International Journal of Computing Science and Information Technology, 2022, Vol.6, 01-10
ISSN: 2278-9669, December 2022 (http://ijcsit.org)

A Review of Machine Learning Techniques

Gokula Krishnan
Department of Computer Science, Global University
Email: gokula2001@gmail.com

Received 01/09/2022

Abstract

Over the past couple of decades, the resurgence of interest in machine learning (ML) can be seen across many disciplines, including the hard sciences and engineering. The application of ML algorithms has spread beyond data mining and processing to the realm of scientific computing. This paper provides a comprehensive overview of the current state of the art in ML applications to computational science and engineering. We talk about how ML may be used to boost the efficiency and accuracy of simulation methods, including CFD, MD, and structural analysis. Here, we investigate ML's potential for creating surrogate models of biological applications that are computationally efficient, eliminating the need for more costly simulation methodologies. Many scientific disciplines, including engineering, medicine, astronomy, and computers, provide as examples of how ML may be utilized to handle massive volumes of data. Finally, we discuss how ML has been implemented in VR apps to make them more lifelike and interactive. Machine learning methods were applied to inform healthcare providers' decision making using a wide range of techniques, algorithms, statistical software, and validation processes. Using several machine learning methods, having a well-defined model selection process, and doing both internal and external validation are all essential steps in ensuring that clinical choices are supported by high-quality data. In the future, it will be common practice to use ensemble techniques, which combine many different ML algorithms.

Keywords: machine learning; artificial intelligence; data-mining; scientific computing; neural networks; deep learning

1. Introduction (*Heading* 1)

Big data and other observational studies have traditionally been analyzed with an eye toward informing policy at the population level. Although population-level implications of real-world study results are clear, the ability to predict or provide meaningful evidence at the patient level is considerably less well established due to the complexity of clinical decision making and the variety of factors considered by the health care provider [1, 2]. Even when incorporating findings from subgroup studies, it is difficult to precisely forecast how each individual patient would do using conventional approaches that provide population estimates and measures of variability. A number of non-linear, interrelated aspects are involved in patient care and must be considered while making decisions. Health care decision making is less well informed when data are provided that are only relevant at the population level.

Clinical prediction models are one strategy for making better healthcare decisions through the incorporation of evidence at the patient level. Health care providers have relied on these types of models for decades [3]. Statistical and mathematical models, such as regression, classification, or neural networks, have traditionally been used to combine patient demographic, clinical, and therapeutic data; however, these models have often dealt with a small number of predictor variables (usually below 25). A prominent case of using longitudinal data to construct a conventional decision-making model is the Framingham Heart Study. Patients' risks of developing atrial fibrillation, coronary heart disease, and other cardiovascular events have been the focus of several risk calculators and estimators [4,5,6]. Multivariate regression is commonly used in these investigations to assess previously established risk variables. These results are used to create a scoring system for each risk factor that may be used to estimate the probability of a negative outcome for a given patient.

Sample sizes and potential predictor variables (such as genomic data) can exceed the tens of thousands with the advent of more complex data collection and readily available data sets for patients in routine clinical care, establishing the need for alternative approaches to rapidly process large amounts of information. Prediction models, pattern recognition, and deep learning approaches used to merge complicated information like

genetic and clinical data [7,8,9] are all examples of how AI is being used in clinical research. These techniques are used in the medical sciences to do tasks that would normally require a great deal of time and effort from a human specialist, as well as the possibility of mistake. The basic idea is that computers may learn to make decisions without being explicitly programmed to do so by experimenting with the data themselves. Simply said, "learning from data" is the best way to conceptualize machine learning. [8].

Learning from data may be done in two ways: unsupervised and supervised. When no labels are attached to the replies in a dataset, an unsupervised learning technique can be used to make conclusions about the data. Cluster analysis is the most widely used unsupervised learning technique, and it is used in exploratory data analysis to reveal previously unseen groupings or patterns. To anticipate an outcome using a pre-defined input and output set is the goal of supervised learning. Supervised learning makes use of a variety of statistical methods. Regression models (e.g., regression splines, projection pursuit regression, penalized regression) are one type of classic statistical prediction approach; they entail fitting a model to data, assessing the fit, and estimating parameters that are then utilized in a predictive equation. Tree-based approaches (such as classification and regression trees [CART] and random forests) are another option, since they use the interdependencies between predictor variables and an outcome variable to gradually divide a dataset. A few more instances include neural networks, discriminant functions and linear classifiers, support vector classifiers, and machines. It is common practice to construct prediction tools by combining models trained on resampled or re-weighted data sets using model aggregation (also known as ensemble learning). Model averaging allows for many models to be fit to the same data.

The statistical sciences and the scientific community that utilizes classical statistical regression methods for predictive modelling thoroughly understand these techniques. Though transparent and hypothesis-driven, these approaches can be rigid and miss subtle correlations when studying a large number of variables. Additionally, it is not easy to pick the 'correct' model when doing traditional regression analysis. In the era of big data, non-traditional machine learning algorithms and machine learning approaches may help overcome some of these limitations of classical regression models [2]. However, these methods are not a panacea and must be evaluated within the context of the data limitations that affect the analysis.

The data, model, and outputs used to inform the care of an individual patient must meet the highest standards of research quality because the decision will likely affect the patient's long- and short-term outcomes. Machine learning methods can be used for both population-based models and informed patient-provider decision making. While some degree of uncertainty is to be expected in population-based estimations, the risk of error in patient-level models must be kept to a minimum to provide good care. Ethicists have voiced worries about the potential dangers of using machine learning to make decisions about specific patients [10]. Lack of openness, insufficient information about the certainty of the findings, and the potential for fostering a more paternalistic model of healthcare are only some of the dangers. These are all legitimate worries, and they highlight the need of holding machine learning in healthcare to the highest standards in order to facilitate collaborative decision making grounded in evidence.

The structure of the paper is as follows:

Recent ML-based approaches that boost computational model speed or accuracy are discussed in Section 2. Additional distinctions are made between computational modeling and simulations, and surrogate models. Here, "simulations" mean computer models that explicitly resolve a system of differential equations that characterizes some physical processes. Surrogate models, on the other hand, are (semi-)empirical models that stand in for the governing equations, reduce them significantly, and nevertheless provide useful forecasting abilities in a much shorter amount of time.

To analyze big and complicated datasets and extract relevant values, scientists and engineers have turned to ML-based approaches, which are discussed in Section 3.

In the fourth section, we'll examine how ML has been used to the field of virtual reality (VR). Despite the fact that the research presented in this section might also be included in Sections 2 and 3, the author still consider virtual reality to be a relatively extensive and one-of-a-kind field of study. Consequently, we think it warrants its own subsection.

Current machine learning (ML) initiatives in engineering are discussed, and their potential future applications are outlined in Section 5.

2. Background

Finding a happy medium between computing load and simulation precision is a major challenge in computational modeling and simulation. For instance, QM-based simulations may be used to precisely resolve complicated, turbulent flows; yet, in practice, this is an unimaginable effort due to computing limitations. Here we describe ML-based studies that try to strike a better balance between precision and speed.

MD is a fantastic place to begin since it is a simulation method that provides atomic-level resolution of the system. To put it simply, MD is founded on the principles of Newtonian physics. To illustrate the atomic structure, we use a point for each atom. However, the quantum mechanical (QM) dynamics of its electronic structure are disregarded. Semi-empirical functions of the atomic coordinates, often generated by regressions using QM-based data, are called potentials, and the forces acting on the particles are computed as the gradient (i.e., spatial derivative) of these potentials. Once the forces are known, the system's evolution through time may be estimated using Newton's second law of motion and the integration of time. Since MD's computing complexity grows exponentially with the number of atoms, it can typically only be used to systems with a few hundred thousand particles or fewer.

However, the MD semi-empirical potentials aren't always reliable. These potentials might be situationally or thermodynamically state-specific, depending on the data utilized to derive them (e.g., specific temperature and pressure). For instance, the interaction between water molecules can serve a variety of purposes, each with its own advantages and disadvantages [21].

Alternatively, quantum mechanical (QM) models resolve the system down to individual electrons, allowing for precise calculations of intermolecular and intramolecular forces and interactions. However, these methods require much more computing power than MD do, because atoms have numerous electrons that are associated with one another (i.e., quantum entanglement). In an effort to simplify such many-body systems while yet giving a quantum mechanical precision, simulation approaches like density-functional theory (DFT) have been developed. However, despite DFT's relative efficacy, it remains extremely computationally costly in comparison to more massive techniques.

To preserve the computational advantage of MD while approximating the accuracy of QM-based approaches, ML-based methods trained on QM data may be utilized to produce accurate force-fields that can then be used by MD simulations. To rebuild more generic and accurate Potential Energy Surfaces, ML techniques like ANNs [22,23,24,25] or Gaussian processes (GP) [26,27] have been trained, utilizing DFT simulations (PES). MD simulations rely on the gradient of these PES to give the interatomic forces. Using the atomic coordinates as input, Chmiela et al. [27] have utilized Kernel Ridge Regression (KRR) to compute the attractive forces between atoms. Recent work has shown that an ANN may be taught to generate potentials that are on par with those generated by the computationally intensive but precise Coupled-Cluster technique [28]. This approach is a numerical system that can yield exact solutions to the time-independent Schrödinger equation.

However, the functional shape of these potentials tends to be stable throughout time. As a result, they have trouble generalizing and are inadequate at modeling complicated changes like chemical reactions and phase transitions. First-principle molecular dynamics (FPMD) models are hybrid computations that use classical MD's integration over time based on Newton's laws but employ quantum mechanical simulations like density-functional theory (DFT) to compute force-fields at each timestep. Alternatives to this method try to perform quantum mechanical computations only when absolutely necessary [22]. The computational burden of such methods can be lightened by employing ML techniques like GP [30] or KRR [23]. The QM simulation will calculate the force-fields when a new state is encountered. However, the data will also be utilized to train an ML algorithm; this way, the computationally expensive QM simulation may be bypassed in favor of the much faster ML component if and when identical conditions are encountered. The computationally intensive QM simulations are only examined when a novel process is encountered.

Complex physical systems have been simulated using these ML-based MD techniques. In addition to simulating phase-change materials, predicting the various phases of silicon, and describing the structural features of amorphous silicon, ANNs have also been used to rapidly and reliably compute infrared spectra of materials.

The performance of traditional MD force-fields for diverse tasks has also been assessed using clustering techniques. Principal Component Analysis (PCA) and K-means classification were used by Ashit et al. [24] to classify and rank the various force-fields implemented in MD for modeling carbs. The PCA transformed the system's high dimensionality (because to its many locations and energies) into a low dimensionality (due to its two orthonormal bases). An overview map showing how various force fields are alike or distinct was the outcome of the preceding simplification. To properly classify the force-fields, the PCA data was clustered using a hierarchical k-means clustering technique. Using k-means clustering on MD simulation data, N.Aljrallah et al. [25] compared the performance of several classical force-fields in modeling the production of particular proteins (i.e., for homology modelling) and proposed solutions to address the found deficiencies.

Alternatively, computational fluid dynamics (CFD) is the go-to method for simulating macroscopic fluid dynamics. The continuity equation, derived from the conservation of mass, the momentum equations, derived from the conservation of momentum, and the energy equation, derived from the conservation of energy, are all solved

using this approach. The fluid molecular structure is neglected in favor of treating material qualities as continuous functions of location and time. These equations need to be discretized so that a computer can solve them. To do this, the physical domain is converted into a grid, and then the equations are solved on individual grid cells using numerical techniques (e.g., grid points, cell centers). More grid points mean a more precise answer, but at the cost of processing time.

Physical issues that occur on a scale between the nanoscale (often addressed with QM-based approaches and MD) and the macroscopic (typically addressed with CFD or a comparable continuous approach) exist. For instance, a microfluidic system, which is a system with characteristic dimensions on the micrometre scale, is often made up of a very large number of molecules (e.g., in liquid water there are around 1010 molecules in a cubic micrometer). Traditional medical practice is currently untenable. However, continuum approaches frequently fail at such tiny sizes. It is very uncommon for models like CFD to fail to capture the complex physics that occur at solid-liquid interfaces, such as changes in the thermodynamic characteristics or the effects of surface roughness.

In its place, hybrid models have been developed, heavily using the more computationally efficient continuum solver while resorting to a molecular solver for the unresolved flow aspects. However, the processing demands are substantially higher since MD simulations must run quite often, possibly at each macroscopic timestep. As an alternative, ANNs have been employed to provide molecular-level information into a CFD solver, allowing for a computationally efficient resolution of such scales. In the same way that the QM-based simulations mentioned above only employ the MD solver when they meet states outside of a pre-defined confidence interval, states inside the confidence interval are not utilized. The ANN becomes increasingly important as the simulation develops and more conditions are encountered. The ANN-based hybrid model accurately represents the physics of the flow.

3. Machine learning and Data Science

To get perspective on this topic, we will explore work that has focused on applying ML to the task of analyzing and making sense of massive datasets generated by virtual experiments and simulations.

There are a number of relevant textbooks that detail how ML techniques have been utilized for some time in biology to handle and evaluate the vast datasets connected with the area. Molecular structures with desired chemical characteristics can be detected using data-driven methods. Protein folding, or the process that gives rise to proteins, is a crucial example of the development of molecular structures.

The identification of such conformational states is greatly aided by clustering techniques. In order to better comprehend conformational states seen in MD simulations of biomolecular processes, the K-means and average-linkage clustering techniques were tested. Both the original multidimensional data (MD) and the principal component analysis (PCA) subspaces (of varying dimensionality) were used in the classification process. This article found that when applied to both raw data and PCA-based subspaces, k-means clustering worked effectively, but the average-linkage approach yielded inconsistent clusters. Similar work identified molecular fingerprints of quiescent stem cells in mammalian brains and state transitions using a combination of hierarchical clustering, principal component analysis, and k-means clustering on a transcriptome dataset collected from RNA sequencing. To find instances of ligand binding in several lengthy MD simulations, studies employed k-medoids, a clustering approach analogous to k-means. The elbow approach was used to determine the optimal amount of medoids for this purpose. In addition, the authors estimated the free energy using Regularized Least Squares using a Gaussian kernel.

Most of the aforementioned approaches to clustering fall under the umbrella term "structure-based methods," in which the time-dependent data generated by MD simulations (or experiments) is treated as either a single large data set or a sequence of independent data through which the clustering is refined. There have been other studies that have employed clustering techniques that are better suited to dynamic data. Taking into account the simulation data in a more organic, time-dependent manner, these techniques are known as dynamic-based clustering algorithms (e.g., by considering the timesteps as steps along a Markov chain). Clustering based on dynamic processes works to divide time-varying data into meta-stable states. Results from structure- and dynamics-based approaches sometimes overlap, but there are also often significant variances. The impact of metastability on the folding dynamics of proteins using the Perron Cluster Cluster Analysis (PCCA) technique. In addition, more robust forms of PCCA have been developed, which were later employed for clustering gene expression data with applications in breast cancer research.

ANNs have been utilized in biology for quite some time, the last decade has seen a significant uptick in the use of such methods . This is especially true for the process of protein-DNA binding, for which existing biophysical models appear to be inadequate. Predictions of RNA splicing have been utilized to assess the impact of various genetic variations, and this has been accomplished by training fully connected ANNs with multiple inputs

matching to genetic characteristics. Additionally, CNNs have been employed to predict protein-binding, where they have shown superior performance to prior similar models. Researchers looked at how different CNN topologies affected their capacity to predict DNA-protein interactions. While the research does a good job of explaining a variety of different CNN factors, it ultimately concludes that training with a higher number of filters will lead to the acquisition of more useful sequence characteristics. In genetics, CNN-based architectures have been employed for tasks like as sequence classification and determining the impact of noncoding genomic changes .

Also, ML has been implemented in CAD systems, or computer solutions used to aid in medical diagnosis. Lung cancer identification is a significant instance of ML's application to CAD. Lung cancer is a prominent cause of mortality worldwide. Despite the fact that lung cancer is curable if caught early, it frequently goes undetected even by highly trained doctors until it has progressed to a more advanced stage. Machine learning algorithms may aid in cancer identification in two ways given the right information, such as scanned images: I first, they can identify the sick region, a process called segmentation; and (ii) second, they can categorize malignancies (benign, aggressive etc.). Possible applications of multiple classifiers for early lung cancer diagnosis are discussed by Krishnaiah et al. Methods including Naive Bayesian classifiers, if-then rules, decision trees, and NNs were all taken into account in the article. Subsequent research employed feed-forward NNs to analyze CT images for signs of lung cancer. Segmentation in the form of thresholding and morphological procedures were first used by the authors to locate the lungs. From the resulting pixel intensities in the segmented pictures, statistical parameters were derived. The mean, standard deviation, skewness, kurtosis, fifth central moment, and sixth central moment were all calculated. Together, these formed the input feature vector to the ANN, which ultimately determined whether or not the instance presented malignant characteristics. Data supplied in the form of imaging studies like X-rays, CT scans, MRIs, etc. have been organized using a hybrid of NNs and genetic algorithms by D'Cruz et al.. Following feature extraction on the raw data, a picture was classified as normal or abnormal using a fully linked, feed-forward NN. The genetic algorithm would further categorize a case into malignant or noncancerous if it were found to be aberrant.

Complex NNs have also been employed for CAD, and they include DNNs, CNNs, and Autoencoders. Lung nodules on CT images have been employed with autoencoders to extract characteristics for classification as malignant or benign. In order to determine if a lung nodule on a CT scan is benign or malignant, Song et al. analyzed the performance of DNNs, CNNs, and Autoencoders. It was found that CNNs performed better than the other two approaches combined. Lung tumors have been classified as adenocarcinoma, squamous cell carcinoma, and small cell carcinoma using a CNN architecture similar to that described above. According to another research paper, a procedure was established to take a CT image as input, segment it using thresholding to identify the lung area, and then classify the found nodules. According to the authors, applying a CNN straight to the thresholded data is inefficient and yields a high rate of false positives. U-net design was employed instead, which is a subset of CNN architecture that may improve input data resolution by upsampling layers rather than lowering it via pooling layers. Lung nodules were detected by the U-net and subsequently labeled as malignant or benign using a traditional 3D CNN architecture. Successive research has taken into account a variety of CAD pipelines, some of which use CNN designs, and has achieved excellent results. An autoencoder was created by using convolutional neural networks (CNNs) in a recent deep learning method for lung cancer diagnosis. A CT scan is "encoded" by a 3D convolutional neural network (CNN) and "decoded" by a second CNN. The produced representation was simultaneously input into a fully connected NN, which was then used to determine whether the tumor was malignant or not. Please note that during training, a single cost function was utilized that incorporated both the autoencoder loss and the classification loss.

cancers, including as breast cancer and prostate cancer, have also benefited from ML's use. Diseases including Alzheimer's and Parkinson's have also benefited from its use.

Microfluidics' growing popularity in the biomedical sector necessitates the development of data-driven methods for analyzing the resulting information . The enzyme-linked immunosorbent assay (ELISA) is the standard for current medical diagnostic testing. Berg et al. developed a portable device that can be attached to a smartphone and includes a 96-well plate for ELISA. Test results are read by the phone's camera and sent to servers, where machine learning (ML) techniques, in this case adaptive boosting, are used to make diagnoses (i.e., AdaBoost).

The anti-malware business has used ML as a tool for protecting computers from infection (viruses, adware, spyware, etc.). Running malicious software in a sandbox and analyzing its behavior at runtime is one method of discovering and classifying it. On the other hand, clustering may be used to discover novel classes of malware , while classification can establish whether or not a given piece of software is malicious . An improvement can be shown

when clustering is used in tandem with classification.
The malware executable's object code (binary) may be transformed into a 2D array and shown as a grayscale picture, providing a static alternative to dynamic analysis for malware detection. According to this theory, it should be possible to use image processing techniques like K-Nearest Neighbors (KNN) to categorize malware based on the similarities between the virus and the images it generates. Recent research has employed principal component analysis (PCA) on greyscale photos, and then utilized artificial neural networks (ANNs), KNNs, and support vector machines (SVMs) to classify the reduced data, with KNNs proving superior.
Also, ML has found widespread use in the area of astronomy , where huge amounts of image data are generated by modern telescopes and then need to be analyzed. While ML approaches like ANN have been in use for about 30 years , newer efforts have centered on CNNs because of their computational efficiency in processing and analyzing pictures. The morphology of galaxies, galaxy cluster detection, gravitational lens identification, photometric redshift prediction, and picture reconstruction are only some of the applications of convolutional neural networks (CNNs) in astronomy.
Advances in ML have also helped to boost the quality of video categorization. CNNs have been utilized by Karpathy et al. to categorize sports footage on YouTube. They looked at a variety of topologies to reduce the computational cost of training the CNN and better integrate video's time frames into it. Several subsequent research proposed various CNN- and RNN-based designs for video categorization.
Knowledge distillation is a method wherein a more complicated ML algorithm (the "teacher") is learned on a big dataset and then used to train smaller ML models (the "students") at a reduced computational cost. Videos were classified quickly and accurately using knowledge distillation by Bhardwaj et al. The instructor has been taught with the help of every single video frame. Instead, the student focused on just a few specific time periods. The goal of the learner's cost function was to provide outcomes that were as close to the teacher's as possible. A variety of convolutional neural network and recurrent neural network topologies were investigated for both the instructor and the learner.
Human behavior prediction is another area of ongoing study in ML. Human interactions , a person's context (i.e., location and surroundings) , and a person's characteristics (e.g., facial expressions) may all be taken into account by ML algorithms (typically involving CNNs or RNNs) when attempting to predict a person's future behavior. More recently, LSTM was applied in a study, and not only did the researchers successfully forecast the person's future direction, but also her future activities.

4. Machine learning and Virtual Environment

A new framework based on data-driven features, automation, and machine learning is enhancing decision-making across industries (including manufacturing, medical, e-commerce, and military). To this aim, ML is facilitating the use of new technologies like VR and AR, as well as virtual and digital prototypes, environments, and models, to enable the immersive visualization and simulation of data pertaining to complex systems [260]. Virtual and augmented reality's capacity for immersive, interactive data visualization has the potential to boost productivity, cut down on development costs, and facilitate the testing of different systems in a simulated, digital setting.
Virtual reality (VR) applications in the field of automotive engineering research, like VR driving simulators, necessitate the user's ability to interact realistically with agent vehicles (i.e., computer-controlled), also known as Non-Player Characters (NPCs), that recreate natural traffic flow conditions and accident scenarios. The NPC vehicles and VR infrastructure (i.e., the road network and wireless communications) in many VR driving simulators are trained using ML algorithms, allowing for more realistic interactions and assessments of both the vehicles and the drivers.
Virtual reality (VR) driving simulators have been developed by Charissis and Papanastasiou, who use machine learning (ML) to train the agent vehicles for optimal operation of traffic flow and for recreation of accident scenarios based on real-world data provided by the relevant traffic monitoring and police authorities. A wide variety of scenarios, including a rear-end accident owing to low visibility weather, and rear-end and side-impact incidents due to driver distraction by in-vehicle multimedia systems, were replicated in simulations using ML-trained agent cars. Incorporating ML into the simulations improves how realistically the Vehicular Ad-Hoc Networks systems are shown (VANETs). To better accurately reflect the current network communication patterns and difficulties, they were replicated using an NS-3 network simulator and integrated into the VR driving experience. Further training, improvements, and enhancements are made to the driving simulator with new scenarios and technologies to create realistic VR simulations appropriate for assessing complicated Human-Machine Interfaces (HMI) such as gesture recognition and augmented reality. Feng and Fei also investigate the use of ML for biomechanical gesture detection in VR applications; they successfully train a gesture recognition system involving

10 distinct gestures by using information fusion theory and ML. The suggested solution was meant to fix the problems with identification and placement that arise from using the old-fashioned methods of gesture segmentation.

The need for agents to respond realistically also prompted the creation of an Intelligent Tutoring System (ITS) that used machine learning (ML) to educate a virtual reality (VR) driving simulator. Natural behavior from the NPC agents is crucial for a realistic driving simulation experience, and this may be achieved with ML training. In this article, Lim et al. explain why it's important to create NPCs that can think for themselves, feel emotions, and adapt to their users' preferences. Specifically, the presence of autonomous, emotionally responsive entities within a virtual reality simulation has the potential to greatly increase the sense of immersion.

However, the role of ML in VR may be reversed: the latter can be utilized to generate realistic data that is then used to train the former. In order to train pedestrian detection systems for vehicle applications, Vazquez et al. explain the creation and utilization of virtual environments that can automatically give precise and detailed annotations. The traditional annotation method is labor-intensive and subjective because it uses taught classifiers, and this is why ML researchers are turning to virtual environments. As a result, realistic virtual environments that recreate faithfully the needed visual information for the annotations might replace the conventional routes of camera- and sensor-based training annotation collection.

Vazquez et al. provide a virtual reality (VR) DL development environment that allows users to install a DL model for picture categorization, expanding the synergy between ML and VR in cross-disciplinary applications. The system provides a virtual reality setting in which the user may make choices and provide input for training purposes. Although this system's primary target audience is the biomedical industry, it has the potential to be used in other fields provided the necessary picture sequences are included.

Virtual reality (VR) is used to generate digital twins of engineering systems, equipment, or whole facilities like oil-rig installations. This allows for more accurate and efficient monitoring, assessment, and identification of possible causes of malfunction. Recently, twin technologies have been employed in conjunction with ML. With the use of ML, Madni et al. describe a method for creating a digital replica of a physical system (in this case, an oil rig) in order to simulate and keep tabs on its performance. This research classifies digital twin systems into four levels: the pre-digital twin, the digital twin, the adaptive digital twin, and the intelligent digital twin. Intelligent Digital Twin is the only level that uses both supervised ML (i.e., operator's preferences) and unsupervised ML to differentiate between patterns and structures/items occurring in different operating contexts. For their part, Jaensch et al. provide a digital twin platform that uses model-based and data-driven approaches to modeling and simulation, expanding on the work of the aforementioned authors. In comparison to conventional ways, theirs can quickly and cheaply fix common problems in production and operations.

Networking is another area where ML may help VR. In today's world, it's crucial that virtual reality software can be used with mobile connections. This opens up new possibilities for creating virtual reality settings where people may work together and have shared, immersive experiences, even if they are physically apart. Such actions may pave the way for remote monitoring and maintenance of systems (i.e., digital twin systems) in the form of wireless virtual reality (VR) environments. But low-quality wifi connections can slow down the transfer of virtual reality data (QoS). Because of the massive amounts of data needed for VR applications, this happens over a wide range of network topologies. To alleviate and lessen this problem, Chen et al. suggest a unique method based on the ML framework of Echo State Networks (ESNs) (SCNs). Chen et al. provide data correlation-aware resource management for wireless VR applications in an effort to reduce these kinds of problems in future studies. The latter opens up virtual reality to teams of workers who can all do their jobs at once. Users in the same virtual reality environment likely share the same virtual areas, differing only in their orientation or activity within the virtual world, and this might lead to correlations in the geographical data they request or communicate. Since this is a machine-learning method, it makes use of ESN and transfer learning, which is how the suggested system learns. The latter ESN algorithm can more quickly adapt to the shifting conditions of a wireless network, allowing for more equitable distribution of virtual reality data to end users. Similarly, the methods for collecting, monetizing, and disseminating 360-degree virtual reality (VR) footage shot by unmanned aerial vehicles (UAVs) to a wide range of end users are investigated. In this research, an unique method is proposed using ESNs and a Liquid State Machine as part of a distributed DL algorithm (LSM). In addition, Alkhateeb et al. [283] look at wireless VR communication and the provision of adequate QoS, with the goal of mitigating the common transmission problems experienced by Millimeter Wave (mmWave) communication systems and Base Stations (BS). A technique for identifying a user

based on factors such as their physical location and the nature of their interactions is detailed in this study. The suggested method makes use of a DL model that is educated to anticipate the user's needs and, as a result, to anticipate the beamforming vectors that deliver enhanced QoS.

It is certain that other sciences and sectors that require timely, flexible, and accessible information for complex systems will embrace the aforementioned combinations of upcoming technologies and ML in the not-too-distant future.

5. Conclusion

At the moment, it appears that supervised learning is being employed by the vast majority of ML algorithms designed to improve computer simulations. In order to reliably forecast the future, the ML part is often trained using high-fidelity, already-existing data. Typical ML practices and factors must be taken into account. Supervised learning entails the following steps: (a) divide the available data into training and test sets to measure the generalization error, and possibly a cross-validation set for testing ML algorithms and hyperparameters are required; (b) use learning curves or other methods to identify under- or over-fitting; (c) test various ML algorithms and architectures on a cross-validation set; and (d) select the most optimal set-up; try different features; etc. The amount of training data and the length of the feature vector will determine which method (such as ANN, GP, or SVM) is used. ML's effectiveness in these contexts has been demonstrated thus far, and we anticipate a proliferation of similar efforts in the years to come.

Surrogate modeling is another use of ML in scientific computing; rather than supplementing an existing method, it may be used to totally replace it with a reduced, computationally beneficial model. The ability to employ surrogate models to forecast how a system will change over time is a hot topic right now. RNNs appear to be ideally suited for dynamic models, at least intuitively. We anticipate a greater role for them in the coming years, despite their low usage in the present. This is likely owing to the difficulty of their training.

Improving the use of ML in scientific computing will rely heavily on physics-informed ML models. We have a strong foundation for making predictions constrained by physical rules thanks to methods like SINDy, automated differentiation, and physics-informed NNs. There is potential for even greater utility from the application of non-parametric algorithms like GPs. They are particularly well suited to physics-informed ML because of their built-in capacity to inject information via the covariance matrix. In addition to providing statistical information (beyond merely the greatest probability estimate), the mathematical beauty of these techniques is another thing that draws people to them. Last but not least, as shown by Raissi and Karniadakis [157], it can model general partial differential equations with "small-data," making this work more manageable for many applications.

In addition, ML is a powerful tool for data processing and mining. Biomedical research is one field that has benefited from these advancements by developing more nuanced algorithms to better understand their massive datasets. On the other hand, we think these tactics may be quite useful in other fields as well. In the case of transitional and turbulent flows, for instance, large-scale DNS simulations reveal transient patterns that might be too complicated to grasp using conventional methods such as eye examination. Such structures are amenable to categorization and correlation using ML algorithms. Similar difficulties arise when trying to extrapolate meaningful, experimentally verifiable numbers from the results of microscopic simulations (e.g., DFT, MD), where the outcome is frequently exceedingly noisy. Averaging across lengthy simulation times is commonly needed for the aforementioned. Again, ML need to be able to discern patterns within this undesirable noisy input.

At last, ML is emerging as an integral part of VR's many potential uses. With proper training, ML can more accurately simulate genuine settings and include smart agents into a virtual setting. Digital twins are digital representations of physical systems that can benefit from ML techniques by having the virtual environment customized to the user's or operator's preferences, or by employing unsupervised learning to recognize items and patterns in the operational environment. Network speed is a major barrier to widespread usage of collaborative VR settings, but NNs, and especially ESNs, can help alleviate this issue. As a final use case, VR may be leveraged to supply ML algorithms with a plethora of realistic, synthetic training data.

Finally, we point out that theory-based computational tools (such as DFT, MD, and CFD) are widely accessible and, in many cases, freely downloadable. In order to gain understanding of novel systems, the use of such models may be too expensive computationally. However, ML model training may be quite resource intensive. In contrast to the standard numerical approaches, a trained model can generate predictions on fresh data fast. We believe that a collaborative effort should aim at developing a centralized library of trained algorithms that can conveniently be utilized for comparable challenges because of the time, money, and resources needed to train ML algorithms like DNNs, as well as the possible envi-

ronmental effect [284]. Then, and only then, will AI and ML be able to contribute significantly to the advancement of science.

Consequently, ML algorithms are rapidly spreading in virtually all areas of research. This article provides a quick overview of recent advances and attempts of ML in scientific and engineering fields. We think that, despite its recent huge success, ML is still in its infancy and will play a vital role in scientific research and engineering in the years to come.

REFERENCES

[1] Chinnasamy, P., P. Deepalakshmi, Ashit Kumar Dutta, Jinsang You, and Gyanendra Prasad Joshi. "Ciphertext-Policy Attribute-Based Encryption for Cloud Storage: Toward Data Privacy and Authentication in AI-Enabled IoT System." Mathematics 10, no. 1 (2021): 68.

[2] Dutta, Ashit Kumar, and Abdul Rahaman Wahab Sait. "An application of intuitionistic fuzzy in routing networks." International Journal of Advanced Computer Science and Applications 3, no. 6 (2012).

[3] Dutta, Ashit Kumar. "Earthquake prediction using artificial neural networks." International Journal of Research and Reviews in Computer Science 2, no. 6 (2011): 1279.

[4] Dutta, Ashit Kumar. "Detecting phishing websites using machine learning technique." PloS one 16, no. 10 (2021): e0258361.

[5] Dutta, Ashit Kumar, Basit Qureshi, Yasser Albagory, Majed Alsanea, Manal Al Faraj, and Abdul Rahaman Wahab Sait. "Optimal Weighted Extreme Learning Machine for Cybersecurity Fake News Classification." COMPUTER SYSTEMS SCIENCE AND ENGINEERING 44, no. 3 (2023): 2395-2409.

[6] A. K. Dutta, T. Meyyappan, B. Qureshi, M. Alsanea, A. W. Abulfaraj, M.Al Faraj, Abdul Rahaman Wahab Sait, "Optimal deep belief network enabled cybersecurity phishing email classification," Computer Systems Science and Engineering, vol. 44, no.3, pp. 2701–2713, 2023. (ISI – Web of Science)

[7] A. K. Dutta, M. M. Alqahtani, Y. Albagory,Abdul Rahaman Wahab Sait and M. Alsanea, "Optimal machine learning enabled performance monitoring for learning management systems," Computer Systems Science and Engineering, vol. 44, no.3, pp. 2277–2292, 2023. (ISI – Web of Science)

[8] A. K. Dutta, N. M. A. Zakari, Y. Albagory and Abdul Rahaman Wahab Sait, "Colliding bodies optimization with machine learning based parkinson's disease diagnosis," Computer Systems Science and Engineering, vol. 44, no.3, pp. 2195–2207, 2023. (ISI – Web of Science)

[9] N. Ali Aljarallah, A. Kumar Dutta, M. Alsanea and Abdul Rahaman Wahab Sait, "Intelligent student mental health assessment model on learning management system," Computer Systems Science and Engineering, vol. 44, no.2, pp. 1853–1868, 2023. (ISI – Web of Science)

[10] A. Kumar Dutta, Y. Albagory, M. Al Faraj, M. Alsanea and Abdul Rahaman Wahab Sait, "Cat swarm with fuzzy cognitive maps for automated soil classification," Computer Systems Science and Engineering, vol. 44, no.2, pp. 1419–1432, 2023(ISI – Web of Science)

[11] A. Kumar Dutta, Y. Albagory, M. Al Faraj, Y. A. M. Eltahir and Abdul Rahaman Wahab Sait, "Optimal sparse autoencoder based sleep stage classification using biomedical signals," Computer Systems Science and Engineering, vol. 44, no.2, pp. 1517–1529, 2023. (ISI – Web of Science)

[12] A. Kumar Dutta, M. Al Faraj, Y. Albagory, M. Zeid M Alzamil and Abdul Rahaman Wahab Sait, "Intelligent smart grid stability predictive model for cyber-physical energy systems," Computer Systems Science and Engineering, vol. 44, no.2, pp. 1219–1231, 2023. (ISI – Web of Science)

[13] A. Kumar Dutta, Y. Albagory, Abdul Rahaman Wahab Sait and I. Mohamed Keshta, "Autonomous unmanned aerial vehicles based decision support system for weed management," Computers, Materials & Continua, vol. 73, no.1, pp. 899–915, 2022. (ISI – Web of Science)

[14] A. K. Dutta, Y. Albagory, M. Alsanea, H. I. Almohammed and Abdul Rahaman Wahab Sait, "Ensemble deep learning with chimp optimization based medical data classification," Intelligent Automation & Soft Computing, vol. 35, no.2, pp. 1643–1655, 2023(ISI – Web of Science)

[15] A. Kumar Dutta, Y. Albagory, M. Alsanea, Abdul Rahaman Wahab Sait and H. Saleh AlRawashdeh, "Fuzzy with metaheuristics based routing for clustered wireless sensor networks," Intelligent Automation & Soft Computing, vol. 35, no.1, pp. 367–380, 2023. (ISI – Web of Science)

[16] A. Kumar Dutta, B. Qureshi, Y. Albagory, M. Alsanea, A. Waleed AbulFaraj, Abdul Rahaman Wahab Sait, "Glowworm optimization with deep learning enabled cybersecurity in social networks," Intelligent Automation & Soft Computing, vol. 34, no.3, pp. 2097–2110, 2022. (ISI – Web of Science)

[17] Dutta, Ashit Kumar, Majed Alsanea, Basit Qureshi, Faisal Yousef Alghayadh, and Abdul Rahaman Wahab Sait. "Intelligent Rider Optimization Algorithm with Deep Learning Enabled Hyperspectral Remote Sensing Imaging Classification." Canadian Journal of Remote Sensing (2022): 1-14.

[18] Sait, Abdul Rahaman Wahab, Irina Pustokhina, and M. Ilayaraja. "Modeling of multiple share creation with optimal signcryption technique for digital image security." Journal of Intelligent Systems and Internet of Things 1 (2019): 26-36.

[19] Sait, Abdul Rahaman Wahab, and Dr T. Meyappan. "Data preprocessing and transformation technique to generate pattern from the web log." In International conference on Computer Science and Information Systems (ICSIS" 2014) Oct, pp. 17-18. 2014.

[20] Sait, Abdul Rahaman Wahab, and T. Meyyappan. "An Automated web page classifier and an algorithm for the

extraction of navigational pattern from the web data." Journal of Web Engineering (2017): 126-144.

[21] Shankar, K., E. Perumal, A. R. W. Sait, I. Pustokhina, and D. A. Pustokhin. "A secure visual share creation model using cauchy mutation based steam cipher for digital image security." International Journal of Advanced Science and Technology 29, no. 8 Special Issue (2020): 676-687.

[22] Dutta, Ashit Kumar, R. Uma Mageswari, A. Gayathri, J. Mary Dallfin Bruxella, Mohamad Khairi Ishak, Samih M. Mostafa, and Habib Hamam. "Barnacles Mating Optimizer with Deep Transfer Learning Enabled Biomedical Malaria Parasite Detection and Classification." Computational Intelligence and Neuroscience 2022 (2022).

[23] Shankar, Kathiresan, Sachin Kumar, Ashit Kumar Dutta, Ahmed Alkhayyat, Anwar Ja'afar Mohamad Jawad, Ali Hashim Abbas, and Yousif K. Yousif. "An Automated Hyperparameter Tuning Recurrent Neural Network Model for Fruit Classification." Mathematics 10, no. 13 (2022): 2358.

[24] Dutta, Ashit Kumar. "Detecting Lung Cancer Using Machine Learning Techniques." INTELLIGENT AUTOMATION AND SOFT COMPUTING 31, no. 2 (2022): 1007-1023.

[25] Shankar, K., Ashit Kumar Dutta, Sachin Kumar, Gyanendra Prasad Joshi, and Ill Chul Doo. "Chaotic Sparrow Search Algorithm with Deep Transfer Learning Enabled Breast Cancer Classification on Histopathological Images." Cancers 14, no. 11 (2022): 2770.

[26] Kumar, K. Vijaya, E. Laxmi Lydia, Ashit Kumar Dutta, Velmurugan Subbiah Parvathy, Gobi Ramasamy, Irina Pustokhina, and Denis A. Pustokhin. "Deep Learning Enabled Object Detection and Tracking Model for Big Data Environment." CMC-COMPUTERS MATERIALS & CONTINUA 73, no. 2 (2022): 2541-2554.

[27] Nadeem, Muhammad, Ali Arshad, Saman Riaz, Syeda Wajiha Zahra, Ashit Kumar Dutta, Abdulrahman Alruban, Badr Almutairi, and Sultan Almotairi. "Two-Layer Security Algorithms to Prevent Attacks on Data in Cyberspace." Applied Sciences 12, no. 19 (2022): 9736.

[28] Dutta, Ashit Kumar. "Managing natural hazards in smart cities in Kingdom of Saudi Arabia using a technique based on interior search algorithm." Electronic Government, an International Journal 16, no. 1-2 (2020): 155-169.