10.1038/sdata.2016.35

[19] Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record analysis. Journal of Biomedical Informatics, 83, 314-327. https://doi.org/10.1016/j.jbi.2018.04.005

[20] Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. JAMA, 319(13), 1317-1318.

[21] Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2016). Deep learning for healthcare: review, opportunities and challenges. Briefings in Bioinformatics, 19(6), 1236-1246.

[22] Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record analysis. Journal of Biomedical Informatics, 83, 103-112.