CURRENT SCIENCE

Volume 122 Number 6 **25 March 2022**

GUEST EDITORIAL

Need for machine learning applications in solid Earth geosciences in India

The art of learning science from data is cultivated through machine learning (ML), which utilizes the knowledge of statistics and computer science for pattern recognition and data mining applications (Hastie, T. *et al.*, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer Series in Statistics, Springer, New York, 2009, 2nd edn, p. 533). The potential of ML in comprehending multivariate data sets of solid Earth involving different observational platforms, resolutions, scales, and interlinkage of physical laws, may play a pivotal role in unravelling hidden patterns in the data, understanding subsurface structures and eventually developing predictive models of present and deep past geo-processes (Bergen, K. J. *et al.*, *Science*, 2019, **363**, eaau0323). Such approach would make a paradigm shift from reductive approach to inductive approach to infer plausible physical models in explaining intricately linked physical, chemical, biological, and geological processes operating at varying spatiotemporal scales within the solid Earth. Four major classifications of ML, viz. supervised, unsupervised, semi-supervised and reinforcement learnings are mainly being used in developing and establishing data interrelationship and complicated models. The supervised ML induces the model from the training data, which is guided by optimizing error or loss function, based on the internal architecture of the learning algorithm, whereas the unsupervised ML predicts model with a learning algorithm having unrestricted possibilities to find natural groups or clusters in the data that best define their inherent relationships (Hastie, T. *et al.*, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer Series in Statistics, Springer, New York, 2009, 2nd edn, p. 533). Supervised ML utilizes labelled data, which means some input data is already tagged with the correct output and unsupervised ML is operated on unlabelled data sets to analyse and cluster. Therefore, supervised ML is a better option. As name indicates, the semi-supervised ML uses both labelled and unlabelled data for training and predicting the model.

The past decade has witnessed tremendous growth in the application of ML in analysing Earth observation data of climate and ocean. However, the ML applications in the solid Earth geoscience started recently, mainly because of the inability of large data handling and inadequacy of satisfactory data for training and testing. The support vector machines (SVM), decision trees (DT), artificial neural networks (ANN), self-organizing maps (SOM), ensemble methods such as random forest, case-based reasoning, neuro-fuzzy (NF), genetic algorithm (GA), multivariate adaptive regression splines (MARS), etc. are some of the commonly used ML algorithms which are applied to solid Earth data sets. A summary of the emerging opportunities of ML in solid Earth geosciences is dotted down with a few selected examples, considering the global availability of huge, diverse geoscience data sets and challenges of computationally expensive physics-based models. However, there exist significant challenges of using ML algorithms. The physics-based models are yet to be incorporated with suitable ground-truthing, supervised or unsupervised, to reliably reveal the Earth's internal structures and processes.

Surface geological mapping is one of the fundamental themes of study in solid Earth geosciences. This is conventionally done by an experienced geologist physically going to the field and recording geological information such as lithology, structural properties, etc. from the outcropping rocks. These maps are prepared based on field observations limited by synoptic view and sometimes extrapolated due to the unapproachability of the terrain. These maps can now be improved by supplementing airborne geophysical observations and satellite imageries. A range of algorithms are tested on different geological terrains using the available remotely sensed and geophysical variables and their applicability is demonstrated in the different geological areas (Cracknell, M. J. *et al.*, *Austr. J. Earth Sci.*, 2014, **61**(2), 287–304). Geodynamic modelling, natural resource exploration, and monitoring of geological processes are inextricably linked to subsurface characterization. Often, subsurface model parameters are estimated from geophysical observations using the inversion approach following complex mathematical relations. Computation with higher-order derivatives of optimization algorithms to obtain the optimum objective function is a challenge in geophysical inversion. In such a situation, a data-driven approach to ML algorithms have become popular in solving inverse problems without having in depth knowledge of physical equations or theories. Some of the examples that demonstrate the applications of ML in the subsurface imaging using geophysical data are joint inversion scheme of marine Controlled Source Electromagnetic (CSEM) data using the Bayesian approach (Ray, A. *et al*., *Geophys. J. Int.*, 2014, **199**(3), 1847–1860); a deep learning method for a 3D sparse inversion of gravity data (Huang, R. *et al*., *J. Geophys. Res.*, 2021, **126**, e2021JB022476); density model of sedimentary basin (Roy, A. *et al*., *Geophysics*, 2021, **86**(3), 1–63); estimation of the subsurface velocity model through seismic inversion by hybrid ML (Chen and Saygin, *J. Geophys. Res.*, 2021, **126**, e2020JB021589). Arrival of first phase waves in seismic exploration is the main component that leads to geophysical imaging and its accurate estimation is needed for several processes like tomography, static correction, velocity analysis, amplitude versus offset (AVO) analysis. Generally, seismologists and other researchers mark this first phase arrival of seismic waves manually. Such manual picking of waves from a large set of seismic data is time-consuming and causes a wide range of errors in determining the arrival times leading to uncertainty in geophysical interpretations. Therefore, AI/ML-based algorithms like traditional ANN, CNN, fuzzy c-means clustering analysis (FCM), and SVM are being developed for the automatic classification of seismic signals. Geological architecture in different terrains may have distinct seismic footprints, and seismic attributes provide a flawless understanding of the geological and geophysical interpretation of the data. Seismic attributes help in extraction of ambiguous sub-surface information from data and further help the interpreter in structural and stratigraphic interpretation by highlighting a range of geological structures of interest like faults, fractures, channels, chimneys, dykes, sills, etc. ML has evolved as a worthy tool to extract seismic attributes from data sets that need significant amount of time, rigorous human effort in a repeatable workflow (Qi, J. *et al*., *Geophysics*, 2019, **85**(2), 1–95).

Earthquakes and other natural hazards are some of the most abstruse areas of geosciences. Most of the researchers are naïve when it comes to understanding of the relation of deep earth dynamics with surface processes. Generally, earthquake generating rupture processes comprise complex interactions of stress, fracture, and frictional properties in the subsurface. AI/ML might help one to choose the plausible geometries along with physical model parameters for the countless multidimensional models of the tectonic processes that geodynamic theory proposes. New ML advanced techniques demonstrate the great potential to reveal the processes and nowcasting of an earthquake (Shreedharan, S. *et al*., *J. Geophys. Res.*, 2021, **126**, e2020JB021588; Rundle, J. B. *et al*., *Earth Space Sci.*, 2021, **8**, e2021EA001757). The geodynamic process modelling is also addressed by means of ML, where SVM is used to understand the mantle convection using temperature observation and mantle density anomalies (Shahnas, M. H. *et al*., *J. Geophys. Res.*, 2018, **123**, 2162–2177).

Global Navigation Sensing System (GNSS) and geodetic observations are extensively used for understanding crustal deformation due to different Earth processes, e.g. earthquake cycle, hydrological loading, coastal subsidence. The foremost challenge in geodetic data from GNSS is to identify with least uncertainty the outliers and to detect the anomalies in the time series. Several ML algorithms like multilayer perceptron neural network, Bayesian regulation, and Gaussian processes are developed and used for the hidden information in time-series-based data in an automated manner. Similar time series data from InSAR (Interferometric Synthetic Aperture Radar) provide a wide range of applications in rural management and natural hazard assessment through continuous data generation. AI/ML-based algorithms may provide a brilliant framework to analyse such huge data sets and simulate the model that allows quantification of uncertainties in time series data sets (Anantrasirichai, N. *et al*., *J. Geophys. Res.*, 2018, **123**, 6592–6606). ML algorithms are found to be very useful for data reconstruction with the aim of transforming an incomplete data set into a corresponding complete data set, which is a cross-cutting research problem in geoscience. Logistic regression of supervised learning is widely used as a classifier to distinguish automatically and to predict the probability of an event through a discrete number of zero and one (Reynen, A. and Audet, P., *Geophys. J. Int.*, 2017, **210**(3), 1394–1409).

From the foregoing discussion, it is evident that there is a need for swift adaptation of ML applications in solid Earth geoscience, education and research. This will offer new potentials through new tools for detailed re-examination of looking of archived data, and also analyse rapidly gathering voluminous new observations. Availability of data is critical for applying ML tools; thus, there should be a comprehensive data sharing policy for solid Earth geoscience as per the globally prevalent FAIR (Findable, Accessible, Interoperable, and Reusable) data principles which should balance the issues of IPR of the originator of data, as well as policies of the government/sponsors. This policy should categorize the data appropriately as novel data, high-value data, experimental data, routine/common data, confidential data and clearly state implementation methods by organizations/researchers (Data policy at CSIR-NGRI, https://www.ngri.res.in/cms/ngri-data-policy.php). Another vital necessity is selection of appropriate ML approaches and developing new architectures; particularly, ML algorithms incorporating physics-based models, which can be realized through collaboration with experts of ML until expertise is developed in the solid Earth geoscience community. For this purpose, workshop, and special session (INSA Annual Report, 2018–19) and international training course (www.ngri.res.in) are being organized, which could help to drive forward the application of ML in solid Earth geoscience.

Virendra M. Tiwari

CSIR-National Geophysical Research Institute,
Hyderabad 500 007, India
e-mail: vmtiwari@ngri.res.in