SOFTWARE FOR AUTOMATED GENERATION AND VERSIONING OF MULTIMODAL DATASETS FOR LABOR COST ESTIMATION IN MANUFACTURING

Maria Yu. Karelina,
State University of Management, Moscow, Russia,
karelinamu@mail.ru

Vladimir S. Makarov,
Nizhny Novgorod State Technical University named after R.E. Alekseev, Nizhny Novgorod, Russia,
makvl2010@gmail.com

Vladimir D. Kutkov,
State University of Management, Moscow, Russia,
kutkovVD@yandex.ru

Vasilisa-Anastasia V. Yudina,
State University of Management, Moscow, Russia,
vv_badakova@guu.ru

DOI: 10.36724/2072-8735-2025-19-10-54-60

Manuscript received 20 July 2025;
Accepted 28 September 2025

This article has been prepared as part of a research project funded by the federal budget (the source of funding is the Ministry of Education and Science of the Russian Federation) on the topic: "Development of scientific, methodological and practical foundations of reverse engineering for solving complex import substitution problems in the agro-industrial complex of the Russian Federation" (code of scientific topic FZNW-2024-0026).

Keywords: *labor cost estimation, data preparation, multimodal data, machine learning, software suite, data versioning, Large Language Models (LLM), manufacturing, MLOps*

The problem of accurate labor cost estimation in manufacturing is complicated by the challenge of preparing heterogeneous data (CAD models, ERP data, textual technical requirements) for machine learning models. This paper presents the architecture and implementation methods of a software suite designed to automate the process of forming and versioning multimodal datasets. The methodology includes the implementation of an ETL pipeline for extracting distributed data, as well as the application of two AI software agents for intelligent data enrichment: a convolutional neural network to calculate a geometric complexity index from CAD models, and a large language model to extract structured technological parameters from textual descriptions. To ensure the reproducibility of research, a data version control system based on DVC and Git has been implemented. The suite's functionality is demonstrated on model examples of parts with varying complexity, showcasing its effectiveness in automatically generating enriched, analysis-ready datasets. It is concluded that the proposed approach significantly reduces data preparation labor, improves data quality and objectivity, and establishes a solid foundation for building high-precision predictive models.

Information about authors:

Maria Yu. Karelina, *Doctor of Technical Sciences, Doctor of Pedagogical Sciences, Professor, Head of the Department of Management of Transport Complexes, State University of Management, Moscow, Russia*

Vladimir S. Makarov, *Doctor of Technical Sciences, Professor, Nizhny Novgorod State Technical University named after R.E. Alekseev, Nizhny Novgorod, Russia*

Vladimir D. Kutkov, *Postgraduate Student, Specialist, Engineering Project Management Center, State University of Management, Moscow, Russia*

Vasilisa-Anastasia V. Yudina, *PhD student, specialist in the Laboratory of Reverse Engineering, State University of Management, Moscow, Russia*

Для цитирования:
Карелина М.Ю., Макаров В.С., Кутков В.Д., Юдина В.-А. В. Программный комплекс для автоматизированного формирования и версионирования мультимодальных наборов данных в задачах оценки трудоемкости // T-Comm: Телекоммуникации и транспорт. 2025. Том 19. №10. С. 54-60.

For citation:
M.Yu. Karelina, V.S. Makarov, V.D. Kutkov, V.-A.V. Yudina, "Software for automated generation and versioning of multimodal datasets for labor cost estimation in manufacturing," *T-Comm*, 2025, vol. 19, no. 10, pp. 54-60.

Introduction

Improving the accuracy of estimating the complexity of design and technological preparation of production is a necessary condition for ensuring the economic efficiency of machine-building enterprises. In conditions of small-scale production and when solving problems of reverse engineering, where there is no accumulated statistics on analogues, a reliable forecast of labor costs in the early stages determines the profitability of the project. The application of modern methods of data analysis and machine learning to solve this problem is fraught with significant difficulties due to the specifics of the source information [1, 2].

Production data is inherently heterogeneous and multimodal. They include structured parameters from regulatory reference databases, semi-structured attribute information from PDM systems, as well as unstructured data in the form of two-dimensional and three-dimensional CAD models and text descriptions in technological maps. This information, as a rule, is distributed across isolated enterprise information systems (ERP, PDM, MES), does not have unified formats and direct logical connections, which forms significant barriers to its automated processing and the construction of adequate mathematical models based on it [3, 4].

Historically, the task of preparing data for labor intensity assessment systems has been solved through expert selection of the dominant design and technological features. This process involved the classification of parts, the formation of a list of potentially significant parameters and their subsequent verification by interviewing process engineers. This approach is characterized by high labor intensity, dependence on the qualifications and subjective experience of experts and, as a result, a low level of reproducibility of the results. The lack of formalized procedures and software tools made it impossible to systematize the process and ensure its scalability.

Overcoming these limitations requires the development of a specialized software package designed to automate procedures for extracting, converting, intelligently enriching and uploading (ETL) multimodal data. The functionality of such a complex should ensure not only the aggregation of information from diverse sources, but also the implementation of versioning practices that guarantee reproducibility and traceability of both the datasets themselves and the models trained on them. The purpose of this work is to present the architectural principles and methods of software implementation of such a complex, as well as to describe approaches to intellectual data enrichment using modern AI tools [4-6].

Materials and methods

The research methodology consists in the design and software implementation of a modular complex designed to automate the process of forming multimodal datasets. The main objective of this complex is to perform sequential operations of extracting, converting, and loading (ETL) distributed and heterogeneous enterprise data into a centralized, structured repository suitable for subsequent analysis and training of machine learning models. The architecture of the complex assumes sequential data processing, starting from connection to primary sources and ending with the creation of versioned, ready-to-use datasets.

The data typical for the production cycle of a machine-building enterprise act as the initial materials for the operation of the complex. These materials differ fundamentally in their structure, presentation formats, and sources of origin, which requires the use of various methods for their extraction and processing. Their conditional classification, as well as the problems solved by the software package at the processing stage, are presented in Table 1.

The software package is based on a modular architecture implemented using the Python programming language and related libraries for data processing and machine learning. This architecture includes a layer of data extraction (connectors), a layer of data transformation and enrichment (transformers), and a layer of loading and versioning (loaders). This work focuses on the first two layers responsible for the formation of the primary, enriched dataset [7].

At the first stage of the complex's operation, a data extraction layer functions, implemented as a set of software connectors. To access structured data from ERP and MES systems, standard libraries such as pyodbc and psycopg2 are used, allowing direct SQL queries to databases. Semi-structured data is extracted from PDM systems via HTTP requests to their REST API, followed by parsing responses in JSON or XML formats. This approach allows you to automatically aggregate disparate information into a single intermediate storage area [8, 9].

Table 1

Classification and characterization of the source data

Data type	A source	Format	Examples of information	The main problem for processing is
Structured	ERP, MES-systems, normalization archives	SQL tables, CSV	Nomenclature part number, operational time standards, technological operation code	The absence of direct links (keys) with design data in PDM systems.
Semi-structured	PDM/PLM systems	API requests, XML/JSON, internal database formats	Part name, material grade, type of workpiece, main overall dimensions, applicability	Inconsistent attribute naming; programmatic restrictions on data access via the API.
Unstructured	KB and TB archives, file storage	Vector (DWG, STEP), Raster (TIFF), Text (PDF, DOCX)	Geometry and topology of the part, special technical requirements, descriptions of operations	The information is implicit and requires the use of computer vision and natural language processing techniques to extract it.

This automated approach to information collection is a direct development of historically established practices where this step was performed manually. Within the framework of the classical approach, an expert analysis of a limited sample of drawings and technological maps was performed in order to manually generate a table of features for the subsequent construction of a regression model. The software implementation of connectors makes it possible to completely eliminate the subjective factor and time-consuming manual operations from the process of collecting primary information, ensuring the completeness and objectivity of the source data.

Since the original geometry of the part, containing information about its complexity, is represented in vector formats (DWG, STEP), transformation is required for its use in deep learning models (Fig. 1).

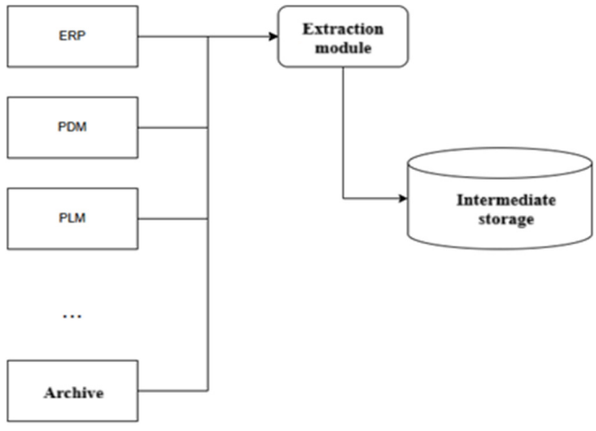

Fig. 1. Diagram of the operation of the data extraction and primary transformation module

The diagram shows the data flows:

1) tables with time standards are extracted from the ERP system.;

2) The attributes of the parts are extracted from the PDM/PLM system via the API;

3) CAD files re extracted from the file archive.

All streams are sent to the central block "Extraction and Transformation Module", at the output of which an intermediate storage is formed containing ID-related parts of SQL tables with metadata and a folder with rasterized images of drawings.

Next to the primary data extraction is their intellectual enrichment. The purpose of this stage is to transform implicit information hidden in unstructured formats into explicit, quantitative features that can be directly used by machine learning models. This process replaces traditional manual analysis and subjective expert assessment with automated, reproducible procedures performed by software agents based on artificial intelligence models. The complex includes two independent agents, each of which specializes in its own type of data modality: visual and textual [10-12].

The first software agent, the Geometric Complexity Estimator, is designed to obtain a quantitative metric that characterizes the topological and geometric complexity of a part. The agent uses rasterized CAD model images obtained at the previous stage as input data. The agent is based on a convolutional neural network (CNN) of the ResNet-50 architecture, pre-trained on a large-scale image dataset. The network is used in feature extraction mode: the input image is converted into a high-dimensional vector representation (embedding) extracted from the penultimate fully connected layer of the network. This vector accumulates generalized characteristics of the shape and structure of an object, invariant to its position and scale [13-14].

The resulting vector of high-dimensional features is not used directly. To obtain a final scalar estimate of complexity, the dimensionality reduction method is used. Embedding goes to the input of a small fully connected neural perceptron (MLP), consisting of two hidden layers and one output neuron with a linear activation function. In the process of pre—training this MLP (using a sample marked up by experts or using indirect metrics such as the number of structural elements), it learns to display a multidimensional feature vector into a single number - the "geometric complexity index". This index is added as a new feature to the dataset for each detail.

The second software agent, the LLM parser of technological requirements, solves the problem of extracting structured data from unstructured text fields, such as notes on drawings or descriptions of operations in technological maps. The agent is built on the basis of a large language model (LLM) operating in the "instrumental agent" mode. An engineering prompt is generated for the model, which contains clear instructions: analyze the input text from the perspective of a process engineer and extract a predefined set of entities (for example, roughness requirements, accuracy standards, special tolerances) [15].

A special feature of the implementation is the requirement for the model to return the result in a strictly defined JSON format (JavaScript Object Notation). This is achieved by including examples (few-shot prompting) in the prompt and explicitly specifying the JSON schema. This approach ensures reliable machine parsing of the result and its subsequent integration into a structured database. The agent's work allows for the automatic digitization and systematization of critical technological information, which in traditional approaches was either ignored or required manual analysis.

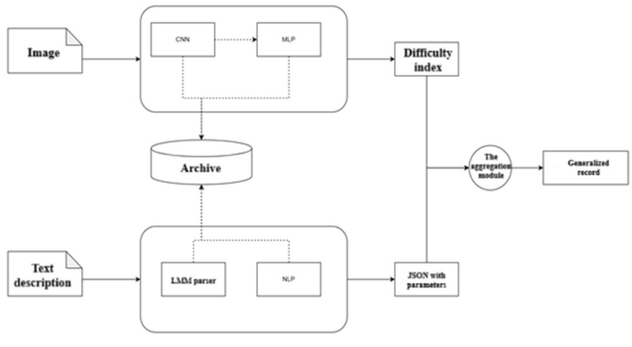

Fig. 2. Architecture of the intelligent data enrichment module

The diagram in Figure 2 illustrates two parallel streams. Stream 1: Rasterized images from the "Intermediate Storage" are sent to the "Agent 1: Geometric Complexity Evaluator (CNN+MLP)" block, at the output of which a "Geometric Complexity Index" is generated. Stream 2: Text data from technological maps is sent to the "Agent 2: LLM parser" block, at the output of which a "Structured JSON with technological parameters" is formed. Both results are sent to the "Aggregation Module" block,

which combines them with data from ERP and PDM, forming an "Enriched Detail Record" [16].

Upon completion of the agents, the enriched data is aggregated and uploaded to the target storage. At this stage, the DVC (Data Version Control) data version control system is used, integrated with the Git code version control system. Each ETL pipeline launch that results in a change in the dataset is recorded. DVC stores checksums (hashes) of data files, while Git stores small text meta files indicating a specific version of the data. This mechanism ensures full reproducibility of any experiment: having access to the Git repository, you can at any time restore the exact version of the code and the corresponding dataset on which a particular model was trained or tested.

The final result of the software package is the creation of a versioned, multimodal dataset, presented in the form of a relational metadata table and an associated repository of binary artifacts. This structure is fully prepared for use in training cycles and evaluation of labor intensity forecasting models, ensuring high-quality source data and transparency of the research process.

Results and discussion

To test and demonstrate the functionality of the developed software package, a computational experiment was conducted on a set of model examples simulating real production data. Three parts representing different classes of design and technological complexity were selected: "Smooth bushing" (simple), "Straight-tooth gear" (medium complexity) and "Gearbox housing" (complex). The initial information on each part was distributed among sources simulating industrial information systems (ERP, PDM, KB archive)[16].

At the first stage of the complex's operation, the extraction and transformation module successfully aggregated attribute data and rasterized CAD models. The resulting bitmap images and text fields were transferred to the input of the intellectual enrichment module. The summary results of the work of AI software agents on data enrichment for all three details are presented in Table 2.

The analysis of the results presented in the table allows you to make several conclusions. First, the "Geometric Complexity Evaluator" demonstrated the ability to adequately differentiate details. The index generated by him increases monotonously from a simple "body of rotation" type part to a complex body part, which corresponds to engineering intuition. This quantitative feature, obtained completely automatically from visual data, is an objective measure that replaces subjective expert assessments of complexity.

Secondly, the "LLM parser of technological requirements" has shown high efficiency in extracting structured information from an informal text. The agent correctly identified and extracted numerical values of accuracy qualities and roughness parameters, as well as highlighted the presence of additional technological requirements. It is important to note that the model has successfully coped with the variability of formulations, which demonstrates its flexibility compared to traditional parsing methods based on regular expressions [17].

The final result of the complex's work is the formation of enriched, versioned records in the target dataset. For each part, all extracted and generated features were aggregated into a single record associated with the primary key. The download process was recorded by the DVC versioning system, which ensures full traceability and reproducibility. This approach allows us to accumulate high-quality, verified data, forming a reliable foundation for the subsequent training of high-precision labor intensity forecasting models [18-19].

During the discussion of the results, it should be recognized that the testing of the software package using model examples, although it demonstrates its fundamental operability, does not cover the full range of problems associated with full-scale industrial implementation. The transition from a controlled environment to working with real production archives numbering tens of thousands of items will require solving the problems of scaling ETL processes and ensuring their fault tolerance. Special attention should be paid to the development of mechanisms for processing "noisy" and incomplete data: CAD models made with deviations from ESCD standards, scanned drawings of poor quality and technological descriptions with variable terminology. Therefore, a promising area for further work is the development and integration into the complex of an additional validation module that implements the concept of "human-in-the-loop". This module will allow the process engineer to verify the results of the work of AI agents, correct abnormal emissions and form a feedback loop for their further training and adaptation to the specifics of the data of a particular enterprise.

Table 2

Summary results of the work of AI agents on data enrichment for various types of parts

Part (ID)	Input data (Text from notes)	Agent 1's result (geom index. difficulties)	Agent 2 Result (LLM) (Extracted by tech. parameters in JSON format)
The sleeve is smooth TUE-40-120-CHAPTER	"Turning of the outer and inner and end surfaces. Accuracy according to IT11. Roughness Ra 6.3."	12.5	{"accuracy_grade": {"IT": 11}, "surface_roughness": {"Ra": 6.3}, "extra_operations": null
The gear is straight-toothed SP-08-15-40	"Op. 025 Zubofrezernaya. Module m=8. To ensure the accuracy of the 8th grade (IT8). The roughness of the working surfaces of the teeth is no more than Ra 1.6. Carry out heat treatment: “HDPE hardening”	68.7	{"accuracy_grade": {"IT": 8},"surface_roughness": {"Ra": 1.6},"extra_operations": ["HDTV hardening "]
Gearbox housing KR-05-250	"Milling of support planes and drilling of mounting holes. Boring of seats for bearings with H7 tolerance. Control of the centerline distance. Roughness of the landing surfaces Ra 1.25."	89.1	{"accuracy_grade": {"H": 7},"surface_roughness": {"Ra": 1.25},"extra_operations": ["center-to-center distance control"]}

Conclusion

This paper presents methods for software implementation of a complex that solves the urgent problem of automated generation and versioning of multimodal datasets for tasks of labor intensity assessment in mechanical engineering. The conducted research has shown that traditional manual approaches to data preparation are characterized by high labor intensity, subjectivity and lack of reproducibility, which is a significant barrier to the effective application of modern machine learning methods.

The main result is the demonstrated functionality of the software package, which successfully automates the full cycle of the ETL process: from extracting heterogeneous data from simulated industrial sources to uploading them to versioned storage. The novelty in the proposed solution is the use of two software AI agents for intelligent data enrichment. The CNN-based Geometric complexity estimator made it possible to transform unstructured visual data into an objective quantitative feature, while the LLM parser effectively extracted structured technological information from non-formalized text descriptions. These methods make it possible to move from manual feature design to the paradigm of automatic representation learning.

The practical significance of the presented results lies in the possibility of drastically reducing the time spent on data preparation and improving their quality by extracting previously unavailable implicit information. The implementation of the DVC versioning system ensures full transparency and reproducibility of experiments, which is a prerequisite for building reliable MLOps solutions in an industrial environment. Further development of this area is associated with the transition to direct analysis of three-dimensional models using graph neural networks and the expansion of the functionality of LLM agents for data validation and purification tasks.

Acknowledgement

This article has been prepared as part of a research project funded by the federal budget (the source of funding is the Ministry of Education and Science of the Russian Federation) on the topic: "Development of scientific, methodological and practical foundations of reverse engineering for solving complex import substitution problems in the agro-industrial complex of the Russian Federation" (code of scientific topic FZNW-2024-0026).

References

[1] E.A. Yakovleva, I.A. Tolochko, A.A. Kim, A.A. Chernyaeva, “Digital Transformation of the Planning System Based on a Digital Twin,” *Creative Economy*, 2021, no.15(7), pp. 2811-2826.

[2] M. V. Ponomarenko, “Automated Production Management Systems in the Context of Product Lifecycle Management,” *XII Congress of Young Scientists: Collection of Scientific Papers*, 2023, pp. 151-155. National Research University ITMO.

[3] A. V. Ryabchenko, “Digital Transformation of the Organizational and Economic Mechanism of Industrial Corporations,” *Bulletin of the Altai Academy of Economics and Law*, 2022, no. 3-1, pp. 120-127. https://doi.org/10.17513/vaael.2106

[4] N. V. Kurganova, M. A. Filin, D. S. Chernyaev, A. G. Shaklein, D. E. Namiot, “The Implementation of Digital Twins as a Key Area of Production Digitalization,” I*nternational Journal of Open Information Technologies*, 2019, no. 7(5), pp. 105-115.

[5] A. A. Vishnevsky, “Problems of Implementing Information Systems for Production Management,” *Construction Materials, Equipment, Technologies of the 21st Century*, 2022, no. 6(275), pp. 15-21.

[6] E. E. Sidorycheva, V. A. Sidorychev, I. P. Efimov, G. V. Dmitrienko, “Methodology for Automated Comparison of Electronic Product Structures in PDM/PLM Systems for Aviation Instrumentation Enterprises,” *Automation of Control Processes*, 2023, no. 1(71), pp. 123-132. https://doi.org/10.35752/1991-2927_2023_1_71_123

[7] K. Yu. Cherkasov, I. N. Khaimovich, “Transformation of IDEF0 Models of Design and Process Planning into Tabular Data Models for PDM/PLM Systems,” *News of the Tula State University. Technical Sciences*, 2025, no. (5), pp. 150-161. https://doi.org/10.24412/2071-6168-2025-5-150-151

[8] I. V. Kosinsky, A. V. Ivanov, A. V. Gavrisheva, et al. Certificate of State Registration of the Computer Program No. 2024661192 Russian Federation. Adapter for Interaction of Client Applications for Automated Creation and Management of Technical Documentation Lists with PLM/PDM Systems* (No. 2024660592). JSC "St. Petersburg Marine Engineering Bureau "Malakhit". 2024.

[9] A. V. Bredikhin, A. A. Meshkov, A. B. Agadzhanyan, Certificate of State Registration of the Computer Program No. 2024689134 Russian Federation. MASIOT. Agent Factory (No. 2024688029). Limited Liability Company "DM Solution". 2024.

[10] O. V. Ulyanin, E. M. Abakumov, “Issues of Improving the Structure of the Electronic Archive of High-Responsibility Technical Documentation and Information Protection,” *Electromagnetic Compatibility Technologies*, 2024, no. 1(88), pp. 35-46.

[11] I. S. Delimova, D. M. Shprekher, “Lifecycle Management Software Essential for Modern Production,” *News of the Tula State University. Technical Sciences*, 2024, no. (12), pp. 437-441. https://doi.org/10.24412/2071-6168-2024-12-437-438

[12] Digital Twins and Digital Transformation of Defense Industry Enterprises. (n.d.). Retrieved August 18, 2025, from https://assets.fea.ru/uploads/fea/news/2019/04_april/15/elibrary_37180048_50837228.pdf

[13] Yu. A. Tempel, O. A. Tempel, “Algorithm for Automated Correction of the Control Program Based on a Modified CAD Model of a Part, Taking into Account Errors,” *News of the Tula State University. Technical Sciences*, 2023, no. (1), pp. 444-448. https://doi.org/10.24412/2071-6168-2023-1-444-448

[14] K. A. Barilo, “Production Organization and Enterprise Management in the AI Era,” *Initiatives of the Youth for Science and Production: Proceedings of the VI All-Russian Scientific-Practical Conference*, 2023, pp. 96-99. Penza State Agrarian University.

[15] A. V. Sapunov, T. A. Sapunova, “The Relevance of Implementing Artificial Intelligence in Production Management at an Enterprise,” *Economics and Business: Theory and Practice*, 2022, no. 5-3(87), pp. 47-50. https://doi.org/10.24412/2411-0450-2022-5-3-47-50

[16] S. Y. Dementev, “Artificial intelligence in the manufacturing arena: innovations, challenges and prospects,” *International Journal of Information Technology and Energy Efficiency*, 2024, no. 9(1(39)), pp. 9-13.

[17] L. A. Gladkov, N. V. Gladkova, S. A. Gromov, “A Hybrid Model for Solving Operational Production Planning Tasks,” *News of the Southern Federal University. Technical Sciences*, 2018, no. 4(198), pp. 99-110. https://doi.org/10.23683/2311-3103-2018-4-99-110

[18] Y. S. Stadnik, T. Y. Kovalenko, T. S. Pushkina, S. Y. Pestova, “Processing Engineering Data from CAD and CAE Software Using SpringBoot,” *Architectural, Construction, Road and Transport Complexes: Problems, Prospects, Innovations: Collection of Materials of the VII International Scientific-Practical Conference*, 2022, pp. 577-581. Siberian State Automobile and Highway University (SibADI).

[19] I. A. Kozlova, R. B. Slavin, B. M. Slavin, “Graphical Disciplines and Informatization of Engineering Education,” *Geometry and Graphics*, 2022, no. 10(4), pp. 35-45. https://doi.org/10.12737/2308-4898-2022-10-4-35-45

[20] A. P. Buslaev, D. A. Kuchelev, M. V. Yashina, “Dynamical systems and mathematical models of information traffic,” *T-Comm*, 2018, vol. 12, no.3, pp. 22-38. (in Russian)

[21] A. S. Bugaev, A. G. Tatashev, M. V. Yashina, O. S. Lavrov, E. A. Nosov, “Reconstruction of traffic flow dynamics based on deterministicstochastic model and data obtained from intelligent transport systems,” *T-Comm*, 2019, vol. 13, no.10, pp. 35-44. (in Russian)

ПРОГРАММНЫЙ КОМПЛЕКС ДЛЯ АВТОМАТИЗИРОВАННОГО ФОРМИРОВАНИЯ И ВЕРСИОНИРОВАНИЯ МУЛЬТИМОДАЛЬНЫХ НАБОРОВ ДАННЫХ В ЗАДАЧАХ ОЦЕНКИ ТРУДОЕМКОСТИ

Карелина Мария Юрьевна, *Государственный университет управления, Москва, Россия,* ***karelinamu@mail.ru***

Макаров Владимир Сергеевич, *Нижегородский государственный технический университет им. Р.Е. Алексеева, Нижний Новгород, Россия,* ***makvl2010@gmail.com***

Кутков Владимир Дмитриевич, *Государственный университет управления, Москва, Россия,* ***kutkovVD@yandex.ru***

Василиса-Анастасия Васильевна Юдина, *Государственный университет управления, Москва, Россия,* ***vv_badakova@guu.ru***

Аннотация

Проблема точной оценки трудоемкости в машиностроении сопряжена со сложностью подготовки гетерогенных данных (САПР-модели, ERP-данные, текстовые технологические требования) для моделей машинного обучения. В статье представлена архитектура и методы реализации программного комплекса, предназначенного для автоматизации процесса формирования и версионирования мультимодальных наборов данных. Методология включает реализацию ETL-пайплайна для извлечения распределенных данных, а также применение двух программных ИИ-агентов для их интеллектуального обогащения: сверточной нейронной сети для вычисления индекса геометрической сложности из САПР-моделей и большой языковой модели для извлечения структурированных технологических параметров из текстовых описаний. Для обеспечения воспроизводимости исследований внедрена система версионирования данных на основе DVC и Git. Функционал комплекса продемонстрирован на модельных примерах деталей различной сложности, показана его эффективность в автоматическом формировании обогащенных, готовых к использованию датасетов. Делается вывод, что предложенный подход существенно сокращает трудозатраты на подготовку данных, повышает их качество и объективность, формируя надежную основу для построения высокоточных предиктивных моделей.

Ключевые слова: *оценка трудоемкости, подготовка данных, мультимодальные данные, машинное обучение, программный комплекс, версионирование данных, большие языковые модели (LLM), машиностроение.*

Литература

1. *Яковлева Е.А., Толочко И.А., Ким А.А., Черняева А.А.* Цицифровая трансформация системы планирования на основе цифрового двойника // Креативная Экономика. Т. 15, вып. 7. С. 2811-2826, 2021.

2. *Пономаренко М.В.* Автоматизированные системы управления производством в разрезе управления жизненным циклом изделия // XII Конгресс молодых ученых : сборник научных трудов, Санкт-Петербург, 03-06 апреля 2023 года. Санкт-Петербург: Национальный исследовательский университет ИТМО, 2023. С. 151-155. EDN WZWBYE.

3. *Рябченко, А. В.* Цифровая траснформация организационно-экономического механизма функционирования промышленных корпораций // Вестник Алтайской академии экономики и права. 2022. № 3-1. С. 120-127. DOI 10.17513/vaael.2106. EDN ZKSOKO.

4. *Курганова Н.В, Филин М.А, Черняев. Д.С, Шаклеин А.Г, и Намиот Д.Е.* Внедрение цифровых двойников как одно из ключевых направлений цифровизации производства // Int. J. Open Inf. Technol. Т. 7, вып. 5. С. 105-115, 2019.

5. *Вишневский А.А.* Проблемы внедрения информационных систем для управления производством // Строительные материалы, оборудование, технологии XXI века. 2022. № 6(275). С. 15-21. EDN JJDZIB.

6. *Сидорычева Е.Е., Сидорычев В.А., Ефимов И.П., Дмитриенко Г.В.* Методика автоматизированного сравнения электронных структур изделий в PDM/PLM-системах для предприятий авиационной приборостроительной отрасли // Автоматизация процессов управления. 2023. № 1(71). С. 123-132. DOI 10.35752/1991-2927_2023_1_71_123. EDN NZXEFW.

7. *Черкасов К.Ю., Хаймович И.Н.* Трансформация IDEF0-моделей конструкторско-технологической подготовки производства в табличные модели данных для PDM/PLM-систем // Известия Тульского государственного университета. Технические науки. 2025. № 5. С. 150-161. DOI 10.24412/2071-6168-2025-5-150-151. EDN XVJGVD.

8. Свидетельство о государственной регистрации программы для ЭВМ № 2024661192 Российская Федерация. Адаптер для взаимодействия клиентских приложений по автоматизированному созданию и управлению ведомостями технической документации с PLM/PDM системами : № 2024660592 : заявл. 16.05.2024 : опубл. 16.05.2024 / И. В. Косинский, А. В. Иванов, А. В. Гавришева [и др.] ; заявитель Акционерное общество "Санкт-Петербургское морское бюро машиностроения "Малахит". EDN SPZYXH.

9. Свидетельство о государственной регистрации программы для ЭВМ № 2024689134 Российская Федерация. MASIOT. Фабрика агентов : № 2024688029 : заявл. 22.11.2024 : опубл. 04.12.2024 / А. В. Бредихин, А. А. Мешков, А. Б. Агаджанян ; заявитель Общество с ограниченной ответственностью "ДМ Солюшн". EDN KSWOSI.

10. *Ульянин О.В., Абакумов Е.М.* Вопросы совершенствования структуры электронного архива технической документации повышенной ответственности и защиты информации // Технологии электромагнитной совместимости. 2024. № 1(88). С. 35-46. EDN BXLETE.

11. *Делимова И.С., Шпрехер Д.М.* Программное обеспечение для управления жизненным циклом, необходимое современному производству // Известия Тульского государственного университета. Технические науки. 2024. № 12. С. 437-441. DOI 10.24412/2071-6168-2024-12-437-438. EDN RAXDBM.

12. "Цифровые двойники и цифровая трансформация предприятий ОПК". Просмотрено: 18 август 2025 г. [Онлайн]. Доступно на: https://assets.fea.ru/uploads/fea/news/2019/04_april/15/elibrary_37180048_50837228.pdf
13. *Темпель Ю.А., Темпель О.А.* Алгоритм автоматизированной коррекции управляющей программы по измененной CAD-модели детали с учетом погрешностей // Известия Тульского государственного университета. Технические науки. 2023. № 1. С. 444-448. DOI 10.24412/2071-6168-2023-1-444-448. EDN JCDMOH.
14. *Барило К.А.* Организация производства и управление предприятием в эпоху ии // Инициативы молодых – науке и производству : Сборник статей VI Всероссийской научно-практической конференции молодых ученых и студентов, Пенза, 29-30 ноября 2023 года. Пенза: Пензенский государственный аграрный университет, 2023. С. 96-99. EDN KBCYKQ.
15. *Сапунов А.В., Сапунова Т.А.* Актуальность внедрения искусственного интеллекта в управлении производством на предприятии // Экономика и бизнес: теория и практика. 2022. № 5-3(87). С. 47-50. DOI 10.24412/2411-0450-2022-5-3-47-50. EDN CPFEDK.
16. *Dementev S.Y.* Artificial intelligence in the manufacturing arena: innovations, challenges and prospects // Международный журнал информационных технологий и энергоэффективности. 2024. Vol. 9, No. 1(39), pp. 9-13. EDN USCXIO.
17. *Гладков Л.А., Гладкова Н.В., Громов С.А.* Гибридная модель решения задач оперативного производственного планирования // Известия ЮФУ. Технические науки. 2018. № 4(198). С. 99-110. DOI 10.23683/2311-3103-2018-4-99-110. EDN PNMAXL.
18. *Стадник Я.С., Коваленко Т.Ю., Пушкина Т.С., Пестова С.Ю.* Обработка инженерных данных из программных средств CAD и CAE с помощью SpringBoot // Архитектурно-строительный и дорожно-транспортный комплексы: проблемы, перспективы, инновации : Сборник материалов VII Международной научно-практической конференции, приуроченной к проведению в Российской Федерации Десятилетия науки и технологий, Омск, 24-25 ноября 2022 года. Омск: Сибирский государственный автомобильно-дорожный университет (СибАДИ), 2022. С. 577-581. EDN DDYYOU.
19. *Козлова И.А., Славин Р.Б., Славин Б.М.* Графические дисциплины и информатизация инженерного образования // Геометрия и графика. 2022. Т. 10, № 4. С. 35-45. DOI 10.12737/2308-4898-2022-10-4-35-45. EDN FXHKNJ.

Информация об авторах:

Карелина Мария Юрьевна, *д.т.н., д.п.н., профессор, заведующий кафедрой управления транспортными комплексами, Государственный университет управления, Москва, Россия*

Макаров Владимир Сергеевич, *д.т.н., профессор, Нижегородский государственный технический университет им. Р.Е. Алексеева, Нижний Новгород, Россия*

Кутков Владимир Дмитриевич, *аспирант, специалист Центра управления инжиниринговыми проектами, Государственный университет управления, Москва, Россия*

Юдина Василиса-Анастасия Васильевна, *аспирант, специалист Лаборатории реверсивного инжиниринга, Государственный университет управления, Москва, Россия*