Machine Learning and Data Analytics for Environmental Science: A Review, Prospects and Challenges

Tharsanee Raman Maganathan¹, Soundariya Ramasamy Senthilkumar², Vishnupriya Balakrishnan³

¹, ², ³ Assistant Professor, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Sathyamangalam.

E-mail: tharsanee@bitsathy.ac.in

Abstract. Innovations in Machine Learning and Data Analytics can possibly affect numerous aspects of Environmental Science (ES). Data Analytics refers to a collection of data resources indicated in terms of variety, velocity, veracity and volume. Big data contributes to the ES arena in applications such as weather forecasting, energy sustainability and disaster management with the advent of techniques such as Remote Sensing, Information and Communication technologies. Though big data is used to accomplish data analysis and interpretation for ES, there are still requirements for efficient ways of data storage, processing and retrieval. Machine Learning and Deep Learning are the sub fields of artificial intelligence which deals with training the models to learn from data without being explicitly programmed. When Machine Learning and Deep Learning are combined together it is possible to unleash the supremacy of data analytics. These techniques show high prospective for process optimization, information-centric decision making and scientific discovery. Scientific developments like these will assist ES to make real time autonomous decisions by extracting useful insights from huge data. These advancements also aid in bridging the gap between the theoretical backgrounds on ES to practical implementation. The primary objective of this survey is to figure out the basic concepts of Machine Learning, Deep Learning, and Data Analytics and find the state-of-the-art applications in ES, and observe the impending benefits of information-centric investigation on ES.

1. Introduction
In the early 1990’s, the environmental science was strived by physical, chemical and different natural resources that makes the living things to interact with themselves. Environmental data is growing faster as it leads to major issues like complexity, size and resolution. In dealing interdisciplinary problems such as the intelligence to process the data, analyzing the large volume of data was experienced by Environmental Scientists. Extracting heterogeneous data from different sources to perform integrated analysis and acquire knowledge, requires a wide wisdom over Data Science. The popularity of Data Science techniques helps in managing the environmental systems. Data Science is used to model different scenarios and protocols, leading to data induced innovation in business sectors. Environmental scientists are forced to think in a way, the various scientific interdisciplinary problems would be solved by the existing or innovative data science approaches. Data Science adds value to the Environmental Sciences as it provides a realistic approach [1].

A stream of data from the environmental sensors has to be analyzed to obtain beneficial information in Data Science. The data are acquired through remote sensing satellites, information from satellites and sensors to measure air quality, water quality through weather and climatic observations, ground based sensors which also provide information to measure the magnitude of earthquake and other geographical events. To extract information from the sensors or from satellite data, different meta-heuristic bio inspired algorithms were used. The formulation approach in processing or extracting the information from remote sensing satellite data or sensor data requires training, methods and efficient tools to proceed with. The significant challenge and the rapid growth of Environmental Science have
paved the way for research by applying different approaches in data science techniques. The following challenges were identified as part of the investigation on complications in environmental data [2].

- Demand for trained/experienced data scientist for the environmental study
- A gap in identifying the proper approach for building a framework
- Lack of knowledge in figuring out the environmental issues.
- Robust way in dealing with environmental data.
- Optimal techniques to extract information.
- Protocols to prefer the right method for analysis.
- Information exchange and preserving of data
- Storage cost and Energy Consumption.

Machine learning techniques, a futuristic approach in data mining emerged from Artificial Intelligence and are now widely used in the field of environmental science. Machine Learning techniques are applied in processing satellite data, climatic predictions, forecasting, designing and examining of environmental data. To extract valuable information from the above said data we need a realistic and modernized approach which includes linear statistical analysis, time series analysis, Feed Forward Neural Network Models, Non - Linear optimization, Generalization Learning, Classification Models, Regression Models, Principal Component analysis and correlation analysis, these models will help in formulating an optimal approach to the various environmental science problems [3]. Deep Learning, a subset of machine learning uses different data structures often termed as Artificial Neural Networks. A different Deep Learning model provides a feasible approach, specialized computational tools for benefits of environmental science, deep learning allows the environmental data scientist to concentrate more on pre-processing tasks and optimized capabilities for real-time large scale integrated environmental monitoring. In this paper, a brief explanation on different machine learning methods and deep learning algorithms were given, that could be suggested for implementing those for various Environmental Data Analysis and its applications [4].

2. Analysis of Machine Learning Algorithms

In the era of machines, these machines are like rewards that follow human instructions. But nowadays machines have started to learn on its own which we name as Machine Learning, a futuristic technology. Machine Learning is a part of Artificial Intelligence that focuses on modelling the system, allowing them to learn and prepare them to predict based on past experiences. In turn ML uses statistical computation that empowers the experiences. As a first and foremost step, the machine learning algorithm is fed with labelled or unlabelled datasets to produce the model, and will input the new data, whereas the machine creates predictions based on the model which is evaluated for accuracy through some performance metrics. ML is subcategorized into three types namely, Supervised, Unsupervised and Reinforcement Learning. Supervised Learning contains the class label, uses algorithms like regression, classifications etc. to make the machine learn, once the model is equipped the machine will predict the new input data correctly. Unsupervised Learning is a self-learning approach where the algorithms like clustering and association that take the input dataset without any type of labels. Reinforcement Learning, a third type of ML, where an agent takes an appropriate action in order to maximize the rewards, Markov Decision Process is one of the techniques. To be more specific, the agent is put inside an unknown environment that uses trial and error methods to evaluate the environment to turn up with an appropriate outcome. The Figure 1 gives a pictorial representation of subcategories of Machine Learning. The working mechanisms of the popular machine learning algorithms that have high potential for influencing environmental sciences is discussed in this section.

2.1. Support Vector Machine

Support vector machines are generally categorized under supervised learning algorithms as it does not involve any conventions regarding the data distribution in datasets. This algorithm is widely adopted
in several classification as well as regression type of problems [2]. In this algorithm, input space in the dataset is assigned to preferably high dimensional feature space to easily solve the classification problems. In the simplest sense, this algorithm can be viewed as a machine which divides the dataset into parts to make it more precise to derive meaningful patterns. Thus, it forms a decision boundary in order to place the data points according to its respective class. This algorithm is more suitable for classification problems containing a linear separation between the data points.

2.2. Random Forest
This is a kind of an ensemble algorithm which is used to solve classification and regression problems. This algorithm is termed as an ensemble algorithm since it involves the process of building decision trees and takes the prediction with the highest votes as final prediction [6]. The decision trees are built by taking training samples in random with chances for some samples to occur frequently and some not occur at all. The decision trees constructed are independent of each other and contain features through random selection. The forest is constructed in this way to achieve lower value in bias and higher value in variance.

2.3. K-Nearest Neighbour
This machine learning algorithm on receiving new data samples finds the similarity of features between the training samples already available and the new data sample. The final class prediction depends on the voting given by the neighbours sharing similar features with the test sample [7]. In order to find the closest training samples, a distance function is used to estimate the distance between the new data sample and the training samples in the dataset. Finally, the new data sample is classified to the specific class nearest with its neighbours. One setback in this technique is the memory usage and cost incurred for computations as it involves the calculation of distance between the samples.

2.4. K-means clustering
In this type of unsupervised learning algorithm, the data points are distributed into different k numbers of clusters. This algorithm tends to find the cluster centres for each and every dataset in order to split it into the appropriate cluster [13]. This cluster centre is assigned randomly during the first phase of the algorithm execution. Later, the Euclidean distance between the data point and the cluster centre is calculated and the data point is assigned to the nearest cluster. Cluster centre is then updated by computing the mean value. Choosing K value is the difficult task in this algorithm, as there are no predefined conditions to select the value of K.

Figure 1. Categories of Machine Learning algorithms
3. Analysis of Deep Learning Algorithms

Deep learning (DL) is considered as the subset of machine learning and further ML is the subset of artificial intelligence. DL uses neural networks (inspired by biological neurons) to train machines to perform certain tasks. Neural networks consist of an input layer, hidden layer and an output layer. The number of hidden layers is not restricted to one, the more hidden layers are nested to the network, and it becomes a so-called deep learning network, which helps in learning features at the higher levels. DL outperforms classical ML algorithms in terms of efficient performance (due to its parallel processing architecture) and handling massive data. Neural networks with its layered architecture, plays a versatile role in many applications. The following section clearly explains the different neural net frameworks that can be helpful for environmental science.

3.1. Artificial Neural Networks

Artificial Neural Networks are one of the fields of deep learning which mimics the operational style of the human brain. These algorithms do not require any straightforward programming to formulate rules for its execution. Similar to the neurons in the human nervous system, ANN consists of neurons to implement parallel operations [13]. These systems produce the value of the final prediction based on the real time input given to the neuron. The input received by the neuron is characterized as a combination of weight and bias added to it. The weight is assigned based on the importance the input carries in the contribution towards making valid predictions.

3.2. Convolution Neural Networks

In a CNN, data is carried from the input to the output in a forward path; hence it can be characterized as a feed forward neural network [4]. These systems are primarily used in image processing applications and work well for all types of data including one dimensional data such as time series input, two dimensional data such as grayscale or color images or multidimensional data such as seismic data. CNN architecture consists of convolutional and pooling layers followed by fully connected layers. Convolutional layers are responsible for feature extraction, which is performed using convolution operation. Convolution operations use a filter across the image to produce a feature map.

Several deep architectures based on CNN are AlexNet, VGGNet, GoogleNet, ResNet used for a variety of applications in image processing.

3.3. Recurrent Neural networks

Real world datasets consist of sequential data inputs containing an inherent relationship between the data samples. Traditional machine learning methods fail to work on this kind of sequential data. RNN is a special field of deep learning which can model the sequential data to learn the meaningful insights in it. This network consists of an input layer and output layer with multiple hidden layers in between. Each layer takes an input sequence and produces the output sequence that should be passed on to the next layer in the network. Long Short Term Memory (LSTM) Networks are a variation of RNN [8], which is composed of gates like input, forget and output to retain the status of the previous input. In this setup, all the essential statuses of previous inputs are retrieved whenever required rather than obtaining the recent outputs alone.

3.4. Back propagation Neural Networks

This is one of the simplest forms of neural network with a minimum one hidden layer existing in the network along with the input and the output layer with numerous neurons present in each layer [8]. There are two types of propagation, one as forward Propagation and the other as Backward Propagation. Input layer consists of neurons which are transmitted to the output layer through the hidden layer. The results obtained in the output layer are verified with the expected results and in case of any deviations reported then the back propagation process is carried out to adjust the weights of the neurons until the error is minimized. This algorithm is widely adopted in most of the remote sensing applications.
4. Data Analytics for Environmental Sciences

Data Analytics is generally the process of deriving meaningful insights from intricate data in order to facilitate the decision-making strategy for real world problems. This emerging field has potential impacts in several application areas such as healthcare, smart cities, transport, and marketing. There are very few works in the literature which focuses on applying data analytics for understanding environmental sciences [14]. The environment processes lead to the production of enormous data which provides more scope for adopting new techniques to analyse and extract useful patterns. Climate is one of the prominent factors which would be an interesting data for performing analysis. As nature faces changes day-by-day due to changing climate or other natural resources, there arises an increasing need to understand the environmental data.

4.1. Importance of Data Analytics for Natural Environment

In order to overcome the struggles involved in the environmental sciences, data analytics must be implemented effectively to analyse the volumes of data produced every day. Though there are several significant works involving data analytics for environmental data, still there is a need for more substantial effort to be carried out in this field. Data scientists are expected to extend their support for the welfare of the community. Data analytics is applied remarkably in other fields of science like physics and bioinformatics [14]. Proper means for integrating complex data to convert the raw data to useful knowledge is essential for applications like ecosystem services. Thus, it is clearly evident that there should be a synergetic association connecting data analytics and environmental sciences.

4.2. Challenges in adopting Data Analytics for Environmental data

Environmental data is collected from different sources about several naturally occurring processes. This data is extremely diverse in nature and varies in the storage and retrieval process. It is possible to have structured data that will be systematized according to the schema of the data model or unstructured data without any systematic organization. The data may be distributed across different geographic regions. Since the environmental changes are expected according to different spaces and time, finding the spatial and temporal insights from environmental data is a key challenge [14]. Thirdly, earth is a composite structure; hence the analysis of the data associated with it involves...
several inputs with different features interconnected with each other which will contribute towards producing the output. So, it is a great challenge to deal with this type of complexity, thus data analytics combined with machine learning and deep learning techniques works well to analyze and derive useful insights from complex environmental data.

5. Applications and use cases

5.1. Environment Water Management (EWM)

The recent studies from machine learning and deep learning in environment and water management relevant to remote sensing problem statements include object tracking, scene classification, object detection and classification, semantic segmentation and many image processing techniques. The objective of image/scene/ object classification (Earth Data Classification) is to fix the input images or objects or scenes to predefined class labels, so as to give a clear picture of the input images. In object tracking or object detection, the idea is to locate the desired objects like vehicle, airplane, and ship from numerous remote sensing data. Deep learning architectures like Convolution Neural Networks (CNN), Deep Belief Network (DBN), Faster Region-based Convolutional Network (Faster R-CNN), Transfer Learning, Stacked auto encoder (SAE) and Long short-term memory (LSTM) are widely utilised in solving problems related to EWM [5]. Also, AI serves as a potentially valuable tool for environment monitoring, where organisations seek to use data analytics for making smart decisions. However Tree based supervised models like random forest algorithms can help in making anticipations (predictions) in detecting the facilities are likely to fail an inspection and utilise the anticipated risk scores to recommend deputy surrogate inspections [6].

5.2. Hydrology and water resources

Conventionally, water resources management relies on physical models derived from experiential equations defining the behaviour and nature of water bodies. Machine learning models serve numerous applications in water resource management pertaining to climatology, water resources and hydrology. Supervised learning models can be used to monitor the water level in various water resources, precipitation level and rainfall management. Circulation models can be applied to monitor atmospheric circulations globally by predicting atmospheric variables. Due to their excellence in outlining the relationship between input and output with sparse interpretation of the physical process, neural networks have been progressively utilized in hydrological processes. Decision tree learning, neural network learning are widely utilized in building reasonably precise and efficient learning models for handling scenarios like downscaling, and rainfall management [7].

5.3. Environmental Remote Sensing

Machine learning algorithms have shown an eventual success in the remote sensing domain. Remote sensing is a branch of science that aids in collecting information about objects or areas from aircraft or satellites. Over a huge amount of data generated every day from earth observation and meteoric advancements in ML, strengthening opportunities for state-of-the-art techniques have come out to aid in monitoring earth environmental conditions. Remote sensing data contain images with complex and varied patterns, which basically differ from the natural images. Hence there is a need for deep learning algorithms, for extracting the low level features, middle level features and high level of features hidden in the remote sensing images. Those extracted features help in retrieval of environmental parameters, land cover survey and information prediction. Image processing and computer vision techniques can be utilised in processing remote sensing data [8]. When the input images are allowed for feature extraction, low level features like edge or curve information can be extracted by performing rotation and scaling operations. With these low level features, high level features are extracted with much more meaningful insights of data.
Figure 3. Deep learning framework for analysis of remote sensing data [9]

Deep learning architectures like Convolution Networks (ConvNets), sparse AutoEncoders (AE) and Deep Belief Networks (DBN) are considered to be hierarchical frameworks which can learn high-level feature representations in the deep layers automatically generated by the features learned in the shallow layers. In case of target recognition, when a new target is given as input, the frameworks can automatically extract the low level features and high-level features and then classification is performed by classification algorithm to find whether the proposal is the target or not [9].

5.4. Soil Science
Soil science basically deals with the development, ecology and taxonomy of soil. AI holds huge benefits for soil science in the form of soil resistance, soil nutrition, soil moisture, soil health and many. When it comes to soil moisture, precise prediction helps in flood and drought surveillance and forecasting. Estimation and classification of soil moisture relies on assorted factors, like vegetation, soil surface, vegetation and roughness [8]. AI supports in spotting the nutrient deficiencies and defects in the soil by applying suitable image processing algorithms on the images captured by the camera. Machine learning is used to analyse vegetation based on soil data and multispectral images captured from images, which then, on further analysis, helps in understanding of the soil nutrients in terms of individual composition of NPK. Three prominent categories of soil data collected in terms of the past (history - used as experience), present (fresh samples / current data that can be used to test) and future are mandatory for such analysis. In an attempt to predict the soil fertility, ML techniques use the past data to learn nutrient management and utilize the new crop samples to calculate the NPK composition of the soil. Also the chronology of crops is one such factor which is also helpful in predicting the NPK requirements for soil.

Soil penetration resistance is considered to be an important soil property that allows identifying areas with restrictions that occur because of compaction that impacts in mechanical impedance for decreased crop yield and spoils the growth of roots. The combination of Artificial Neural Networks (ANN) and Support Vector Machines (SVM) can be used to predict soil penetration resistance based on the composition of clay and sand, soil and particle density, sand and silt and soil volumetric moisture The ANN model is further connected to a multilayer perceptron (MLP) and suitable back propagation algorithms can be applied to predict resistance penetration [10].

5.5. Agriculture
It is highly evident that agriculture plays a significant role in improving the country’s economy. Automation in agriculture is one such emerging area that helps people to improve farming in terms of irrigation, crop establishment and monitoring, weeding to decrease the use of pesticides and disease identification in plants. CNN has a huge role in agriculture applications like plant disease identification. The images of infected plants and healthy plants are captured in a camera and stored in a database with labels. CNN is trained to learn the features of infected plants and healthy plants. The various phases involved in discovery of plant disease are: Data collection, Data cleaning (pre-
processing), image segmentation, feature removal and classification. When a new image is given as input, CNN can predict whether the new sample belongs to a healthy or unhealthy plant. In addition, CNN and its frameworks like AlexNet, LeNet, CaffeNet, GoogleNet are widely utilised in leaf disease prediction, land cover and crop type categorisation, soil and root segmentation, fruit counting and crop yield estimation [11].

![DL architecture for plant disease detection and classification](image)

**Figure 4.** DL architecture for plant disease detection and classification [12]

Chatbots or Digital Assistants are another milestone of AI that helps people in establishing communication by messaging in natural languages. Chatbots are widely employed in various services in websites and mobile applications. Similarly, agri chatbots are designed to help farmers by resolving their queries in terms of advice and recommendations. AI enabled Agricultural drones are effectively used in monitoring crops, spraying pesticides and herbicides, health assessment and soil analysis. AI based agriculture controls the surplus of water, herbicides and pesticides, preserves soil fertility and helps the farmers in improving the quality and increased productivity of crops [13].

It is inferred that Artificial Intelligence can be well utilized for several environmental science use cases. ES deals with huge amount of data in image formats; hence Convolution Neural Network can be the suitable deep learning framework to extract the hidden features and classification algorithms like Support Vector Machine can be used to classify the extracted features. Other frameworks like Recurrent Neural Network, Autoencoders, Boltzmann machines can be used to extract useful informal from data sources represented in formats other than images. The deep learning models are expected to perform more sophisticated as in when the data increases. Numerous datasets like geodata, environment, geospatial, water, and climate are available online, that can be tested across different DL frameworks and ML classification techniques. The self-learning capabilities of deep learning and visualization techniques of data analytics, acts as a foundation for automatic categorization of spatial and pictorial data involved in various perspective of environmental science.

6. Conclusion

The proposed paper gives an informative survey on environmental science and reveals that the dynamic impact of environmental change would be a great challenge and those could be faced and would be able to obtain the optimal solution with the emerging fields of Data Science, Machine Learning and Deep Learning. Environmental Data Science offers a profound understanding of how the
sensed environment data has to be processed and helps us to overcome the different environmental hazards. Gathering of data will continue to increase in accordance with our ability to analyse the data. As a part of the survey, the existing machine learning and deep learning algorithms were discussed, among those algorithms a few were implemented to overcome diverse environmental issues. Marching towards the smart world, there is an increasing need to ratify modern technologies, such as Artificial Intelligence (AI) that reforms the environmental risks. AI supports society in handling the high impact environmental risks which have a diverse effect on industries and human health. The most essential is to bring the various research regimens (Machine Learning, Deep Learning, Environmental Data Science, Health etc.) into a circle of discussion to frame the policies to address and resolve the environmental challenges.

References
[1] Karina Gibert, Jeffery S. Horsburgh, Ioannis N. Athanasiadis and Geoff Holmes 2018 Environmental Data Science
[2] Hsieh, William. 2009 Machine learning in the environmental sciences. Neural networks and kernels.
[3] Hsieh 2009 Cambridge: Cambridge University Press. pp. 274-317.
[4] Aakash Lamba, Phillip Cassey, Ramesh Raja Segaran Lian Pin Koh 2019 Deep learning for environmental conservation
[5] Alexander Y Sun, Bridget R Scanlon 2019 IOP Publishing Ltd, Environmental Research Letters 14(7).
[6] Hino, M., Benami, E. & Brooks, N. 2018 Nat Sustain 1 pp.583–588.
[7] Manish Kumar Goyal, Chandra S. P. Ojha, and Donald H. Burn 2017 Sustainable water resources management pp.165-178.
[8] Qiangqiang Yuan, Huanfeng Shen, Tongwen Li, Zhiwei Li, Shuwen Li, Yun Jiang, Hongzhang Xu, Weiwei Tan, Qianqian Yang, Jiwen Wang, Jianhao Gao, Liangpei Zhang 2020 Remote Sensing of Environment 241.
[9] L. Zhang, L. Zhang and B. Du 2016 IEEE Geoscience and Remote Sensing Magazine 4(2) pp. 22-40.
[10] Pereira, Tonismar dos S. et al 2018 Eng. Agríc. [online]. 38, pp.142-148.
[11] Kamilaris A, Prenafeta-Boldú F X 2018 the Journal of Agricultural Science pp.1–11.
[12] Saleem, Muhammad Hammad and Potgieter, Johan and Arif, Khalid Mahmood 2019 Plants 8(11) pp. 468.
[13] Tanha Talaviya, Dhara Shah, Nivedita Patel, Hiteshri Yagnik, Manan Shah 2020 Artificial Intelligence in Agriculture 4, pp. 58-73.
[14] Blair Gordon S., Henrys Peter, Leeson Amber, Watkins John, Eastoe Emma, Jarvis Susan, Young Paul J. 2019 Frontiers in Environmental Science 7 pp. 121.