Article

A Study on Data Scaling Methods for Machine Learning

(IJGASR) International Journal For
Global Academic & Scientific Research
ISSN Number: 2583-3081
Volume 1, Issue No. 1, 31–42
© The Author 2022
journals.icapsr.com/index.php/ijgasr
DOI: 10.55938/ijgasr.v1i1.4

IJGASR

Vinod Sharma

Abstract:

Machine learning (ML), a computational self-learning platform, is expected to be applied in a variety of settings. ML, on the other hand, uses a model built with a learning structure rather than traditional code that is written line by line in a continuous pattern. These models are created and equipped to determine the results of training using historical data. Scalability is a major challenge in real machine learning programs. Many ML-based technologies are essential to quickly analyze new data and create forecasts, as forecasts become meaningless after a few ticks (think real-time methods such as stock markets and clickstream data). Many machine-learning programs, on the other hand, need to be able to scale and train with gigabytes or terabytes of data during model training (As is found in the model from a web-scale image corpus). High-dimensional challenges pose new obstacles to machine learning professionals who are increasingly interested in scalability as well as algorithm quality. Against the backdrop of the current situation, this overview article on the scope of scalability in machine learning platforms collects, investigates, and analyzes the current state, aspects, and perspectives of scalability that can be added to machine learning platforms in a variety of ways to improve efficiency. The purpose is to do. Reliability when processing large amounts of data.

Keywords:

Scalability in Machine Learning, Data Scaling For Machine Learning, Utility of Scalability in Machine-Learning, Scalability Methods of Machine-Learning

1. Introduction

Modern technological breakthroughs are rapidly moving toward automated self-learning processes, in which sophisticated digital technologies are used to speed up a variety of complex real-time processes that require little or no human participation.

Research Scholar, Jiwaji University, Gwalior.

Corresponding Author: Vinod Sharma Research Scholar, Jiwaji University, Gwalior India.
Email: computerscience.scholar2022@gmail.com

Machine Learning is one such approach that is widely utilised to address the challenge of learning from experience in the context of certain activities and performance metrics. Users can utilise machine learning techniques to deduce underlying structure and generate predictions from huge datasets. ML survives on strong computer systems, appropriate learning approaches (algorithms), and rich and/or huge data. As a result, machine learning has a lot of potential and is an important aspect of big data analytics [17].

When it comes to machine learning algorithms, if the values of the features are closer together, the methodology has a better chance of being trained well and faster, whereas when the data points or feature values are far apart, it will take longer to understand the data and the correctness will be lessened.

As a result, if the data contains data points that are located farther in any way, scaling is a method for bringing them closer together. To put it another way, scaling reduces the distance between data units by making them more generic. Min max scaler, Max Abs Scaler,Quantile Transformer Scaler are the various feature components of scalability that help for different classification and extraction of scalabilty,

As a result, if the data under any circumstance comprises data points that are far apart, scaling is a technique for bringing them closer together. To put it another way, scaling is used to make data units more general so that the distance between them is reduced. For example, when the distance range between feature values rises, machine-learning algorithms such as logistic regression or linear regression that use the Gradient descent algorithm tend to produce erroneous results. The movement will then grow, and the function will not operate correctly. In this case, having a well-rescaled data set is essential so that the function can assist in the creation of the machine learning model [16]
.dif attribute of rescaling testing are max min [0, 1]

$$\mathbf{asx'=x-min(x)/max(x)-min(x)}$$

As previously said, machine learning methods learn from data as the learning model maps the data as it is supplied from input to output. And the distribution of data points can vary depending on the data attribute. The model's results are more ambiguous when there are larger differences between the data points of input variables.

As a result, scalability is a critical factor to consider when developing and comparing machine-learning-based strategies in order to produce the most reliable and long-term results. This review-based study on Data Scaling Methods for Machine Learning is thus thoughtfully composed to recognise the industrial need for scalability, as well as their shortcomings and future scope, in order to improve machine-learning approaches, particularly those that use large and diverse datasets.

The background study addressed below addresses the fundamental components of scalability, including the requirement for and application of scalability so that machine-learning algorithms can be optimally upgraded and deliver authentic findings with the fewest possible flaws/uncertainties.

1.1 Types of Commonly Used Machine-Learning Based Approaches

The kind of learning feedback, the objective of learning tasks, and the time of data availability are all characteristics of ML methods [17]. ML can be classed into three forms based on the nature of the input available to a learning types followed: supervised, unsupervised mode, and reinforcement type of learning. The purpose of supervised learning is to teach the learning system a function that maps inputs to outputs by presenting it with samples of input-output pairings. The purpose of unsupervised learning is to identify patterns in the data without providing explicit feedback or desired result.

A reinforcement learning system, like unsupervised learning, does not have input-output pairings. Reinforcement learning, like supervised learning, receives feedback on its prior actions. In contrast to supervised learning, reinforcement learning provides feedback in the form of incentives or punishments associated with actions rather than intended output or selective corrections of inferior actions. The process of semi-supervised way of learning is a hybrid of supervised and unsupervised learning in which a small number of input-output samples are fed to the model also added with a large number of un-annotated resources.

ML can be divided into two types: representational learning and task learning, depending on whether the learning goal is to learn the pre-decided functions based on the data that is fed or to learn the features themselves. When developing classifiers or other predictions, representational learning seeks to learn new representations of data that make it easier to extract meaningful information. Task learning, on the other hand, usually involves intended outcomes and is classified as classification, regression, or clustering.

ML can be separated into batch learning and online learning based on the schedule of making training data available (i.e., whether the training data are available all at once or one at a time). Batch learning creates models by learning on all of the training data at once, whereas online learning updates models as new data is added. The premise of the batch learning method is that the data are independent and are similarly shared or derived from the probability distribution taken from same source. This is little met with real data. In most cases, online learning does not rely on statistical interpretations about the data.

1.2 General Machine-Learning Based Model Development Stages

Data preprocessing, learning, and evaluation are the three primary steps of machine learning development. Data preprocessing aids in the transformation of raw data into the "correct shape" for further learning procedures. It’s likely that the raw data is unstructured, noisy, incomplete, and inconsistent. Through data refinement, extraction, modifications, and mixing, the preprocessing stage turns such data in the form as may be suitable in inputs to learning. Using the preprocessed input data, the process of learning selects learning steps and modifies model aspects to yield desired outputs. After then, the learnt models are evaluated to see how well they work. The evaluation of a classifier's performance, for example, entails choice of datasets, performance score, error sorting, and statistical tests. The findings of the evaluation could lead to changes in the parameters of chosen learning schemes and/or the selection of new algorithms

1.3 Importance of Data Preprocessing, its Scope and Challenges

In fact, data preprocessing techniques used in machine learning approaches typically lower the size of the data set supplied to algorithms by orders of magnitude [18]. Before the data can be used for next processes, it must be transformed and scaled up or down. Data redundancy, noise, inconsistency, labelling (done in semi-supervised ML processes), heterogeneity, transformation, data imbalance, and feature depiction/selection are among the difficulties it seeks to address.

Deep learning neural networks learn how to map inputs to outputs from examples in a training dataset.

The weights of the model are initialized to small random values and updated via an optimization algorithm in response to estimates of error on the training dataset.

Given the use of small weights in the model and the use of error between predictions and expected values, the scale of inputs and outputs used to train the model are an important factor. Unscaled input variables can result in a slow or unstable learning process, whereas unscaled target variables on regression problems can result in exploding gradients causing the learning process to fail.

The requirement for data reduction as a pre-process may, however, be more of a constraint in terms of learning algorithms than a fundamental constraint on data mining. With the need for human labour and lots of available choices of varied types, data preparation and preprocessing is frequently expensive.

1.4 Scalability Tools and Their Purpose in Machine-Learning Based Models

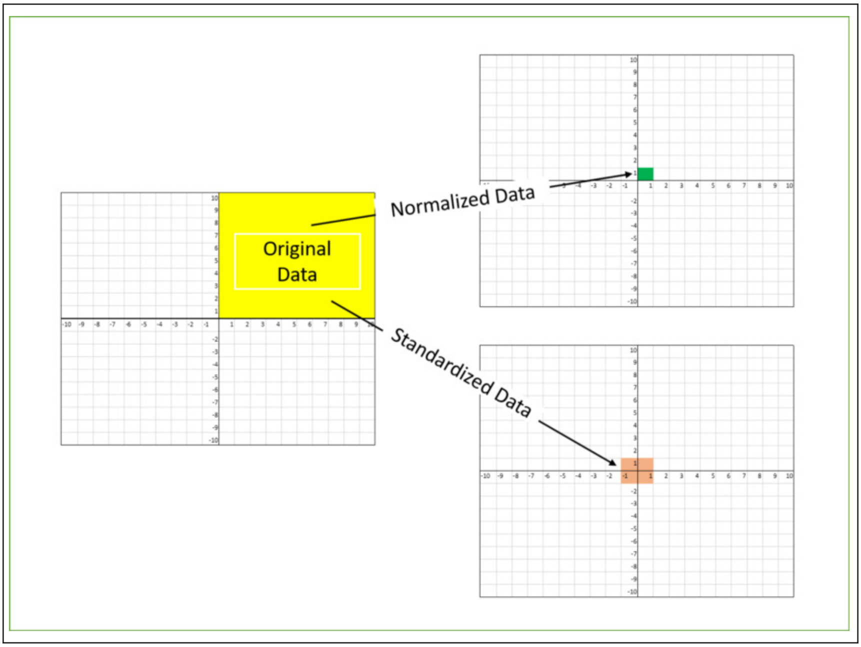
The two major approaches for scaling data are normalisation and standardisation, which are commonly employed in algorithms that need scaling. In statistics, normalisation is a rescaling procedure in which we strive to fit all of the data points into a range of 0 to 1 such that they are closer to each other [16]. In many algorithms, when we desire **faster convergence**, scaling is a MUST like in Neural Network.

Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions do not work correctly without normalization. For example, the majority of classifiers calculate the distance between two points by the distance.

It is a relatively common method of data scaling. The smallest value of any feature is transformed to 0 in this method of data scaling, while the maximum value of the feature is turned to 1. Basically, normalisation divides the difference between any value and the minimum value by the difference between the maximum and minimum values. The Z-score and Min-Max are two of the most frequent normalising methods.

The standardisation function's core premise is to centre data points around the mean of all data points displayed in a feature with a unit standard deviation. The data point's mean will be zero, and the standard deviation will be one [16].

This method similarly tries to scale the data point from zero to one, but it does so without using the terms maximum or minimum. This is where the mean and standard deviation come into play. The mean is said to be the average value of any considered part in a set of numbers, and the standard deviation is a measurement of the data points' dispersion from the mean value of the data points in statistics. As a result, the data points are rescaled to ensure that they are in a curve shape after scaling.



Credit : towardsdatascience.com

2. Objective

The Review Article on Data Scaling Methods for Machine Learning aims to locate the current machine learning areas where scalability plays major role and should be correctly implemented to reduce uncertainty, erroneous result or increase in cost/processing time.

3. Methodology

The present review article on purpose and utility of scalability in machine-learning platforms is entirely composed based on the existing authentic computer science and machine learning analyses and articles that are published in acknowledged online portals, such as, IEEE, Google Scholar, Science Direct, etc.. Genuine facts and informations are gathered from trusted web libraries and institute portals are thoroughly studied, sorted and selected based on their relevance.

4. Literature Review

Hang Cao et al. (2016) [1] suggested a Generalized Logistic (GL) rule that could scale data overall to an acceptable gap using histogram equalisation techniques. The technique was used in conjunction with Machine Learning, which learned a generalised logistic function to suit the data's empirical cumulative distribution set up. The platform validation was done with by carrying out 16 tasks on binary classification with a variety of variable values, covering larger spectrum of scenarios. The area under the receiver operation characteristic curve (AUROC) and the percentage value of correct classification were measured to confirm the model performance. Their findings revealed that their proposed GL method has more potential than the commonly used rules, such as, Min-Max and Z-score, which are currently accessible data scaling tools. Furthermore, their method demonstrated its worth in reducing outliers, obviating the need for further data pretreatment steps such as denoising or outlier identification.

Elrafey et al. (2017) [2] proposed a unified approach of scaling-down sampling algorithms suitable for Machine Learning by combining random and progressive sample components. They used suitable references from published literatures submitted in this field of study to develop their concept.

Hestness et al. (2017) [3] developed a large-scale empirical pattern of mistake generalisation framework and applied it to four machine learning aspects as: machine translation, image processing, language modelling, and speech recognition. The investigation looked at how beneficial their notion was as the model size rose as the training sets grew. Their empirical findings validated power-law generalisation error scaling across a wide range of parameters, finalizing the power-law exponents—the "steepness" of the learning curve, a part of which was based on limited theoretical considerations. Improvements to the model also helped avoid the issue without compromising the power-law exponent.

They discovered that scaling occurs in a sublinear relationship with data size. The ramifications of these scaling connections on deep learning-based analysis, practise, and systems were profound. The tool might be used for model validation, accuracy standard set up, and making data set expansion decisions. They also boosted the design of computing systems and emphasised the significance of continuing computational scalability.

For privacy-preserving machine learning approaches, **Mohassel et al. (2017)** [4] used the stochastic gradient descent method as a training tool for their unique linear regression, neural network, and logistic regression based protocols. The protocol used in their investigation was a two-server approach, which allowed data owners to spread their private information between two non-colluding servers. These protocols could use secure two-party computation to train multiple models on the combined data (2PC). These protocols were shown to be effective at multi-staged magnitude than many advanced implementations n terms of tightly secured linear and logistic regressions in their research. They could develop about more than a million data samples and lots of features.

R. Masegosa et al. (2017) [5] developed AMIDST Toolbox, a publicly available Java toolbox licensed under the Apache Software License version 2.0. The programme was created with a focus on (large) streaming data and was designed to do scalable probabilistic machine learning. The toolbox contained necessary elements to let users create versatile modelling languages constructed with time-dependent probabilistic graphical principles tang latent variables and temporal dependences. By using a parallel or distributed application of the Bayesian learning technique with either streaming or batch facts for training a particular scheme from a large dataset. By integrating with software tools such as Flink, HUGIN, Weka, Spark, R, and MOA, the AMIDST tool is able to incorporate current functionality and methods.

Wan (2019) [6] used algorithm simulation to investigate the effects of normalisation and interval scaling on stochastic gradient descent convergence. He did his research using UCI's computer hardware data collection. The study indicated that feature scaling is valid for machine learning and that normalisation should be used consistently. It aided in achieving desirable convergence in the case of a big set of data. He recommended using a normalisation tool for small datasets, depending on the setup requirements.

G. Camilla et al. (2019) [7] used Apache Spark, an in-memory distributed application, to solve scalability issues with machine learning models that must deal with large amounts of data. The tool was chosen because it offered a large number of machine learning libraries. The study aimed to quantify scalability by analysing the execution time attained by a classifier with greater workloads. Experiments on logistic regression and random forest rule were used to validate classifier models, as well as their adaptation to the Apache Spark framework. They conducted a comparative classifier analysis to back up their claim. According to the model's usability score, logistic regression had the quickest execution time and the best scalability. The research intended to cover the domains of big data and machine learning in terms of scalability, as well as the use of optimization approaches, cache, and persist.

Karimipour et al. (2019) [8], in attempt to meet the demand for scalable anomaly detection engines suitable for large smart grids created unsupervised anomaly detection built up on statistical correlation rule. Their strategy was to distinguish real problems from confusion and advanced cyberattacks. To reduce the computational load while revealing the causal factors between the subsystems, the proposed method used feature extraction using symbolic dynamics filtering (SDF). The team used PSS/E software to validate the model's applicability in a variety of operational settings, including IEEE measures such as 39, 118, and 2848 bus systems. The results showed a 99% correctness, a 98% true positive measure, and a false positive value of less than 2%.

Arafatur et al. (2020) [9] proposed two ways, semi-distributed and distributed, to address the scalability issues in centralised intrusion detection systems (IDS) for resource-constrained devices. Their method blended high-end function extraction and choice with fog-edge coordinated measures to achieve best performance standard. They created parallel machine-studying fashions matching to a partitioned attack dataset to disperse the computing tasks. In the semi-disbursed example, the edge-based parallel processes had been similarly used for function alternatives earlier than appearing a one multi-layer perceptron class at the fog portion. In the disbursed condition, each of the parallel processes accomplished feature choice and multi-layer perceptron sorting on their own, and then the outputs had been joined collectively with the aid of using a adjusted edge or fog for the very last choice making. The numerical findings, primarily based totally on a evaluation of preceding works, indicated the potency of developed scheme, demonstrating matching accuracy to the advanced centralised IDS at the same time as additionally acknowledging the inherent discrepancies among accuracy and constructing time performance.

D. Singh et al. (2020) [10], to ensure scalability, evaluated the impact of 14 data normalization methods applicable on classification efficacy, taking into account the entire feature batch, their selection, and feature relevance. The creation and evaluation of a modified Ant-Lion optimization that could locate feature pattern subsets and best characteristic weights and nearest neighbour classifier attributes was part of this study. Test accuracy, feature reduction level, and action time were all taken into account. Because no one approach could be scored as an ensured scaling scheme in the test, the team decided to develop a model using a combination of normalising tools based on their observed performance and empirical analysis results. For the whole feature batch and feature selection process, -Score and Pareto Scaling outperformed *tanh* and its variation, and for feature weighting, tanh and its variant outperformed tanh and its variant. Mean Centered, Median and Median Absolute Deviation techniques, Variable Stability Scaling, as well as unnormalized data, were the worst performers.

When the spatiotemporally chaotic system was very large and complicated, **Wikner et al. (2020)** [11] developed combinational approaches including parallel machine learning processes used for prediction and hybrid methodology to facilitate scalability of machine-learning based weather prediction systems. As a result, they aimed to create a composite prediction rule built with knowledge-influenced and a machine learning-centred aspects. They discovered that their proposed approach could be scaled well, resulting in great usability for very large systems. In addition, the amount of time series training info required for varied, coinciding machine learning feaures was far lower than it would have been without parallelization. Furthermore, in cases where the knowledge-oriented features computational realisation did not resolve subgrid-scale processes, the suggested scheme was able to incorporate the consequences of unresolved short-scale dynamics on the resolved longer-scale dynamics using training data.

Wang et al. (2020) [12] built their study on a current topic that they discovered from existing literature. It was shown that most Machine Learning (ML) based Intelligent Transportation Systems (ITS) require a greater cost of implementation in terms of time spent training and predicting due to their complicated structure. As a result, they assessed not only the accuracy of several state-of-the-art ML-models, but also their efficiency and scalability. They also developed Desensitization, an off-line optimization strategy, to improve the scalability of their model.

M. Ahsan et al. (2021) [13] tested eleven machine learning (ML) algorithms—Logistic Regression (LR), K-Nearest Neighbours (KNN), Gradient Boost (GB), Support Vector Machine (SVM), Classification and Regression Trees (CART), Naive Bayes (NB), XG Boost (XGB), Ada Boost (AB), Random Forest Classifier (RF), Linear Discriminant Analysis (LDA), Extra Tree Classifier (ET)—and data scaling methods of six variations—Robust Scaler (RS), Stand scale (SS), Min-Max (MM), Normalization (NR), Max-Abs (MA), and Quantile Transformer (QT) on a set of data that had information of heart affected client info. The study tried to solve the issues connected with datasets that disrupt medical advice in this field, such as facts that were not found, mismatching data, and unorganized data, by using appropriate scaling methods (containing mismatched/absent info both both of figures and classification types by nature). CART, paired with RS or QT, surpassed all other ML processes with 100 percent accuracy, 100 percent precision, 99 percent recall, and 100 percent F1 score, according to their research. The results of the investigation showed that the usefulness of the appropriate ML-based model varied depending on the data scaling strategy.

Krittanawong et al. (2021) [14] gave a study of device-based diagnostics now performed using machine learning-based cardiovascular analysis/prediction systems that are extensively used to facilitate automated diagnosis, as well as difficulties such as scalability and ethical considerations. Their paper described the current days' cardiovascular monitoring processes and their condition, from the capture of biological signals to the creation of new biosensors, the planning of analytical techniques, and finally structural and ethical concerns. The report also included an explanation of the new cardiovascular monitoring paradigm.

Henghes et al. (2021) [15] proposed a standard to compare the capacity and scalability of various machine learning approaches that were of supervised type for photometric red shift (photoz) measurement used in space machine learning techniques. Researchers used the Sloan Digital Sky Survey (SDSS-DR12) dataset to investigate various machine learning techniques. They were able to obtain many metrics that proved the usefulness level and scalability of the method for this task. They set one million as the number of galaxies that were used to train and test the algorithm to. In addition, they were able to understand how a small error concession can lead to a significant increase in efficiency done with a new optimization technique, time-based optimization. They found their model to work best with random forest method giving a mean square error of 0.0042. The model created meets the needs of future research, such as the Vera C. Rubin Observatory's Space-Time Legacy Survey (LSST), aimed at capturing billions of galaxies that require metered red shifts. is needed.

5. Findings and Conclusions

The researches that are studied and discussed as given above showcases the need and vitality of scalability in the machine-learning based approaches that depend on large scale datasets for their computation and decision making. Most of the studies have acknowledged the usefulness of scalability and they have extended their studies by developing an optimized scalability tool for their platform. This condition ensures a lack of universal scalability tool/method that can work under the principles of every/most frequently used machine learning platforms. Even, in multiple literatures, the use of combinational scalability approach is suggested due to incompatible performance of a single scaling tool. The scalability methods of machine-learning platforms are also found to be changing subject to changes made in selection of feature selection and extraction algorithms. Datasets also play a major role in implementing the correct scalability tool and ensure its usability. At present, most of the researches rely on selecting scalability tool based on comparative studies on performance measures. Also, the projects suffer faulty outputs and increase of time/cost based on wrong selection of scalability tool. Therefore, scalability is still an assessment dependent method where more focussed researches should be done to ensure uniformity and reliability.

6. Recommendations and Suggestions

Since the need and areas of usages of Scalability tools are increasing fast, the researches should pay deeper attention on planning up methodologies that encompasses overall machine-learning system and build up an optimal scalability concept workable for most of the available ML platforms. Most importantly, scalability approach should ensure the system adaptability so as to produce best output that help the ML platform in its better performance. Dataset design should also be given importance as most of the ML systems that need scalability require Big Datasets. So, dataset development schemes should subsequently be evolved and assessed so that it minimizes the effort of scalability stage.

ORCID iD

Vinod Sharma https://orcid.org/0009-0006-5885-0456

7. References

1. Cao Xi Hang, Stojkovic Ivan & Obradovic Zoran: A robust data scaling algorithm to improve classification accuracies in biomedical data. *BMC Bioinformatics*. *Vol-17*. Article number-359. 2016.
2. Elrafey Amr, wojtusiak janusz: Recent advances in scaling-down sampling methods in machine learning. *WIREs Computational Statistics*. 2017.
3. Hestness Joel, Narang Sharan, Ardalani Newsha, Diamos Gregory, Jun Heewoo, Kianinejad Hassan, Patwary Md. Mostofa Ali, Yang Yang, Zhou Yanqi: Deep Learning Scaling Is Predictable, Empirically. *Baidu Research*. 2017.
4. Mohassel Payman, Zhang Yupeng: SecureML: A System for Scalable Privacy-Preserving Machine Learning. *IEEE Symposium on Security and Privacy*. 2017.
5. Masegosa Andres R., Martınez Ana M., Ramos-Lopez Darıo, Caba˜nas Rafael, Salmeron Antonio, Nielsen Thomas D., Langseth Helge, Madsen Anders L.: AMIDST: a Java Toolbox for Scalable Probabilistic Machine Learning. *Arxiv.org*. 2017.
6. Wan Xing: *Influence of feature scaling on convergence of gradient iterative algorithm*. *Journal of Physics: Conference Series*. *IOP Publishing*. Number-1213. 2019.
7. Camilla Anna Karen Garatees, Hassani Amir Hajjam El, Andres Emmanuel: Big data scalability based on Spark Machine Learning Libraries. *Conference Paper*. 2019.
8. Karimipour Hadis, Dehghantanha Ali, Parizi Reza M., Choo Kim-Kwang Raymond, Leung Henry: A Deep and Scalable Unsupervised Machine Learning System for Cyber-Attack Detection in Large-Scale Smart Grids. *Special Section On Digital Forensics Through Multimedia Source Inference*. 2019
9. Arafatur Md, Taufiq Rahmana A., Leonga Asyharib L.S., Satryac G.B., Taod M. Hai, Zolkipli M.F.: Scalable machine learning-based intrusion detection system for IoT-enabled smart cities. *Sustainable Cities and Society*. *Vol-61*. 2020
10. Singh Dalwinder, Singh Birmohan: Investigating the impact of data normalization on classification performance. *Applied Soft Computing*. *Vol-97*. 2020.
11. Wikner Alexander, Pathak Jaideep, Hunt Brian, Girvan Michelle, Arcomano Troy, Szunyogh Istvan, Pomerance Andrew, and Ott Edward: Combining machine learning with knowledge-based modeling for scalable forecasting subgrid-scale closure of large complex spatiotemporal systems *Chaos: An Interdisciplinary Journal of Nonlinear Science*. *Vol-30*. Issue-5. 2020.
12. Wang Jiahao, Boukerche Azzedine: The Scalability Analysis of Machine Learning Based Models in Road Traffic Flow Prediction. *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*. 2020.
13. Ahsan Md Manjurul, Mahmud M. A. Parvez, Saha Pritom Kumar, Gupta Kishor Datta, Siddique Zahed: Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. *Technologies*. *Vol-9*. *52*. 2021.
14. Krittanawong Chayakrit, Rogers Albert J., Johnson Kipp W., Wang Zhen, Turakhia Mintu P., Halperin Jonathan L. & Narayan Sanjiv M.: Integration of novel monitoring devices with machine learning technology for scalable cardiovascular management. *Nature Reviews Cardiology*. *Vol-18*. pp-75–91. 2021.
15. Henghes 15. Ben, Pettitt Connor, Thiyagalingam Jeyan, Hey Tony, Lahav Ofer: Benchmarking scalability of machine-learning methods for photometric redshift estimation. *Monthly Notices of the RoyalAstronomical Society*. *Vol-505*. Issue-4. pp-4847–4856. 2021.

16. Verma: Yugesh Why Data Scaling is important in Machine Learning & How to effectively do it. *Developers Corner*. 2021
17. Zhou Lina: Machine Learning on Big Data: Opportunities and Challenges. *National Science Foundation*. 2017
18. Provost Foster: A Survey of Methods for Scaling Up Inductive Algorithms. Kluwer Academic Publishers. 1997.