Dynamic Data Linkage Keys Generation Technology for Process Flow Data Analysis using User's Criteria

1st Amit Kumar
Digital Platform Innovation Center, Research and Development Group Hitachi Ltd. Tokyo, Japan
amit.kumar.en@hitachi.com

2nd Hideya Yoshiuchi
Government and Public Corporation Information System Division Public Digital Transformation Department, Hitachi Ltd. Tokyo, Japan
hideya.yoshiuchi.fq@hitachi.com

Abstract— **The digitalization of chemical processes, which use advancement of AI for data management and analysis, offers significant potential to boost operational efficiency and drive innovation in the pharmaceutical, chemical, and materials science industries. Process such as Single-Screw Extruder, Co-rotating extruder and Twin-Screw Extruders process, can benefit from digitalization, which could enhance the operation efficiency of mixing, compounding, processing materials, such as plastics and polymers. However, the digitalization of chemical processes could be challenging, as it requires managing independent diverse and complex process flow data analysis, especially when there is lack of critical Linkage Keys (LKs) for data linkage. In the extrusion process, new processing steps or materials often result in entity name changes. User-defined criteria rely on these changes because they require complete data linkage that matches the updated entity names. Existing data linkage methods like Re-Link[1] match words using lexical and semantic similarity. However, they often fall short when linking relationships based on knowledge, facts, and theories, as they primarily focus on mathematical similarity. Bit Vector [2] further enhances record matching and entity resolution by utilizing patterns. Despite these advancements, these methods struggle to achieve the knowledge-based linkage based on user criteria when entity names change, and when the same equipment name is referred to differently across datasets. To address these challenges, we propose a novel data linkage technique using Dynamic Multiple LKs (DMLKs), driven by user-defined criteria derived from historical data & expert knowledge. Our approach utilizes a two-stage ensemble ML model. In the first stage, the model generates numerous DMLKs by learning from meaning, facts, patterns, theories, principles, and procedures, which helps in recognizing entity name change based on diverse patterns and terminologies. In the second stage, these LKs are optimized using user-defined criteria, such as process name, material name, and sensing parameters, ensuring the LKs align with the most desired criteria, focusing on the most relevant data for effective linkage. Our technique achieved a maximum 83.4% precision score and a 12.6% improvement over the Re-Link. This improvement demonstrates our model's ability to prioritize the most relevant data features based on user criteria. By simplifying the linkage of complex chemical processes, the DMLKs approach enhances the management of independent process flow data, significantly boosting operational efficiency in the chemical industry, particularly in extrusion processes.**

Keywords— Data Linkage, Linkage Keys, User Criteria, Dynamic Multiple Linkage Keys

I. INTRODUCTION

In recent years, digitalized chemistry (DC), also known as Chemistry 4.0, has become a key force driving change in the chemical material industries[3]. This involves using digital tools like big data, artificial intelligence, and smart sensors to data collection, analyses of how materials are made and developed. As a result, areas like pharmaceuticals, food processing, and materials science have become much more efficient. By using advanced data analysis, these industries can run experiments faster, make production processes smoother, and come up with new and innovative products. Market analysis shows that the global Chemistry 4.0 market was valued at USD 64.61 billion in 2022 and is expected to reach USD 124.32 billion by 2029, with a Compound Annual Growth Rate (CAGR) of 9.8% from 2023 to 2029[4]. This notable growth highlights the increasing dependence on digital technologies to revolutionize chemical research, production, and material development. The development of new materials often required fine-tuning numerous process parameters, such as temperature, pressure profiles, screw speeds, and feed rates, to achieve desired properties. Digitalization tackles this complexity with precise instead of managing manual parameters. By leveraging DC, it becomes possible to integrate data across various stages of the process from material characterization to process monitoring. Utilizing this integrated data allows industries to implement real-time monitoring, predictive maintenance, optimized operational workflows, and new material development analysis. These advancements lead to higher product quality, and efficient operation in the chemical sector. An application of DC is observed in complex industrial processes like extrusion machines. An extrusion is a continuous material processing system equipped with two intermeshing; co-rotating screws housed within a cylindrical barrel. It performs multiple functions, including conveying, mixing, shearing, heating, cooling, and shaping materials under precise process controls [5]. Screw extrusions are representative and important because they play a key role in advanced material processing. They are widely used in industries like polymer compounding, pharmaceuticals, and food processing for tasks such as mixing, kneading, and extruding materials[6]. However, for the development of new material, user define the criteria based on industries expert knowledge and historical data for efficient production such as process setting, equipment or sensor it is

important to analyse specific data of the complete process to know new process effect and cause. It leads to the linkage of data based on user criteria such as based on new process name, equipment name or sensor. Each stage of the extrusion process generates its own independent data. However, integrating this data requires a common key, which is often missing. Without this common linkage, it becomes difficult to analyze the process of material characterization. To address this challenge, it is crucial to implement robust data integration frameworks that connect different data sources based on user criteria. This approach enables comprehensive analysis and valuable insights that drive process integration. Integrating complex and diverse data types such as process sensing data, materials information, equipment information, simulation outputs, and feature evaluations is essential to maximizing the full potential of DC.

Managing diverse and complex process flow data in extrusion is challenging for data linkage. In process flow, when new materials or processing stages are introduced, users often need to link data specific to these new or desired entities. However, in a extrusion, each step of the process generates independent data, and the absence of LKs makes it difficult to link this data. Current data linking methods, like Re-Link and Bit Vector, often fail to manage complex data based on user criteria and knowledge-based learning. This limits the ability to link independent datasets and derive insights and optimize processes. To overcome these issues, we propose a method for creating DMLKs based on user-defined criteria from expert knowledge and historical data. Our solution uses a two-stage ensemble ML and BERT model. In the first stage, the model generates initial DMLKs by learning from the data. This helps identify relationships and variations in the datasets. In the second stage, these keys are refined using user-defined criteria, such as equipment name or process value, to tailor them to specific datasets. It can revolutionize material flow and product development. By dynamically creating and refining LKs, we ensure robust data integration even with diverse datasets. This supports effective research and development and scalable chemical production, bridging the gap between lab experiments and large-scale manufacturing. In summary, the deployment of DMLKs within DC represents a significant advancement, enhancing data management, driving innovation, and ensuring operational excellence across industries. This development underscores the transformative potential of DC, leading to a new era of efficiency and innovation. By leveraging data from sources like material composition [8], process parameters, simulations, and evaluations, we aim to unlock the full potential of chemistry data digitalization, improving efficiency, quality, and innovation in industries reliant on extrusion and related processes.

The literature survey on Data linkage will be summarized in section II. Problem overview, challenges of data linkage and methodology used for input dataset for DMLKs in III and IV. Section V briefly explains the result and discussion of our research. Section V presents the conclusion of this research.

II. Related Work

Data Linkage has been widely studied in literature over the past decades. We have classified literature based on technology used in the linkage process such as Re-Link using lexical and semantic similarity measure[9] for the record linkage, and Bit vector record linkage use entity resolution by using temporal patterns.

A. ReLink: Complete-Link Industrial Record Linkage Over Hybrid Feature Spaces

Record Linkage (Re-Link) is the task of identifying records from different databases that refer to the same real-world entity. This is crucial for organizations to integrate data across silos, improving efficiency in data engineering, analytics, and business applications like personalized marketing. State-of-the-art (SOTA) ML techniques for Re-Link. However, these methods often struggle with industrial data due to schema heterogeneity, the need to leverage data structure, and the lack of training data. Re-Link is a proposed system designed to incorporate both lexical and semantic similarity measures[7]. This processes records by selecting attribute pairs across databases, applying value transformations, and extracting features. It then trains a combination of traditional and deep learning models on these features. Key Technical contributions:

1) Complete-Linkage: Similarity across related but not identical attribute pairs when databases do not share the same schema.
2) Hybrid Feature Spaces: Semantic similarity measures. Feature augmentation to handle data sparsity and repetitive attributes.
3) End-to-End Solution: Achieve high F1 scores on both benchmark and real-world datasets.

In conclusion, Re-Link's hybrid approach, combining lexical and semantic features and addressing industrial challenges, provides a robust solution for record linkage tasks, demonstrating superior performance over existing SOTA.

B. Bit Vector Record Linkage

This work describes an advanced system and method designed to improve the accuracy and efficiency of record matching [13]and entity resolution, particularly in healthcare. This system enhances traditional record linkage methods by incorporating power-spectrum-based temporal patterns and phenotypic bit-vector fingerprints. The process begins with retrieving candidate health records, each associated with multiple encounters. The likelihood of these spectra determined using Bayesian Chain Monte Carlo simulation, helping to identify temporal patterns in the data. Concurrently, a record linkage scoring weight is calculated using methods such as the Fellegi-Sunter(F-S) approach. These weights are combined with the power spectrum likelihood using methods like root-mean-square(RMS) transformation or cosine transformation to form a composite score for each record. These composite scores are ranked, and a threshold is applied to identify matching records, with records exceeding the threshold considered probable matches. Records identified as matches through the composite score and fingerprint similarity are then merged to form a unified health record. This methodology significantly enhances the precision of record linkage by leveraging time-series analysis, making it particularly valuable in healthcare settings where accurate patient data matching is crucial for health record and improving data accuracy and utilization in healthcare delivery and research

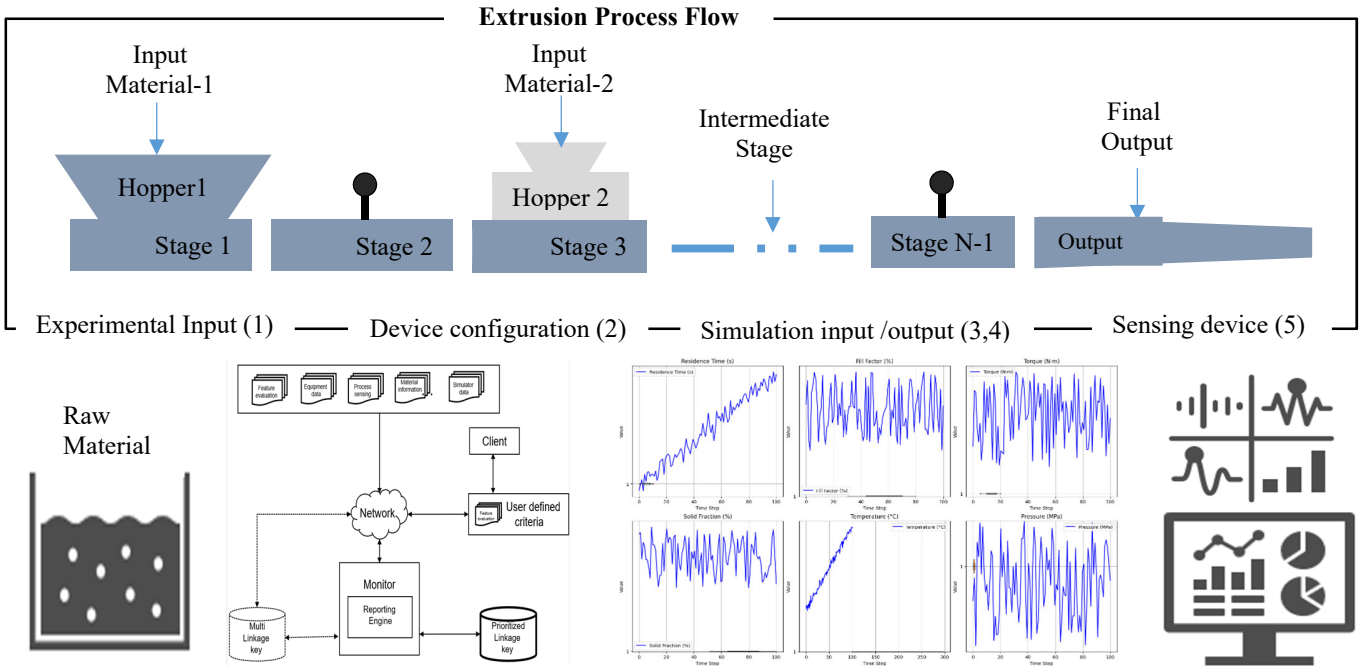

Fig.1 Overall view of extrusion process example data management system

III. Problem Overview

In this section we introduced the problem statement and challenges in material flow data linkage. We also explained the user criteria and their importance in dynamic data LKs. A critical component of DC is the efficient management and integration of complex data streams, which include device configuration, raw materials, device configuration specifications, process sensing, material properties and simulation.

A. Data Managemnet System of extrusion

In the realm of chemical manufacturing, extrusion process is a sophisticated piece of equipment used for continuous mixing, compounding, or processing of materials. The extrusion operates through a series of stages where materials are conveyed, mixed, and transformed under controlled conditions.

The digitalization of extrusion processes allows for systematic collection and analysis of data at each stage such as (Stage1: Input material-1, experimental input, Device configuration of the extrusion machine, Stage 2 to Stage N-1 : Which is extrusion process consider as intermediate stage for the sensing device such as Temperature, Pressure and humidity and viscosity of material, And Final stage is the output where the final material ready for the evaluation output. Thus, enabling better process control and optimization. When new materials or processing stages are introduced, to identify the effect of new process users often need to link data to these new or desired processes. The complete overview is shown below in Fig.1 which has the following data management system. This diagram outlines a process and data management system for an experimental and simulation setup involving materials processing. Experimental Conditions (1): Define the recipe and conditions for the raw materials. Device Configuration (2): Start by setting up the equipment for the experiment. Simulation Input (3): Input the experimental conditions and parameters into the simulation software. Simulation Output (4): Run the simulation and collect the output data. Process Sensing (5): Use sensors to collect data during the material processing. Continuously monitor the equipment to ensure it operates correctly and measure the temperature, pressure, and humidity parameters. Characteristic Evaluation (6): Analyze the properties of the processed materials.

B. Problem Statement

The primary technical challenge lies in the integration of heterogeneous and large volume data generated at distinct stages of the extrusion process as per user requirement.

TABLE I. Extrusion Process Data Type

Data Type	**Dataset**
Experimental Input	(1), (2)
Experimental output	(4), (5)
Simulation Input	(3)
Simulation Output	(4)
Evaluation result	(6)

TABLE II. Data Need to Linked

Tables that need to be linked together as analysis data	(1)Experimental input raw material (2) Device configuration (5) Process sensing data (6) Characteristic evaluation

Each operational step produces distinct datasets with name variable with has different name in other stages, encompassing variables such as material feed rates, screw speed, temperature profiles, and torque measurements. The lack of standardized LKs complicates the task of correlating these datasets, which is essential for comprehensive process understanding and optimization. Current methods like Re-Link and Bit Vector struggle to manage changes in the name of materials, stages, and naming conventions, leading to poor data integration. This limits the ability to improve efficiency and innovation in extrusion operations. The chart emphasizes the need to link all data tables using LKs. This linkage allows researchers to compare and contrast experimental results with simulation outcomes, ensuring comprehensive data analysis and validation of the findings.

C. Challenges Faced in Material Flow Data Linkage

- In digitalization processes, each stage of operation produces independent datasets, due to which it has no direct linkage and without proper LKs, accurately integrating this information based on user criteria becomes challenging.
- Existing methods exhibit significant limitations when faced with dynamic and evolving datasets. For instance, the introduction of new materials, variations in processing stages, or inconsistent nomenclature for equipment across different datasets can lead to inaccurate or incomplete data linkage.

D. User Criteria

The user criteria are based on the industry expert knowledge to optimize the LKs initiate with the user criteria and followed by the historical data for the achieving the maximum output in optimized steps. Considering the extrusion process where the user is curious about the new added steps or the material. And how the outcome value will be changed. User criteria are customized and can be based on expert knowledge or historical data, depending on factors like yield efficiency or vendor specifications. The key parameters that can be used to define user criteria for linking data in the extrusion process. These criteria help generate the appropriate LKs, User-based criteria, such as process or equipment name, are added as new features. Key Parameters for User-Defined Criteria is a set of rules pre-defined for maximum yield such as in case of extrusion for high quality material design. Examples such as Temperature Range Process Name: Distillation, Equipment Name: Heater, Pressure Range, Reactor Type Preference, Flow Rate Range, Concentration Range. These parameters are derived from expert knowledge or based on historical performance data and used to identify appropriate dynamic LKs, ensuring the system meets high yield expectations. These user-defined rules enhance the model's ability to classify whether a product batch meets quality standards. In conclusion, expert knowledge is critical in setting criteria for LKs generation. By focusing on features like temperature, pressure, and equipment type, the system becomes more effective in analysis and decision-making, ensuring that the extrusion process data is accurately linked and optimized.

IV. METHODOLOGY

In this chapter, we propose data linkage technology for the data without having LKs. We have designed the architecture based on the issues faced by the existing technologies. To overcome such issues, we come up with novel ideas. So, the idea to select the two-stage ML model and system architecture as explained in section 4A and section 4B in detail, respectively.

A. Proposed DMLK technology

To tackle the data-linking challenges discussed earlier, we created a dynamic data linkage framework that adjusts to user-defined criteria, making it easier to connect related information even when the data is complex or lacks obvious connections. This approach addresses two main issues: First similar word but different meanings, sometimes words look the same but mean different things depending on the context, which makes linking tricky. Our first stage, using a BERT model classifier, focuses on understanding these nuanced meanings. It captures patterns, facts, theories, and other contextual details that relate to specific topics. Second similar context, different datasets, Datasets may contain related information but don't directly link to each other. Our second stage leverages both the first model's results and the user's specific requirements to identify new LKs. It processes unconnected information by interpreting it within context, drawing knowledge from previously unseen data to create new linkages key. DMLKs aims to identify the LKs for data link dynamically. Utilizes BERT for knowledge identification from multiple tables. Overall, the proposed approach aims to address the challenges in DC data utilization by developing a solution that leverages NLP and ML techniques for efficient data linkage using LKs. To identify linking keys across different datasets that do not share a common key.

B. Proposed Architecture

The architecture is divided into the detailed components and functionalities of the model architecture diagram shown in Fig. 2. Each part has been marked as a number and presented in a diagram.

I. Feature Engineering Techniques: The data linkage technology incorporates advanced feature engineering techniques tailored for numerical, categorical, and textual features. This helps to pull out key facts, meanings of words, and theories from each type of data. It is important because it helps solve issue one.

II. NLP Model: NLP models BERT are trained using historical data. The architecture adopts a diverse set of algorithms, including deep learning, decision trees, and ensemble methods. Issue one is solved using the first stage ML which generates the LKs.

III. DMLK Generation: To get data linkage across disparate datasets, BERT dynamically generates multiple LKs based on the extracted features. These keys are like labels that help connect similar records from various data sources, like equipment details, process data, and material info. And tackles issue two using BERT classifier [10].

IV. User-Defined Criteria and LKs Prioritization: The architecture incorporates user-defined criteria for data matching, allowing user to iteratively refine matching rules based on domain expertise and specific requirements. Additionally, multiple LKs are generated and prioritized to optimize the efficiency and accuracy of data linkage processes.

V. Model Integration and Analysis: Based on user criteria the architecture emphasizes the integration of newly generated features into existing models. Integration, thorough analysis is conducted to assess the distribution of the new features and their relationships with existing features.

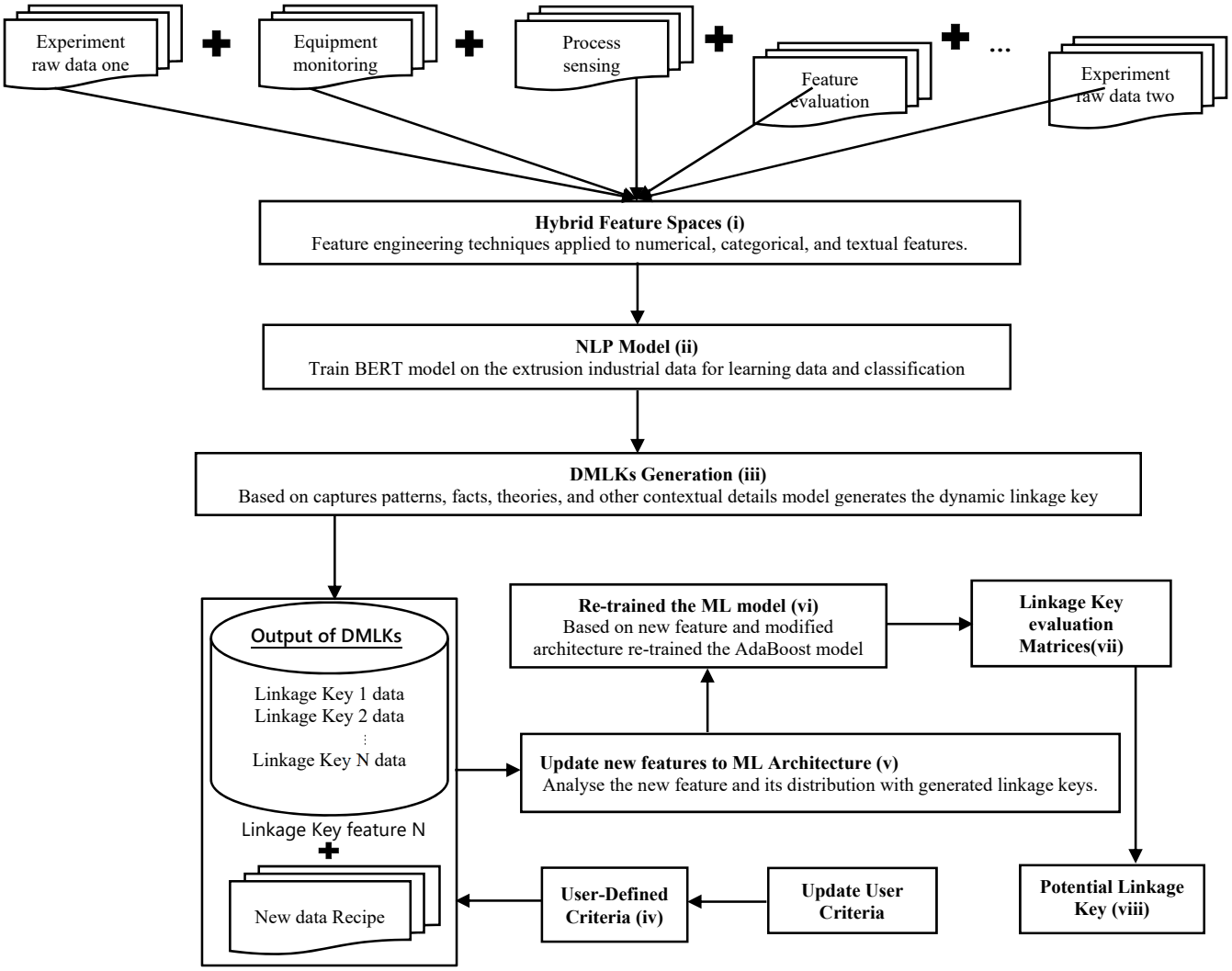

Fig. 2. Model architecture

VI. Re-Training Tree-Based Models: Tree-based models, including AdaBoost and XG-Boost, are re-trained iteratively based on updated datasets and refined feature sets. This iterative re-training process enhances model accuracy and adaptability to evolving data patterns and distributions.

VII. Feature Importance Analysis: A critical aspect of architecture involves evaluating the importance of features in the data through techniques such as precision. Analyzing feature importance analysis is conducted to determine the significance of each feature. Precision is used by the fitted attribute. This involves using feature attributes precision plots, such as heatmap and network graph, to identify the top LKs. Typically, by identifying the highest values of the precision score and choosing those that meet an acceptable limit.

VIII. Potential LKs: Once the potential LKs are identified based on the user criteria. If the user criteria are not satisfied, the process iterates, and the user criteria are redefined, potentially considering other records to meet the requirements. This iterative approach allows for continuous refinement of the LKs selection process until satisfactory results are achieved.

Dynamic key present in large in number, So, user defined criteria come in picture to identify the potential LKs. I have represented three steps as shown in Fig. 3, such as multiple LKs data which is output of ML1 used as input of the ML2 for generating the linkage based on user criteria. The architecture of ML2 as shown in Fig. 3 has been modified and tuned using the user criteria which is a novel approach we have used here. And by turning the ML2 with the new feature architecture we have generated potential LKs. To make the analysis of the raw data, it is important to manage data management in hybrid feature space shown in Fig. 4 for combining data based on one hot encoding. Hybrid feature spaces are the integration of distinct types of features, such as numerical, categorical, and textual features, in a unified framework. This approach is commonly used in ML and data analysis to leverage the strengths of diverse types of data for better model performance. Hybrid feature space can be applied to numerical, categorical, and textual features. Then the ML1 model as shown in Fig. 4 (ensemble decision-tree classifier methods) has been used for generating the DMLK which is (solution of issue 1). The classification model is trained by industrial chemical data, and it is used with the BERT for the knowledge-based learning of the classification model. Once the test data which is in tabular form is given to the classifier, it is classified based on the knowledge. And generate a large number of DMLKs.

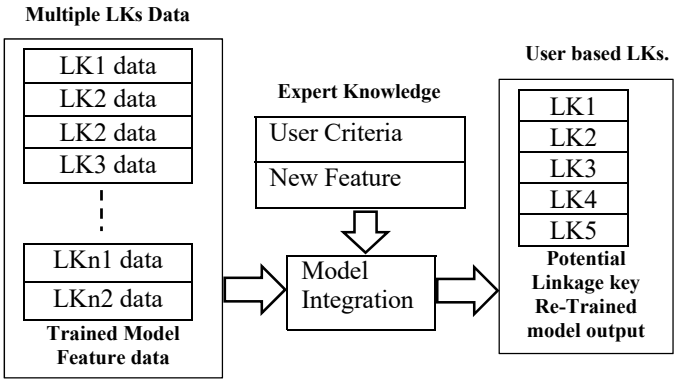

Fig. 3. Model integration using user defined criteria

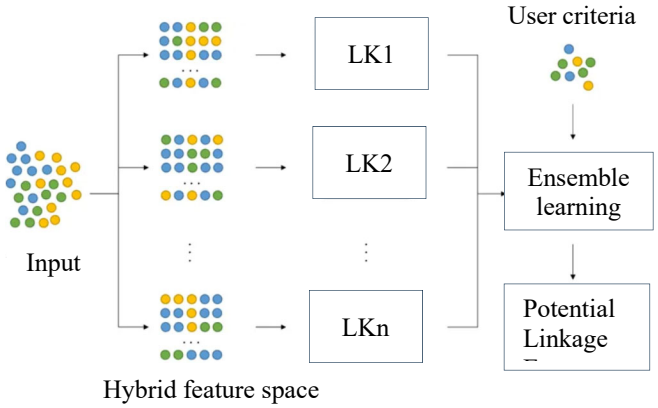

Fig. 4. Linkge key classifier using user defined criteria

C. DMLKs Learning Function

a) Data Preprocessing and Integration: The preprocessing pipeline effectively concatenated text data from six distinct dataframes, representing various aspects of mineralogical data. This ensured compatibility with the downstream NLP tasks. In the equation (1) creating a unified corpus of text T, where each text T_i is a combination of column entries. Mathematically,

$$T_i = \bigcup_{i=1}^{n} concat(C_{ij}) \quad (1)$$

where C_{ij} represents the j^{th} column in the i^{th} dataframe.

b) Embedding Generation Using BERT: Each T_i is tokenized into X_i = tokenizer (T_i), resulting in a sequence of tokens X_i. In the equation(2), tokens were encoded into embeddings $E_i \in R^d$ using the pre-trained BERT model:

$$E_i = f_{BERT}(X_i) \quad (2)$$

where d is the embedding dimension, and f_{BERT} denotes the BERT encoding function.

The pooler output $E = [E_1, E_2, \ldots, E_m]$ formed the matrix of embeddings for m texts.[11]

c) DMKLs Matching Calculation: We introduced a DMLKs matching metric M(i,j) in the equation(3), which evaluates the similarity between texts T_i and T_j dynamically based on their alignment in terms of semantic meaning, facts, patterns, theories, principles, and procedures. This accounts for variations in terminologies and entity naming conventions. The metric is defined as:

$$M(i,j) = \frac{\sum_{k=1}^{d} 1(E_{ik}, E_{jk})}{d} \quad (3)$$

where align(E_{ik} , E_{jk}) is a function that evaluates semantic alignment between corresponding dimensions E_{ik} and E_{jk}, and $1(\cdot)$ is an indicator function returning 1 if aligned and 0 otherwise.

d) Agglomerative Clustering: Clustering has been performed using the precomputed matching matrix M. The agglomerative clustering algorithm minimized linkage distances E(i,j) in equation(4), C_k calculated as in equation(5):

$$E(i,j) = 1 - M(i,j) \quad (4)$$

$$C_k = arg\,min_k \left(\frac{\sum_{i,j \in k}^{n} D(i,j)}{|k|} \right) \quad (5)$$

where assigned cluster labels C_k and|k| represents the size of cluster k. [12]

e) LKs Assignment based on user criteria: For each feature i in cluster k, the linkage key L(i) is calculated as in equation(6):

$$L(i) = C_k + U_k \quad (6)$$

where: C_k is the cluster label for i. U_k is a user-defined adjustment factor applied to cluster k, enabling prioritization or weighting of specific clusters.

D. Input Data

To make the analysis of independent industry raw data , it is important to make it useful for the learning model preference. In case of data linkage, As the chemical industrial data are available in vast amounts. To make it useful for the analysis and get the useful inference to increase production, automation and new material development, data needs to be linked so that it can be analyzed easily and applied to researchers for the application of new age cutting technologies. We have used the data explained below for the input of our model. The presented data is open-source data for our model analysis, For the actual analysis we have used the IMA database [14] of mineral properties and Walmart-Amazon[15] dataset has been used. The IMA database is for creating a complete set of high-quality spectral data forms as well characterized minerals and is developing the technology to share the information. Datasets include attributes names mineral name, unique formula (RRUFF and IMA) in plain text (Abellaite is $NaPb^{2+}{}_2(CO_3)_2(OH)$), IMA number and RRUFF ids, chemistry elements, structural group. RRUFF is an acronym for Raman, X-ray, and Infrared.

TABLE III. INTERNATIONAL MINERALOGICAL ASSOCIATION (IMA) DATABSE

(a)

S. N.	Mineral Name	Chemistry Elements
1.	Abellaite	Na Pb C O H
2.	Abelsonite	Ni C H N
3.	Abhurite	Sn O H Cl
4.	Abernathyite	K U O As H
…	…	…

(b)

S. N.	RRUFF Chemistry(Plain)	RRUFF IDs
1.	NaPb2+2(CO3)2(OH)	NaN
2.	Ni2+C31H32N4	R070007
3.	Sn2+21O6(OH)14Cl16	NaN
4.	K(U6+O2)As5+O4A.3H2O	NaN
…	…	…

(c)

S. N.	IMA Chemistry (Plain)	IMA Number
1.	NaPb2(CO3)2(OH)	2014-111
2.	NiC31H32N4	1975-013
3.	Sn2+21O6(OH)14Cl16	1983-054
4.	K(UO2)AsO4A.3H2O	NaN
…	…	…

(d)

S.N.	Mineral Name	Chemistry Elements
1.	Abellaite	Na Pb C O H
2.	Abelsonite	Ni C H N
3.	Abhurite	Sn O H Cl
4.	Abernathyite	K U O As H
…	…	…

(e)

S.N.	Mineral Name	Chemistry Elements
1.	Abellaite	Na Pb C O H
2.	Abelsonite	Ni C H N
3.	Abhurite	Sn O H Cl
4.	Abernathyite	K U O As H
…	…	…

There are four independent tabular data which is shown in Table III(b, c, d, e) and Table III (a) is the user defined dataset which need to link based on the DMLKs approach, but these four data are internally correlated to each other which need to determine using the LKs. And potential LKs are derived using User defined data Table III (a). As shown tabular data in Table III(b, c, d, e) , there is no common key present. So for the linkage there need to derive the linkage key among the column of datasets, Because all these data are internally correlated, So correlation has to derived based on the DMLKs.

V. RESULTS AND DISCUSSION

Based on the experiment run we have evaluated the IMA database for the generating DMLKs and then potential LKs based on the user criteria. The evaluation of result is resented in the two stages. In the experiment we used the matrices such as precision and recall values for the evaluation. Firstly the DMKLs learning of the IMA properties and then evaluate the percentage matching based on the learning from meaning, facts, patterns, theories, principles, and procedures, which helps in recognizing entity name change based on diverse patterns and terminologies. And then second stage, these LKs are optimized using user-defined criteria, such as equipment type, process name, material name, and sensing parameter range, ensuring the LKs align with the most desired criteria, focusing on the most relevant data for effective linkage. The evaluation metrics precision is used to evaluate the potential linkage.

A. Linkage key Precision Value

The LKs are calculated of the given tabular data presented in Table III (a, b, c, c, d, e). We calculated the linkage key pairs and evaluated based on the precision value. These data needs links and to generate the LKs, the precision of the LKs has been derived using a two-stage learning algorithm using user criteria.

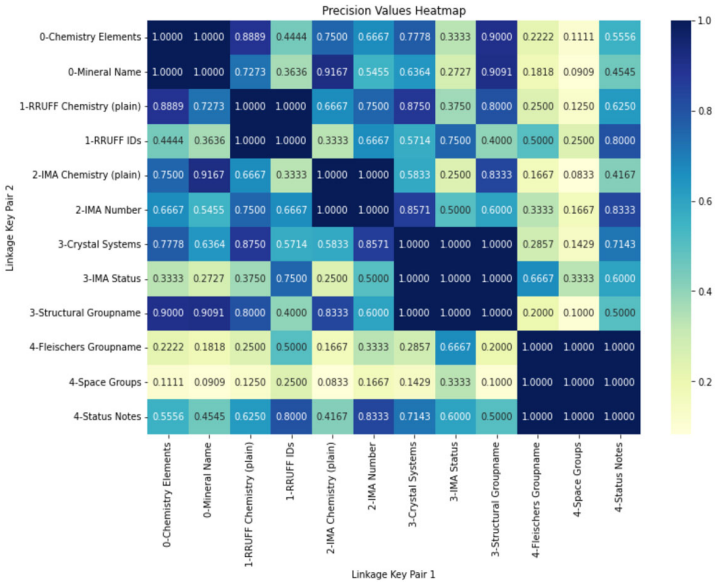

Fig. 5. Precison heatmap of linkage keys

The heatmap presented in Fig. 5 visualizes the precision values obtained from evaluating various linkage key pairs which indicates the accuracy of the linkage between datasets. The elements of the heatmap show a precision value of 1, indicating the same tabular data and their attributes. The highest potential linkage key based on user criteria Table III (a) the precision values are: Between "Mineral Name" and "RRUFF Chemistry (plain)" with a precision of 0.9167 are highly related based on the chemical formula of compound. Between "IMA Number" and "Crystal Systems" with a precision of 0.8571. Between "Crystal Systems" and "RRUFF Chemistry (plain)" with a precision of 0.8750. These values suggest a strong correlation and effective linkage between these particular pairs of datasets, indicating their potential utility for precise data integration. Conversely, lower precision values are evident in pairs such as "Space Groups" with "Chemistry Elements" (0.1111) and "Fleischers Group name" with "Mineral Name" (0.1818). These lower values suggest weaker correlations where linkage possibility is least. The network graph depicted in figure 6 illustrates the precision values between different linkage key pairs, providing a visual representation of the correlations within the dataset. Each node represents a dataset category, while the edges connecting them are labeled with precision values, indicating the strength of their linkage. Consistent with the heatmap analysis, certain pairs exhibit high precision, highlighted by thicker edges in the graph. The dataset with their same data attributes, with precision values of 1. Overall, the network graph provides a comprehensive view of the precision values, highlighting both strong and weak linkage data integration.

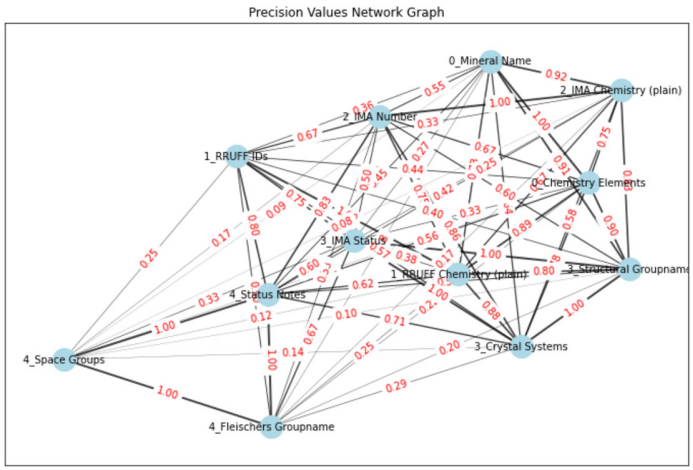

Figure. 6 Network graph of precison linkage keys

This visualization aids in identifying key areas for improvement in data integration strategies, ensuring more accurate and reliable synthesis across datasets.

TABLE IV. PRECISION OF AGAINST BASELINES DATASETS

Dataset	Precision	Method
Industry Data	0.778	Re-Link
Walmart-Amazon	0.708	Re-Link
Walmart-Amazon	0.834	DMLKs
IMA Data	0.9167	DMLKs

The Re-Link method has precision score of 0.778 on Industry data, and to the Walmart-Amazon dataset, yielding a precision of 0.708 over Re-link. In contrast, the proposed DMLKs methodology was utilized for the IMA dataset and Walmart-Amazon dataset, achieving a significantly higher precision score of 0.9167 and 0.834 respectivelly of as illustritated in Table IV. These results demonstrate the superior performance of the DMLKs approach in accurately linking complex datasets compared to traditional methods like Re-Link.

There is significant imprevemt of 0.128 precision score for Walmart-Amazon dataset.

VI. CONCLUSION AND FUTURE WORK

The research demonstrates the effectiveness of the proposed DMLKs methodology in enhancing data linkage within digitalized chemical processes, particularly for the industrial process flow data. We performed experiment over the IMA data and Walmart-Amazon data by leveraging user-defined criteria, the experiment achieved an maximum 91.67% precision score for IMA and 83.4% for Walmart-Amazon data, marking a significant improvement 12.8% over traditional methods like Re-Link. This advancement underscores the potential of DMLKs to address the challenges posed by complex and diverse datasets, facilitating more accurate and efficient data integration. The results highlight the DMLKs' ability to dynamically generate and optimize linkage keys, ensuring precise data alignment and contributing to the operational efficiency and innovation in chemical industries. In the future we will focus on diverse domain data linkage such as healthcare, logistics where complex data linkage is crucial.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to Qi Xiu and Mika Takata for their invaluable contributions to the improvement of this research paper. Their insightful reviews and constructive feedback were instrumental in refining our methodology and enhancing the overall quality of the work.

REFERENCES

[1] Joshi, Salil Rajeev, Arpan Somani, and Shourya Roy, Relink: Complete-link industrial record linkage over hybrid feature spaces. In 2021 IEEE 37th International Conference on Data Engineering (ICDE), pp. 2625-2636. IEEE, 2021.

[2] D. S. McNair, L. Kailasam, and J. C. Murrish, "System and method for record linkage," U.S. Patent 10,580,524 B1, Mar. 3, 2020.

[3] Pietrasik, Marcin, Anna Wilbik, and Paul Grefen. "The enabling technologies for digitalization in the chemical process industry." *Digital Chemical Engineering* (2024): 100161.

[4] Maximize Market Research Pvt. Ltd., "Chemistry 4.0 Market: Uptake in the Deployment of Digital Infrastructure in Chemical Plants to Unleash New Growth Opportunities.

[5] Jieya Extruder. "Exploring the Four Main Types of Twin Screw Extruders." *Jieya Extruder*, 1 Dec. 2023.

[6] Thermo Fisher Scientific. Compounding and extrusion equipment overview. https://www.thermofisher.com

[7] Devlin, J., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[8] Singh, Balwant, Raman Kumar, and Jasgurpreet Singh Chohan. "Polymer matrix composites in 3D printing: A state of art review." Materials Today: Proceedings 33 (2020): 1562-1567.

[9] Kumar, A. and Yoshiuchi, H. Graph Similarity-Based Data Management Platform for Advancing Material Informatics. In 2024 9th International Conference on Big Data Analytics (ICBDA) (pp. 85-91). IEEE

[10] Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. "Bertscore: Evaluating text generation with bert." *arXiv preprint arXiv:1904.09675* (2019).

[11] Sasirekha, K. and Baby, P., 2013. Agglomerative hierarchical clustering algorithm-a. *International Journal of Scientific and Research Publications*, *83*(3), p.83.

[12] Lafuente, Barbara, Robert T. Downs, Hexiong Yang, Nate Stone, Thomas Armbruster. "The power of databases: the RRUFF project." *Highlights in mineralogical crystallography* 1 (2015): 25.

[13] S. Mudgal, H. Li, T. Rekatsinas, A. Doan, Y. Park, G. Krishnan, et al., "Deep learning for entity matching: A design space exploration", *Proceedings of the 2018 International Conference on Management of Data*, pp. 19-34, 2018.

[14] Chemkaeva, Daria. n.d. IMA Database of Mineral Properties. Kaggle 2024. https://www.kaggle.com/datasets/dariahemkaeva/ima-database-of-mineral-properties.

[15] Hasso-Plattner-Institut, "Amazon-Walmart dataset," 2025. https://hpi.de/naumann/projects/repeatability/datasets/amazon-walmart-dataset.html.