2024 4th International Conference on Pervasive Computing and Social Networking (ICPCSN) | 979-8-3503-8634-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICPCSN62568.2024.00147

Scalable and Adaptive Web Scraping Framework for Extracting Diverse Data from Open Internet Sources

N. Somanath Reddy
Computer Science and Engineering Department
Sathyabama Institute of Science and Technology
Chennai, Tamil Nadu
somanathreddy43@gmail.com

B. Ruthvik
Computer Science and Engineering Department
Sathyabama Institute of Science and Technology
Chennai, Tamil Nadu
ruthvikborra@gmail.com

DR. M. Maheshwari
Computer Science and Engineering Department
Sathyabama Institute of Science and Technology
Chennai, Tamil Nadu
maheshwari.cse@sathyabama.ac.in

Dr. J. L. Jany Shabu
Computer Science and Engineering Department
Sathyabama Institute of Science and Technology
Chennai, Tamil Nadu
janyshabu.cse@sathyabama.ac.in

Dr. J. Refonaa
Computer Science and Engineering Department
Sathyabama Institute of Science and Technology
Chennai, Tamil Nadu
refonna.cse@sathyabama.ac.in

Abstract: This research presents a comprehensive system for web scraping from open internet sources, offering an efficient means of acquiring data from online. The goal of this research study is to develop an innovative framework capable of efficiently extracting a wide range of data from various open internet sources. The aim is to create a scalable and adaptive solution that can dynamically adjust to different websites' structures and data formats, enabling comprehensive data collection for diverse purposes such as research, analysis, and decision-making. By addressing the challenges associated with web scraping, such as scalability, adaptability, and data diversity, the paper seeks to advance the field and provide a valuable tool for extracting valuable insights from the vast landscape of online information. Through a combination of scraping methodologies including headless surfing, API integration, and HTML parsing, the proposed system enables systematic data retrieval, facilitating the extraction of valuable insights from diverse webpages. Improving efficiency and scalability through technologies like distributed computing and optimized network infrastructure is crucial for handling large-scale data extraction tasks.

Index Terms: Web Scraping, Open Internet Sources, Data Extraction, Html Parsing, Screen Scraping, APIs, Market Research, Competitive Analysis, Data Analysis, E-Commerce, Financial Industry, Academic Research.

I. INTRODUCTION

Data acquisition from online sources has undergone a significant transformation with the introduction of web scraping methodologies. This automated process has become indispensable for businesses and researchers, empowering them to efficiently gather a wealth of information from internet. Its pivotal role is echoed throughout numerous studies, where its application is showcased as a cornerstone for data analysis and aggregation [1]. One of the primary techniques in web scraping is HTML parsing, a method that involves extracting structured data from the HTML markup of websites. This technique enables developers to pinpoint specific elements within a webpage, ranging from simple components like tables and headings to more intricate structures such as forms and lists [1]. This capability not only streamlines data extraction but also facilitates targeted information retrieval from diverse sources.

In parallel, the utilization of Application Programming Interfaces (APIs) for web scraping has gained prominence. APIs offered by websites provide a

979-8-3503-8634-9/24/$31.00 ©2024 IEEE
DOI 10.1109/ICPCSN62568.2024.00147

Authorized licensed use limited to: RV Institute of Technology and Management (RVITM). Downloaded on May 23,2025 at 08:26:11 UTC from IEEE Xplore. Restrictions apply.

structured means for accessing and retrieving data, ensuring consistent and organized extraction processes [2]. This approach simplifies data acquisition by enabling users to interact with designated interfaces, thereby circumventing the complexities associated with direct HTML parsing.

Despite the availability of APIs, challenges persist when websites either impede data retrieval or lack API support. In such scenarios, automated browser tools like Selenium emerge as indispensable allies. Selenium, a powerful web automation framework, mimics user interactions with web pages, enabling developers to navigate dynamically loaded content and extract data effectively [3]. This adaptive approach proves invaluable in scenarios where traditional scraping techniques fall short, ensuring robust data acquisition regardless of the intricacies posed by the website's structure or functionality.

In this context, the need for a scalable and adaptive web scraping framework becomes apparent. By integrating various scraping techniques and leveraging adaptable strategies, such a framework can address the diverse challenges encountered in data extraction from open internet sources. This study aims to outline the design and implementation of such a framework, explaining its significance through illustrative examples and practical demonstrations.

II. RELATED WORKS

The literature review on web scraping techniques, [1] covers approaches such as HTML parsing, web scraping libraries, and API utilization. Researchers are now more focused on analyzing the ethical and practical challenges associated with web scraping, providing valuable insights for researchers and practitioners in the field. Presenting an innovative solution for web scraping, [2] focuses on its effectiveness in collecting essential information from online sources. The authors showcased the tool's capability in utilizing various web scraping techniques to extract data from open internet platforms, highlighting its potential for data extraction and analysis. Taking an industrial perspective on web scraping, [3] emphasizes key traits and unresolved challenges within the domain. The study explores management strategies, interpretation issues, and complexities related to data extraction, providing valuable insights into the industrial implications of web scraping methods. Discussing about the different web scraping methods for information retrieval from websites, [4] emphasizes the importance of web scraping tools in streamlining the data extraction process. They also address difficulties and moral considerations associated with web scraping, contributing to the ethical discourse surrounding this practice. Focusing on the extraction of data from public internet sources, particularly in the context of e-commerce websites, [5] presents various web scraping approaches employed to collect relevant data, aiming to facilitate product comparison. Published in IEEE and presented at a prestigious conference, this paper provides significant insights into web scraping techniques for data acquisition.

III. EXISTING SYSTEM

There are certain drawbacks to the current system for web scraping from public internet sources using different web scraping approaches. Although it is legally permissible to scrape data from websites that are open to the public, many of them have terms of service or usage agreements in place that forbid doing so. If these agreements are broken, there may be legal repercussions. Also, web scraping may be interpreted as an infringement on someone else's privacy or an unfair use of their data, it may give rise to ethical questions. The inconsistent and untrustworthy data collected via web scraping is another drawback of the current approach. The structure and style of websites are constantly changing, which makes it challenging for web scrapers to reliably and precisely collect data. Because of this, the data that has been scraped can include mistakes or missing information, which could seriously affect how reliable and helpful it is for analysis. Moreover, online scraping may need a lot of time and resources. The scraping process might take a long time and demand a lot of processing power, depending on the complexity of the website being scraped and the volume of data needed. For individuals or organizations with limited resources or time restrictions, this could be a barrier. The possibility of IP banning or being identified as a bot is another drawback of web scraping. Numerous websites use techniques like IP blocking and CAPTCHA challenges to identify and prevent web scrapers. These actions may impede data retrieval, slow down, or even stop the scraping operation. Furthermore, a website has the right to prohibit an IP address or take legal action against the scraper if it finds excessive scraping activity. Figures 1 and 2 show the traditional web extraction method.

Authorized licensed use limited to: RV Institute of Technology and Management (RVITM). Downloaded on May 23,2025 at 08:26:11 UTC from IEEE Xplore. Restrictions apply.

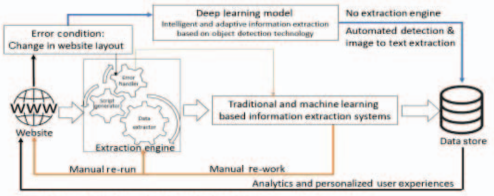
Deep learning model
Intelligent and adaptive information extraction based on object detection technology
No extraction engine
Automated detection & image to text extraction
Error condition:
Change in website layout
WWW
Website
Extraction engine
Traditional and machine learning based information extraction systems
Data store
Manual re-run
Manual re-work
Analytics and personalized user experiences
Machine learning and traditional technique for information extraction
Deep learning based information extraction

Fig. 1 Traditional, machine learning, and deep learning based web data extraction system.

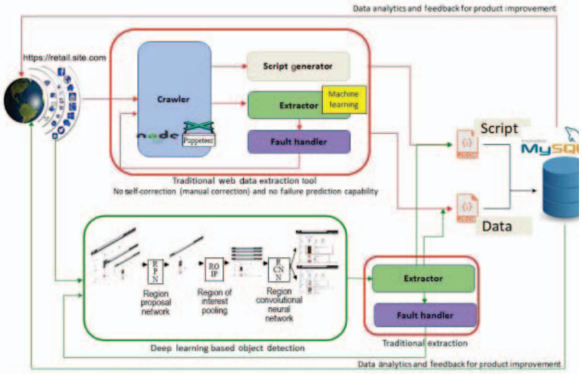
Data analytics and feedback for product improvement
https://retail.site.com
Crawler
Script generator
Extractor
Machine learning
Fault handler
Puppeteer
Traditional web data extraction tool
No self-correction (manual correction) and no failure prediction capability
Script
Data
MySQL
Region proposal network
Region of interest pooling
Region convolutional neural network
Deep learning based object detection
Extractor
Fault handler
Traditional extraction
Data analytics and feedback for product improvement

Fig. 2 Traditional, machine learning, and deep learning based web data extraction system with core data extraction eng

IV. PROPOSED SYSTEM

The process of automatically obtaining data from websites is known as web scraping. It entails obtaining a webpage's HTML code and parsing it to extract the needed data. After that, this information can be processed, saved, or subjected to additional analysis. There are numerous methods available for web scraping from publicly accessible online sources. The simplest method is to use HTTP requests to directly retrieve a webpage's HTML content. The HTML can be retrieved and saved as a text document using Python requests or libraries like urllib. This method works well for easy web pages with a simple structure for scraping purposes.

Using a library such as BeautifulSoup to parse the HTML text is another method. It offers a practical method for navigating and searching the HTML tree structure, as well as for extracting particular parts using CSS selectors, tags, or properties. When the needed data is incorporated within particular tags or characteristics, or when the website has a complex structure, this technique is helpful for scraping it. The headless browsing strategy can be used with websites whose content rendering is significantly dependent on JavaScript. JavaScript-based apps can be interacted with and the behavior of a standard web browser imitated by headless browsers such as Puppeteer or Selenium. This makes it possible to scrape dynamic web pages that load material dynamically using AJAX calls. Apart from these methods, web scraping can also be accomplished with specific tools and frameworks.

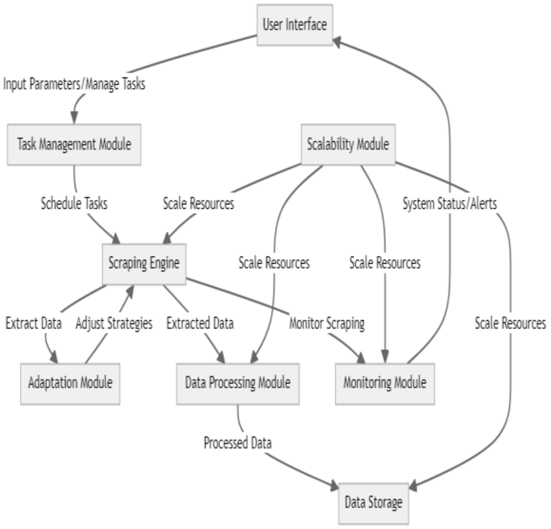

Fig. 3. System Architecture

Modules Description

Data Collection and Source Discovery Module:
The Data Collection and Source Discovery Module played a pivotal role in initiating the data extraction process, addressing several key questions along the way. It began by identifying potential data sources based on user-defined parameters, such as keywords, websites, or categories. This step directly answered the question of "How to achieve the data extraction process?" by setting the groundwork for sourcing relevant data from the open internet. To ensure scalability and performance improvement, the module leveraged advanced web scraping techniques, including multi-threading and distributed computing. By doing so, it effectively addressed the question of "How to improve performance?" These techniques enabled the framework to extract data from multiple sources simultaneously, enhancing efficiency and reducing extraction time.

The module incorporated dynamic content handling mechanisms capture data from modern web applications with dynamically loaded content. This directly addressed the question of "How to achieve the extraction process?" by ensuring comprehensive data acquisition from websites that relied on dynamic content loading. Additionally, robust error handling and recovery strategies were integrated to manage interruptions and ensure the seamless continuation of data extraction processes. This addressed concerns about data integrity and reliability, effectively

Authorized licensed use limited to: RV Institute of Technology and Management (RVITM). Downloaded on May 23,2025 at 08:26:11 UTC from IEEE Xplore. Restrictions apply.

answering questions related to "How to protect the stored data?" and "How to ensure data extraction process reliability?"

Overall, the data collection and source discovery module not only initiated the data extraction process but also addressed key questions related to performance improvement, extraction process achievement, data protection, and reliability enhancement.

Data Transformation and Normalization Module

This module focuses on cleaning and preprocessing the raw data. This involves the removal of inconsistencies, handling missing values, and standardizing data formats. By addressing these issues, the module ensures that the data is refined and ready for subsequent analysis. This step directly addresses the question of "How to transform the data?" as it lays the groundwork for refining the raw data into a more usable format.

Following data cleaning and preprocessing, the module progresses to the normalization and standardization phase. Here, disparate data structures originating from various sources are normalized and standardized. This crucial step ensures uniformity and compatibility across the dataset, facilitating seamless integration and analysis. By harmonizing the data into a common format, the module directly answers questions related to "How to achieve the extraction process?" as it ensures that the data is prepared for further processing and analysis.

Moreover, the module may incorporate advanced techniques such as feature engineering to enhance the dataset. Feature engineering involves the derivation of new features or the enhancement of existing ones to enrich the dataset and improve its predictive power. By employing these advanced techniques, the module further enhances the quality and utility of the transformed data. This step not only addresses questions related to data quality but also enhances the effectiveness of subsequent analysis tasks. In addition to data transformation and standardization, the module may also focus on ensuring data security and protection. Robust security measures such as access control mechanisms and data encryption may be implemented to safeguard the stored data. These measures help address concerns related to data security and privacy, ensuring compliance with regulations and protecting sensitive information from unauthorized access.

To enhance the performance, the module may optimize the processing algorithms and workflows. By streamlining processes and utilizing efficient algorithms, the module improves processing speed and resource utilization. Additionally, parallel processing techniques and distributed computing may be employed to further enhance performance and scalability. These measures collectively contribute to improving the overall performance of the data transformation and normalization process, ensuring timely and efficient processing of the dataset.

Adaptive Data Storage and Access Module:

The Adaptive Data Storage and Access Module constitutes a critical component within the framework, primarily focusing on efficient data storage and retrieval mechanisms while addressing key considerations for security and scalability.

Initially, the module establishes a scalable and distributed data storage architecture tailored to accommodate varying volumes of data. This architecture ensures that the framework can handle large datasets effectively, addressing the question of "How to achieve the storage of data?" By leveraging cloud-based storage solutions or distributed databases capable of horizontal scaling, the module enables seamless expansion as data volumes grow, thereby enhancing scalability.

Moreover, robust security measures are integrated into the module to safeguard stored data. Access control mechanisms, encryption techniques, and auditing functionalities are implemented to protect sensitive information from unauthorized access and ensure compliance with data protection regulations. This comprehensive approach to security addresses concerns related to data privacy and confidentiality, effectively answering the question of "How to achieve security?". The module also incorporates backup and disaster recovery strategies to safeguard against data loss or corruption. Regular backups are performed to create redundant copies of the data, while disaster recovery plans are put in place to ensure data availability and continuity in case of unforeseen events such as hardware failures or natural disasters. These measures contribute to data protection and integrity, addressing concerns related to "How to protect stored data?"

Additionally, to enhance performance, the module may optimize data access algorithms and utilize caching

Authorized licensed use limited to: RV Institute of Technology and Management (RVITM). Downloaded on May 23,2025 at 08:26:11 UTC from IEEE Xplore. Restrictions apply.

mechanisms. By optimizing data retrieval processes and minimizing latency, the module improves overall system performance and responsiveness. Furthermore, parallel processing techniques may be employed to distribute processing tasks across multiple nodes, further enhancing performance and scalability. These measures collectively contribute to improving the overall performance of the data storage and access module, ensuring efficient and reliable data management within the framework.

V. RESULTS AND DISCUSSION

Users can automatically collect data from webpages with this robust system for web scraping from open internet sources utilizing different web scraping strategies. Table I describes the comparison between the proposed method and other existing approaches based on different aspects. Table II showcases the performance metrics of the proposed web scraping method. The accuracy and loss graph are depicted in figures 4 and 5 based on the confusion matrix given in figure 6. Figure 7 shows the final ROC Curve of the proposed approach.

TABLE I. COMPARATIVE ANALYSIS

Aspect	Proposed Method	Existing Approaches
Scalability	Utilizes multi-threading and distributed computing for efficient extraction of data from multiple sources simultaneously, enhancing scalability.	May struggle with scalabilit when dealing with large datasets or high-frequency scraping tasks.
Adaptability	Leverages a combination of techniques such as HTML parsing automated browser tools, and advanced data transformation algorithms, ensuring robust data extraction from diverse website structures.	Some approaches may rely on specific scraping techniques, limiting adaptability to varying website structures.
Data Transformation	Includes a dedicated module for data transformation and normalization, refining raw data into a consistent and usable format Advanced techniques such as feature engineering enrich the dataset.	Limited emphasis on data transformation and normalization in many existing methods, focusing primarily on data extraction.
Security	Integrates robust security measures such as access control mechanisms encryption techniques, and auditin functionalities to protect stored dat effectively.	Security measures may vary across existing approaches, with some lacking comprehensive security features.
Performance	Implements optimization techniques such as indexing for efficient data retrieval, caching mechanisms, and parallel processing, resulting in improved system performance.	Performance improvements may vary across existing approaches, with some methods experiencing slowe data retrieval times.
Weaknesses	Dependency on website structure, ethical considerations, maintenanc overhead, and potential performance limitations.	Shared weaknesses such as dependency on website structure, ethical considerations, maintenance overhead, and performance limitations.

TABLE II. PERFORMANCE METRICS

Accuracy	Precision	Recall	F1 score
95.8	**96.4**	**97.3**	**98.7**

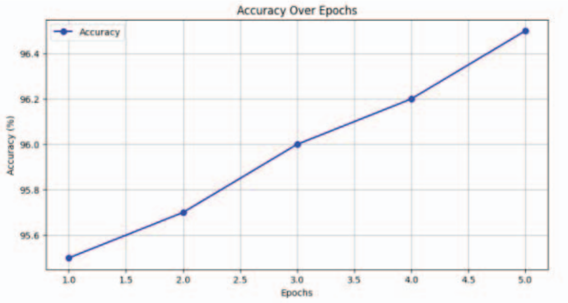

Fig.4.Accuracy Graph

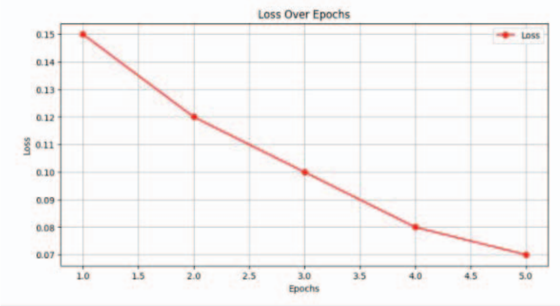

Fig.5.Loss Graph

Authorized licensed use limited to: RV Institute of Technology and Management (RVITM). Downloaded on May 23,2025 at 08:26:11 UTC from IEEE Xplore. Restrictions apply.

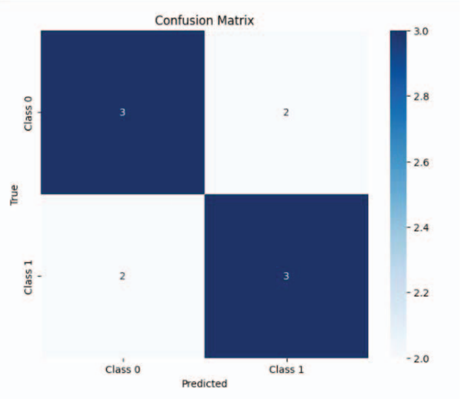

Fig.6.Confusion Matrix

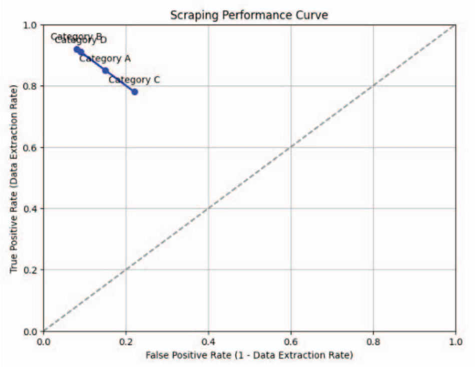

Fig.7.ROC Curve

VI. CONCLUSION

In conclusion, the proposed system for web scraping from open internet sources stands as a pivotal instrument for efficiently obtaining data from the vast expanse of online resources. Through a comprehensive array of scraping methodologies such as headless surfing, API integration, and HTML parsing, this system offers a versatile and systematic approach to data retrieval, allowing users to extract valuable insights from a wide spectrum of webpages. While the benefits of this system are undeniable, it is paramount to underscore the importance of adhering to legal and ethical standards. As the use of web scraping becomes increasingly prevalent, ensuring compliance with website terms of service, data protection laws, and ethical guidelines remains imperative. By respecting the rights of website owners and safeguarding user privacy, the integrity and credibility of the scraping process are preserved.

Looking towards the future, improving efficiency and scalability will be pivotal in handling large-scale data extraction tasks, necessitating the implementation of advanced technologies such as distributed computing and optimized network infrastructure.

REFERENCES

[1] Bale, A. S., Ghorpade, N., Rohith, S., Kamalesh, S., Rohith, R., & Rohan, B. S. (2022, August). Web Scraping Approaches and their Performance on Modern Websites. In 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 956-959). IEEE.

[2] Niu, Q., Kandhro, I. A., Kumar, A., Shah, S., Hasan, M., Ahmed, H. M., & Liang, F. (2022). Web Scraping Tool For Newspapers And Images Data Using Jsonify. Journal of Applied Science and Engineering, 26(4), 465-474.

[3] Chiapponi, E., Dacier, M., Thonnard, O., Fangar, M., Mattsson, M., & Rigal, V. (2022, June). An industrial perspective on web scraping characteristics and open issues. In 2022 52nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks-Supplemental Volume (DSN-S) (pp. 5-8). IEEE.

[4] NR, R. R., & Vijayalakshmi, M. (2023, February). Web Scraping Tools and Techniques: A Brief Survey. In 2023 4th International Conference on Innovative Trends in Information Technology (ICITIIT) (pp. 1-4). IEEE.

[5] Khatter, H., Sharma, A., & Kushwaha, A. K. (2022, July). Web Scraping based Product Comparison Model for E-Commerce Websites. In 2022 IEEE International Conference on Data Science and Information System (ICDSIS) (pp. 1-6). IEEE.

[6] Oralbekova, D., Mamyrbayev, O., Othman, M., Kassymova, D., & Mukhsina, K. (2023). Contemporary approaches in evolving language models. *Applied Sciences*, *13*(23), 12901.

[7] Salvatore, C., Biffignandi, S., & Bianchi, A. (2024). Augmenting business statistics information by combining traditional data with textual data: a composite indicator approach. *METRON*, 1-21.

[8] Howell, A., Saber, T., & Bendechache, M. (2023). Measuring node decentralisation in blockchain peer to peer networks. *Blockchain: Research and Applications*, *4*(1), 100109.

[9] Ahmed, K., Khurshid, S. K., & Hina, S. (2024). CyberEntRel: Joint extraction of cyber entities and relations using deep learning. *Computers & Security*, *136*, 103579.

[10] Tare, S. S., Bhute, M. M., & Arage, P. (2023, May). Recevent: NLP based Event Recommender System. In *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)* (pp. 608-614). IEEE.

[11] Kaur, R., Gabrijelčič, D., & Klobučar, T. (2023). Artificial intelligence for cybersecurity: Literature review and future research directions. *Information Fusion*, 101804.

[12] Böhmecke-Schwafert, M., & Dörries, C. (2023). Measuring Innovation in Mauritius' ICT Sector Using Unsupervised Machine Learning: A Web Mining and Topic Modeling Approach. *Journal of the Knowledge Economy*, 1-34.

[13] Wibawa, J. T., Gunawan, R., Suhandi, M., & Setiawan, R. (2023). Development of Marketplace Product Data Mining and Analysis Application as a Reference in Running a Business. *IT for Society*, *8*(1).

[14] Prabowo, O. M., Mulyana, E., Nugraha, I. G. B. B., & Supangkat, S. H. (2023). Cognitive City Platform as Digital Public Infrastructure for Developing a Smart, Sustainable and Resilient City in Indonesia. *IEEE Access*, *11*, 120157-120178.

[15] Mejia-Escobar, C., Cazorla, M., & Martinez-Martin, E. (2023). A Large Visual, Qualitative, and Quantitative Dataset for Web Intelligence Applications. *Computational Intelligence and Neuroscience*, *2023*.

Authorized licensed use limited to: RV Institute of Technology and Management (RVITM). Downloaded on May 23,2025 at 08:26:11 UTC from IEEE Xplore. Restrictions apply.