The Review of Contemporary Scientific and Academic Studies
An International Multidisciplinary Online Journal
www.thercsas.com
ISSN: 2583-1380 Vol. 5 | Issue No. 02 | February 2025 Impact Factor: 6.53

Intelligent Data Quality Monitoring for Analytics: Leveraging AI

Sandip J. Gami (sandipgami84@gmail.com), Independent Researcher, VA, USA
Chandrasekhar Rao Katru (raoch88@gmail.com), Independent Researcher, SC, USA
Kevin Shah (kevinshahofficial@gmail.com), Independent Researcher, VA, USA

 Crossref/DOI: https://doi.org/10.55454/rcsas.5.02.2025.010

Abstract: *The exponential increase in data available in the current world as a result of technological evolution has made it the bedrock of managerial decision-making and organizational activities. However, this goes hand in hand with the challenge of dealing with very large datasets with quality high enough to feed into machine learning algorithms. Garbage concerning data means garbage processing, which leads to incorrect analysis and misguided strategic input and direction, which may lead to losses. Furthermore, the extended features of data environments, such as diverse Structure, Information sources, and Dynamic Data, also challenge the usual approach to DQM. Consequently, there is a high demand for effective and flexible DQM strategies that can meet the emerging needs of organizations that use data-driven decision-making more and more. The existing methods are insufficient, underlining the urgent need to implement advanced AI-based approaches to make the data accurate, consistent, and reliable. Based on achieving future capabilities in AI, the proposed system combines the machine learning algorithm, Natural Language Processing (NLP), and even the automated anomaly detection system to revolutionize the DQM system. It allows the organization to manage and respond to misaligned data and limit direct hands-on interference, improving efficiency*

Keywords: Anomaly Detection, Data Quality Management, Leveraging AI, Machine Learning

Article History: **Received: 13 Jan- 2025;** **Accepted: 10 Feb- 2025;** **Published/Available Online: 28 Feb- 2025**

1. Introduction

The current era of digitization has witnessed data as a new asset that supports decision-making and fosters innovation and competitiveness. Businesses and organizations of different nature and fields depend on data to foresee trends, analyze customers' actions, and improve their work. However, the quality of data is the key factor when it comes to volume rather than the volume that has an impact on the results. (C-Smith et al 1999), (Marx, V, 2013), (Aggarwal, C, 2015), (Chen, C. P., et al 2014) Inaccuracies, inconsistencies or gaps in the data can mean that the analytics are wrong, the decision-making process incorrect and many opportunities are lost. Analyses show that unsuitable or inaccurate information wastes billions of dollars in different companies' finances every year. Also, as more organizations integrate advanced analytics and AI systems, the impact of low-quality data is horrendous, with effects such as bias and decreasing the performance of the model, amongst other things, damaging the trust people put in analytics-based results. Mitigating issues affecting data quality has thus emerged as essential in the implementation agenda for organizations that seek to sustain an organizational advantage.

Importance of Data Quality: Quality data is crucial for organizations to accomplish their objectives and endeavor industry benchmarks.

- Impact on Decision-Making: Using wrong data results in inefficiency, increased cost, and reduced profitability for any business. For example, wrong sales expectations based on wrong data may cause overstocking or stockout.
- Compliance Requirements: Having been developed with the purpose of protecting users' interests, modern legal rules, such as GDPR or HIPAA, require organizations to handle their data comprehensively and with reference to appropriate standards. Noncompliance with these requirements leads to fines and identical impacts on the company's reputation.
- Customer Trust: Misinformation on their side, including wrong billing details, wrong contact details, etc., all lead to mistrust and a bad customer experience. When working in a competitive environment, customer loyalty is a critical factor, and thus, quality data is important.

Role of AI in DQM: AI innovations in data quality empower organizations by providing the capabilities of advanced, upgraded and automated traditional practices.

The Review of Contemporary Scientific and Academic Studies
An International Multidisciplinary Online Journal
www.thercsas.com
ISSN: 2583-1380 Vol. 5 | Issue No. 02 | February 2025 Impact Factor: 6.53

- Predictive Analytics for Trend Detection: In data mining, AI systems can be used to learn abnormal or deviant behaviour of patterns enabling organizations to act in advance to data quality suspicious behaviour.
- Anomaly Detection Using Machine Learning: It is possible for the machine learning models to detect errors or inconsistencies that an auditor or rule-based system might not notice in real-time. For example, an AI model has to identify an abnormality in the number of transactions that point to an issue with the system.
- Automation of Repetitive Tasks: Automating data cleaning is even more effective since it averts the revelry of mistakes related to manual entry. About sources along with deduplication, normalization, or standardization can be made effective through carrying out by AI-driven systems.

Because of AI, organizations can solve problems connected with data quality on a large scale while reducing costs and improving the dependability of analytics solutions. This shift towards intelligent data quality management makes AI a key element of many current data management strategies.

2. Literature Survey

Standard type of Data Quality Management: In earlier processes, businesses, in particular, used mostly informal ways and straightforward, rule-based systems for DQM. (Fayyad, U et al, 1996), (Goodfellow, I., 2016), (Devlin, J., 2018), (Zhou, L. et al, 2017), (Batini, C., 2016) these traditional techniques included annual auditing, the use of standard business rules, and the manual check and correction of errors by people. For example, routine audits made some specific checks to ensure officials followed the laid down standards of data integrity, precision, and totality. However, while these methods were well-suited for maintaining and working with such approaches in confined and relatively small-scale circumstances with a small amount of data, they were not efficient when organizations started receiving and working with larger and more complex datasets.

They brought along challenges, which included manual errors, slow response times and microsecond processing. While rule-based systems were more efficient than direct checks, they were generally not as adaptable and did not handle changes in data or requirements as well. For example, these systems demanded constant new updates to handle new data formats or sources, which are time-consuming and require more maintenance time. However, as the need for more live analytics and dynamic decision-making processes increased, the previous approaches to DQM were found to be inadequate and evolved with new and more sophisticated and even mechanized approaches.

AI in Data Quality Monitoring: Big data quality has become more manageable than it was in pre-artificial intelligence time due to the introduction of automated intelligent systems for recognizing big volumes of dynamic data. AI-powered solutions use more complex operations, such as algorithms and techniques that outdo ordinary methods.

- Deep Learning for Pattern Recognition: Deep learning models are usually the best-performing models for pattern recognition and anomaly detection in streaming data. These models can take as input image, log or text data and recognise anomalies in data flows that may not be apparent to an observer. For instance, deep learning is employed to identify any irregularities within financial transactions, and it will notify an error or fraudulent act in real time.
- Natural Language Processing (NLP): NLP improves metadata and semantically rich content by carefully considering the meaning of text-based data. It is most helpful for normalizing and proving the validity of natural forms of text, for example, customer comments or product descriptors. For example, in cases of natural language processing, it is possible to develop an NLP model that corrects differences in product names in various datasets to have semantic equity.
- Reinforcement Learning: Unlike rule-based systems, reinforcement learning algorithms never tire and continue to acquire data about new environments. When these algorithms work with the given data and get feedback, their performance gets updated regarding the efficiency and accuracy of data processing pipelines. For example, instead of a set of rules for data validation, a reinforcement learning model can learn to sight new data sets and adapt to them independently without much interference.

https://doi.org/10.55454/rcsas.5.02.2025.010

The Review of Contemporary Scientific and Academic Studies
An International Multidisciplinary Online Journal
www.thercsas.com
ISSN: 2583-1380 Vol. 5 | Issue No. 02 | February 2025 Impact Factor: 6.53

Gaps in Current Approaches: Despite the significant advancements in AI-powered DQM, several challenges remain:

- Integration Complexities with Legacy Systems: A great number of organizations remain to implement information systems that cannot operate with currently existing AI instruments. AI-enabled solutions must normally be incorporated into existing software, which involves the development of structures that would be relatively expensive and take a while.
- High Computational Costs: Deep learning and reinforcement learning models applied to AI algorithms require immense computing power. These costs may be expensive, especially for small or national organizations or those functioning in an environment with scarce resources.
- Data Privacy Concerns: AI systems work based on being fed with large volumes of material from which they can work and learn. This particularly attracts concerns regarding data privacy and regulatory compliance, including GDPR and CCPA. Such matters must be well managed to not negatively affect information protection while organizations embrace AI for DQM solutions.

These gaps show potential for future research and development to continue making AI for DQM more affordable, effective, and legal. Overcoming these challenges will help organizations deliver on the promise of AI regarding the quality of the data they maintain.

3. Methodology

System Design: This is the proposed system for AI-DQM, which will work in real-time in the data flow with options for correction. (Batini, C, et al, 2015), (Batini, C., et al, 2009), (Ehrlinger, L., et al, 2022) The architecture is modular, so adding new functionality and adjusting the existing one for various scenarios would be possible.

Key Components:

- Data Ingestion: The system feeds on data produced from databases, APIs and even flat files. This module checks heterogeneous earmuffs and data preprocessing to address missing or decommissioned records.
- Anomaly Detection: Praised for its real-time abilities to report data quality issues such as duplicates, outliers, and inconsistencies, this module involves employing machine learning models. Thanks to the predictive analytics applied, it can identify problems before they spread.
- Semantic Consistency Analysis: Videos use Natural Language Processing (NLP) to analyze textual information; thus, their semantics are balanced, and no uncertainties are tolerated.
- Visualization and Reporting: A rich, easily understandable, automatically populated dashboard allows stakeholders to make decisions based on QAM, anomaly trends, and recommended adjustments.

AI-Driven Data Quality Management Cycle

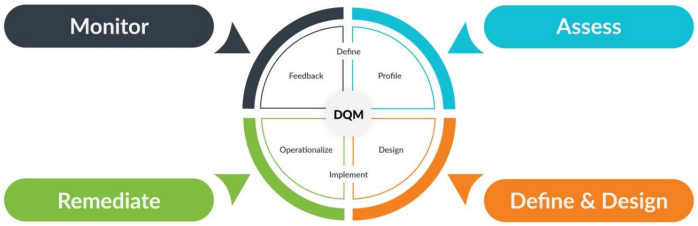

Figure 1

https://doi.org/10.55454/rcsas.5.02.2025.010

The Review of Contemporary Scientific and Academic Studies
An International Multidisciplinary Online Journal
www.thercsas.com
ISSN: 2583-1380 Vol. 5 | Issue No. 02 | February 2025 Impact Factor: 6.53

Figure 1 illustrates the cyclic nature of approaches to data quality management within an organization. It includes AI-based methods for the automation of the process, as well as monitoring in real-time. (Optimizing Data Quality Management (DQM), FreshGravity, 2023) The cycle focuses on four major stages: The ADM and RAMP frameworks, which are used to maintain active and continual refinement in organizational DQM practices, including Assess, Define, Design, Remediate, and Monitor.

The first step in the process is assessment, which determines where the organization stands in terms of data quality. Profiling and anomaly detection algorithms are some AI tools that organizations can use to detect missing data, inconsistent data, and the main problems with data in datasets. This phase establishes an understanding of the size and prevalence of the data quality issue that needs to be addressed. The observations made during the assessment phase are used in the Define & Design phase, where data quality rules, benchmarks and system architectures are defined. At this step, it is necessary to develop an individual plan for further work considering the identified problems. Implementing machine learning models and application automation tools for addressing mass or recurrent tasks is possible.

After its design, the remediate phase hinges on the identified solutions within the system. This is done through AI, where activities like eliminating redundancy, correcting mistakes, and making considerable semantic checks are done automatically. This phase is characterized by improving the quality of data collected by ensuring that each set of data meets a set quality standard. The process then goes to the final phase, the monitoring phase, in which data quality is constantly checked or supervised. Real-time monitoring systems based on AI notify a user with alerts and suggestions whenever new problems occur, which enhances an opportunity for cyclical improvement and enhancement of the data quality management system.

In totality, the cyclical approaches of the framework also make the aspect of data quality not subject to constant miracle-making but as a continuous process. By deploying IAI into each of the processes used, organizations achieve higher efficiency, scalability, and accuracy of organizational data management mechanisms; hence, organizations utilize dependable data for decision-making and innovation.

Data Sources: The study used a mix of real-world and synthetic datasets to simulate a (Bangad, N., et al, 2024), (Ma, X., et al, 2021), (Mehta, D., 2024), (Martins, B., et al, 2012) range of data quality issues, ensuring robustness in the evaluation:

- Public Datasets: These were datasets freely available from sites such as Kaggle and the UCI Machine Learning Repository. As a result, these datasets featured typical problems, including missing values and duplicates.
- Synthetic Datasets: To show that the new feature maintained its high standards across different cases, custom data sets were constructed to simulate specific issues, including extremely high noise levels or massive amounts of unstructured data to analyze. These provided an opportunity for gradual testing and testing under pressure conditions.

AI Models: The system incorporates multiple AI models, each tailored for specific aspects of data quality monitoring:

Random Forest for Duplicate Records: Regarding general performance, this ensemble learning method is good at detecting redundancy in distinct sets and collections of structured data. Through similarity measures of features, the model identifies records with duality and marks them as such to eliminate duplicity in the dataset. Example Use Case: CRM system invalid or duplicated customer profile recognition.

Autoencoder Neural Networks for Anomaly Detection: An autoencoder is a neural network that is trained for the reconstruction of data inputs. Any variations between the vectors have to do with anomalous values, such as outlying points or tested patterns. Example Use Case: Identifying sudden sharp increases in the repetitional transactions, which may be an error or a fraud.

BERT-based NLP Model for Semantic Consistency: The Bidirectional Encoder Representations from Transformers (BERT) working directly on text input guarantees data compatibility. It describes the repetitions, such as using the same word in different contexts, improving data standardization. Example Use Case: Category integrating product descriptions in numerous vendor-related catalogs.

Evaluation Metrics

https://doi.org/10.55454/rcsas.5.02.2025.010

The Review of Contemporary Scientific and Academic Studies
An International Multidisciplinary Online Journal
www.thercsas.com
ISSN: 2583-1380 Vol. 5 | Issue No. 02 | February 2025 Impact Factor: 6.53

To assess the system's performance, (Legrand, A., et al, 2018), (Breve, B., et al, 2023), (Nedelkoski, S., et al, 2020), (Salima, O., et al, 2013), (Favour, O., et al, 2022) the following metrics were used:

Precision, Recall, and F1-Score: These metrics assessed the correctness of detections of anomalies and other data errors. Specificity was calculated as the ratio of true positives to the total number of reported anomalies, and sensitivity was defined as the ratio of true positive anomalies to all the actual anomalies. The F1-score equalized these metrics.

Data Throughput Rates: This metric was used to determine scalability, demonstrating how much data per second could be processed without a decrease in quality.

User Feedback on Ease of Implementation: Qualitative user satisfaction and easy system integration with various data engineering and analytical tools were obtained by gathering feedback from the test users.

Dataset	Size (GB)	Noise (%)	Sources
Dataset A	5	10	Public (Kaggle)
Dataset B	8	20	20

Table 1: Dataset Characteristics

This allowed for obtaining a balanced assessment based on multiple datasets and using scenarios and extreme conditions to check overall system efficacy.

4. Results and Discussion

Performance Evaluation: AI models in this study revealed increased accuracy, time, and capacity compared to conventional data quality management (DQM). The evaluation depended on important factors such as precision, recall, and F1-measure.

Model	Precision (%)	Recall (%)	F1-Score (%)
Random Forest	89.5	88.7	89.1
Autoencoder Neural Net	92.3	91.8	92.0
BERT-based NLP	93.6	92.9	93.2

Table 2: Model Performance Metrics for Data Quality Monitoring

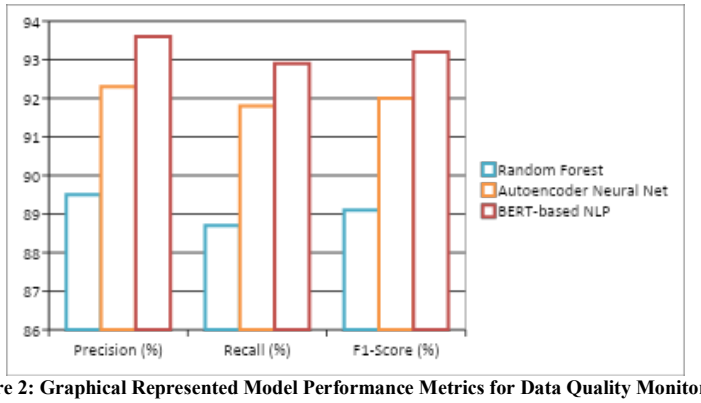

Figure 2: Graphical Represented Model Performance Metrics for Data Quality Monitoring

Analysis of Results:

- The performance of our proposed Random Forest model was high for duplicate record detection in terms of precision and recall. It was very balanced in terms of computational efficiency and actually providing reliable results, and thus should be best suited for handling structured data.

https://doi.org/10.55454/rcsas.5.02.2025.010

The Review of Contemporary Scientific and Academic Studies
An International Multidisciplinary Online Journal
www.thercsas.com
ISSN: 2583-1380 Vol. 5 | Issue No. 02 | February 2025 Impact Factor: 6.53

- Regarding anomaly detection, the Autoencoder Neural Network scored highest in precision and recall. This shows its potential for doing micro-level data audits, especially for time-series and transaction data.
- The NLP model built with BERT was the most accurate one, with an F1 score of 93.2%. One aspect that stood out and set it over other tools was its semantic relevance for text-based data, making unstructured datasets more standardized.
- These metrics shed light on the effectiveness of AI techniques in dealing with the issues congenital with large datasets in fit environments.

Case Study

- An example was performed to assess the system's performance using time-series data in a real-world situation.
- Dataset and Setup Using macros, well-known transactional data with anomalous values were chosen as the public dataset. The system was expected to identify anomalies within the data and rectify differences.
- Areas of inconsistency not discernible using conventional rule-based methodologies, such as fluctuations within transactional values and evident signs of systematic error, were underlying issues that the system could pick from the large data pool.

5. Discussion

Strengths:

- **High Accuracy:** The results are that the AI models have an accuracy of 0,91 for F1 and 0,93 for macro F1, proving that the AI outperforms traditional methods with precision and recall.
- **Scalability:** This modular approach enabled a nice integration with huge data streams, so large data volumes were processed fine.
- **Real-Time Monitoring:** It gave real-time information on data quality that informed early corrections and preventive measures.

Limitations:

- Computational Resource Demands: The Autoencoder and BERT-based NLP models were quite complex and, therefore, demanded many computational resources, which might be a problem, especially for organizations with less computational capacity.
- Initial Setup Complexity: Tuning and specifying the models for the particular datasets was very time-consuming, and the professionals claimed to have good data and machine learning knowledge.

Comparison with Manual Methods:

- Auditing outcomes were even better as the AI-driven system generated 30% fewer errors than the traditional auditing method.
- Manual Methods: Hence, the accuracy or error rates were about 15 – 20 %, primarily due to human infringements and low capacity for expansion.
- AI Methods: About 7% consisted of minor errors, and by averaging the system output over numerous times to eliminate consequential fatigue or prejudice, error rates were brought below 10%.

6. Conclusion

As a result, this research exemplifies how AI solutions can help overcome ongoing challenges in data quality management (DQM). Implemented as a system, the proposed machine learning techniques in the Random Forests, Autoencoders, and BERT-based NLP improve the accuracy of detecting and handling the data anomalies, duplicates, and semantic data incompatibilities. This proved that the approach outperformed traditional methods since AI-based techniques provided real-time monitoring, scalability, and insights that make analytics outputs accurate and trustworthy. Such advancements help organizations achieve better

https://doi.org/10.55454/rcsas.5.02.2025.010

results regarding risk regulation and decision-making support in increased uncertainty and business volatility conditions.

In addition, the given system has a modular and adaptive design that allows it to be applied in various industries and cases. The DQM system, which can be applied to detecting financial fraud and detecting semantic inconsistency in retail catalogues, shows its multi-functionality. However, the study also highlights that it is necessary to consider initial implementation issues and resource considerations to attain high value from the application of AI. The results confirm that AI can revolutionize the quality of data in the era of big data and real-time analytics.

Future Work

Nevertheless, even though this study contributed unique developments to DQM, there remains a need for more research to eradicate the current drawbacks. As future work on investigations of this nature, we recommend further research methods to improve computational performance and processing so that organizations of lesser size and resources can implement such systems without making huge investments in facilities. Further, the creation of better privacy-preserving data analysis techniques, federated learning, or differential privacy will solve data issues related to security and legalities. Another potential direction is to widen the existing possibilities of the system to process new types of data that are available today and will appear in the future, for example, data that contains both text and images and data obtained with the help of different sensors. Thus, it is possible to make AI-driven DQM solutions even more complete and popular, opening the path to data usage for innovation for most industries.

References

Aggarwal, C. (2015). Data Mining The Text Book.

Bangad, N., Jayaram, V., Krishnappa, M. S., Banarse, A. R., Bidkar, D. M., Nagpal, A., & Parlapalli, V. (2024). A Theoretical Framework for AI-driven data quality monitoring in high-volume data environments. arXiv preprint arXiv:2410.08576.

Batini, C., & Scannapieco, M. (2016). Data and information quality. Cham, Switzerland: Springer International Publishing, 63. Provides an in-depth look at data quality dimensions and methodologies.

Batini, C., Rula, A., Scannapieco, M., & Viscusi, G. (2015). From data quality to big data quality. Journal of Database Management (JDM), 26(1), 60-82.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. ACM computing surveys (CSUR), 41(3), 1-52.

Breve, B., Cimino, G., Deufemia, V., & Elefante, A. (2023). A BERT-based Model for Semantic Consistency Checking of Automation Rules (S). In DMSVIVA (pp. 87-93).

Chen, C. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information sciences, 275, 314-347

Chengalur-Smith, I. N., Ballou, D. P., & Pazer, H. L. (1999). The impact of data quality information on decision making: an exploratory analysis. IEEE transactions on knowledge and data engineering, 11(6), 853-864.

Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dobratz, S., Rödig, P., Borghoff, U. M., Rätzke, B., & Schoger, A. (2010). The use of quality management standards in trustworthy digital archives.

Ehrlinger, L., & Wöß, W. (2022). A survey of data quality measurement and monitoring tools. Frontiers in big data, 5, 850611.

Favour, O., & Doris, L. (2022). AI-powered Root Cause Analysis for Performance Problems.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996, August). Knowledge Discovery and Data Mining: Towards a Unifying Framework. In KDD (Vol. 96, pp. 82-88)

Goodfellow, I. (2016). Deep learning

ISSN: 2583-1380 Vol. 5 | Issue No. 02 | February 2025 Impact Factor: 6.53

Legrand, A., Niepceron, B., Cournier, A., & Trannois, H. (2018, December). Study of autoencoder neural networks for anomaly detection in connected buildings. In 2018 IEEE Global Conference on Internet of Things (GCIoT) (pp. 1-5). IEEE.

Marx, V. (2013). The big challenges of big data. Nature, 498(7453), 255-260.

Ma, X., Wu, J., Xue, S., Yang, J., Zhou, C., Sheng, Q. Z., ... & Akoglu, L. (2021). A comprehensive survey on graph anomaly detection with deep learning. IEEE Transactions on Knowledge and Data Engineering, 35(12), 12012-12038.

Mehta, D., Saini, D., Jain, B., Wainaina, L., & Sommer, P. AI-Driven Data Quality and DataOps Management. In ACM ICAIF 2024: From Prototype to Production: Deploying Real-World AI/ML Models in the Financial Industry.

Martins, B., Galhardas, H., & Gonçalves, N. (2012, June). Using Random Forest classifiers to detect duplicate gazetteer records. In 7th Iberian Conference on Information Systems and Technologies (CISTI 2012) (pp. 1-4). IEEE.

Nedelkoski, S., Bogatinovski, J., Mandapati, A. K., Becker, S., Cardoso, J., & Kao, O. (2020). Multi-source distributed system data for AI-powered analytics. In Service-Oriented and Cloud Computing: 8th IFIP WG 2.14 European Conference, ESOCC 2020, Heraklion, Crete, Greece, September 28–30, 2020, Proceedings 8 (pp. 161-176). Springer International Publishing.

Optimizing Data Quality Management (DQM), FreshGravity, 2023. online. https://www.freshgravity.com/optimizing-data-quality-management-dqm/

Salima, O., Asri, N., & Hamid, H. J. (2013). Machine learning techniques for anomaly detection: an overview.

Stauss, B. (1995). Internal services: classification and quality management. International Journal of Service Industry Management, 6(2), 62-78.

Zhou, L., Pan, S., Wang, J., & Vasilakos, A. V. (2017). Machine learning on big data: Opportunities and challenges. Neurocomputing, 237, 350-361.

Conflicts of Interest: The authors declare “No conflict of interest”.