International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 14, Number 7 (2021), pp. 700-707

Utilization of big data and Machine learning in image processing

Mohamed Ali Mohamed*, Ibrahim Mahmoud El-henawy, Ahmed Moustafa

Faculty of Computers and Informatics, Zagazig University, Zagazig, Egypt

Abstract

We are surrounded by a torrent of data from sensors, mobile devices, social media, transactions, and online data, among other sources. This massive volume of data is rising at a breakneck pace as the internet, e-commerce, and technology, most notably the Internet of Things, advance (IoT). IoT is a term that refers to the process through which computers, smart devices, and other things that generate data are linked through a network in order to transport the data. Thus, data is produced and updated on a continuous basis to reflect changes in all human fields and activities. This exponential growth of data has spawned a new phrase and idea known as big data. Big data is required to provide insight into the relationships between objects, to forecast future patterns, and to provide more information to decision makers. However, the primary challenge currently is how to collect and evaluate massive amounts of diverse and complicated data in a timely manner. This article presents an overview of the big data ecosystem and the issues that have arisen that should be addressed when designing or implementing big data systems. Machine learning and deep learning are the most widely utilized ways of comprehending and analyzing large amounts of data and obtaining the essential knowledge in relevant sectors or applications. We concentrate on three applications in this paper, namely, astronomical images, medical images, and satellite images; all of these three topics are considered big data application areas.

Keywords: Big data, machine learning, image processing

I. INTRODUCTION

Big data is a phrase that refers to big data sets with a huge, diverse, and complicated structure that provides challenges in terms of storing, processing, and displaying them for subsequent processes or outcomes. Statista estimates that 74 zettabytes of data will be generated in 2021 [1]. According to IDC, the cumulative amount of the world's data will reach 175ZB by 2025, representing a compounded annual growth rate of 61% [2]. The technique of analyzing enormous volumes of data in order to uncover hidden patterns and connections is referred to as big data analytics. The study of big data is at the heart of contemporary research and industry. This data is created as a result of online transactions, emails, movies, audio files, photos, click streams, logs, postings, search queries, health records, social networking interactions, scientific data, sensors, and mobile phones and their apps. They are kept in databases, which rapidly develop in size and complexity, making them challenging to collect, create, store, manage, distribute, analyze, and display using standard database software tools [3].

The term "big data" refers to datasets that cannot be viewed, gathered, managed, or processed in a reasonable amount of time using typical IT and software/hardware methods. The exponential growth of data in the big data era creates enormous hurdles in terms of data capture, storage, administration, and analysis. Historically, data management and analysis solutions have been built on relational database management systems (RDBMS). However, such RDBMSs are not applicable to semi-structured or unstructured data. Additionally, RDBMSs are consuming an increasing amount of pricey hardware. It seems that standard relational database management systems (RDBMSs) are incapable of handling the massive amount and heterogeneity of big data [4].

Researchers, companies, and people have come up with a variety of definitions for big data. The three V's (Volume, Velocity, and Variety) are the most frequent definitions of big data. Organizations and scholars have introduced additional qualities throughout time. These features provide researchers and practitioners with a research horizon to work with in order to efficiently handle big data. The whole field of big data study centers on these features in order to successfully handle and exploit massive data. The following short list defines the majority of big data features, which may be added or altered as thorough knowledge of big data advances and new concerns arise [5]:

- Volume: The size of the data, the quantity of data gathered and kept. The data size units are TB and PB.
- Velocity: Data Transfer Rate, the rate at which data is sent between a source and a destination.
- value: simply refers to the commercial value that may be extracted from big data.
- Variety: Variation in the Types of Data, At the receiving end, various types of data such as images, videos, and audio are received.
- Veracity: Data Quality, Accurate analysis of acquired data is almost useless unless it is accurate.
- Validity: The correctness or accuracy of the data used to extract the outcome in the form of information.
- Volatility: Refers to the amount of data saved and how long it will be relevant to the consumer.
- Visualization: is a method of expressing abstract data (data process/data act).
- Virality: Is the amount of data that a user broadcasts and shares with others, and then receives for use by them.

- Viscosity: Event Lag, the time interval between the occurrence of the event and the description of the event.
- Variability: Differentiation of data, data is continually arriving from a variety of sources, and the efficiency with which it distinguishes between irrelevant and critical data is critical.
- Venue: Various Platforms, Numerous sorts of data come from a variety of sources through a variety of platforms, including personnel systems, private and public clouds, and so on.
- Vocabulary: Data Terminology includes terms such as data model, data structure, and so on.
- Vagueness: Is the indeterminacy of existence inside a data set. It refers to the actuality included within information that implies little or no consideration for the message each may communicate.
- Verbosity: The duplication of information accessible from several sources.
- Versatility: The capacity of big data to be utilized in a variety of ways depending on the context.
- Voluntary: The willful availability of large amounts of data to be utilized appropriately.
- Complexity: Data Correlation Data arrives from a variety of sources, and it is required to identify changes in data, whether minor or huge, in comparison to previously arriving data, so that information may be accessed promptly.

Big Data Analytics is the act of analyzing massive data sets, including a wide variety of data types (big data), in order to identify any hidden patterns, unexplored correlations, market trends, customer preferences, and other critical business information. Following the findings, these results might then be used to help promote marketing, generate extra income, provide improved customer service, raise operational efficiency, hold up to rivals, and provide other business benefits. Incorporating data science and data-mining techniques, the big data analytics program's main goal is to empower data scientists, predictive modelers, and other analytics professionals to analyze large volumes of transactional data as well as other types of data that could be untouched by more traditional Business Intelligence (BI) applications to help organizations make better business decisions. Most Internet activity (whether using your computer, mobile device, or any other device that connects to the Internet) is logged and can be captured with help of web server logs and Internet click stream data, social media content and social network activity reports, text from customer emails and survey replies, mobile phone call detail records, and machine data acquired by sensors and linked to the Internet of Things [6].

II. BIG DATA LIFECYCLE

The framework for big data lifecycle management consists of four phases: data collection, data storage, data processing and analytics, and knowledge generation. Data from many sources is collected in a variety of forms, including structured, semi structured, and unstructured. The obtained data is saved and prepared for use in the following step during the data storage phase (data analytics phase). Data processing analysis is conducted after data is collected and stored in safe storage systems to provide relevant information. Data mining techniques including clustering, classification, and association rule mining are utilized in this phase. Data miners collect sensitive data using advanced machine learning and deep learning algorithms. Finally, the analytics step generates fresh data and valuable insights for decision-makers to employ [7].

The data life cycle management process is critical in big data analytics, since it entails multiple processes. Big data analytics encompasses a variety of technological features, including searching, mining, analysis, and usability, all of which are necessary for the development of any application. Today, data life cycle management under the umbrella of big data analysis is critical to identifying and streamlining marketing tactics for popularizing their products in contrast to other sectors. Apart from big data analytics, augmented analysis is a growing trend these days, in which IoT and machine learning (ML) are combined to create, develop, and share analyses. Several data-driven techniques for corporate application, such as customer service and document management, as well as smarter city planning, conversational analysis, and natural language programming (NLP), have been proven effective [8] and image analysis which will be presented in the next sections.

Machine learning (ML) is stated as an instance of a computer program that learns from experience, when used for a specific task and measure. This happens when a program's performance, as assessed by a measure, improves over time as a result of having been applied to that task. Machine learning is concerned with developing methods or programs that allow computer systems to learn on their own from data. Data, models, and learning are the three main components of a machine learning system. Fitting data to a model and training a function approximation technique (Hypothesis) based on performance criteria are at the heart of the procedure. A machine learning system's fundamental design may be broken down into four steps: (1) selecting the training experience; (2) selecting the target function; (3) selecting a representation for the target function; and (4) selecting a function approximation methodology. The training data offers the training experience from which the ML system will learn in a normal ML procedure. The goal function that determines precisely what kind of information will be learnt and how it will be utilized by the performance program is referred to as the model performance. A model representation (Learning algorithm) will be presented to explain the goal function once the target function has been defined (Model performance). Following the instructions and performing according to a set of performance criteria, a function approximation approach (Hypothesis) will be learnt from the training cases. Machine learning consists of three major categories: supervised learning (classification or regression), unsupervised learning (clustering or association), and reinforcement learning (reward-based) [9].

Deep learning is a family of machine learning algorithms that: (1) extract and manipulate features through a cascade of

numerous layers of nonlinear processing units. Each subsequent layer takes the output of the preceding layer as input, and (2) learns numerous layers of representations that correspond to various degrees of abstraction; the layers constitute a hierarchy of ideas [10]. Neural network with convolutional layers (CNN) The most extensively used deep learning techniques are the recurrent neural network (RNN), denoising autoencoder (DAE), deep belief networks (DBNs), and Long Short-Term Memory (LSTM). Deep learning techniques are rapidly developing. Several of them have evolved to become highly specialized in a single application area [11].

Deep learning has been proved to outperform previous machine learning approaches for a certain class of problems, enabling the building of models that perform as well as or better than humans. This revolution was fueled by the rising availability of massive datasets and the computational power of graphics processing units (GPUs), even as neural network learning algorithms and mathematics advanced. Deep learning is a term that refers to a deep neural network, a particular configuration in which neurons are grouped in consecutive layers. Increased layering improves the expressive capability and performance of these approaches and may result in a greater degree of abstraction. Deep learning is the most sophisticated machine learning technology available today for a wide range of high-level tasks and applications, particularly those requiring huge structured datasets [12].

III. CHALLOENGES OF BIG DATA FOR IMAGE ANALYSIS

Astrophysics and cosmology are data-rich fields that provide excellent candidates for image analysis. Data quantities formerly required to conduct full surveys may now be obtained in a single night, and real-time analysis is often needed. For example, the Sloan Digital Sky Survey telescope generates 200 GB of data per night, and so far, almost a million field photos have been recorded, revealing about 200 million galaxies and countless more stars. Upcoming polls will collect far more data. Thus, contemporary astronomy needs an understanding of huge data, most notably extremely efficient machine learning and image analysis methods. However, scalability is not the only issue: Astronomy applications address a number of current machine learning research challenges, including learning from biased data and coping with label and measurement noise [13].

Over the last 20 years, developments in medical image analysis have boosted the usage of image analysis algorithms in commercial products. Simultaneously, new industrial difficulties emerge. To begin, there is a need for more general image analysis tools that are easily adaptable to a particular therapeutic purpose. The second, closely connected difficulty is efficient ground truth annotation production in order to meet rising needs for robustness and dependability in commercial systems, as well as growing needs when applying machine learning. Thirdly, techniques for interpreting more diverse image data will allow a broader adoption of MRI post-processing applications as well as new applications, such as those involving big data and analytics. The fourth difficulty is in developing comprehensive anatomy and organ models with little user interaction. Apart from their immediate use in a wide variety of applications, these models are critical for advancements such as the Virtual Physiological Human, which might connect medical image analysis to semantic and reasoning technology. These difficulties are in addition to the continuous need for more accurate, trustworthy, and quicker algorithms, as well as for algorithms tailored to particular applications [14].

III.I Image analysis of big data methods for astronomy images

The term "photometric redshift" refers to a redshift determined only via the use of medium or wide band photometry or imaging. Photometric redshifts are most often computed using the colors of galaxies in three or more filters. Artificial neural networks were utilized to estimate photometric redshift. The relationship between photometry and redshift is learned by an ANN using a suitable training set of galaxies for which the redshift is known [15].

For fine-grained galaxy morphology prediction, a convolutional neural network was utilized, with a new design that takes use of rotational symmetry in the input pictures. The network was trained using Galaxy Zoo 2 data and is capable of consistently predicting many features of galaxy shape straight from raw pixel data, without the need for any type of customized feature extraction. It is capable of autonomously annotating enormous quantities of pictures, allowing unparalleled quantitative investigations of galaxy morphology [16].

The autohrs in [17] employed an unsupervised Kohonen-maps approach to interpret enormous quantities of galaxy data automatically, generating a series of prototypes. This generated model may be used to view both the supplied galaxy data and previously undiscovered pictures. Kohonen-maps are a simple but efficient approach for reducing the dimension of data by projecting it to a two-dimensional map. On $p \in P$ in the latent space has a derived prototype after mode training. Astronomers may use this approach to evaluate massive volumes of data. Apart from getting a classification scheme, one can also identify outliers effectively. These are the things that need professional analysis [17].

In astronomy, where datasets are often vast and feature space dimensionality is generally low, sample selection bias is a prevalent concern. A closest neighbor density ratio estimator paired with a model selection criteria that is unbiased under covariate shift is used to choose the neighborhood size in this case. Because of the rigorous hyperparameter selection, the final method is simple and resilient [18].

Two major astronomical issues have been addressed using a mix of genetic algorithm (GA) and support vector machine (SVM) machine learning methods: star–galaxy separation and photometric redshift estimates of galaxies in survey catalogs. GA was used in the first step to choose significant features, followed by SVM parameter optimization in the second stage to arrive at the optimal set of parameters for classifying or regressing while avoiding overfitting. This approach was used to partition the Pan-STARRS1 data into stars and galaxies. Additionally, this method is used to derive photometric

redshifts for galaxies included in the COSMOS bright multi-wavelength data collection. This illustrates that the combination of GA and SVM is very efficient for solving problems with a high number of dimensions [19].

Two major astronomical issues have been addressed using a mix of genetic algorithm (GA) and support vector machine (SVM) machine learning methods: star–galaxy separation and photometric redshift estimates of galaxies in survey catalogs. GA was used in the first step to choose significant features, followed by SVM parameter optimization in the second stage to arrive at the optimal set of parameters for classifying or regressing while avoiding overfitting. This approach was used to partition the Pan-STARRS1 data into stars and galaxies. Additionally, this method is used to derive photometric redshifts for galaxies included in the COSMOS bright multi-wavelength data collection. This illustrates that the combination of GA and SVM is very efficient for solving problems with a high number of dimensions [19].

Support vector machines (SVM), hybrid SVM–K closest neighbor (SVM–KNN), AdaBoost, and asymmetric AdaBoost have all been employed to address the quasar-star classification issue. SVM and SVM–KNN had been attempted earlier, but their performance (and rationale for use) have been enhanced by including bias management approaches. According to [20], neither AdaBoost nor asymmetric AdaBoost has been attempted to tackle this issue before. In comparison to SVM, asymmetric AdaBoost has a higher processing efficiency. In comparison to SVM, asymmetric AdaBoost is an excellent option for a classifier because of its high accuracy, rapid speed, and ease of parameter modulation [20]. The summary of image processing analysis of big data methods for astronomy images is shown in Table 1.

Table 1. Summary of image processing analysis of big data methods for astronomy images

Reference	Problem	ML/DL techniques
[15]	Photometric redshift estimation	Artificial Neural Networks
[16]	fine-grained galaxy morphology prediction	convolutional neural network
[17]	Visualize the given galaxy data and classify so far unseen images.	Kohonen-maps
[18]	Sample selection bias	Nearest neighbor density
[19]	star–galaxy separation and photometric redshift estimation of galaxies	genetic algorithm & support vector machine
[20]	star-quasar classification	SVM, SVM–KNN, AdaBoost and asymmetric AdaBoost

III.II Image analysis of big data methods for medical images

On approximately 150,000 images of chest radiographs (x-rays), a convolutional neural network deep learning methodology was used to accomplish high-fidelity clinically relevant image categorization. Manual annotations classified the source images as frontal or lateral and separated them into training, validation, and test sets at random. Standard image alterations were used to enrich both the training and validation sets. Following that, the trained networks were fine-tuned using the original and enhanced radiological images. The Youden Index was used to establish a binary cutoff for frontal or lateral categorization and its accuracy was determined. Nearly 100% of the test dataset was categorized properly using the binary classification approach [21].

To detect images containing lung nodules, a huge quantity of medical image data was analyzed using a convolutional autoencoder deep learning methodology based on data-driven feature learning. This network was trained unsupervised using a huge quantity of unlabeled patch data, with just a tiny quantity of labeled data utilized to fine-tune the network topology. This model is used to recognize, classify, and compare lung nodules, and it effectively addresses the concerns of time-consuming ROI (region of interest) labeling and inadequately labeled data [22].

The developed approaches and structures for the entire processing of biological image data, based on big data technologies, are shown. Two image categorization architectures are proposed, based on the Hadoop and Spark frameworks. These systems provide a comprehensive big data workflow for biomedical image processing, including stages for data storage and analysis. Support vector machines were selected for the classification stage, which is one of the phases of big data analysis [23].

As per the two primary data formats for medical image and text data. Two models of deep learning were utilized. The first AutoEncder deep learning model was created, and it is capable of pre-training the network in advance, hence reducing computational time and resource usage. This model was applied to an image of brain MRI medical images. Due to this, the strategy may be easily applied to many kinds of medical image analysis and processing, which is crucial for boosting the accuracy of disease identification. Second, a deep learning model was created that incorporates a 3D convolutional neural network with spatial pyramid pooling (3D CNN-SPP). The three-dimensional convolution structure of the model preserves the timing characteristics of different data while also extracting internal features, while the spatial pyramid pooling structure can process input data of any length, and this is essential for effective data analysis for the patient's future in identifying and controlling diseases based on their risk of becoming intelligent [24].

To address the difficulty of learning representative visual features in medical image analysis, an unsupervised deep learning framework called the convolutional sparse kernel network (CSKN) was devised to learn semantic high-level features from unlabeled medical images. CSKNs employ kernel maps to describe image characteristics rather than the

traditional hierarchical design of CNNs. By modeling invariance, a kernel map is utilized to comprehend the local geometry of image data. The CSKNs competed with supervised CNNs by using three major public datasets (IRMA X-ray dataset, Image CLEF dataset, and ISIC dataset). CSKN's methodology established the capability of characterizing medical image characteristics using vast datasets of unlabeled medical data and provides access to the huge number of unannotated data accessible in medical imaging archives [25].

On skin image datasets, the following five classifiers were used to identify and classify skin tissue (melanoma and nevus): Linear Model Logistic Regression, Gradient Boosting (Stochastic Gradient Boosting), SVM Linear SVC (Support Vector Machine Linear—Support Vector Classification), Linear Model Stochastic Gradient Descendent (Linear Model SGD), and SVM SVC (Support Vector Machine—Support Vector Clustering). The logistic regression model was the most accurate at 97.00 percent [26]. The summary of image processing analysis of big data methods for medical images is shown in Table 2.

Table 2. summary of image processing analysis of big data methods for medical images.

Reference	Problem	ML/DL techniques
[21]	Chest radiographs(x-ray) clinically relevant image classification	convolutional neural network
[22]	Classification of pulmonary nodules for the medical images.	Convolutional Autoencoder Neural Network
[23]	Biomedical image classification.	support vector machines
[24]	Diagnosis of diseases.	AutoEncder neural networks & 3D CNN-SPP
[25]	Characterize medical image features.	Convolutional Sparse Kernel Network
[26]	Analyzing medical images of skin with melanoma and nevus.	Linear Model Logistic Regression Gradient Boosting SVM Linear Model Stochastic Gradient Descendent.

III.III Image analysis of big data methods for satellite images

The subsequent study includes the application of a Random Forests machine learning model to remote sensing and satellite images. The process is used to assist with a typical challenge of identifying and planning rural development locations, as well as prioritizing communities for various sorts of rural development. In this scenario, Random Forests models were employed as both a classifier and a regressor to analyze the image. To begin, the authors in [27] trained the random forest classifier on the distribution of colors in a small area surrounding the training set's roof locations in order to distinguish between metal and thatch roofs, and then used it to obtain soft classifications for each roof location obtained via template matching. A second random forest regression model was trained to predict the total number of roofs in the image patch and the percentage of metal roofs. Separate classifiers and regression functions for distinct satellite collection situations (i.e., foggy, wet, and dry) produced higher accuracy than using the same model for all circumstances [27]. This is because the visual appearance of images varies significantly in various situations.

Two methodologies were utilized to analyse and categorize huge satellite images acquired from the SPOT-5 satellite. The first way is to use a k-means model. The second way is to utilize the Hadoop MapReduce framework and HDFS (Hadoop distributed file system) to store and analyse massive satellite images by integrating remote sensing image processing tools from OTB (The Orfeo ToolBox library) into MapReduce. The results demonstrate that when images are grouped into five thematic categories (urban, water, forest, bare soil, and vegetation), the suggested approach can provide a more accurate result even when dealing with large amounts of data [28].

Remote sensing is a technique for gathering information from a for large remote-sensing image categorization, the distributed convolutional neural networks (RS-DCNN) technique was developed. To minimize the time, it takes to analyze large images, the suggested framework was built on top of the Apache Spark framework for parallel processing. The suggested method consists of two steps: 1) preparing the training dataset by separating large satellite images into tiny images and then using a supervised classification technique such as Maximum Likelihood, 2) classifying large satellite images using a distributed CNN. Asynchronous Distributed Stochastic Gradient Descent (ADSGD) is used to spread the CNN algorithm execution across the large data cluster to guarantee parallelism for image classification. Experimentation is carried out using a number of large RS images obtained from SPOT-6/7 sensors and relating to five different areas in Saudi Arabia. Six types of land cover have been identified: water, urban, soil, plants, mountains, and roads. The suggested RS-DCNN exhibits good performance, notably in terms of training time and classification accuracy [29].

To minimize artifacts, a tree-based CNN is employed to classify semantically segmented satellite images. Various abnormalities caused by atmospheric conditions often impact satellite images during the image capturing process. The

removed artifacts from the satellite images aid in the process of feature extraction. The removed artifacts from the satellite images aid in the process of feature extraction. The NWPU-VHR-10 dataset was used to classify objects using a tree-based CNN. This collection contains around 60,000 images [30].

Random Forest (RF) Machine Learning Algorithms were taught using a huge quantity of field-collected reference training data as well as photographs with spatial resolutions ranging from sub-meter to five meters (VHRI). The Random Forest classifier was used to create crop and non-cropland categories. For five agro-ecological zones in South Asia, accuracy was assessed using independent reference validation data and error matrices [31].

MLCs (maximum likelihood classifiers) are statistical supervised learning techniques that are mostly binary classifiers. To calculate the damage assessment index for satellite images, MLCs are selected. The optical images of land, built up land, sand, water body, fallow land, degraded scrub, and land with scrub are mostly used to identify and identify these MLCs. The images were taken from Digital Globe's free repository. The damage assessment index clearly shows the extent of the destruction and may give significant information to planners in guiding rescue and research efforts. This kind of scale-based quantification facilitates straightforward comprehension of the extent of harm and may benefit in decision-making. The importance of the index may be seen in the fact that it provides a comprehensive perspective of the damage by comparing pre-and post-event images [32]. The summary of image processing analysis of big data methods for satellite images is shown in Table 3.

Table 3. summary of image processing analysis of big data methods for satellite images.

Reference	Problem	ML/DL techniques
[27]	satellite image analysis for selecting poor villages to rural development	Random Forests
[28]	classification of large-scale remote sensing image related to Spatial big data	K-Means
[29]	classify large volumes of satellite images speedily and efficiently	Distributed Convolutional-Neural-Networks
[30]	Diminish the artifacts of satellite images create by atmospheric conditions.	Tree-based convolutional neural network
[31]	Identify croplands versus non-croplands of satellite images from NASA's LP DAAC	Random Forest
[32]	determine damage assessment index for satellite images in disaster management	Maximum likelihood Classifier

IV. CONCLUSION

Daily, corporate demands and requirements in all spheres of life get more complex, resulting in an enormous volume of raw data in a variety of forms, sources, kinds, and sizes, together referred to as big data. Big data contains a variety of qualities or issues that may grow or alter as lifeareas evolve. The framework for big data lifecycle management consists of four phases: data collection, data storage, data processing and analytics, and knowledge generation. The big data framework is applicable to a wide variety of disciplines, including image processing, education, health, finance, telecommunications, and social media. The article illustrates the junction of three fields of study (big data – image processing – machine learning). This article provides an overview of the combination of big data and machine learning and how they are used in a variety of image analysis applications to address a variety of difficulties.

REFERENCES

[1] Nisioti A, Mylonas A, Yoo PD, Member S, Katos V. From Intrusion Detection to Attacker Attribution : A Comprehensive Survey of Unsupervised Methods. IEEE Commun Surv Tutorials. 2018;PP(c):1.

[2] Hafizah S, Ariffin S, Muazzah N, Latiff A, Khairi MHH, Ariffin SHS, et al. A Review of Anomaly Detection Techniques and Distributed Denial of Service (DDoS) on Software Defined Network (SDN). Technol Appl Sci Res [Internet]. 2018;8(2):2724–30.

[3] https://www.statista.com/statistics/871513/worldwide-data-created

[4] https://www.idc.com/getdoc.jsp?containerId=prUS47560321

[5] Sagiroglu, S., & Sinanc, D., 2013. Big data: A review. 2013 International Conference on Collaboration Technologies and Systems (CTS). doi:10.1109/cts.2013.6567202

[6] Min Chen, Shiwen Mao,Yunhao Liu "Big Data: A Survey", Springer Science+Business Media New York 2014.

[7] G. Kapil, A. Agrawal and R. A. Khan, "A study of big data characteristics," 2016 International Conference on Communication and Electronics Systems (ICCES), 2016, pp. 1-4, doi: 10.1109/CESYS.2016.7889917.

[8] Kuchipudi Sravanthi et al," Applications of Big data in Various Fields" / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (5) , 2015, 4629-4632.

[9] Yazan Alshboul, Raj Nepali, Yong Wang," Big Data LifeCycle: Threats and Security Model", Twenty-first Americas Conference on Information Systems, Puerto Rico, 2015 (AMCIS 2015) (6 VOLS). ISBN: 9781510814936.

[10] Rahul, K., & Banyal, R. K. (2020). Data Life Cycle Management in Big Data Analytics. Procedia Computer

Science, 173, 364–371. doi:10.1016/j.procs.2020.06.042

[11] Tengxiang Su, Haijiang Li, Yi An,A BIM and machine learning integration framework for automated property valuation, Journal of Building Engineering, Volume 44,2021,102636, ISSN 2352-7102, https://doi.org/10.1016/j.jobe.2021.102636.

[12] Zhang, W. J., Yang, G., Lin, Y., Ji, C., & Gupta, M. M. (2018). On Definition of Deep Learning. 2018 World Automation Congress (WAC). doi:10.23919/wac.2018.8430387

[13] Mosavi A., Ardabili S., Várkonyi-Kóczy A.R. (2020) List of Deep Learning Models. In: Várkonyi-Kóczy A. (eds) Engineering for Sustainable Future. INTER-ACADEMIA 2019. Lecture Notes in Networks and Systems, vol 101. Springer, Cham. https://doi.org/10.1007/978-3-030-36841-8_20

[14] Chassagnon, G., Vakalopolou, M., Paragios, N., & Revel, M.-P. (2019). Deep learning: definition and perspectives for thoracic imaging. European Radiology. doi:10.1007/s00330-019-06564-3

[15] Kremer, J., Stensbo-Smidt, K., Gieseke, F., Pedersen, K. S., & Igel, C. (2017). Big Universe, Big Data: Machine Learning and Image Analysis for Astronomy. IEEE Intelligent Systems, 32(2), 16–22. doi:10.1109/mis.2017.40

[16] Weese, J., & Lorenz, C. (2016). Four challenges in medical image analysis from an industrial perspective. Medical Image Analysis, 33, 44–49. doi:10.1016/j.media.2016.06.023

[17] A.A. Collister and O. Lahav, "ANNz: Estimating Photometric Redshifts Using Artificial Neural Networks," Publications of the Astronomical Society of the Pacific, vol. 116, no. 818, 2004, p. 345

[18] Dieleman, S., Willett, K. W., & Dambre, J. (2015). Rotation-invariant convolutional neural networks for galaxy morphology prediction. Monthly Notices of the Royal Astronomical Society, 450(2), 1441–1459. doi:10.1093/mnras/stv632

[19] Polsterer, K. L., Gieseke, F., and Igel, C., "Automatic Galaxy Classification via Machine Learning Techniques: Parallelized Rotation/Flipping INvariant Kohonen Maps (PINK)", in "Astronomical Data Analysis Software an Systems XXIV (ADASS XXIV)", 2015, vol. 495, p. 81.

[20] J. Kremer, F. Gieseke, K. Steenstrup Pedersen, C. Igel, Nearest neighbor density ratio estimation for large-scale applications in astronomy, Astronomy and Computing, Volume 12, 2015, Pages 67-72, ISSN 2213-1337, https://doi.org/10.1016/j.ascom.2015.06.005.

[21] Heinis, S., et al. "Of genes and machines: application of a combination of machine learning tools to astronomy data sets." The Astrophysical Journal 821.2 (2016): 86.

[22] Viquar, M., Basak, S., Dasgupta, A., Agrawal, S., & Saha, S. (2018). Machine Learning in Astronomy: A Case Study in Quasar-Star Classification. Emerging Technologies in Data Mining and Information Security, 827–836. doi:10.1007/978-981-13-1501-5_72

[23] Rajkomar, A., Lingam, S., Taylor, A.G. et al. High-Throughput Classification of Radiographs Using Deep Convolutional Neural Networks. J Digit Imaging 30, 95–101 (2017). https://doi.org/10.1007/s10278-016-9914-9

[24] Chen, M., Shi, X., Zhang, Y., Wu, D., & Guizani, M. (2017). Deep Features Learning for Medical Image Analysis with Convolutional Autoencoder Neural Network. IEEE Transactions on Big Data, 1–1. doi:10.1109/tbdata.2017.2717439

[25] urelle Tchagna Kouanou, Daniel Tchiotsop, Romanic Kengne, Djoufack Tansaa Zephirin, Ngo Mouelas Adele Armele, René Tchinda, An optimal big data workflow for biomedical image analysis, Informatics in Medicine Unlocked, Volume 11, 2018, Pages 68-74, ISSN 2352-9148, https://doi.org/10.1016/j.imu.2018.05.001.

[26] H. Sun, Z. Liu, G. Wang, W. Lian and J. Ma, "Intelligent Analysis of Medical Big Data Based on Deep Learning," in IEEE Access, vol. 7, pp. 142022-142037, 2019, doi: 10.1109/ACCESS.2019.2942937.

[27] Ahn, E., Kumar, A., Fulham, M., Feng, D., & Kim, J. (2019). Convolutional Sparse Kernel Network for Unsupervised Medical Image Analysis. Medical Image Analysis. doi:10.1016/j.media.2019.06.005

[28] Almeida MAM, Santos IAX. Classification Models for Skin Tumor Detection Using Texture Analysis in Medical Images. Journal of Imaging. 2020; 6(6):51. https://doi.org/10.3390/jimaging6060051.

[29] Varshney, K. R., Chen, G. H., Abelson, B., Nowocin, K., Sakhrani, V., Xu, L., & Spatocco, B. L. (2015). Targeting Villages for Rural Development Using Satellite Image Analysis. Big Data, 3(1), 41–53. doi:10.1089/big.2014.0061

[30] Chebbi, I., Boulila, W., & Farah, I. R. (2016). Improvement of satellite image classification: Approach based on Hadoop/MapReduce. 2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP). doi:10.1109/atsip.2016.7523046

[31] Boulila, Wadii & Mokhtar, Sellami & Driss, Maha & Al-Sarem, Mohammed & Safaei, Mahmood & Ghaleb, Fuad. (2021). RS-DCNN: A novel distributed convolutional-neural-networks based-approach for big remote-sensing image classification. Computers and Electronics in Agriculture. 182. 106014. 10.1016/j.compag.2021.106014.

[32] Robinson, Y. H., Vimal, S., Khari, M., Hernández, F. C. L., & Crespo, R. G. (2020). Tree-based convolutional neural networks for object classification in segmented satellite images. The International Journal of High

Performance Computing Applications, 109434202094502. doi:10.1177/1094342020945026

[33] Murali Krishna Gumma, Prasad S. Thenkabail, Pardhasaradhi G. Teluguntla, Adam Oliphant, Jun Xiong, Chandra Giri, Vineetha Pyla, Sreenath Dixit & Anthony M Whitbread (2019): Agricultural cropland extent and areas of South Asia derived using Landsat satellite 30-m time-series big-data using random forest machine learning algorithms on the Google Earth Engine cloud, GIScience & Remote Sensing, DOI: 10.1080/15481603.2019.1690780

[34] Siva D, Bojja P. MLC based Classification of Satellite Images for Damage Assessment Index in Disaster Management. Int. J. Adv. Trends Comput. Sci. Eng. 2019;8(1)