IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

APPLICATIONS OF MACHINE LEARNING – REVIEW

R. Kiruthiga, T. Aruna Jyothi

[1]*Assistant Professor, IT, Matrusri Engineering College, Saidabad, Telangana State, India*

[2]*Assistant Professor, IT, Matrusri Engineering College, Saidabad, Telangana State, India*

Abstract:

Machine Learning has been exponentially growing in the past with applications in almost every field of science and engineering. It is a very useful and efficient technique to make sensible interpretations of the vast amount of data which is very difficult to carry out but the human mind. There are many fields in science where huge amounts of experimental data are systematically developed over several decades. It can be utilized to make new interpretations and understandings with the help of machine learning, which was otherwise not possible. This article reports advancements made in the fields of genetics, material science, and agriculture with the successful use of machine learning.

Introduction:

Machine learning, fundamentally is a method of using algorithms to understand the data and then further based on the understanding make predictions about the future where the outcome is unknown. These algorithms can learn from the existing data and then effectively perform tasks and make decisions. These algorithms can perform more efficiently and make accurate predictions even when the quantity of data is very large which is very difficult for human mind to do.

Machine learning can be classified into three broad categories, supervised and unsupervised and reinforcement learning [1]. In supervised learning, the algorithm generates a function which maps the input variables with output data. Unsupervised learning relates the input data with the unknown output and the learning proceeds with the availability of new data and the predictions improve. There are no labeled examples available with the unsupervised learning. Reinforcement learning algorithms learns a strategy of how to respond provided an observation of the world [2].

Supervised learning methods are the most widely used machine-learning methods. Supervised learning systems include spam classifiers of e-mail, face recognizers over images, and medical diagnosis systems for patients etc. These methods embody the function approximation problem, where the training data is collection of (x, y) pairs and the goal is to generate a prediction y* in reaction to a query x*[3]. The input x can be classical vectors or can be complex items such as documents, images, DNA sequences, or graphs. Similarly, many different kinds of output y have been studied. A great deal advancement has been made by concentrating on the simple binary classification problem where y takes one of two values (for example, "0" or "1").

Considerable progress has also been made on problems such as multiclass classification, multilabel classification, ranking problems, and general structured prediction problems [4].

Application of machine learning has become very successful in many areas and it is transforming many of the branches of science and engineering in ways that was never thought of in the past. The aim of this paper is to review the developments in the fields of genetics, material science and agriculture due to the assistance of machine learning.

1. Genetics and genomics

In genomics machine learning can be used to learn to recognize the locations of transcription start sites (TSSs) of a genome sequence [5]. The process usually takes place in three steps. First step is development of an algorithm that will lead to successful learning. Second step is to provide a large collection of TSS sequences and a list of sequences that are not TSSs to the algorithm. The annotation that represents whether a sequence is a TSS or not is called as label. The algorithm generates a model using the labeled sequences. Third step is to provide a novel, unlabeled sequences to the algorithm. The algorithm predicts labels (“TSS” or “not TSS”) for each sequence with the help of the model. The results can be verified using a test data set with known results or can be validated experimentally in the lab.

Machine learning methods were successfully employed in so many areas of genetics. It has been utilized widely for genomic sequencing. In addition, the algorithms can identify splice sites, promoters, enhancers, positioned nucleosomes etc [6-9]. Combining individual models that can identify the individual elements, it is also possible generate algorithms that can learn the location of the elements and annotate genes, including their full UTR/intron/exon structure [10]. Machine learning has also been used successfully to assign functional annotations to genes [11]. Predictive algorithms are capable of taking broad array of data types like genomic sequence, gene expression profiles, phenotypes, protein-protein interactions, synthetic lethality information etc as input.

Machine learning methods are capable of understanding the mechanisms of gene expression. Some techniques attempted to predict the expression of a gene only using DNA sequence as input [12]. However, others have also considered ChIP-seq histone modification [13] or TF binding [14] profiles. Advanced machine learning methods have attempted to model the expression of all genes together in a cell with the help of a neural network model [15].

2. Material Science

Rapid developments in automobile, energy, defence, and other fields have caused diverse requirements for high performance materials. There are three potential methods that can be used for new material development. They are empirical trial and error method, thermodynamic based computer simulation, and big data-driven machine learning. Conventional material development and analysis techniques, such as the empirical trial and error method is not viable because of the long development cycles, and high costs. New methods based on machine learning have emerged as a very useful technique for development of materials and property predictions with huge reduction in time and cost of development. It is one of the most efficient ways of replacing the recurring laboratory experiments. It has been utilized successfully to invent new materials, to investigate material properties, to understand quantum chemistry, and to design drugs. [16-20] Resources and tools for machine learning are abundantly available and are easy to access and hence its use in material science is increasing exponentially.

2.1 Material property analysis

2.1.1 Damage detection

Machine learning is very useful in complex problems in materials engineering. It gives precise predictions in material analysis and detection of material defects and flaws. Agrawal et al used machine learning methods to predict the fatigue strength of steel by understanding the relationship of various properties of the alloy and its composition and process parameters [21]. Twenty five features associated with fatigue strength have been used as input data in the study and tempering temperature was identified to be one that most influences fatigue strength. Machine learning algorithms, such as artificial neural networks, Support vector machine, and linear regression have resulted in high prediction accuracy with error rates <4%. Pavement cracks were detected in asphalt surface using three-dimensional images employing a neural network algorithm with an accuracy of 90.13%. A model for railway track inspection based on a fully convolutional network was developed [22]. An artificially illuminated car has been used to collect images of the track and 80% of the data was divided as training data set and the remaining 20% as testing data set. The model has been proved to have high accuracy.

2.1.2 Nano materials

Study of nano materials is an important field where machine learning has been used successfully [23-25]. Solubility of C60 has been successfully analyzed using machine learning techniques [16]. Machine learning has been successfully used to predict the toxicity of nano materials, to develop new nontoxic nano particles, to develop multi-structure/single-property relationships of nano particles, to understand quantum-mechanical properties of molecular systems, to study chemical reactions of nano materials and to explain kinetic systems [23-28].

2.1.3 Material properties

Conventionally, density function calculations are used for molecular property predictions. However this method is very time-consuming and hence researchers explored other methods such as machine learning with a motivation to obtain faster results. . Machine learning has been successfully used for the quick prediction of properties of molecules, compounds and materials with good accuracy [29-31]. A deep neural network model has been developed to automatically extract the physical, chemical interactions and similarities between elements and offer accurate prediction of material properties [32]. Another model based on a convolutional neural network that transforms raw data of compounds into 2D images and calculates properties such as toxicity, activity etc [33].

2.2 Discovering new materials

2.2.1 Element-oriented design

Machine learning is capable of predicting new compounds and their structures from the input composition data [34-36]. In a study, 209 new ternary oxides were identified using a probabilistic model which was prepared based on experimental crystal structure database [34]. A machine learning model was developed with the help of a large data base of DFT calculations and used to predict thermodynamic stability of arbitrary compositions. They were able to identify 4500 new stable materials through combined use of a heuristic and a machine learning based predictive model [36]. In another work, binary compounds were studied using machine learning techniques [35]. As a first step, 67 octet compounds were classified into distinct classes according to their crystal structures with the help of an unsupervised learning algorithm. In second step, crystal structures of

55 compounds were correctly identified by a supervised learning algorithm. Finally, melting points of 44 AB suboctet compounds were predicted with the help of a regression algorithm where a combination of 16 properties of the atoms of each binary compound has been used as input. Machine learning has been proved to be capable of discovering new compounds with good accuracy.

2.2.2 Drug development

Drug design is an important field where machine learning has been used successfully [37, 38]. Machine learning has been proved to be capable of handling the structural complexity of proteins which facilitates the invention of target drugs [37]. Modern strategies for new drug development include ligand-based drug design (LBDD) and structure-based drug design (SBDD) techniques [38]. These techniques make use of machine learning to simplify calculations and to build statistical validation models. LBDD use similarity search (2D or 3D information from one or more compounds for calculating similarity index which is then used for sorting compounds in the unknown database) and creation of a classification or regression model for predicting biological activity. LBDD is easy to carryout and has lower computational cost when compared to SBDD. SBDD is preferred when information about biological target structures is known. In SBDD, machine learning is used to predict the tertiary structures of receptors, solvent accessibility, disordered regions etc [38]. Machine learning algorithms like artificial neural network, support vector machine, decision tree, random forest, and k-nearest neighbor are widely used in both the drug discovery strategies.

3. Agricultural

Increasing population, reducing water resources, raising fertilizer and pesticide usage is putting huge pressure on agriculture. Precision farming is the new area that uses data intensive approaches to enhance agricultural productivity while minimizing input cost and environmental damage. A variety of different sensors are used to generate the data (Weather, Crop & Soil) which helps to get a better understanding of the operational environment enabling faster decision making. Machine learning (ML) with big data and high-performance computing has emerged as a potential technology to exploit in precision farming.

3.1 Yield Prediction

Yield prediction is a vital area in precision agriculture which plays a crucial role in yield mapping and crop management to get increased productivity. Machine learning has been successfully applied in yield prediction in several studies. A low cost method using support vector machine algorithm has been developed which automatically counts coffee fruits on a branch and classify them in to harvestable, not harvestable, and fruits with disregarded maturation stage [39]. An automated low cost harvesting system with Gaussian naive bayes algorithm has been developed with a machine vision system for shaking and catching cherries during harvest [40]. Yield mapping system for finding immature green citrus in a citrus grove has been developed using support vector machine algorithm and the results showed an accuracy of 80.4% [41]. Artificial neural network based model for estimating grassland biomass using multi-temporal remote sensing data has provided results with reasonable accuracy[42], Another study employing artificial neural network attempted yield prediction using satellite imagery, crop growth characteristics and soil data to get a more precise prediction of wheat yield [43]. Detection of tomatoes using images captured by an unmanned aerial vehicle (UAV) has been attempted using expectation maximization algorithm [44].

3.2. Disease Detection

Effective pest and disease control is serious concern in agriculture which affects crop quality and yield. Spraying pesticides uniformly over the cropping area is the most commonly used method in pest and disease control. It causes elevated financial loss and considerable environmental damage. Machine learning plays an important role in precision agriculture management, where chemical fertilizers and pesticides are given precisely to the affected areas instead of overall application. An artificial neural network based method has been successfully used to discriminate Silybum marianum plants infected by smut fungus Microbotyum silybum with an accuracy of 95.16% [45]. A new method based on image processing with the help of support vector machine algorithm has been proposed for the categorization of parasites and detection of thrips in strawberry greenhouse environment [46]. Detection of pathogen Fusarium fujikuroi (Bakanae disease) in rice has been attempted using support vector machine algorithm [47]. This method was less time-consuming compared to visual examination with naked eye and increased grain yield. A CNN-based method to classify healthy and diseased leaves in various plants for the disease detection has been devised with an excellent accuracy of 99.53% [48].

3.3. Weed Detection

Weed detection is an important problem in agriculture. It is very difficult to precisely identify the weeds. If left undisturbed, weeds pose high threat to the crop production because they not only consume the nutrients that are meant to be used by the crop plant, but also grow rapidly and covers the area leading to suppressed growth of crop plant. Hence weed management is a significant problem in agriculture. Machine learning can be used for accurate identification and discrimination of weeds and help sustainable agriculture. It can help destroy weeds and minimize the need for herbicides. A counter propagation (CP)-ANN method has been used on multispectral images captured by unmanned aerial vehicle for the identification of a weed, Silybum marianum [49]. An accuracy of 98.87% has been achieved. In another study, using ML techniques and hyperspectral imaging, an active learning system has been developed for weed discrimination in Maize. Ranunculus repens, Cirsium arvense, Sinapis arvensis, Stellaria media, Tarraxacum officinale, Poa annua, Polygonum persicaria, Urtica dioica, Oxalis europaea, and Medicago lupulina were considered as weed species. Very good prediction accuracy has been reported.

3.4. Crop Quality

Detection of crop quality characteristics precisely will help to improve product price and trim down waste. A new method for the identification and categorization of botanical and non-botanical foreign matter in cotton lint during harvesting has been devised [50]. The authors used support vector machine algorithm and aimed at improving the quality and reducing the fiber damage. Classification accuracy of 95% has been reported. In another study, a machine learning method was presented for the identification and differentiation of pears into deciduous-calyx or persistent-calyx categories [51]. They have used hyperspectral reflectance imaging and support vector machine algorithm in the study. An accuracy of 93.3% and 96.7% for deciduous-calyx and persistent-calyx respectively has been achieved.

4. Summary

Machine learning has proved to be a potential technique to make better understanding and predictions in many fields of science and technology. The use of various machine learning techniques in the fields of genetics, material science and agriculture has been reviewed. Machine learning is a very successful and useful technique for analysis and interpretation of data. It was used successfully in gene sequencing, material development, material defect identification, agriculture disease control, yield prediction etc. There is going to be ever increasing use of machine learning in all the fields with the development of advanced and more matured machine learning algorithms/techniques.

REFERENCES

[1] M. I. Jordan1, T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. Science Magazine. 349(2015):255-260.

[2] Leslie Pack Kaelbling , Michael L. Littman, Andrew W. Moore. Reinforcement Learning: A Survey. Journal of Artificial Intelligence Research 4 (1996):237-285.

[3] Konstantinos G. Liakos, Patrizia Busato, Dimitrios Moshou, Simon Pearson, Dionysis Bochtis. Machine Learning in Agriculture: A Review, Sensors 18(2018):2674.

[4] Maxwell W. Libbrecht William Stafford Noble. Machine learning in genetics and genomics Nat Rev Genet. 2015 June ; 16(6): 321–332.

[5] Jing Wei, Xuan Chu, Xiang-Yu Sun, Kun Xu, Hui-Xiong Deng, Jigen Chen, Zhongming Wei, Ming Lei. Machine learning in materials science. InfoMat. 1(2019):338–358.

[6] Degroeve S, Baets BD, De Peer YV, Rouz P. Feature subset selection for splice site prediction. Bioinformatics. 18(2002):S75–S83.

[7] Bucher P. Weight matrix description of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. Journal of Molecular Biology. 4(1990):563–578.

[8] Heintzman N, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature Genetics. 39(2007):311–318.

[9] Segal E, et al. A genomic code for nucleosome positioning. Nature. 44(2006):772–778.

[10] Picardi E, Pesole G. Computational methods for ab initio and comparative gene finding. Methods in Molecular Biology. 609(2010):269–284.

[11] Gene Ontology Consortium. Gene ontology: tool for the unification of biology. Nature Genetics. 25(2000):25–29.

[12] Beer MA, Tavazoie S. Predicting gene expression from sequence. Cell. 117(2004):185–198.

[13] Karlic R, R.Chung H, Lasserre J, Vlahovicek K, Vingron M. Histone modification levels are predictive for gene expression. Proceedings of the National Academy of Sciences of the United States of America. 107(2010):2926–2931.

[14] Ouyang Z, Zhou Q, Wong HW. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. Proceedings of the National Academy of Sciences of the United States of America. 106(2009):21521–21526.

[15] Friedman N. Inferring cellular networks using probabilistic graphical models. Science. 303(2004):799–805.

[16] Ruoff RS, Tse DS, Malbotra R, Lorents DC. Solubility of fullerene (C60) in a variety of solvents. Phys Chem. 97(1993):3379-3384.

[17] Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. Nature. 559(2018):547-555.

[18] Nantasenamat C, Isarankura-Na-Ayudhya C, Prachayasittikul V. Advances in computational methods to predict the biological activity of compounds. Expert Opin Drug Discov. 5(2010): 633-654.

[19] Nash W, Drummond T, Birbilis N. A review of deep learning in the study of materials degradation. npj Mater Degrad.2(2018): 37-49.

[20] Wang M, Wang T, Cai P, Chen X. Nanomaterials discovery and design through machine learning. Small Method.1900025(2019):1-7.

[21] Agrawal A, Deshpande PD, Cecen A, et al. Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters. IMMI. 2014 3:8

[22] Gibert X, Patel VM, Chellappa R. Deep multitask learning for railway track inspection. IEEE Trans Intell Transport. 18(2017): 153-164.

[23] Oh E, Liu R, Nel A, et al. Meta-analysis of cellular toxicity for cadmium-containing quantum dots. Nat Nanotechnol. 11(2016):479-486.

[24] Sun B, Fernandez M, Barnard AS. Machine learning for silver nanoparticle electron transfer property prediction. J Chem Inf Model. 57(2017):2413-2423.

[25] Zhu Q, Samanta A, Li B, Rudd RE, Frolov T. Predicting phasebehavior of grain boundaries with evolutionary search and machine learning. Nat Commun.9(2018):467-476.

[26] Pyrgiotakis G, Kundakcioglu OE, Pardalos PM, Moudgil BM. Raman spectroscopy and support vector machines for quick toxicological evaluation of titania nanoparticles. J Ranman Spectrosc.42(2011):1222-1231.

[27] Chandana Epa V, Burden FR, Tassa C, et al. Modeling biological activities of nanoparticles. Nano Lett.12(2012):5808-5812.

[28] Zhou Z, Li X, Zare RN. Optimizing chemical reactions with deep reinforcement learning. ACS Cent Sci.3(2017):1337-1344.

[29] Pilania G, Wang C, Jiang X, Rajasekaran S, Ramprasad R. Accelerating materials property predictions using machine learning. Sci Rep.3(2013):2810-2816.

[30] Schütt KT, Glawe H, Brockherde F, et al. How to represent crystal structures for machine learning: towards fast prediction ofelectronic properties. Phys Rev B.89(2014):205118.

[31] Pilania G, Mannodi-Kanakkithodi A, Uberuaga BP, Ramprasad R, Gubernatis JE, Lookman T. Machine learning bandgaps of double perovskites. Sci Rep.6(2016):19375.

[32] Jha D, Ward L, Paul A, et al. ElemNet: deep learning the chemistry of materials from only elemental composition. Sci Rep. 2018;8:17593.

[33] G.B. Goh, C. Siegel, A. Vishnu, et al. Chemception:a deep neural network with minimal chemistry knowledge matches the performance of expert-developed QSAR/QSPR models.

[34] Hautier G, Fischer CC, Jain A, Mueller T, Ceder G. Finding nature's missing ternary oxide compounds using machine learningand density functional theory. Chem Mater.22(2010):3762-3767.

[35] Saad Y, Gao D, Ngo T, Bobbitt S, Chelikowsky JR, Andreoni W. Data mining for materials: computational experiments with AB compounds. Phys Rev B. 85(2012):104104.

[36] Meredig B, Agrawal A, Kirklin S, et al. Combinatorial screening for new materials in unconstrained composition space with machine learning. Phys Rev B. 89(2014):094104.

[37] Wallach, I., Dzamba, M. Heifets, A. AtomNet a deep convolutional neural network for bioactivity prediction in structure based drug discovery.

[38] Lima AN, Philot EA, Trossini GH, et al. Use of machine learning approaches for novel drug discovery. Expert Opin Drug Discov.11(2016):225-239.

[39] Ramos, P.J.; Prieto, F.A.; Montoya, E.C.; Oliveros, C.E. Automatic fruit count on coffee branches using computer vision. Comput. Electron. Agric. 137(**2017**): 9–22.

[40] Amatya, S.; Karkee, M.; Gongal, A.; Zhang, Q.; Whiting, M.D. Detection of cherry tree branches with full foliage in planar architecture for automated sweet-cherry harvesting. Biosyst. Eng. 146(**2015):** 3–15.

[41] Sengupta, S.; Lee,W.S. Identification and determination of the number of immature green citrus fruit in a canopy under different ambient light conditions. Biosyst. Eng. 117(**2014)**: 51–61.

[42] Ali, I.; Cawkwell, F.; Dwyer, E.; Green, S. Modeling Managed Grassland Biomass Estimation by Using Multitemporal Remote Sensing Data—A Machine Learning Approach. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 10(**2016)**: 3254–3264.

[43] Pantazi, X.-E.; Moshou, D.; Alexandridis, T.K.; Whetton, R.L.; Mouazen, A.M. Wheat yield prediction using machine learning and advanced sensing techniques. Comput. Electron. Agric. 121(**2016**): 57–65.

[44] Senthilnath, J.; Dokania, A.; Kandukuri, M.; Ramesh, K.N.; Anand, G.; Omkar, S.N. Detection of tomatoesusing spectral-spatial methods in remotely sensed RGB images captured by UAV. Biosyst. Eng. 146(**2016)**:16–32.

[45] Pantazi, X.E.; Tamouridou, A.A.; Alexandridis, T.K.; Lagopodi, A.L.; Kontouris, G.; Moshou, D. Detection of Silybum marianum infection with Microbotryum silybum using VNIR field spectroscopy. Comput. Electron. Agric. 137(**2017):** 130–137.

[46] Ebrahimi, M.A.; Khoshtaghaza, M.H.; Minaei, S.; Jamshidi, B. Vision-based pest detection based on SVM classification method. Comput. Electron. Agric. 137 (**2017**): 52–58.

[47] Chung, C.L.; Huang, K.J.; Chen, S.Y.; Lai, M.H.; Chen, Y.C.; Kuo, Y.F. Detecting Bakanae disease in rice seedlings by machine vision. Comput. Electron. Agric. 121(**2016):** 404–411.

[48] Ferentinos, K.P. Deep learning models for plant disease detection and diagnosis. Comput. Electron. Agric. 145(**2018):** 311–318.

[49] Pantazi, X.E.; Tamouridou, A.A.; Alexandridis, T.K.; Lagopodi, A.L.; Kashefi, J.; Moshou, D. Evaluation of hierarchical self-organising maps for weed mapping using UAS multispectral imagery. Comput. Electron. Agric. 139(**2017):** 224–230.

[50] Zhang, M.; Li, C.; Yang, F. Classification of foreign matter embedded inside cotton lint using short wave infrared (SWIR) hyperspectral transmittance imaging. Comput. Electron. Agric. 139(**2017):** 75–90.

[51] Hu, H.; Pan, L.; Sun, K.; Tu, S.; Sun, Y.; Wei, Y.; Tu, K. Differentiation of deciduous-calyx and persistent-calyx pears using hyperspectral reflectance imaging and multivariate analysis. Comput. Electron. Agric. 137(**2017):** 150–156.