International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Machine Learning Algorithms for Predictive Quality Assurance in Healthcare in DW ETL Processes

Arun Kumar Ramachandran Sumangala Devi

Architect II - Software Testing, UST Global Inc.
www.linkedin. com/inarunkumarramachandransumangaladevi

Abstract: ***The healthcare sector constantly contains significant amounts of data from various sources, including patient records and clinical trials. Since the early 1980s, data warehousing and Extract, Transform, and Load (ETL) frameworks have been constantly utilized in predictive analytics to make treatment options and personalized patient care decisions. However, the challenge of the continuously increasing amount of data being generated on a daily basis has called for more efficient techniques to analyze data. The focus of this article is to analyze how the advent of machine learning (ML) has revolutionized predictive analytics by developing models and algorithms that allow for the automation of data warehousing and ETL processes. The various ML applications in predictive analytics, from predicting disease occurrence to predicting surgery outcomes, are further discussed. The benefits of ML include the potential to personalize patient care, accelerate drug development, and identify at - risk patients, among others. ML has the potential to revolutionize predictive analytics in the healthcare sector due to the constant evolution and innovation in the technological world.***

Keywords: Machine learning, healthcare, predictive analytics, ETL, data warehousing

1. Introduction

The healthcare industry has seen significant changes over time due to the growing need for data - informed decision - making and the increasing volume of patient data. As healthcare sectors strive to gain significant insights from this data, data consistency, integrity, and dependability become very critical. However, deriving insights from these data has regularly proved challenging because different data sources—such as electronic health records (EHRs), laboratory information systems, and billing software—complicate the integration process. According to Yangui et al. (2017), the several formats, standards, and nomenclature of every system impede the analysis of patient data. Dhaouadi et al. (2022) inferred that the large amounts of
data generated daily often go beyond the capacity of traditional data management solutions. The healthcare sector constantly looks for innovative approaches to address these challenges. The two strategies that have been widely applied to address these challenges are ETL (Extract, Transform, Load) operations and data warehouses for data storage and analysis.

A data warehouse (DW) is an organized form of collection of data entered into a format that makes it easy to analyze and draw insights. DW's most often used definition is “a subject - oriented, integrated, nonvolatile, and time - variant collection of data in support of management's decisions, ” as Inmon (2005) stated. This data is entered into a data warehouse using a three - phase approach—extract, transform, and load—ETL. Data is extracted from several sources and transformed by this procedure through preparation, conversion, cleaning, filtering, joining, aggregation, and loading into a DW (Dhaouadi et al., 2022). Data extraction from several sources and its conversion into a suitable format to support processing during the transformation process constitute the extracting phase. Extensive data extractions tailored to the database used for handling the applications for decision - making constitute the transformation phase. Activities at this level consist of joining, converting, aggregation, and filtering. The data is then loaded into a data warehouse, which can then be used to draw significant insights.

While data warehousing and ETL processes have been effective in predictive analytics, the challenge is the continuous increase in the amount of data collected daily. Conventional data warehousing and ETL frameworks often call for human input, but this becomes inefficient when there is a need to analyze significantly large amounts of data. ML, therefore, emerges as a solution to this challenge, allowing for the automation of data analysis approaches and, consequently, the simplification of predictive analytics.

2. Solution

Machine learning, a type of artificial intelligence, has emerged as an important tool for predictive analytics. It entails creating algorithms that allow algorithms to learn from vast volumes of patient data, from demographics and drug history to diagnostic tests and treatment outcomes, enabling systems to produce predictions. machine learning algorithms can generate more accurate prediction models than traditional approaches for patient outcome projection. Furthermore, ML systems are designed to be constantly learning and flexible enough to adapt to new data, allowing their evolution over time.

The process of predictive analytics using machine learning involves data collection from various sources, including test results, patient histories, and electronic health records. This data is then cleaned to eliminate any noise and ensure that only quality data is kept (Sakib et al., 2022). Based on the three phases of the ETL framework, the data is then transferred to a data warehouse where the relationships between variables being tested are determined. For instance, the risk of patients developing cancer and their family history is compared. Based on the results of this analysis, a predictive model is developed, and its efficiency is tested using various

data sources. Once a reliable model is created, it is included in everyday operations to assist in decision - making and provide a data - driven analysis of efficient patient care. By using machine learning (ML), data quality is improved, data analysis is automated, and efficiency is maximized. The various ways in which ML is applicable in data warehousing and ETL processes include automation of analysis of structured and unstructured data formats in the data extraction phase (Dahiwade et al., 2019). ML enables data cleansing, error detection, repair, pattern, and relationship recognition during data transformation. ML models continuously change the processes to optimum loading times and methodologies depending on system demands, therefore enhancing data loading into DW.

One of the cases of effective ML application in predictive analytics is a platform utilized by the University of California (UC) San Diego Health System—which incorporates predictive analytics algorithms into daily healthcare operations. The PARAllel predictive MOdeling (PARAMO) platform also examines electronic health record (EHR) data for early disease diagnosis, including sepsis. This platform maximizes patient prediction for at - risk patients using ML models (Grampurohit & Sagarnal., 2020; Ng et al., 2014). Another example is Biome Diagnostics GmbH (Biome DX®) which is transforming cancer treatment by creating microbiome - based precision medical solutions utilizing machine learning (Al - Tashi et al., 2023; Cammarota et al., 2020). Biome DX investigates the link between microbiomes and a patient's medical history and genetic composition using ML models to determine whether cancer treatment would help a patient.

3. Applications of the solution

Machine learning has been applied in healthcare in various ways:

a) Disease occurrence prediction

One of the most important applications of machine learning in predictive analysis is the prediction of disease. Machine learning algorithms are designed to evaluate patient data, including genetic information, medical history, and family history, and produce predictions on specific high - risk illnesses. For instance, it can be quite challenging to identify an ailment like diabetes mellitus early on and to diagnose it correctly. Many techniques have been used to predict and diagnose diabetes in people, even in its early stages: artificial neural networks (ANN), decision trees (DT), logistic regression, K - nearest neighbors (KNN), random forest (RF), and extreme gradient boosting (XGB) (Mujumdar & Vaidehi., 2019; Soni & Varma., 2020; Zou et al., 2018). Additionally, ML algorithms have been used to track the transmission of infectious diseases and predict disease outbreaks. In order to enhance the prediction of communicable diseases like meningitis, hepatitis B, and malaria in laboratory tests, Park et al. (2021) developed machine learning models. The J48 decision tree, RF, k - NN, MLP, NB, XGBoost, and LR were used by Moulaei et al. (2023) to predict the severity of COVID - 19 from heterogeneous data. It has been demonstrated that every model accurately predicts these illnesses.

b) Recovery rate prediction and predicting complications

ML algorithms have been used to predict patients' recovery rates and whether or not they would develop complications after treatment. For example, ML prediction models have been employed in the ICU to predict recovery rates and renal function reversibility in patients suffering from acute kidney injury (AKI). There have been promising outcomes with the application of machine learning prediction models to improve the prognosis of AKI patients and assist health practitioners in developing timely therapies. ML models have also been employed in the prognosis of pediatric Traumatic Brain Injuries (TBI), which can subsequently be used to decide the best course of treatment.

c) Prediction of immunotherapy effectiveness from histopathology

Although ML - driven digital pathology can accurately detect cancer cells, measure the results of immunohistochemical labeling, and interpret challenging images, standard histopathological procedures are labor - intensive and often insufficient for precision medicine. Tumor - infiltrating lymphocytes (TIL), microsatellite instability, and PD - L1 expression are significant indicators that aid in the prediction of immunotherapy responses. Research indicates that machine learning (ML) models beat pathologists in assessing PD - L1 scores and can predict those who are likely to benefit from immune checkpoint medications with accuracy (Davenport & Kalakota, 2019; Safa et al., 2023). Machine learning (ML) techniques are demonstrated in Figure 1 below for the purpose of improving immunotherapy patient categorization and predicting immune responses from histology pictures. In order to predict how the immune system would react to immunotherapy, machine learning models employ graph neural networks to analyze radiomic images, pathology images, genetic information, epigenetic information, microbiological data, hematological values, proteomics data, and multi - omics data.

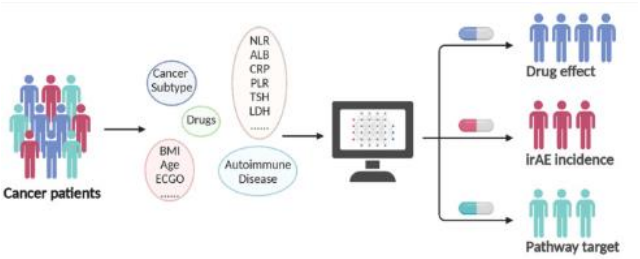

Figure 1: ML model to predict cancer immunotherapy adverse effects

d) Prediction of surgery outcomes

ML models are also revolutionizing predictive outcomes in surgery. In hepatobiliary and colorectal surgery, researchers predicted complications in colorectal, hepatic, and pancreatic operations using a model trained on 15, 657 patient records from the National Surgical Quality Improvement Program (NSQIP). The model outperformed the American College of Surgeons Surgical Risk Calculator (ACS - SRC), attaining an AUC of 0.76 for predicting surgical site infections and 0.98 for stroke predictions (Dixon et al., 2024; Ehlers et al., 2017). In implant - based breast reconstruction, ML models have also been applied to determine complications (Bakas et al., 2018). Significantly improving over conventional logistic regression in identifying important risk factors, models trained on

perioperative data from 481 patients predicted periprosthetic infections with an AUC of 0.73 and device explanation with an AUC of 0.78 (Elfanagely et al., 2021). Predicting complications in brain tumor surgery using an extreme gradient boosting model produced an AUC of 0.74 with a 70% accuracy rate in a dataset comprising 668 cases (Kunze et al., 2022). With AUCs of 0.97 for infections, 0.91 for complications within 12 months, and 0.88 for needing a second surgery, machine learning also showed strong predictive accuracy for complications following deep brain stimulation surgery.

e) Risk Calculators

ML models have advanced the use of risk calculators to deliver individualized treatment to patients and determine risks associated with certain procedures. Two noteworthy ML models that have been developed are MySurgeryRisk and POTTER. With an area under the curve (AUC) ranging from 0.77 to 0.94, MySurgeryRisk—developed from data on 51, 457 patients—predicts eight major postoperative complications including acute renal damage, sepsis, and mortality within two years after a given procedure (Bihorac et al., 2019; Gordon et al., 2019; Hassan et al., 2023). Likewise, trained on data from 382, 960 patients in the National Surgical Quality Improvement Program (NSQIP), the POTTER risk calculator uses decision tree algorithms to estimate mortality and morbidity more precisely than conventional models, including the ASA score and Emergency Surgery Score (Maurer et al., 2023; Ribeiro Junior et al., 2023). These instruments improve the surgery options and preoperative planning.

Benefits of the Solution

a) Providing clinical decision support

ML models in predictive analytics drive clinical decision support, which uses risk scoring to stratify risks at both individual and group levels and detect risk variables inside patient populations improving decision - making. Predictive models—such as a blood test for HPV for throat cancer patients created by the University of Michigan Rogel Cancer Center—can evaluate therapy efficacy earlier than conventional imaging scans, enabling health practitioners to modify their treatment plans quickly (Aljohani., 2023). Predictive analytics helps enhance patient outcomes and save time.

b) Improving value - based healthcare delivery

Predictive modeling is used by health organizations such as MVP Health Care and Elevance Health to identify high - risk patients and link them with required services, therefore addressing care obstacles like social determinants of health (SDOH). Using predictive analytics, accountable care organizations (ACOs) such as Buena Vida y Salud concentrate on actionable information that guides care teams without overloading them, improving care management (Gordon et al., 2019). Using data and involving physicians helps healthcare companies maximize patient outcomes and improve treatment delivery

c) Population health management

ML models play a crucial part in population health management. Children of Alabama, for instance, forecast patient deterioration in their cardiovascular ICU using predictive modeling as part of the ICU Liberation project, which seeks to minimize post - ICU problems and thereby enhance treatment. Umpqua Health in Oregon also employs predictive analytics to find high - risk patients susceptible to climate - related events such as wildfires (Delahanty et al., 2019). It offers preventive support such as air purifiers, enhancing patient outcomes and care coordination.

d) Accelerated drug development

Machine learning significantly speeds up drug development by evaluating massive databases of chemical compounds, biological interactions, and patient data to identify potential novel drugs and predict their efficacy (Maurer et al., 2023). By simulating how molecules interact with biological targets, machine learning (ML) algorithms streamline the drug discovery process. This reduces the need for laborious lab trial - and - error research and speeds up the development of promising new drug candidates.

e) Improved prescription accuracy

Machine learning improves prescription accuracy through medical history analysis, including allergies, past treatments, and possible drug interactions. These tools guide decisions by alerting medical practitioners to errors such as incorrect dosages or dangerous drug combinations (Bakas et al., 2018). ML also uses genetic data to personalize medications for specific people, reducing human mistakes and the risk of side effects. From this, better patient outcomes and safer, more effective treatments follow.

f) Improved patient care

Machine learning enables customized patient care by analyzing individual data, including genetic information, past treatments, and medication responses, to identify patterns linked to specific outcomes. In oncology, for instance, ML can predict how patients with the same cancer type might respond to various treatments based on their unique profiles (Kunze et al., 2022). This allows doctors to create personalized treatment plans, enhancing therapy effectiveness while minimizing side effects, leading to more precise and effective care.

Conclusion

This paper has closely examined how predictive healthcare analytics utilizes machine learning in ETL systems and data warehouses. ML models have been used to forecast success rates and complications and to predict disease occurrences such as COVID - 19 and malaria, among other aspects made possible by machine learning. ML has allowed for automating processes such as data extraction, cleaning, transformation, and loading into a data warehouse. Benefits include enhanced patient care, accelerated medication development, patient at - risk identification, population health management, and individualized medicine. The constant development of machine learning algorithms promises great possibilities for creating more models to improve healthcare provision.

References

[1] Aljohani, A. (2023). Predictive analytics and machine learning for real - time supply chain risk mitigation and agility. Sustainability, 15 (20), 15088.
[2] Al - Tashi, Q., Saad, M. B., Muneer, A., Qureshi, R., Mirjalili, S., Sheshadri, A.,. . . & Wu, J. (2023). Machine learning models for the identification of prognostic and predictive cancer biomarkers: a systematic review. International journal of molecular sciences, 24 (9), 7781.
[3] Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A.,. . . & Jambawalikar, S. R. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv preprint arXiv: 1811.02629.
[4] Bihorac, A., Ozrazgat - Baslanti, T., Ebadi, A., Motaei, A., Madkour, M., Pardalos, P. M.,. . . & Momcilovic, P. (2019). MySurgeryRisk: development and validation of a machine - learning risk algorithm for major complications and death after surgery. Annals of surgery, 269 (4), 652 - 662.
[5] Cammarota, G., Ianiro, G., Ahern, A., Carbone, C., Temko, A., Claesson, M. J.,. . . & Tortora, G. (2020). Gut microbiome, big data and machine learning to promote precision medicine
[6] Cheng, K. Y., Pazmino, S., & Schreiweis, B. (2022). ETL Processes for Integrating Healthcare Data–Tools and Architecture Patterns. In pHealth 2022 (pp.151 - 156). IOS Press.
[7] Dahiwade, D., Patle, G., & Meshram, E. (2019, March). Designing disease prediction model using machine learning approach. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp.1211 - 1215). IEEE.
[8] Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. Future healthcare journal, 6 (2), 94 - 98.
[9] Delahanty, R. J., Alvarez, J., Flynn, L. M., Sherwin, R. L., & Jones, S. S. (2019). Development and evaluation of a machine learning model for the early identification of patients at risk for sepsis. Annals of emergency medicine, 73 (4), 334 - 344.
[10] Dhaouadi, A., Bousselmi, K., Gammoudi, M. M., Monnet, S., & Hammoudi, S. (2022). Data warehousing process modeling from classical approaches to new trends: Main features and comparisons. Data, 7 (8), 113.
[11] Dixon, D., Sattar, H., Moros, N., Kesireddy, S. R., Ahsan, H., Lakkimsetti, M.,. . . & Hassan, M. J. (2024). Unveiling the Influence of AI Predictive Analytics on Patient Outcomes: A Comprehensive Narrative Review. *Cureus*, *16* (5).
[12] Ehlers, A. P., Roy, S. B., Khor, S., Mandagani, P., Maria, M., Alfonso - Cristancho, R., & Flum, D. R. (2017). Improved risk prediction following surgery using machine learning algorithms. eGEMs, 5 (2).
[13] Elfanagely, O., Toyoda, Y., Othman, S., Mellia, J. A., Basta, M., Liu, T.,. . . & Fischer, J. P. (2021). Machine learning and surgical outcomes prediction: a systematic review. Journal of Surgical Research, 264, 346 - 361.
[14] Gordon, L., Austin, P., Rudzicz, F., & Grantcharov, T. (2019). MySurgeryRisk and machine learning: a promising start to real - time clinical decision support. Annals of Surgery, 269 (1), e14 - e15.
[15] Grampurohit, S., & Sagarnal, C. (2020, June). Disease prediction using machine learning algorithms. In 2020 international conference for emerging technology (INCET) (pp.1 - 7). IEEE.
[16] Hassan, A. M., Rajesh, A., Asaad, M., Nelson, J. A., Coert, J. H., Mehrara, B. J., & Butler, C. E. (2023). Artificial intelligence and machine learning in prediction of surgical complications: current state, applications, and implications. The American Surgeon, 89 (1), 25 - 30.
[17] Inmon, W. H. (2005). Building the data warehouse. John wiley & sons.
[18] Kunze, K. N., Krivicich, L. M., Clapp, I. M., Bodendorfer, B. M., Nwachukwu, B. U., Chahla, J., & Nho, S. J. (2022). Machine learning algorithms predict achievement of clinically significant outcomes after orthopaedic surgery: a systematic review. Arthroscopy: The Journal of Arthroscopic & Related Surgery, 38 (6), 2090 - 2105.
[19] Maurer, L. R., Chetlur, P., Zhuo, D., El Hechi, M., Velmahos, G. C., Dunn, J.,. . . & Kaafarani, H. M. (2023). Validation of the Al - based Predictive OpTimal Trees in Emergency Surgery Risk (POTTER) calculator in patients 65 years and older. Annals of Surgery, 277 (1), e8 - e15.
[20] Moulaei, K., Shanbehzadeh, M., Mohammadi - Taghiabad, Z., & Kazemi - Arpanahi, H. (2022). Comparing machine learning algorithms for predicting COVID - 19 mortality. BMC medical informatics and decision making, 22 (1), 2.
[21] Mujumdar, A., & Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms. Procedia Computer Science, 165, 292 - 299.
[22] Ng, K., Ghoting, A., Steinhubl, S. R., Stewart, W. F., Malin, B., & Sun, J. (2014). PARAMO: a PARAllel predictive MOdeling platform for healthcare analytic research using electronic health records. Journal of biomedical informatics, 48, 160 - 170.
[23] Nithya, B., & Ilango, V. (2017, June). Predictive analytics in health care using machine learning tools and techniques. In 2017 International Conference on Intelligent Computing and Control Systems (ICICCS) (pp.492 - 499). IEEE.
[24] Park, D. J., Park, M. W., Lee, H., Kim, Y. J., Kim, Y., & Park, Y. H. (2021). Development of machine learning model for diagnostic disease prediction based on laboratory tests. Scientific reports, 11 (1), 7567.
[25] Ribeiro Junior, M. A. F., Smaniotto, R., Gebran, A., Zamudio, J. P., Mohseni, S., Rodrigues, J. M. D. S., & Kaafarani, H. (2023). The use of POTTER (Predictive Optimal Trees in Emergency Surgery Risk) calculator to predict mortality and complications in patients submitted to Emergency Surgery. Revista do Colégio Brasileiro de Cirurgiões, 50, e20233624.
[26] Safa, M., Pandian, A., Gururaj, H. L., Ravi, V., & Krichen, M. (2023). Real time health care big data analytics model for improved QoS in cardiac disease prediction with IoT devices. Health and Technology, 13 (3), 473 - 483.
[27] Sakib, N., Jamil, S. J., & Mukta, S. H. (2022, July). A novel approach on machine learning based data warehousing for intelligent healthcare services. In 2022

IEEE Region 10 Symposium (TENSYMP) (pp.1 - 5). IEEE.

[28] Soni, M., & Varma, S. (2020). Diabetes prediction using machine learning techniques. International Journal of Engineering Research & Technology (IJERT), 9 (09), 2278 - 0181.

[29] Thantilage, R. D., Le - Khac, N. A., & Kechadi, M. T. (2023). Healthcare data security and privacy in Data Warehouse architectures. Informatics in Medicine Unlocked, 39, 101270.

[30] Yangui, R., Nabli, A., & Gargouri, F. (2017). ETL based framework for NoSQL warehousing. In Information Systems: 14th European, Mediterranean, and Middle Eastern Conference, EMCIS 2017, Coimbra, Portugal, September 7 - 8, 2017, Proceedings 14 (pp.40 - 53). Springer International Publishing.

[31] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics, 9, 515.