International Journal of Engineering & Scientific Research
Vol.11 Issue 04, April 2023
ISSN: 2347-6532 Impact Factor: 6.660
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Recommendations to Advance the Cloud Data Analytics and Chatbots by Using Machine Learning Technology

Arjun Reddy Kunduru

Abstract

The selection of machine learning tools for data analytics might be challenging due to the ever-growing number of alternatives. The various tools each have benefits and limitations, and many of their applications overlap. The amount of data in the globe is expanding quickly, and as we transition to distributed and real-time processing, conventional machine learning methods are becoming inadequate.This study is designed to help researchers or professionals who are knowledgeable about machine learning but have no background in data analytics and chatbots. This essay provides a quick overview of machine learning and related tools.The majority of machine learning technologies are now more sophisticated and effective thanks to recent improvements. By employing a training set that accurately and effectively predicts the output, the different tools learn the machine. Various applications, including agriculture, data quality, information retrieval, financial market analysis, etc., use machine learning. Several tools, including Scikit Learn, Pytorch, Tensor Flow, Amazon Machine Learning, KNIME, Rapid Miner, Keras, and Shogun, have been covered in this essay along with their features and benefits.

Keywords:machine learning tools, data analysis, tensor flow, chatbot, KNIME

About Author: A seasoned Software Development Expert in the field of Cloud Computing and Enterprise Resource Planning systems with having more than 11 Years of experience. Author researching how modern technologies like Artificial Intelligence and Machine Learning can help Healthcare and Consumer Clients Cloud and Enterprise Infrastructureand holds expert knowledge on Enterprise BPM and ERP Applications.

1. Introduction:

Machine learning (ML) has been used by a variety of sectors, including banking, law enforcement, entertainment, commerce, and healthcare, as the cost of data storage has decreased and high-speed computers have been more readily available. Machine learning technologies are becoming more and more recognized as being not just practical but also essential to many corporate processes as theoretical research is applied to real-world problems [1]

Machine learning's goal is to give a computer the capacity to learn from its own history or from the world around it right now, and then to use that knowledge to predict or choose how to act in unforeseeable future situations. The process for a supervised machine learning activity may be broken down into three stages: constructing the model, evaluating, and optimizing the model, and

lastly putting the model to use in the task at hand. [2] An illustration of an example of this procedure may be seen in Fig. 1.

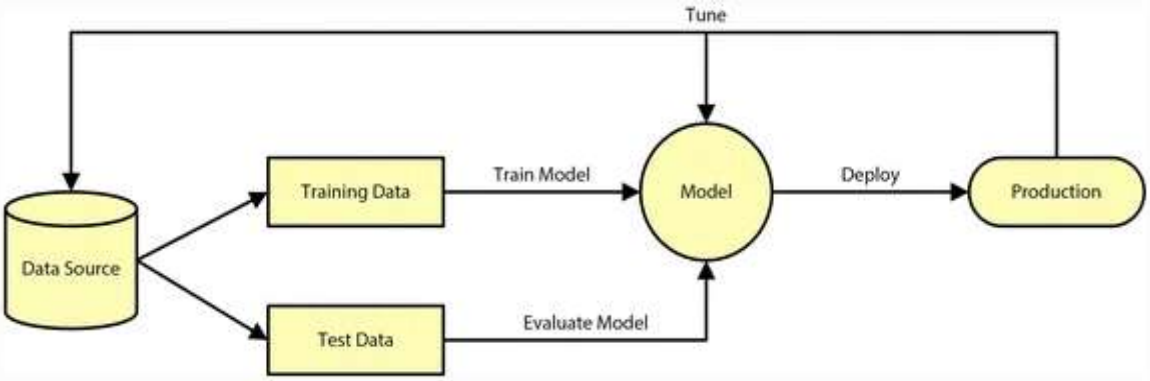

Figure 1: Process of Machine Learning

The data that are used to fuel the models constitute the core component of machine learning, and the advent of the age of big data is propelling machine learning to the forefront of academic and industrial applications worldwide. The phrase "big data" refers to information that is either too complicated or too extensive to be processed by a single system.Data is expanding orders of magnitude faster than previously in the modern era. The quantity of data on our globe is predicted to increase by ten times by 2020 to 44 zettabytes (4.4 1022 bytes), according to International Data Corporation's annual Digital Universe research. While no one organization is dealing with data of this size, numerous industries continue to produce data that is too big to be handled effectively using conventional methods. For instance, Ancestry.com has billions of entries with a combined 10 petabytes of data [3].The problem for the machine learning community is how to analyze and learn from large data as effectively as possible given the growing pace of data output. Weka and R, two widely used machine learning toolkits, weren't designed for this type of job. Weka includes distributed versions of several algorithms, but it is not as advanced as tools that were created from the ground up for terabyte-scale data. This paper's main objective is to describe the tools which offers a highly versatile platform that enables several machine learning projects and applications [4].

The increasing use of data analysis has compelled us to revaluate machine learning algorithm implementations in addition to data processing frameworks. For two reasons, selecting the right tools for a certain activity or setting may be challenging. First, several sorts of solutions may be needed due to the growing difficulty of the machine learning project needs as well as the data itself. Second, developers often start their own open source projects rather than contribute to already established ones because they believe the available tool choices to be unacceptable. As a result, there is a lot of fragmentation across the current data analysis systems. Both of these problems might make it more challenging to create a learning environment since many solutions have similar

use cases but differ in crucial ways. Since no one tool or framework can handle all or even the majority of frequently occurring tasks, it is important to weigh the benefits and drawbacks of different usability, performance, and algorithm selection options. when evaluating different solutions. Despite being extensively used at the business level and having no current industry standard, many of them lack thorough study [5].

This article's goal is to simplify these possibilities and make them more manageable by doing an in-depth analysis of the most recent and cutting-edge developments in scalable open-source software machine learning tools for data analytics and chatbots. To accomplish this, the post will provide a detailed examination of this topic. In this article, comparisons of several open source data processing engines, machine learning libraries, and frameworks are presented, along with suggestions for criteria that should be used while evaluating the various available options. It is expected that the reader of this text is familiar with the fundamental concepts and methods of machine learning. This article will be beneficial to anybody who is interested in big data and machine learning, including researchers, engineers, scientists, and software product managers.

2. Machine learning:

The field of research known as machine learning focuses on statistical models and algorithmic processes that perform certain tasks using patterns and inference as opposed to explicit instructions. According to a formal definition of algorithms that was offered by Tom M, it is stated that a computer programme can learn from its experiences. E with reference to a certain class of tasks T and performance measure P if its performance at tasks in T, as measured by P, increases with experience E. Mitchell. Long-standing Machine Learning Algorithms have the capacity to do intricate mathematical computations with massive data and provide quicker and more accurate solutions [6].

Three major categories may be used to broadly classify machine learning,

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Unsupervised learning determines its data structure from observation, whereas supervised learning requires guidance. Reinforcement learning, however, interacts with the environment directly and uses the hit-and-miss approach. The artificial intelligence-algorithmic applications known as "machine learning tools" provide systems the capacity to comprehend and advance without a great deal of human input.

Machine Learning application consists of

•Data gathering and preparation.

•Developing models

•Deployment of the application and training.

Platforms and libraries may be used to categorize machine learning tools in general. A machine learning tool lets you finish a project from start to finish, while machine learning packages let you finish a piece of a project.

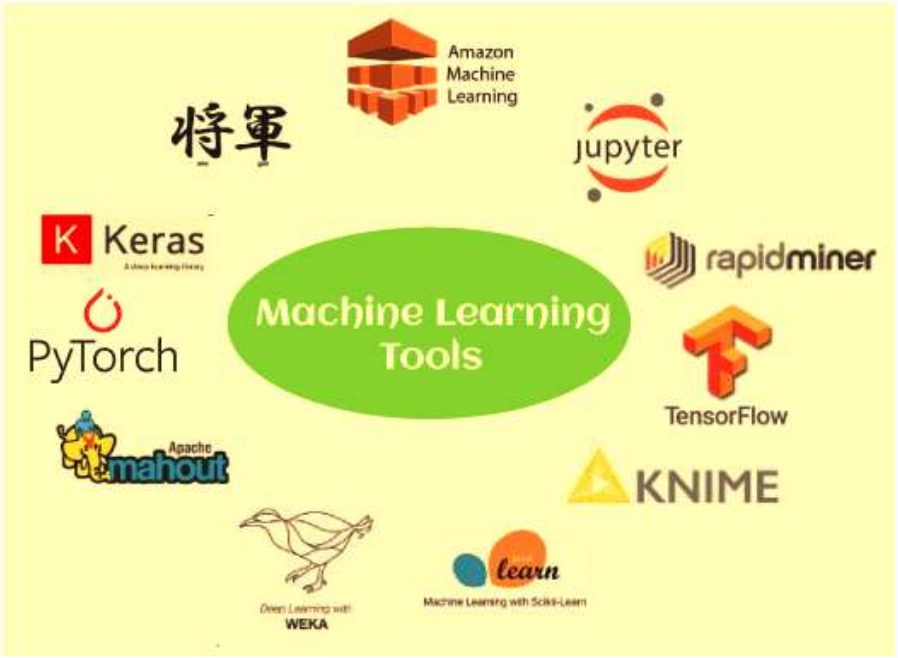

Figure 2: Types of Machine Learning tools

The past data may be used in supervised learning to produce predictions. These forecasts are more precise. Algorithms for regression and classification come under supervised learning. The hidden patterns were discovered via unsupervised learning. This kind of learning includes techniques like clustering and association rule mining. Reinforcement learning may be used when a system's efficiency has to be raised [7].

The reaction that is supplied by virtue of the role of decision-makingin order to describe the link that exists between the variables that are input into the ML model and the variables that are output from the model is referred to as the model interpretation. Interpretability is important for users because it increases their trust in the system, yet in many sectors it cannot be strengthened due to legal limitations or because it results in impartial conclusions [8]. Interpretability has a number of advantages, including the ability to:

a) patterns that are interpretable based on pre-trained machine learning models

b) pinpoint the causes of inaccurate predictions.

c) boost confidence in model predictions.

d) identify bias in ML models; and

e) develop safety measures to guard against overfitted models.

The most recent interpretable machine learning (IML) tools, IML techniques, and related open-source software resources will all be examined in this article.

3. Various Tools in Machine Learning

3.1 Scikit-learn [9]

Scikit-learn is used to construct Python machine learning algorithms. It provides a library for the Python programming language.

Features:

- It facilitates data mining and analysis.
- For classification, regression, clustering, dimensionality reduction, model selection, and pre-processing, it offers models and methods.

Pros:

- Documentation that is clear and well written is provided.
- While invoking objects, parameters for each particular algorithm may be altered.

3.2 PyTorch [10]

Torch is the foundation for PyTorch, a machine learning library built in Python. Based on the computer language Lua, the torch is a computational framework, scripting language, and machine learning library.

Features:

- The Autograd Module facilitates neural network construction.
- It offers several optimization strategies for constructing neural networks.
- Cloud platforms can utilize PyTorch.

Pros:

- It aids in the construction of computational graphs.
- Usability due to the hybrid front-end.

3.3 TensorFlow [11]

TensorFlow is one of the most prominent open-source libraries that is employed to train and create machine learning and deep learning models. It was originally developed by Google. The Google Brain Team was responsible for developing this product, which offers a JS library. It has a lot of popularity among those who are interested in machine learning, and those people utilize it to construct various ML applications.

It provides a robust library, tools, and resources for numerical computing, in particular for large-scale applications including machine learning and deep learning. It gives data scientists and machine learning developers the ability to rapidly construct and deploy machine learning applications. TensorFlow offers customers a high-level Keras application programming interface (API) for use in the training and construction of machine learning models. This simplifies the process of getting started with TensorFlow and machine learning.

Features:

- Aids in developing and training your models.
- TensorFlow.js, a model converter, may be used to execute your current models.
- The neural network benefits from it.
- You have two options for using it: script tags or NPM installation.
- It may even be useful for estimating human stance.

3.4 Weka [12]

Weka is a machine learning tool that runs on open-source software and has a graphical user interface. It is a user-friendly program that is used in both academic instruction and scientific investigation. There are many different uses for Weka in the industrial sector. In addition to this, it gives users access to a wide variety of additional machine learning technologies. Among them are Scikit-learn and R, among others.

Features:

- Preparing data
- Categorization
- Regression
- Clustering
- Visualization and
- Mining of association rules.
- Offers online training classes.
- Simple to comprehend algorithms.
- It benefits students as well.
- The amount of accessible online help and documentation is limited.

3.5 KNIME [13]

Knime is a graphical user interface (GUI) based machine learning application that is open source. It does not need any previous knowledge of coding to use. You are still able to carry out operations by making use of the capabilities that Knime provides.

Knime is often used for tasks that are concerned with data. Data mining, manipulation, and other similar activities are examples of this. The processing of data in Knime involves the creation of a variety of workflows and their subsequent execution. It contains repositories that are made up of a large number of nodes. The Knime portal is then opened with these nodes being dragged and dropped within. After that, a process consisting of nodes is established and run.

Features:

- It may combine the source code from many programming languages, including C, C++, R, Python, Java, and JavaScript.
- It may be used for CRM, financial data analysis, and business intelligence.
- It may serve as an SAS substitute.

- Installation and deployment are simple.
- Simple to learn.
- Complex models are challenging to construct.
- Limited capability for visualization and export.

3.6 Colab

Google now offers a workspace called Colab or Colab notebook. The foundation of this ecosystem is Jupyter Notebook. It is one of the market's most effective ML systems. The main difference is that Colab will be entirely cloud-based. On the Colab, you may use a variety of technologies, including TensorFlow, Pytorch, and Keras. Your Python abilities can be enhanced using Colab. For further processing, we may make use of a complimentary GPU supplied by Colab. Here, a storage option is Google Drive.

Features:

- It supports the teaching of machine learning.
- Aids in the study of machine learning.
- It's accessible from your Google Drive.

3.7 Apache Mahout[15]

The Apache Software Foundation's open-source project Apache Mahout is used to create machine learning programs with a primary emphasis on linear algebra. With its networked linear algebra architecture and mathematically expressive Scala DSL, the programmers may quickly put their own algorithms into practice. Additionally, it offers Java/Scala libraries for mathematical operations that are mostly focused on statistics and linear algebra.

Features:

- It offers methods for distributed linear algebra, regression, clustering, recommenders, and pre-processors.
- Common math operations may be performed using Java libraries.
- It works with big data sets.

3.8 Accord.Net [16]

A machine learning framework for scientific computing called Accord.Net is built on the.Net programming language. It is integrated with C#-written libraries for image and audio processing. For numerous machine learning applications, including pattern recognition, linear algebra, and statistical data processing, this framework offers a variety of libraries.

Features:

- Calculated linear algebra.
- Numerical improvement
- Statistics analysis
- Artificial Intelligence.
- Image, audio, & signal processing.
- It also offers assistance with libraries for graph plotting and visualization.

- Libraries are made accessible using the NuGet package management, executable installer, and the source code.

3. 9 Shogun [17]

Shogun is a machine learning software library that is free and open-source. It was developed in 1999 by Gunnar Raetsch and Soeren Sonnenburg. This C++ software library uses SWIG to offer interfaces for several languages, including Python, R, Scala, C#, Ruby, etc. (Simplified Wrapper and Interface Generator). Shogun's primary focus is on various kernel-based techniques for regression and classification issues, including Support Vector Machine (SVM), K-Means Clustering, etc. Additionally, it offers a full implementation of hidden markov models.

Features:

- Support vector machines for regression and classification are offered.
- It facilitates the use of Hidden Markov models.
- It can handle big data collections.
- Simple to use.
- Offers competent client service.
- Offers features and functions that are excellent.

3.10 Keras.io [18]

A nice machine learning tool for beginners is Keras. On top of Theano, TensorFlow, and CNTK runs Keras. It may produce RNN, CNN, or a hybrid of the two. The library has excellent usability and accessibility. Its design focuses on creating an API for people rather than machines.Keras is one of the well-liked machine learning tools for beginners.

Features:

- It may be used to quick and simple prototyping.
- Convolutional networks are supported.
- Recurrent networks benefit from it.
- It enables the fusion of two networks.
- Both the CPU and GPU can execute it.
- Extensible

3.11 Rapid Miner [19]

A platform for data science is Rapid Miner. It has a wonderful user interface and is quite beneficial to non-programmers. The operating systems supported by this machine learning tool are cross-platform. Businesses and industries often utilize it for fast testing of data and models. A user-friendly platform is provided via the fast miner interface. You may test your own model here using your own data. The object may simply be dropped into the UI by dragging it there. It is often used by non-programmers because of this.

3.12 Amazon Machine Learning [20]

Amazon provides a platform called Amazon ML. For machine learning applications, it offers a variety of services including Sagemaker and Redshift. When it comes to ML, Amazon is a significant participant on the global stage. Its research program is among the best in the world. Sagemaker, a tool offered by Amazon, aids in the creation and testing of models. Additionally, Amazon's deepracer is helpful in understanding reinforcement learning.

3.13 Caffe [21]

The Deep Learning library Caffe is widely utilized in the market right now. It offers effective expressiveness, quickness, and modularity. Berkeley University is where Caffe was developed. Caffe's foundational language is C++. But Python was used to create the interface. It is often used in the segmentation and categorization of images. CPU and GPU are supported by Caffe. Caffe2, which was developed by Facebook, also contains recurrent neural network capabilities.

3.14 Pandas [22]

One of the most fundamental and straightforward ML libraries to use is pandas. Python is often the application. Pandas do data processing and other tasks using data. It offers data structures that are quick and effective. Working with structured and time-series data is quite simple because to these data structures. Its objective is to become the world's most sophisticated tool for data manipulation.

3.15 H2O [23]

It is a further open-source machine learning framework with a business orientation. In order to make judgments based on detailed data, it facilitates the use of predictive analytics with mathematics. With open-source Breed technology, it supports database independence and trains robots using data insights. The REST API makes it possible to embed or access the Java-based core of H2O from any other source code or script. H2O may be expanded to operate with current programming languages and tools by machine learning experts. In addition to assessing insurance, advertising technology, risk, healthcare, and fraud, ML developers also use it in consumer intelligence.

Benefits of H2O

- H2O is adaptable.
- Using H2O for automatic ML is efficient.

- It is really simple to use for programmers with a variety of programming experience.

4. Chatbot Technology:

A chatbot is a piece of programme designed to mimic human conversation on a computer by deciphering customer enquiries employing artificial intelligence (AI) and natural language processing (NLP). Chatbots used to be text-based and programmed to provide pre-written replies to a limited set of simple inquiries. The creators failed when confronted with a challenging or unexpected issue. Although they did well for the specific questions and answers for the material they had been taught, they operated like an interactive FAQ. Over time, chatbots have included more rules and natural language processing to enable conversational interactions with users [24].

In fact, given their contextual awareness, contemporary chatbots could pick up new language as they come across more and more human speech. Modern AI chatbots employ natural language understanding (NLU) as a method to determine the user's demands. These technologies rely on machine learning and deep learning, two subtly different types of artificial intelligence (AI), to create an ever-more-detailed knowledge base of user interactions in the form of queries and responses. This improves their ability to correctly predict customer demands and respond over time [25].

For example, a classic chatbot might simply respond to a user's inquiry about the weather for tomorrow by saying whether it would rain or not. The user may also be asked by an AI chatbot whether they want to set an earlier alarm to take into account the longer morning commute. (because of rain).

Common chatbot uses:

Consumers may connect with mobile apps and utilize devices created especially for the purpose, including smart thermostats and smart kitchen appliances, by using AI chatbots for a range of purposes. Use for commercial purposes also varies. Marketers employ AI chatbots to customize customer experiences, IT teams use them to provide self-service, and customer contact centres use them to speed incoming messages and lead customers in the proper path. There are many different conversational interfaces [26]. AI chatbots are often utilized in internet applications, independent messaging platforms, and social media messaging apps. Typical usage scenarios include:

- Locating nearby eateries and giving instructions
- Specifying fields in financial applications and forms
- Obtaining information about healthcare issues and making appointments
- Getting general assistance from a preferred brand's customer service
- Setting a time or location-based reminder for a task
- Presenting current weather conditions and suggestions for appropriate attire

5. Conclusion:

To characterize cryptic and ambiguous ML models, we have provided an overview of machine learning techniques and the software tools like chatbot that go along with them. These tools make it easier to build models that can be understood. This study came to the conclusion that each ML tool has its limits and that not every kit for developing ML has a single interface. Additionally, we have included an introduction of ML concepts and instances of their use in various programming languages. We want to use a range of tools to improve our present supervised [27] and unsupervised [28] ensemble ML architecture, starting with this tools survey as a guide.Additionally, we want to broaden our investigation into DDoS assaults that have been discovered, choose crucial characteristics in comparison to EnFS approach [29], and monitor secure networks (both offline and in the cloud [30] using a variety of ML tools like alex, chatbots.

References:

1. Li, R., Li, L., Xu, Y., & Yang, J. (2022). Machine learning meets omics: applications and perspectives. Briefings in Bioinformatics, 23(1), bbab460.
2. Alpaydin, E. (2020). Introduction to machine learning. MIT press.
3. Moubayed, A., Injadat, M., Nassif, A. B., Lutfiyya, H., & Shami, A. (2018). E-learning: Challenges and research opportunities using machine learning & data analytics. IEEE Access, 6, 39117-39138.
4. Dong, G., & Liu, H. (Eds.). (2018). Feature engineering for machine learning and data analytics. CRC Press.
5. Mittal, R., Arora, S., Kuchhal, P., & Bhatia, M. P. S. (2021). An Insight into Tool and Software Used in AI, Machine Learning and Data Analytics. AI and Machine Learning Paradigms for Health Monitoring System: Intelligent Data Analytics, 45-64.
6. Subasi, A. (2020). Practical machine learning for data analysis using python. Academic Press.
7. Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., & DATA, M. (2005, June). Practical machine learning tools and techniques. In Data Mining (Vol. 2, No. 4).
8. Hopkins, E. (2022). Machine learning tools, algorithms, and techniques. Journal of Self-Governance and Management Economics, 10(1), 43-55.
9. Kramer, O., & Kramer, O. (2016). Scikit-learn. Machine learning for evolution strategies, 45-53.
10. Imambi, S., Prakash, K. B., & Kanagachidambaresan, G. R. (2021). PyTorch. Programming with TensorFlow: Solution for Edge Computing Applications, 87-104.

11. Pang, B., Nijkamp, E., & Wu, Y. N. (2020). Deep learning with tensorflow: A review. Journal of Educational and Behavioral Statistics, 45(2), 227-248.
12. Holmes, G., Donkin, A., & Witten, I. H. (1994, November). Weka: A machine learning workbench. In Proceedings of ANZIIS'94-Australian New Zealnd Intelligent Information Systems Conference (pp. 357-361). IEEE.
13. Jagla, B., Wiswedel, B., & Coppée, J. Y. (2011). Extending KNIME for next-generation sequencing data analysis. Bioinformatics, 27(20), 2907-2909.
14. Hoyos-Rivera, G. J., Gomes, R. L., Willrich, R., & Courtiat, J. P. (2006). Colab: A new paradigm and tool for collaboratively browsing the web. IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans, 36(6), 1074-1085.
15. Anil, R., Capan, G., Drost-Fromm, I., Dunning, T., Friedman, E., Grant, T., ... & Yilmazel, Ö. (2020). Apache mahout: machine learning on distributed dataflow systems. The Journal of Machine Learning Research, 21(1), 4999-5004.
16. de Souza, C. R. (2012). A tutorial on principal component analysis with the accord. net framework. arXiv preprint arXiv:1210.7463.
17. Sonnenburg, S., Rätsch, G., Henschel, S., Widmer, C., Behr, J., Zien, A., ... & Franc, V. (2010). The SHOGUN machine learning toolbox. The Journal of Machine Learning Research, 11, 1799-1802.
18. Ketkar, N., & Ketkar, N. (2017). Introduction to keras. Deep learning with python: a hands-on introduction, 97-111.
19. Hofmann, M., & Klinkenberg, R. (Eds.). (2016). RapidMiner: Data mining use cases and business analytics applications. CRC Press.
20. Herbrich, R. (2017). Machine Learning at Amazon. WSDM, 535.
21. Kovalev, V., Kalinovsky, A., & Kovalev, S. (2016). Deep learning with theano, torch, caffe, tensorflow, and deeplearning4j: Which one is the best in speed and accuracy?.
22. Snider, L. A., & Swedo, S. E. (2004). PANDAS: current status and directions for research. Molecular psychiatry, 9(10), 900-907.
23. Candel, A., Parmar, V., LeDell, E., & Arora, A. (2016). Deep learning with H2O. H2O. ai Inc, 1-21.
24. Adamopoulou, E., & Moussiades, L. (2020). An overview of chatbot technology. In Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part II 16 (pp. 373-383). Springer International Publishing.
25. Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. Machine Learning with Applications, 2, 100006.

26. Hristidis, V. (2018, September). Chatbot technologies and challenges. In 2018 First International Conference on Artificial Intelligence for Industries (AI4I) (pp. 126-126). IEEE.

27. Das, Saikat, Deepak Venugopal, and Sajjan Shiva. "A Holistic Approach for Detecting DDoS Attacks by Using Ensemble Unsupervised Machine Learning." Future of Information and Communication Conference. Springer, Cham, 2020.
28. Das, Saikat, et al. "Empirical Evaluation of the Ensemble Framework for Feature Selection in DDoS Attack." 2020 7th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2020 6th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom). IEEE, 2020.
29. Das, Saikat, and Sajjan Shiva. "CoRuM: collaborative runtime monitor framework for application security." 2018 IEEE/ACM International Conference on Utility and Cloud Computing Companion (UCC Companion). IEEE, 2018.
30. Das, Saikat, Ahmed M. Mahfouz, and Sajjan Shiva. "A Stealth Migration Approach to Moving Target Defense in Cloud Computing." Proceedings of the Future Technologies Conference. Springer, Cham, 2019