International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 3
Article Received: *28 January 2023* ***Revised:*** *05 February 2023* ***Accepted:*** *12 March 2023*

Optimizing Data Collection Processes with AI: Practical Applications and Efficiency Analysis

[1]Samartha.G.Naik, [2] Dr.Kailash Patidar
[1]Research Scholar, Department of Computer Science and Engineering,
Dr. A.P.J. Abdul Kalam University, Indore
[2]Research Guide, Department of Computer Science and Engineering,
Dr. A.P.J. Abdul Kalam University, Indore
1 samarthanaik32@gmail.com, 2 kailashpatidar123@gmail.com

ABSTRACT

The extraction of information from multimedia files is a critical component in the search for new multimedia information. This study delves into the construction of multimedia systems, examining various data mining techniques and the broader concept of data mining. It provides a comprehensive analysis of the latest approaches, algorithms, and methodologies employed in multimedia data mining. Recognized as one of the most challenging and significant fields in computer science, multimedia data mining addresses numerous complex research questions. This paper explores innovative algorithms and strategies to uncover hidden information within multimedia data sources, discussing the key principles, structures, models, and applications. Additionally, it highlights recent advancements and identifies ongoing challenges in the field.

Keywords: Data mining, Artificial intelligence, multimedia, Gan, SEARs, Proposed MR-DFL

INTRODUCTION

Synthetic labeled data is inexpensive and versatile, making it popular in machine learning. Scikit-learn facilitates sampling from probability distributions, while Generative Adversarial Networks (GANs) and application-specific methods offer advanced generation techniques. GANs involve two competing networks: a discriminative network that differentiates between real and generated data, and a generative network that learns to produce data indistinguishable from the real distribution. GANs are widely used to create realistic synthetic images and videos, and recent advancements include MEDGANs for patient records and TABLE-GAN for privacy-preserving table generation. Techniques for generating synthetic data in computer vision include 3D model updates for object recognition and enhancing image datasets with clear images from fuzzy ones. In natural language processing, methods for generating synthetic text include paraphrasing and semantically equivalent adversarial rules (SEARs) to improve model robustness by varying sentence structures while retaining meaning.

DATA LABELING

Finding and grouping examples follows when adequate information is available. In smart factories, workers may annotate photos of industrial parts to identify defects, often labeling data during collection. This process assumes the data is genuine, making labeling literature more straightforward than data collection. Data labeling can be categorized into several methods: using premade labels, semi-supervised learning to predict missing labels, and crowdsourcing, which involves people classifying data, sometimes with advanced techniques to improve accuracy. Due to the high cost of perfect labeling, creating multiple "weak labels" has become common, compensating for their shortcomings. Labeling methods include classification and regression, with most research focusing on classification. Different data types, such as text, images, or graphs, require distinct labeling approaches, like identifying objects in photos versus extracting information from text.

Utilizing existing

labels A human must spend time mixing a little quantity of labelled data with a huge amount of unlabeled data for machine learning. Semi-supervised learning methods predict using labeled and unlabeled data. Transductive learning

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 3
***Article Received:** 28 January 2023 **Revised:** 05 February 2023 **Accepted:** 12 March 2023*

requires predictions based on new data because inductive learning can only use a small quantity of unlabeled data. Semi-supervised learning includes self-labeled techniques. Predictions in this subfield enable more labels. You can read the complete study here, so we'll just summarize the results. Choose between classic or graph-based label propagation techniques.

9

Category	Approach	Machine learning task	Data types	Techniques
Use Existing Labels	Self-labeled	classification	all	[92]–[96]
		regression	all	[97]–[99]
	Label propagation	classification	graph	[100]–[102]
Crowd-based	Active learning	classification	all	[103]–[109]
		regression	all	[110]
	Semi-supervised+Active learning	classification	text	[111], [112]
			image	[113]
			graph	[114]
	Crowdsourcing	classification	all	[50], [54], [115]–[122]
		regression	all	[123]
Weak supervision	Data programming	classification	all	[3], [124]–[127], [127]–[130]
	Fact extraction	classification	text	[131]–[142]

Table 1 lists all data labeling methods. Different methods can be utilized depending on the scenario. For graph data, label propagation or self-labeling can be used. Each method has its pros and cons. In a gradual approach, two models predict the unlabeled instances, and only cases with identical predictions are used with initially labeled examples to train the model. The loop ends if the model doesn't change. Majority voting is applied only when two models agree on identifying unlabeled data. Democratic Co-learning trains classifiers using different methods (e.g., Naive Bayes, C4.5, and 3-nearest neighbor) on the same training data, combining their predictions via weighted voting. Classifiers that disagree with the majority receive additional labels for further learning until their training data is complete. Co-training involves training models on different conditionally independent feature sets. One model uses the first set of features, while the second model uses the other set, reducing errors by increasing agreement on unlabeled data. These algorithms demonstrate similar transductive or inductive accuracy across 55 datasets from UCI and KEEL repositories. However, co-training requires that features can be divided into two independent subsets, which isn't always possible. Democratic Co-learning may also face limitations if only one machine learning algorithm is available.

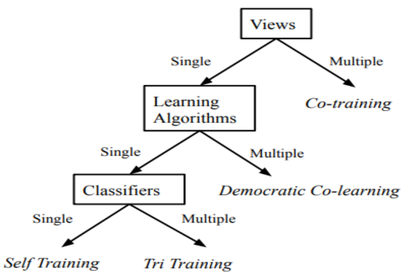

Fig. 1: Self-labeling classification of semi-supervised learning strategies based on survey results, using the best-performing strategies whether they were inductive or transudative learning.

Regression Relatively

Semi-supervised learning for regression, predicting real numbers from examples, includes coregularized least squares and frameworks with redundant views. Another method uses two k-nearest neighbor regressors with different distance measures, averaging their predictions. Co-training for regressors can extend to other base models, improving accuracy.

Graph-based Label

Propagation methods start with a few labeled examples and use the graph structure and similarities between examples to infer labels for others. Graph-based label propagation, used in fields like computer vision and natural language processing, involves creating a weighted graph from labeled and unlabeled examples. Techniques like the Gaussian random field model and algorithms like MAD-Sketch and EXPANDER reduce space and time complexity, making these methods more efficient.

Crowd-based techniques

Hand-labeling is highly accurate, as demonstrated by ImageNet's photo categorization using WordNet and Amazon Mechanical Turk. Active learning, which selects the most informative examples for labeling, has gained popularity, especially with modern tools like Microsoft Custom Vision, Amazon SageMaker, and Google Cloud AutoML. Methods like uncertainty sampling, margin sampling, and entropy measurement enhance active learning by focusing on uncertain examples. Combining active and semi-supervised learning, such as using McCallum and

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 3
***Article Received:** 28 January 2023 **Revised:** 05 February 2023 **Accepted:** 12 March 2023*

Nigam's QBC with Expectation-Maximization, improves model performance by leveraging both techniques to identify and label the most informative data points efficiently.

Crowdsourcing

Crowdsourcing uses non-experts for tasks, focusing on ease of participation, evaluating trustworthiness, eliminating bias, and combining results. Teaching workers to label accurately is challenging, often requiring comprehensive instructions. The Revolt approach addresses this by having workers vote, explain their labels, and review others' explanations. Effective tools and interfaces enhance performance, such as CrowdER for entity resolution and Qurk for record matching. Quality control is crucial, using methods like majority voting and tools like Get Another Label. Scalability of crowdsourced labeling is addressed through concurrent active learning techniques, optimizing performance for large datasets. Less research exists on crowdsourcing for regression tasks compared to classification.

Weak supervision

Due to the vast applications of machine learning, tagged data can be scarce, especially for new products in smart factories. Weak supervision methods are increasingly used to address this challenge, despite semi-automated labels being less accurate than manual ones. However, these labels are still sufficient to train models to a reasonable accuracy level, making them a valuable alternative when manual labeling is impractical. We'll explore a new concept in data programming and various methods to extract insights from data.

Data Programming

Data programming has evolved to utilize multiple labeling functions instead of weak labels, crucial for labeling large datasets, especially in deep learning contexts like Sally's smart factory software. These functions contribute to a generative model or majority voting system, improving accuracy despite individual label inaccuracies. Snorkel, a prominent data programming framework, efficiently integrates weak labels from various sources for quick results, particularly suited for multi-task learning challenges. Extraction methods like fact extraction from sources like the Internet aid in labeling by generating seed labels for distant monitoring, although accuracy may vary based on the knowledge source and extraction technique used.

Fact Extraction

Extracting information from sources can sometimes lead to mislabeling. Knowledge bases compile data from various sources, including the Internet, where facts are positively marked and can serve as weak seed labels for remote monitoring. Fact extraction retrieves structured web data, with approaches like YAGO and open-source programs employing patterns to extract information. Human-curated knowledge bases like Freebase and Google's Knowledge Graph are preferred for accuracy. However, extractions used remotely may lead to mislabeling.

REVIEW OF LITERATURE

Y. Roh et al(2021): At the moment, research in the field of machine learning is focused on how hard it is to get data. There are two main reasons why collecting data has become an important issue all of a sudden. As machine learning becomes more popular, there is a clear rise in the number of new apps that don't have enough tagged data. Second, unlike traditional methods, deep learning methods for machine learning develop features on their own. To save money, this gets rid of the need for feature engineering, but it could mean that more labelled data is needed. Data gathering research is interesting because it comes from the fields of machine learning, natural language processing, computer vision, and data management. This is because managing huge amounts of data is becoming more and more important. We do a thorough investigation when it comes to gathering data for data management. In the process of gathering data, there is a lot of data capture, data labelling, and model tweaking. We give an overview of the research landscape, including suggestions for which methods to use when and interesting research challenges that need to be solved. A growing trend is to combine big data with artificial intelligence (AI), which opens up a lot of new ways to look into things and creates a lot of opportunities.

Z. Hanyang et al(2019): The Automatic Identification System (AIS) gives almost real-time information about ships moving through the world's waters. Because of this, it is used a lot for ocean surveillance, maritime situation awareness (MSA), and preventing collisions. A satellite-based AIS has increased the AIS's range and made it easier to collect AIS data from more places at once. So, there is less of a gap in data for AIS on the high seas. When it comes to analysing maritime traffic, AIS data is a huge treasure trove. The "TianTuo-3" satellite from the National University of Defense Technology used the clustering algorithm K-means to collect S-AIS data. This is a common method used by scientists. For clustering, we use the Elbow

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 3
***Article Received:** 28 January 2023 **Revised:** 05 February 2023 **Accepted:** 12 March 2023*

Rule to figure out the best number of clusters and the normalised standard deviation of vessels' COG (Course Over Ground) and SOG (Speed Over Ground) as their characteristic values. This method is used to figure out how stable a ship is while sailing and to look for strange behaviours or situations that don't usually happen on ships.

K. S. Aggour et al(2019): As time goes on, the gap between where technology is now and where it needs to be for everyone to use it is likely to get bigger. In almost every industry, additive technologies are expected to change the way things are made. To improve the repeatability and reliability of additive manufacturing, you need advanced analytics, and these analytics need data. During the whole process of additive manufacturing, from designing the material to making the item, from printing the part to inspecting it when it's done, a huge amount of multimodal data must be made and used. A distributed polyglot storage and analysis tier with different repositories for different data structures, a metadata knowledge graph for modelling the data across the different repositories, and a user interface tier for visualising and calling analytics on that data. Big Data were made by our team, who also came up with the idea. This was done so that we could gather everything and put it all together. Using the platform, a number of processes that used to be separate have been put together to make a new additive material. Because of this, people who don't know much about software, like materials scientists, can now collect, visualise, and analyse the data they need.

METHODOLOGY

Data Collection using AI

Machine learning enables learning from past data for improved future performance, primarily through automated methods like self-learning, where algorithms autonomously improve based on past experiences. For instance, forensic machine learning proved that J.K. Rowling authored books under the pseudonym Robert Galbraith by analyzing writing styles. Multimedia data mining, a subset of data mining, extracts valuable insights from diverse digital media types such as text, images, audio, and video. With the rise of social media platforms, content-based retrieval systems have become crucial for accessing multimedia files effectively. The evolution of multimedia databases, incorporating structured and unstructured data, has revolutionized multimedia content management and retrieval, facilitated by algorithms and multimedia database management systems (MMDBMS).

DATA MINING AND MACHINE LEARNING

Data Mining:

Data mining (DM) is a modern method for sifting through vast data volumes to extract valuable insights. It involves analyzing diverse datasets, including the vast unstructured data found on social media platforms like Facebook and Twitter, to uncover new information and patterns. While traditional data mining techniques may not always suit dynamic and unstructured datasets, the goal remains the same: to glean insights and connections from various data sources. By discovering patterns and connections that might be challenging to discern manually, data mining not only validates theories but also generates novel ideas. Machine learning thrives on the patterns extracted from previously mined data, driving innovation and discovery. Multimedia data mining spans static media (e.g., text, images) and dynamic media (e.g., audio, video), seeking statistical relationships across diverse multimedia datasets to uncover meaningful patterns and insights.

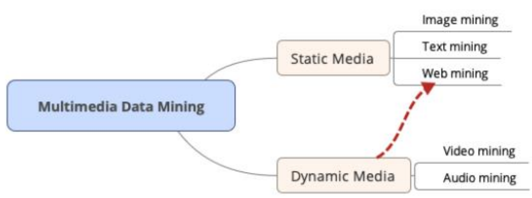

Figure 2: Multimedia Data Mining

Machine Learning

Machine learning elevates processes by enabling computers to learn and adapt with new information, advancing their capabilities over time. Training entails extracting insights from extensive datasets to enhance understanding and yield more accurate outcomes using statistical techniques to detect patterns. Fine-tuning parameters from the outset is crucial for optimal results. Artificial intelligence hinges on machine learning, recognizing the importance of learning from past experiences to enhance performance.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 3
***Article Received:** 28 January 2023 **Revised:** 05 February 2023 **Accepted:** 12 March 2023*

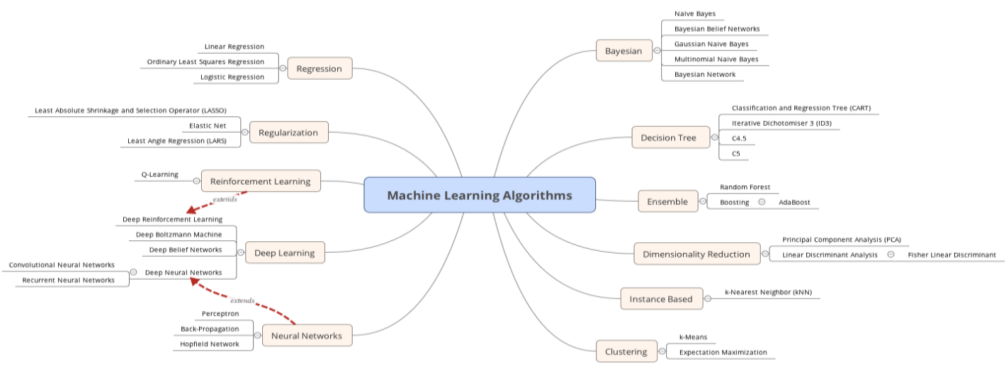

Figure 3: Machine Learning Algorithms used in Multimedia Mining

MACHINE LEARNING TECHNIQUES USED IN MULTIMEDIA SYSTEM

Algorithms for Regression In predictive analytics, regression analysis tries to make use of how the dependent (target) and independent variables are related to each other. Linear Regression, Logistic Regression, Stepwise Regression, Ordinary Least Squares Regression (OLSR), Multivariate Adaptive Regression Splines (MARS), and Locally Estimated Scatterplot Smoothing are all types of regression models. Some of the other models are Ordinary Least Squares (OLSR), Stepwise Regression, and Multivariate Adaptive Regression Splines (MARS) (MARS).

Linear Regression model

Simple linear regression is a statistical technique used to analyze and summarize relationships between two continuous variables, assuming a linear relationship between input and output variables. It's applicable when there's only one input variable. Multiple linear regression extends this to analyze multiple variables simultaneously. Widely used in various fields, linear regression predicts outcomes like traffic ETAs, financial portfolios, wages, and real estate values. Figure 4 illustrates a typical linear regression model.

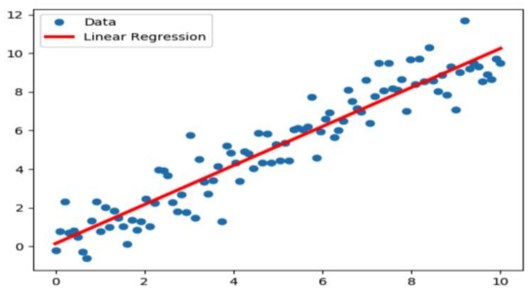

Figure 4: Linear Regression model sample illustration

Logistic regression: Logistic regression, a widely-used regression technique, offers significant advantages in various business applications such as credit card scoring and fraud detection, along with clinical trials. Particularly useful for binary response scenarios, it can handle multiple dependent variables, whether continuous or binary. Despite its popularity, logistic regression has drawbacks, including inconsistency and model dependence. Nevertheless, it remains a valuable tool, extensively employed in industries like real estate and insurance to predict customer lifetime value and purchasing behavior.

RESULT AND DISCUSSION

The study evaluated various classification algorithms to predict personalized news from newspaper databases (Interest news). It utilized a Fusion-level Deep Learning system based on a backpropagation neural network and binary neural networks for classification, incorporating backpropagation and resampling techniques. Evaluation was conducted using Weka 3.8.1 software, allowing for learning rates from -1 to 1. Assessment involved estimating true positive and false positive events using diverse methods to gauge the method's effectiveness and determine its practical application.

There are four different things that can affect how precise a metric is.

The following things are:

- true: "True positives" are documents that are good enough for their class and are correctly identified as belonging to that class.
- There are some papers that don't belong in a certain class, and they have been marked as such. The documents in question are true negatives.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 3
***Article Received:** 28 January 2023 **Revised:** 05 February 2023 **Accepted:** 12 March 2023*

- False positives are documents that shouldn't be accepted but are mistakenly labelled as such. Because of this, there is only one type of error to think about.
- When articles are put into the wrong category by mistake, this is called a "false negative." Because of this, this mistake fits into a different category.

After taking these four features into account, there are two different ways to figure out the accuracy metric: The score for F1 has already been set.

Then we have the following criteria:

The accuracy parameter gives the number of documents that are used to decide if something is correct or not.

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}}$$

This is the parameter that tells the search results how many documents to bring back.

The number called "recall" is calculated by dividing the rate of true positives by the sum the rates of true positives and false negatives.

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}$$

Even though this is one of the best ways to figure out results based on measurements, using it to make the right choice is not easy. The results of the F-measure can now be judged by a single metric. The F-measure method has the best ratio of accuracy to recall of any method.

$$\text{F-measure} = \frac{2\text{True positives}}{2\text{True positives} + \text{False negatives} + \text{False positives}}$$

The evaluation analysis by using RMSE similar
where n is the amount of data patterns, λ_z,designa
value of single data fact m and $\xi_{\beta,}$, β is the me
points. To write a Root Mean Squared Error (RM

$$\text{RMSE} = \sqrt{\sum_{\beta=1}^{\alpha}(\lambda_z,\beta - \xi_{\beta,}\beta)/\alpha}$$

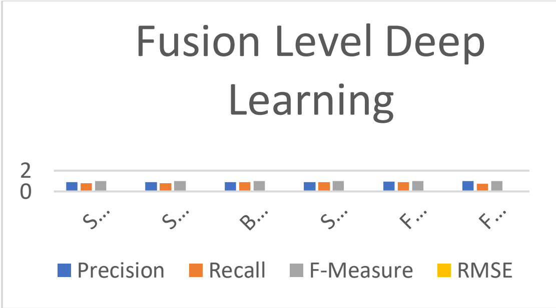

Figure 5: Fusion Level Deep Leaning

Table 2. Fusion Level Deep Leaning (multilayer perceptron) perceptron with resample

Name Algorithm	Precision	Recall	F-Measure	RMSE
SVM	0.900	0.867	0.988	0.0431
SVMG-RBF	0.920	0.788	0.988	0.0441
BPNN	0.920	0.895	0.988	0.0432
S3VM	0.930	0.855	0.988	0.0452
FDL-BPNN	0.940	0.887	0.988	0.0435
DFL-S3VM	0.989	0.897	0.988	0.0452

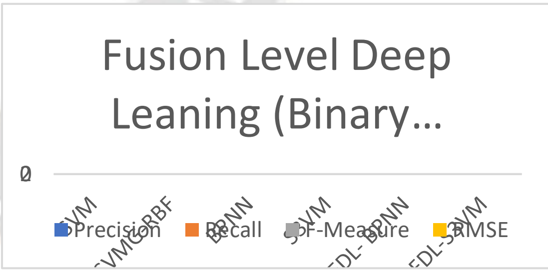

Figure 6: Fusion Level Deep Leaning

The table displays performance metrics of the proposed method using 10-fold cross-validation and various learning rates (0.1% to 0.8%). The study compared four dataset classification toolkits, evaluating six algorithms including SVM, SVMG-RBF, BPNN, S3VM, FDL-BPNN, and DFL-S3VM. Results suggest no single superior tool, emphasizing the importance of dataset and classifier selection.

Name Algorithm	Traditional	Proposed MR-DFL
SVM	Prerequisite huge memory and huge in dataset and high computation time in large dataset High computation cost	Using our proposed model reduce computation time ,boost accuracy
SVMG-RBF	High computation cost	Less computation cost
BPNN	Error evaluation very slow	Error evaluation very high
S3VM	Data classification very slow	Data classification very high response
FDL-BPNN	Less accuracy more computation in big dataset	high accuracy less computation in big dataset
FDL-S3VM	Less accuracy more computation in big dataset , memory utilization high	high accuracy less computation in big dataset , memory utilization less

Table 3: show the comparative analysis Traditional and Proposed MR-DFL

Here are a few of the many good things about this design:

- Using our method, it's easier and faster to classify a lot of data, which saves time.
- So, it is possible to use this in a variety of real-world situations. This will help with the classification of online data.
- Our proposed model improves the accuracy of how big datasets are grouped.
- In order to keep the dataset whole, the new method uses duplicate data. This makes the computations much simpler.
- The classification will do better now that the changes have been made.

CONCLUSION

In conclusion, multimedia data mining remains a pivotal and intricate area of research within computer science, drawing significant interest from researchers. The field presents numerous challenges, particularly in developing new algorithms and methodologies to effectively extract information from diverse multimedia datasets. This study has reviewed essential principles and recent advancements, underscoring the complexity and potential of multimedia mining. Future research should focus on refining these techniques and addressing unresolved questions to further advance the capabilities and applications of multimedia data mining.

REFERENCES

[1]. Y. Roh, G. Heo and S. E. Whang, "A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective," in IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 4, pp. 1328-1347, 1 April 2021, doi: 10.1109/TKDE.2019.2946162.

[2]. Z. Hanyang, S. Xin and Y. Zhenguo, "Vessel Sailing Patterns Analysis from S-AIS Data Dased on K-means Clustering Algorithm," 2019 IEEE 4th International Conference on Big Data Analytics (ICBDA), 2019, pp. 10-13, doi: 10.1109/ICBDA.2019.8713231.

[3]. I. Fujino, C. Claramunt and A. -O. Boudraa, "Extracting route patterns of vessels from AIS data by using topic model," 2017 IEEE International Conference on Big Data (Big Data), 2017, pp. 4744-4746, doi: 10.1109/BigData.2017.8258528.

[4]. K. S. Aggour et al., "Federated Multimodal Big Data Storage & Analytics Platform for Additive Manufacturing," 2019 IEEE International Conference on Big Data (Big Data), 2019, pp. 1729-1738, doi: 10.1109/BigData47090.2019.9006495.

[5]. M. Friedman, C. Ford, M. Elings PI, V. Singh and T. Tan, "Using AI/Machine Learning to Extract Data from Japanese American Confinement Records," 2021 IEEE International Conference on Big Data (Big Data), 2021, pp. 2210-2219, doi: 10.1109/BigData52589.2021.9672076.

[6]. X. Huang et al., "Toward Efficient Data Trading in AI Enabled Reconfigurable Wireless Sensor Network Using Contract and Game Theories," in IEEE Transactions on Network Science and Engineering, vol. 9, no. 1, pp. 98-108, 1 Jan.-Feb. 2022, doi: 10.1109/TNSE.2020.3013064.

[7]. P. Bethaz, K. Belhajjame, G. Vargas-Solar and T. Cerquitelli, "DS4ALL: All you need for democratizing data exploration and analysis," 2021 IEEE International Conference on Big Data (Big Data), 2021, pp. 4235-4242, doi: 10.1109/BigData52589.2021.9671883.

[8]. Z. A. Hasibuan, "Towards Using Universal Big Data in Artificial Intelligence Research and Development to Gain Meaningful Insights and Automation Systems," 2020 International Workshop on Big Data and Information Security (IWBIS), 2020, pp. 9-18, doi: 10.1109/IWBIS50925.2020.9255497.

[9]. D. Arslan and G. Eryiğit, "Evaluation of Wizard-of-Oz and Self-Play Data Collection Techniques for Turkish Goal-Oriented Dialogue Agents," 2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), 2021, pp. 1-6, doi: 10.1109/INISTA52262.2021.9548636.

[10]. X. Wu and M. Cheng, "Computer-aided Research on Buddhist Philosophy Estimation Based on Subconscious Data Collection and Mining," 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), 2022, pp. 1142-1146, doi: 10.1109/ICSSIT53264.2022.9716573.

[11]. F. A. Fayaz, A. Malik and A. A. Yatoo, "Cognitive Internet of things (CIoT) a success for data collection," 2021 Sixth International Conference on Image Information Processing (ICIIP), 2021, pp. 284-287, doi: 10.1109/ICIIP53038.2021.9702706.

[12]. P. G. Flikkema and B. Cambou, "When things are sensors for cloud AI: Protecting privacy through data collection transparency in the age of digital assistants," 2017 Global Internet of Things Summit (GIoTS), 2017, pp. 1-4, doi: 10.1109/GIOTS.2017.8016284.

[13]. O. V. Kluchko and R. R. Savchuk, "Implementation of a Hardware and Software Complex for Monitoring Industrial Equipment (Using the Example of AIS "Dispatcher")," 2022 Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus), 2022, pp. 1671-1674, doi: 10.1109/ElConRus54750.2022.9755561.

[14]. J. Xu, L. Yan, F. Wang and J. Ai, "A GitHub-Based Data Collection Method for Software Defect Prediction," 2019 6th International Conference on Dependable Systems and Their Applications (DSA), 2020, pp. 100-108, doi: 10.1109/DSA.2019.00020.

[15]. Y. Li, "AI-based Data Analyses for Text Classification and Document Summarization," 2020 17th International Joint Conference on Computer Science and Software Engineering (JCSSE), 2020, pp. xvii-xvii, doi: 10.1109/JCSSE49651.2020.9268287.

[16]. J. Proctor and R. Marciano, "An AI-Assisted Framework for Rapid Conversion of Descriptive Photo Metadata into Linked Data," 2021 IEEE International Conference on Big Data (Big Data), 2021, pp. 2255-2261, doi: 10.1109/BigData52589.2021.9671715.

[17]. W. Yu, Y. Liu, T. Dillon, W. Rahayu and F. Mostafa, "An Integrated Framework for Health State Monitoring in a Smart Factory Employing IoT and Big Data Techniques," in IEEE Internet of Things Journal, vol. 9, no. 3, pp. 2443-2454, 1 Feb.1, 2022, doi: 10.1109/JIOT.2021.3096637.

[18]. A. Kaur, R. Garg and P. Gupta, "Challenges facing AI and Big data for Resource-poor Healthcare System," 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), 2021, pp. 1426-1433, doi: 10.1109/ICESC51422.2021.9532955.

[19]. Fairfield, J., & Shtein, H. (2014). Big data, big problems: Emerging issues in the ethics of data science and journalism. Journal of Mass Media Ethics, 29(1), 38-51.

[20]. Gustafson, D. L., & Woodworth, C. F. (2014). Methodological and ethical issues in research using social media: A metamethod of Human Papillomavirus vaccine studies. BMC Medical Research Methodology, 14(1), 127.

[21]. Hammer, M. J. (2017). Ethical considerations for data collection using surveys. Oncology Nursing Forum, 44(2), pp. 1-2.