doi:10.20944/preprints202103.0425.v1

Data Science in Healthcare- Current Challenges and Opportunities

Pankaj Khurana[1]*, Rajeev Varshney[1]

[1]Defence Institute of Physiology and Allied Sciences (DIPAS), Defence R&D Organization (DRDO), Timarpur, Delhi-110054

*Correspondence: pkhurana08@gmail.com

Abstract

The rise in the volume, variety and complexity of data in healthcare has made it as a fertile-bed for Artificial intelligence (AI) and Machine Learning (ML). Several types of AI are already being employed by healthcare providers and life sciences companies. The review summarises a classical machine learning cycle, different machine learning algorithms; different data analytical approaches and successful implementation in haematology. Although there are many instances where AI has been found to be great tool that can augment the clinician's ability to provide better health outcomes, implementation factors need to be put in place to ascertain large-scale acceptance and popularity.

Keywords: Artificial intelligence (AI), Machine Learning, Healthcare, Haematology

Introduction

AI and ML are the talk of the research community. AI is a broad term that refers to the use of machines to be able to do decision-making and ultimately carry out complex tasks like voice recognition, image processing and other complex decisions. ML is a field that gives ability to the computer systems to learn from data and also improve itself with additional data. Both have the ability to revolutionise the Health Care and become an integral part of disease diagnosis, prognosis, prediction and management [1]. Artificial intelligence has seen unprecedented progress in analyzing biological data in the last decade. Thanks to improvements in computing algorithms, computing ability and structured access to repositories of data. The healthcare sector enjoys a significant proportion of Gross Domestic Product (GDP) in developed countries. Today as the volumes of data generated from biomedical research and clinical medicine reach astronomical values, it is a fertile bed for AI research. Currently, the biomedical field can be aptly referred as "data rich, information poor". The varsity, volume and variety of data generated surpasses human ability to make use of it. Though still in its infancy, AI holds a great promise for medical sciences.

Suthesh Sivapalaratnam [2] highlights the use of AI and ML in haematology and outlines their use in 3 broad are namely – a) decision support for incoming referrals to the haematologist b) automated blood film reporting c) aid in prediction and risk stratification. Analysing this increasing amount of data leads to a possibility of more knowledge, better understanding and ultimately better patient care.

A machine learning life cycle:

There are four major steps in the machine learning life cycle, all of which have equal importance and go in a specific order (Figure 1) .

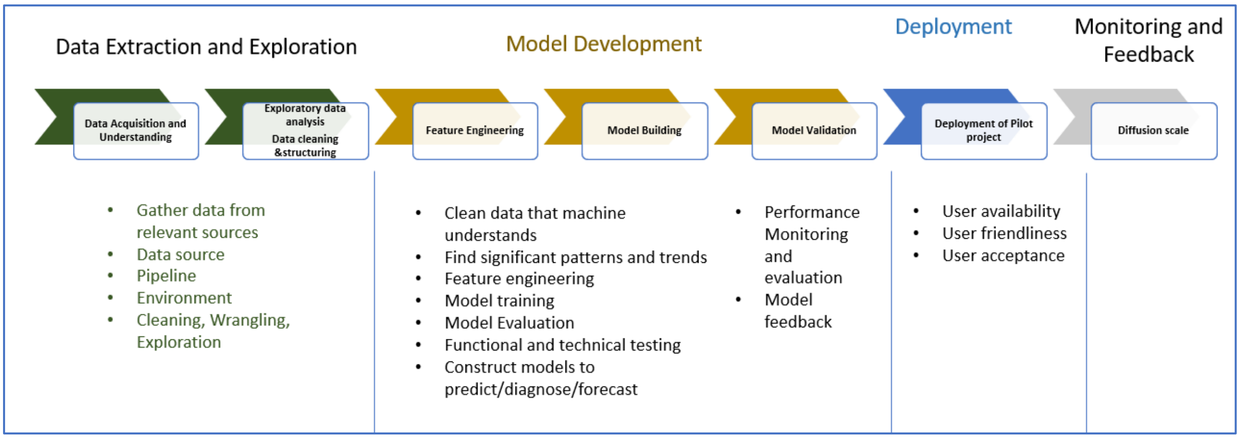
Data Extraction and Exploration
Model Development
Deployment
Monitoring and Feedback
Data Acquisition and Understanding
Exploratory data analysis Data cleaning &structuring
Feature Engineering
Model Building
Model Validation
Deployment of Pilot project
Diffusion scale
Gather data from relevant sources
Data source
Pipeline
Environment
Cleaning, Wrangling, Exploration
Clean data that machine understands
Find significant patterns and trends
Feature engineering
Model training
Model Evaluation
Functional and technical testing
Construct models to predict/diagnose/forecast
Performance Monitoring and evaluation
Model feedback
User availability
User friendliness
User acceptance

Figure 1: A machine-learning life-cycle

1) Data Extraction and Exploration

(i) Data acquisition and understanding:

The first step is to collect and prepare all of the relevant data for use in machine learning. It may include consulting medical domain experts to determine the data might be relevant in predicting disease risk. It involves gathering data from patient records, and getting it into a format suitable for analysis.

In healthcare sector, a major source data is Electronic Health Record (EHR). Data stored in medical records are often incomplete, complex, messy, and can be biased [3]. The use of raw medical records for machine learning models may sometimes lead to false conclusions. Eg. A study on patients who died in emergency department due severe community-acquired pneumonia had very scare information stored in medical records. As a result, some patients who died appeared healthier than those who survived [4]. The most concerning limitation of extracting data related to variables that were stored in medical records as unstructured data in the form of physicians notes (e.g., descriptions of electrocardiograms). To overcome this, it is important to design classifiers that use features that are available in the medical records as single measurements. This reduces the ambiguity and false data collection.

(ii) **Exploratory data analysis (EDA), Data cleaning and structuring**

EDA is an technique to analysing data-sets to summarize their main features and descriptors. EDA can reveal insights about data that can be used to build a model as well as help the scientist identify necessary data cleaning steps. It is a critical process of performing initial investigations on data so as to discover patterns, to spot anomalies and usually involved graphical analysis and representations.

Data cleaning involves identifying, modifying and ensuring that the given data set is free from error, consistent and compatible. It helps in identifying any errors or corruptions in the data, correcting or removing them, or manually processing them to prevent the error from corrupting the final analysis. It is always advisable to create a classifier that could be built into medical records software and automatically identify the patients with a high risk of an unfavourable outcome.

2) Model Development

(i) Feature Engineering

Feature engineering is the process of using domain expertise to extract features from raw data and then preparing the proper input dataset in a format that is compatible with the machine learning algorithm requirements. Scientists spend maximum amount of their time on data preparation.

Various techniques like imputation (removing missing value rows/columns), handling outliers, binning, log transform, grouping, feature splitting, scaling ect are used depending on the data.

(ii) Model Building

Various supervised, unsupervised and reinforcement learning machine learning teachniques may be used for model-building. available data is subjected to data splitting whereby it is split to 2 train-test datasets. The first portion is the larger data subset that is used as the training set (60-80%) and the second is normally a smaller subset and used as the testing set (20-40%).

In order to make the most economical use of the available data, an N-fold cross-validation (CV) is normally used whereby the dataset is partitioned to N folds (i.e. commonly 5-fold or 10-fold CV are used). In such N-fold CV, one of the fold is left out as the testing data while the remaining folds are used as the training data for model building.

Two common machine learning tasks in supervised learning includes classification and regression.

A trained classification model takes as input a set of variables (either quantitative or qualitative) and predicts the output class label (qualitative) whereas a regression model predicts a continuous output (quantitative). Evaluation of the performance of regression models are performed to assess the degree at which a fitted model can accurately predict the values of input data.

A number of common machine learning algorithms include: Linear Regression, Logistic Regression, Decision Tree, SVM, Naive Bayes, kNN, K-Means, Random Forest, Dimensionality Reduction Algorithms, Gradient Boosting algorithms (GBM, XGBoost, LightGBM, CatBoost).

(iii) Model Validation

Using proper validation techniques helps you understand the model and also estimate an unbiased generalization performance. Various model validation techniques include: Train/test split, k-Fold Cross-Validation, Leave-one-out Cross-Validation, Leave-one-group-out Cross-Validation, Nested Cross-Validation, Time-series Cross-Validation and Wilcoxon signed-rank test

3) Deployment of Pilot project

This involves implementing the developed machine learning solution in real-world to solve actual problems for predictive analysis. It involves cloud deployment for user-convenience, its management

4) Monitoring and Feedback

It involves continuous development and improvement of the developed algorithm to increase the specificity and sensitivity of the algorithm. Automated pipelines are built for continuous monitoring and so that the AI capabilities are valuable and productive.

Data Analytics types

There are four types of data analytics Descriptive, Diagnostic, Predictive and Prescriptive. Descriptive Analytics tells what happened in the past. Diagnostic Analytics helps understand why something happened in the past. Predictive Analytics predicts what is most likely to happen in the future. Prescriptive Analytics recommends actions to be taken to affect those outcomes.

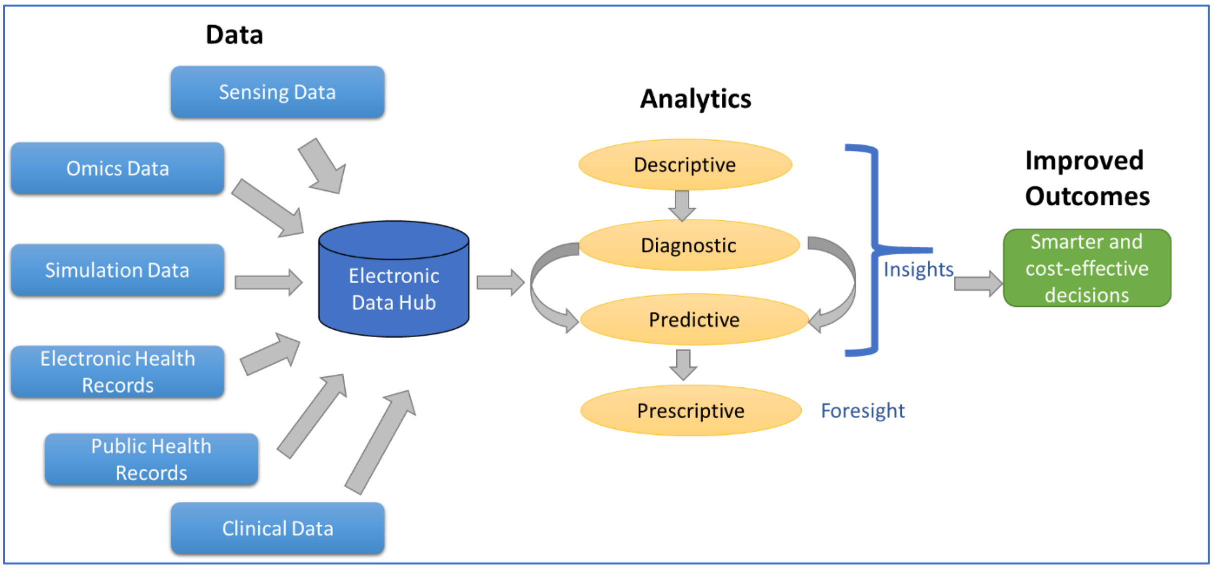

Figure 2: The data science encompasses descriptive, diagnostic, predictive and prescriptive analytics

Machine learning techniques have previously been applied for various clinical decision-making process such as metabolic syndrome [5]; radiology [6], diabetes mellitus [7], cancer [8] and hypertension [9]. The techniques have sometimes performed on par with trained clinicians like detecting diabetic retinopathy in eye fundus images [10], automatically classify skin lesion

doi:10.20944/preprints202103.0425.v1

images [11], detect hip fractures from frontal pelvic X-rays [12]. AI has also made notable progress in the field of haematology [13].

Researchers have also used machine learning approaches such as support vector machine, artificial neural network and association rule analysis to explore the relationship between haematological parameters and glycemic status in the establishment of quantitative population-health relationship (QPHR) model for identifying individuals with or without diabetes mellitus [14]. The machine learning approaches (i.e., SVM and ANN) employed in this study have been shown to be capable of correctly classifying the DM status affording accuracies of more than 98%.

The haematological indices like neutrophil count, neutrophil to lymphocyte ratio (NLR), red cell distribution width (RDW), platelet to lymphocyte ratio (PLR), mean platelet volume (MPV), and platelet distribution width (PDW) have been used to predict significant coronary lesion (SCL) and in-hospital mortality.[15]

Table 1 provides a brief overview of various studies in haematological sciences; the algorithm used, the broad efficiency and outcomes.

	Author	Application	Method	Results	Conclusions
1	Konrad Pieszko etal (2018) [16]	Prediction of short-term acute coronary syndrome (ACS) outcomes by using the predictive value of haematological indices	Ensembles of decision trees and decision rule models	neutrophil count, age, systolic and diastolic pressure and heart rate achieved the high feature importance scores as well as the positive confirmation measures	Association between the elevated inflammatory markers and the short-term ACS outcomes to provide accurate predictions.
2	Chanin Nantasenamat etal (2013) [14]	Explore the relationship between haematological parameters and glycemic status in the establishment of quantitative population-health relationship (QPHR) model for identifying individuals with or without	SVM, ANN	Relationship amongst haematological parameters and glucose level indicated that the glycemic status (normal, Pre-DM and DM) was well correlated with WBC, RBC, Hb and Hct.	Such predictive models provided high classification accuracy as well as pertinent rules in defining DM.

		diabetes mellitus (DM).			
3	Konrad Pieszko etal (2019) [15]	Haematological Markers for Predicting Long-Term Mortality After Acute Coronary Syndrome	The gradient-boosted tree algorithm	Haematological markers, such as neutrophil count and red cell distribution width have a strong association with all-cause mortality after acute coronary syndrome	ML techniques can provide long-term predictions of accuracy comparable or superior to well-validated risk scores.
4	Wellington Pinheiro dos Santos etal (2020)[17]	Heg.IA: An intelligent system to support diagnosis of Covid-19 based on blood tests	Bayesian networks	Haematological parameters can be indicators of the risk factors and degree of severity of Covid-19.	achieved high diagnosis performance and developed as a desktop application
5	Jaewoo Song etal (2020)[18]	Haematological Parameters for Sepsis Screening in Patients with Fever	Complementary model	The complementary model showed statistically better performance	The developed model provided statistically better performance than models for the existing sepsis-related clinical score.
6	Ruth M Ayling [19]	Hb, haemoglobin concentration; MCH, mean corpuscular haemoglobin; MCV, mean corpuscular volume.	Array of decision trees and gradient boosting machines	Prioritisation of Colonoscopy in Anaemia	The ColonFlag™ score identifies patients at risk of colorectal cancer using blood count parameters, age and sex,
7	Hennek etal (2016) [20]	Standard machine learning tools were used to analyze images	Image analysis tools	Improve diagnosis of Iron Deficiency Anemia (IDA)	Use of aqueous multiphase systems (AMPS) combined with machine learning provides an approach to developing

					point-of-care haematology.
8	Go etal (2018) [21]	Automatic Identification of Erythrocytes	decision tree model, support vector machine, linear discriminant classification, and k-nearest neighbour classification.	Digital mircoscopy combined with ML based Image analysis could successfully identify erythrocyte cell types	sensing abnormal erythrocytes and diagnosis of haematological diseases in clinic.

Diabetes mellitus (DM) is a complicated disease, defined as a group of metabolic disorders portrayed by increasing blood glucose or hyperglycemia that occur as a result of defects in insulin secretion, insulin action or insulin resistance (IR). It can be classified as type 1 (i.e. non-secretion of insulin) and type 2 (i.e., defined by IR). DM affects multiple tissues in the body and impairs organ function or biochemical parameters and, without proper diagnosis or treatment, morbidity or mortality can occur. Researchers have used ANN and SVM to explore the relationship between haematological parameters and glycemic status in the establishment of quantitative population-health relationship (QPHR) model for identifying individuals with or without diabetes mellitus (DM) [14].

Acute coronary syndrome (ACS) is inadequate blood flow to the myocardium which can be related to acute cholesterol plaque rupture or erosion and thrombus formation. An increased systemic and local inflammation plays a crucial role in the pathophysiology of ACS. Various haematological indices are nonspecific markers of the inflammatory response and can improve discriminative capabilities of the GRACE score. Researchers have successfully used ML techniques to ascertain that haematological markers of inflammation show strong correlation with the outcomes of ACS, and they can be successfully incorporated into numerical models designed to support clinical decisions [16]. As these parameters are easily available in the electronic medical record system; the ML program could provide risk assessment without any additional input from the physician. It can easily trigger relevant alerts, helping to identify the highest risk patients.

Some researchers have used haematological parameters as indicators of the risk factors and degree of severity of Covid-19. [17] . They further developed an intelligent system in a desktop application format. The authors used a database containing the results of more than one hundred laboratory exams, such as blood count, tests for the presence of viruses such as influenza A, and urine tests, of 5644 patients. Among these patients, 559 of them are infected with SARS-Cov2. They used Particle Swarm Optimization (PSO) and Evolutionary Search (ES) algorithms, since they are well established methods for feature selection. They used classical Bayesian networks using 24 blood tests results to achieve high diagnosis performance: 95.159% ± 0.693 of overall accuracy, kappa index of 0.903 ± 0.014, sensitivity of 0.968 ± 0.007, precision of 0.938 ± 0.010 and specificity of 0.936 ± 0.011.

Some researchers have used haematological parameters for sepsis screening in patients with fever [18]. Sepsis refers to a dysregulated host response to infection and an early intervention is important for patients with sepsis to increase the survival rates. Many non-specific responses makes the early diagnosis of sepsis difficult. A comprehensive data analysis on the 36 haematological parameters was performed and then the optimal combination of the parameters was identified. The machine learning model developed provided statistically better performance than models for the existing sepsis-related clinical score.

Researchers have proposed a machine learning based system for analysing haematological profiles to predict haematological diseases [22]. The authors used data from the University Medical Center of Ljubljana, which were collected between the years 2005 and 2015. A first model (SBAHEM181) was developed using 43 diseases and 181 parameters or attributes. Additionally a second model (SBA-HEM061) was developed with 61 parameters. The missing values were filled with median values for each attribute. Classic classifiers such as Support Vector Machines, Naive Bayes and Random Forest were tested. The model could predict with an accuracy of 59% and 57% considering all the diseases chosen. Therafter by restricting the prediction to five classes, the first and second model could predict with an accuracy of 88% and 86% respectively. The study clearly pointed towards the possibility of detecting diseases through blood tests using classic classifiers. The authors also showed that the accuracy of their predictive models was at par with that of haematology specialists.

Ruth M Ayling (2019) [23] have used blood count parameters, age and sex as an input to an array of decision trees and gradient boosting machines to derive a ColonFlag™ score. This score could potentially be generated as part of every full blood count report. It could facilitate appropriate referral of patients with a suspicion of colorectal cancer, to prioritise endoscopy and reduce invasive investigation of those who are at low risk.

Hennek etal [20] used standard machine learning tools of image analysis to improve diagnosis of Iron Deficiency Anemia (IDA). The authors developed a low-cost and rapid method to diagnose IDA using aqueous multiphase systems (AMPS). Standard machine learning tools were used to analyse images of the tests. The system was robust and could potentially be a suitable candidate for point-of-care haematology. Go etal [21] have used machine learning for automatic identification of erythrocytes

Conclusion

AI and Machine learning holds unprecedented possibilities in the medical field. AI has the potential to revolutionise the healthcare. It is an empowering tool for the clinicians and it’s time to welcome revolution in healthcare. It is indeed clear that AI will not be able to replace human medical professionals; but it would definitely augment their efforts to provide better and superior healthcare to the patients. Human medical professionals may move to more human-centric skills like convincing, empathy, advice, encouragement, morale-building. Overtime AI would be an integral part of the healthcare systems and enhance the capabilities of healthcare providers.

REFERENCES

1. Davenport, T. and R. Kalakota, *The potential for artificial intelligence in healthcare.* Future Healthc J, 2019. **6**(2): p. 94-98.
2. Sivapalaratnam, S., *Artificial intelligence and machine learning in haematology.* Br J Haematol, 2019. **185**(2): p. 207-208.
3. Hripcsak, G. and D.J. Albers, *Next-generation phenotyping of electronic health records.* J Am Med Inform Assoc, 2013. **20**(1): p. 117-21.
4. Hripcsak, G., et al., *Bias associated with mining electronic health records.* J Biomed Discov Collab, 2011. **6**: p. 48-52.
5. Kim, T.N., et al., *A decision tree-based approach for identifying urban-rural differences in metabolic syndrome risk factors in the adult Korean population.* J Endocrinol Invest, 2012. **35**(9): p. 847-52.
6. van Ginneken, B., *Fifty years of computer analysis in chest imaging: rule-based, machine learning, deep learning.* Radiol Phys Technol, 2017. **10**(1): p. 23-32.
7. Quentin-Trautvetter, J., et al., *Assessing association rules and decision trees on analysis of diabetes data from the DiabCare program in France.* Stud Health Technol Inform, 2002. **90**: p. 557-61.
8. Nahar, J., et al., *Significant cancer prevention factor extraction: an association rule discovery approach.* J Med Syst, 2011. **35**(3): p. 353-67.
9. Shin, A.M., et al., *Diagnostic analysis of patients with essential hypertension using association rule mining.* Healthc Inform Res, 2010. **16**(2): p. 77-81.
10. Gulshan, V., et al., *Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs.* JAMA, 2016. **316**(22): p. 2402-2410.
11. Esteva, A., et al., *Dermatologist-level classification of skin cancer with deep neural networks.* Nature, 2017. **542**(7639): p. 115-118.
12. Gale W., O.-R.L., Carneiro G., Bradley A. P., Palmer L. J., *Detecting Hip Fractures with Radiologist-Level Performance Using Deep Neural Networks.* https://arxiv.org/abs/1711.06504., 2017.
13. Radakovich, N., M. Nagy, and A. Nazha, *Artificial Intelligence in Hematology: Current Challenges and Opportunities.* Curr Hematol Malig Rep, 2020.
14. Worachartcheewan, A., et al., *Machine learning approaches for discerning intercorrelation of hematological parameters and glucose level for identification of diabetes mellitus.* EXCLI J, 2013. **12**: p. 885-93.
15. Pieszko, K., et al., *Machine-learned models using hematological inflammation markers in the prediction of short-term acute coronary syndrome outcomes.* J Transl Med, 2018. **16**(1): p. 334.
16. Pieszko, K., et al., *Predicting Long-Term Mortality after Acute Coronary Syndrome Using Machine Learning Techniques and Hematological Markers.* Dis Markers, 2019. **2019**: p. 9056402.
17. Valter Augusto de Freitas Barbosa, J.C.G., Maira Araujo de Santana, Jeniffer Emidio de Almeida Albuquerque, Rodrigo Gomes de Souza, Ricardo Emmanuel de Souza, View ORCID ProfileWellington Pinheiro dos Santos, *Heg.IA: An intelligent system to support diagnosis of Covid-19 based on blood tests.* https://www.medrxiv.org/content/10.1101/2020.05.14.20102533v1, 2020.

18. Choi, J.S., et al., *Implementation of Complementary Model using Optimal Combination of Hematological Parameters for Sepsis Screening in Patients with Fever.* Sci Rep, 2020. **10**(1): p. 273.
19. Goshen, R., et al., *Computer-Assisted Flagging of Individuals at High Risk of Colorectal Cancer in a Large Health Maintenance Organization Using the ColonFlag Test.* JCO Clin Cancer Inform, 2018. **2**: p. 1-8.
20. Hennek, J.W., et al., *Diagnosis of iron deficiency anemia using density-based fractionation of red blood cells.* Lab Chip, 2016. **16**(20): p. 3929-3939.
21. Go, T., H. Byeon, and S.J. Lee, *Label-free sensor for automatic identification of erythrocytes using digital in-line holographic microscopy and machine learning.* Biosens Bioelectron, 2018. **103**: p. 12-18.
22. Guncar, G., et al., *An application of machine learning to haematological diagnosis.* Sci Rep, 2018. **8**(1): p. 411.
23. Ayling, R.M., S.J. Lewis, and F. Cotter, *Potential roles of artificial intelligence learning and faecal immunochemical testing for prioritisation of colonoscopy in anaemia.* Br J Haematol, 2019. **185**(2): p. 311-316.