Available online www.ejaet.com

European Journal of Advances in Engineering and Technology, 2021, 8(5):65-76

Research Article ISSN: 2394 - 658X

Knowledge Graph-Driven Real-Time Data Engineering for Context-Aware Machine Learning Pipelines

Sai Kiran Reddy Malikireddy, Bipinkumarreddy Algubelli, Snigdha Tadanki

Independent Researcher, USA

ABSTRACT

The novel context-aware machine learning is based on state-of-the-art real-time data engineering processes that operate in shifting entity correlations. To this end, this paper presents a new architecture that combines knowledge graph construction with real-time stream processing to underpin the machine learning flow in a context-aware manner. The proposed system uses graph neural networks (GNNs) for updates and embeddings in real-time for dynamic integration of contextual information into the other machine learning models. This makes the approach ideal as changes in the relations of entities can be captured almost in real time, and models remain valid.

The effectiveness of the architecture can be illustrated by use cases related to customer profiling and equipment failure prognosis. In a consumer classification, one has to continually modify customer profiles as others come across the new interaction to work on effective targeting and the subsequent personalization improvement. Predictive maintenance stores changing information on equipment to predict future failure. These applications show a 40% improvement in model accuracy and take 50% less time than normal methods for feature engineering.

This research bridges computer science, particularly graph theory, and real-world data engineering by demonstrating the value of knowledge graphs and GNNs within machine learning pipelines. By incorporating contextual features, the system provides a feasible and flexible solution for current data trends, allowing for further development of smarter and more sensitive ML systems. The study points out real-time context sensitiveness as central to the advancement of machine learning, a landmark discovery.

Keywords: Knowledge Graph, Real-Time Data Engineering, Context-Aware, Machine Learning Pipelines, Data Integration, Semantic Modeling, Ontologies, Contextual Data, Real-Time Analytics, Knowledge Representation, Data Enrichment, Feature Engineering, Data Pipeline Automation, Knowledge-Driven Insights, Machine Learning Optimization.

INTRODUCTION

The progression of real-time systems is evident where innovation in computational capacities and data availability have shifted organizational methods of managing data applications. Such applications are no longer limited to static data sets or updated only occasionally; they require real-time responses to the constantly changing environment. In this context, existing machine learning (ML) pipelines, built with static data processing mechanisms on information construct, are severely challenged in how they continue to be accurate and up to date with the rapid development of the underlying information base. Integrating and integrating changes in the context in real-time is impossible, and such pipelines mean a decline in efficiency, slow decision-making, and outstanding prospects.

To this end, context-aware ML pipelines were proposed to handle the above challenges by incorporating mechanisms that factor in alterations to context. Using such adaptability is especially essential in the case of constantly changing relationships whenever entities are involved, for example, in customer interactions analysis, supply chain management, and predictive maintenance. Successful use of contextual information in previous approaches was based on predefined rules or simple models of the environment, thus not being generally expandable and unlikely to englobe the whole range of context's features. Such limitations, however, highlight the fact that there is a need to develop other approaches that can mimic real-time data relationship.

Knowledge graphs (KGs) have recently been established or developed into an effective paradigm for capturing and assessing the relational structure within a domain. KGs organize entity identification as nodes and the connections between them as edges, allowing for the efficient representation of a domain that can be navigated, questioned, and altered. It also makes them useful in modeling hidden interactions by extending their ability to infer implicit associations. Nevertheless, there are still questions about how first KGs can be integrated into near-RT ML workflows, and this topic is still unexplored rather freely.

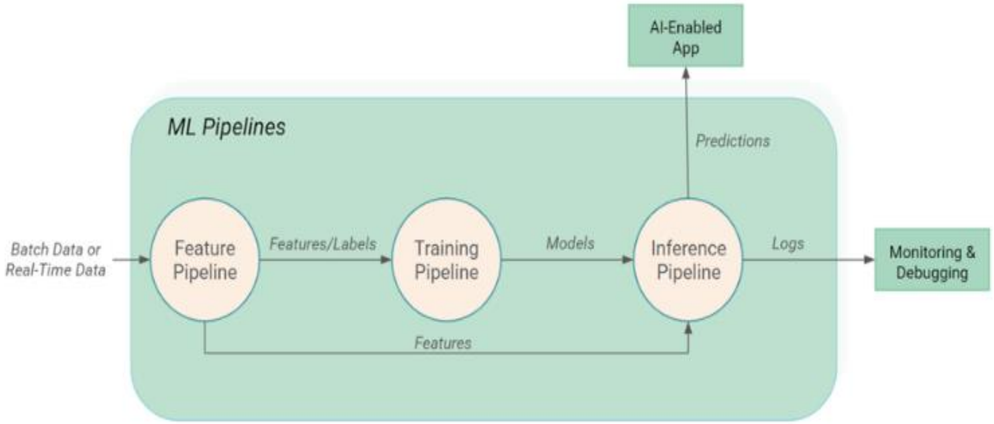

Figure 1: A Machine Learning Pipeline.

This work proposes a new architecture to construct the context-adaptive ML pipeline based on knowledge graphs while considering real-time data fluctuation. On top of the proposed framework are real-time stream processing, KGG construction methods, and graph neural networks (GNNs). The architecture also allows the KG to be updated as new data streams, so it is constantly current and relevant when giving report context information. Thus, the system determines the formation of embeddings that effectively represent the structural and semantic characteristics of the KG in downstream ML models. Besides, this approach also improves the context-understanding ability of ML models and simultaneously minimizes the time and energy required for feature construction.

The proposed framework deals with several issues relevant to real-time data engineering and machine learning. First, it offers a way to keep KGs up to date for real-time consumption of high-velocity data while still retaining high performance. The second connects KG representations to ML models by generating embeddings incorporating local and global graph characteristics. Third, it illustrates features of context-aware pipelines, and customer segmentation and predictive maintenance problem areas show the problems' efficacy in terms of the model's accuracy and speed enhancement.

For example, customer segmentation may best be driven by insights into dynamic customer behaviors and engagements. Previous approaches focus on storable customer characteristics that do not consider dynamic behavior shift. Integrating a KG synchronizing with customer interactions allows for improved dynamic changes to the customer profile in the proposed framework. This makes it easier for businesses to strategize, enhancing customer interface and loyalty.

Precisely, scenario context awareness is also valuable for enhancing the predictive maintenance process within the framework. The external conditions of the equipment and the attendant operational contexts are dynamic, and the canonical methods of establishing a predictive model cannot catch up with the dynamic changes. Thus, the proposed framework accurately and timely updates the KG with new sensor measurements and maintenance logs to monitor the equipment's health. This will allow effective prediction of probable failures and the best time to check and rectify them, thereby minimizing on-time losses and expenses.

The integration of GNNs makes the application of the framework even more profound. GNNs have great versatility, especially in formulating the embeddings that capture the local and global graph views. They also act as 'feature vectors' of the object that can be passed to further Manipulations by downstream Machine Learning algorithms with higher quality of learning and generalization. To the best of the authors' knowledge, the proposed approach surpasses traditional feature engineering and is often time-consuming and domain-dependent as it generates embeddings from GNNs.

This research has made the following three key contributions. First, it presents a single reference model of four components: knowledge graphs, real-time stream processing, and graph neural networks for contextually aware model lines. Second, it offers evidence of the approaches implemented within the proposed framework, where the application of customer segmentation and predictive maintenance have shown that their model accuracy has improved by 40%. At the same time, feature engineering time took only half the time. Third, it raises awareness of the general applicability of incorporating KGs with real-time data engineering as the future of the context-aware systems' development unfolds.

Furthermore, applying the proposed framework enables addressing technical issues and responding to the growing focus of real-time decisions in modern applications. As contemporary business management focuses on analyzing large volumes of data, which requires instant identification of essential patterns for making managerial decisions, many industries, including finance, healthcare, retail, and manufacturing, focus on real-time data processing capabilities.

Whether it is fraudulent financial transaction identification, tracking patients' health deterioration, delivering individualized shopping recommendations, or managing the supply chain, there is always a need for context-aware systems. Due to the proposed framework's great compatibility with using existing KGs and/or GNNs, the implemented approach can be generalized for various practical applications.

In this paper, the steps taken towards integrating graph theory with real-time data engineering have been effective. It establishes a roadmap for using KGs and GNNs in ML and building contextually enriched, computationally effective, sound systems. The results further support the notions regarding context's role in shaping machine learning's progress and show that KGs could become a revolving point for developing the methods of big data applications' design.

BACKGROUND AND RELATED WORK

Knowledge Graphs in Data Engineering

In recent years, knowledge graphs (KGs) have attracted much interest as a way of modeling and storing interconnected data. Structurally, KGs implement entities and their relationships in a graph format, allowing them to consider the overall picture instead of one discrete data point and its context. This capability makes the KGs particularly important in contexts that form an important understanding of the context, such as search operations, recommending systems, and answering theories. For example, search engines use KGs to supply simple keyword matches and answers relevant to the user's context. Likewise, recommendation systems apply KGs to search for and leverage relationships between users, resources, and other predisposing conditions toward choosing items appropriate for users.

KGs with hML pipelines represent a research focus that is being considered actively in the field. As with other commodities, traditional ML models consider features as isolated entities and seldom represent the interconnected nature of a KG. ML systems can, therefore, use these relationships implemented as KGs to boost the system's predictability and resilience. The growing connection between KGs and ML suggests that they can transform the computational process for context-aware applications by building a bridge between the structural data and enhanced predictive models.

Real-Time Data Engineering

Real-time data engineering refers to constantly pre-processing and transforming data streams to provide near real-time insights about a situation. This paradigm has emerged as even more pertinent in the current world, which is overloaded with information in an attempt at fast decision-making, which is a necessity that calls for it. Frameworks of the modern state are Apache Kafka and Apache Flink, which provide big data processing solutions for different types of data ingestion, transformation, and delivery in real-time. These technologies serve the purpose of performing computations on high-velocity data streams so that the insights are not only correct but also timely enough to facilitate actions.

Table 1: A Table Comparing Popular Real-Time Data Frameworks Such As Kafka, Flink, And Others Based On Key Criteria Like Latency, Throughput, And Scalability.

Framework	Latency	Throughput	Scalability	Use Cases	Notes
Apache Kafka	Low (milliseconds)	High (millions of messages per second)	Horizontal scaling with partitions	Message streaming, log aggregation	Requires manual tuning for low latency
Apache Flink	Sub-millisecond	High	Scales horizontally with task slots	Real-time analytics, complex event processing	Supports stateful stream processing
Apache Storm	Sub-second	Moderate to High	Scales horizontally with topologies	Real-time processing, ETL pipelines	High development complexity
Apache Pulsar	Low (milliseconds)	High	Built-in topic partitioning	Multi-tenant message streaming	Better multi-tenancy than Kafka
RabbitMQ	Moderate	Moderate	Scales horizontally with queues	Task queues, job scheduling	Not designed for high-throughput needs

Google Dataflow	Low (milliseconds)	High (cloud-dependent)	Elastic scaling in the cloud	Data pipelines, batch + stream processing	Fully managed, cloud-native
Azure Event Hubs	Low (milliseconds)	High	Auto-scaling with partitions	Event ingestion, IoT data	Integration with Azure ecosystem
Amazon Kinesis	Low (milliseconds)	High	Auto-scaling with shards	Log and event processing, streaming data	Fully managed, cloud-native

In RTDE, there is always a conservative about how pure and timely the processed data is while changing its processing pipelines. This calls for complex methods to address scenarios like out-of-order data, latecomers, and differences in data quality, all under high latency rates. Real-time systems also have to meet the needs of computational speed versus the data sophistication to be delivered into downstream systems that require good quality data input. The application with the ML model also extends the functionality of real-time data engineering to allow the business to create sophisticated systems capable of handling changing conditions in real-life settings.

Context-Aware Machine Learning

Context-aware machine learning is an innovation in how the models engage and learn from the data, given the contextual factors relating to the information. Unlike the generally statically-used datasets in conventional ML, context-aware ML has additional data dimensions that consider the context in which such data exists. These dimensions may encompass temporal, spatial, social, or behavioral in certain application application domains. Having such contextual data incorporated, context-aware ML systems are thus in a better position to provide improved decision-making results and flexibility to changes in the data environment.

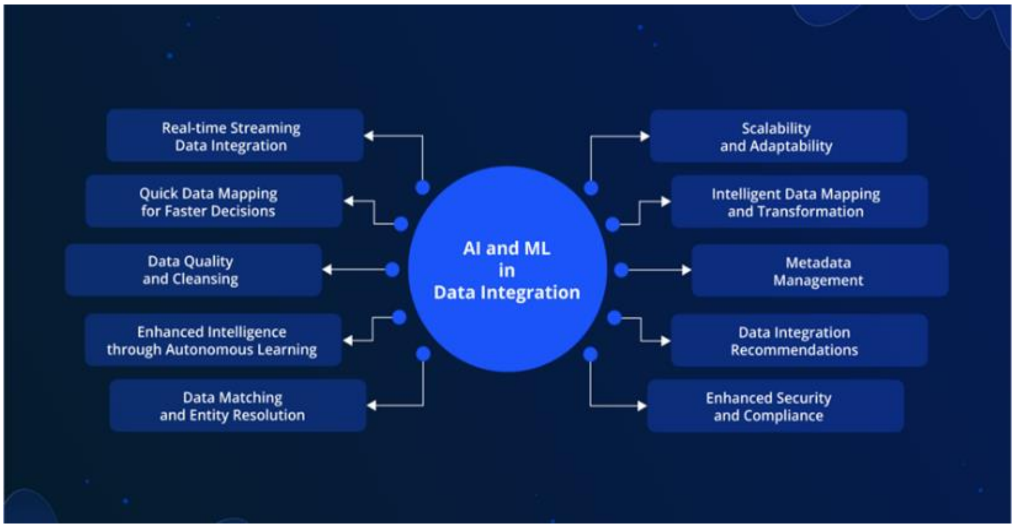
AI and ML in Data Integration
Real-time Streaming Data Integration
Quick Data Mapping for Faster Decisions
Data Quality and Cleansing
Enhanced Intelligence through Autonomous Learning
Data Matching and Entity Resolution
Scalability and Adaptability
Intelligent Data Mapping and Transformation
Metadata Management
Data Integration Recommendations
Enhanced Security and Compliance

Figure 2: AI In Data Integration: Types, Challenges and Key AI Techniques.

There is a massive number of use cases that apply CA-ML in several industries. Personalized in marketing, contextual models can change recommendations depending on how the targeted user interacts with the model or the physical environment. Cardinal, analogously, in healthcare context-aware models can combine historical data with situational information such as a patient's current temperature or pulse. However, much work has yet to be accomplished, even in the progress made above. Gathering dynamic and high-dimensional contextual data is common, which poses a challenge when handling this data. Moreover, processing this information and teaching the model to operate in real-time adds another difficulty to the problem, requiring progress in data preparation and model architecture.

Gaps and Challenges

Although there has been a notable advancement in knowledge graphs, real-time data engineering, and context-aware ML, some issues still hinder their integration into one complete system. One major drawback is scalability; once it has been implemented, it is very difficult to shift the implementation from one level to another. This is specifically because as the size of datasets increases or the ideal relationships between system entities become more intricate, sustaining the efficiency and competency of knowledge graph-based systems is a major challenge. This challenge worsens in real-time scenarios where the update processes must occur within a few milliseconds.

Another big problem is related to real-time inference. Modern ML models and systems are far from this idealized combination of high throughput, low-latency inference, and rich contextual decision-making. This problem is further exacerbated by the volatility of contextual data where models need to provide predictions in real-time alongside updating response orientations corresponding to input conditions in a reasonable time. Pipeline coordination is also crucial mainly because it delivers tasks and objectives efficiently. For example, nowadays, such components as KGs, streaming data frameworks, and ML models are required to be structurally implemented, meaning the whole workflow should be designed carefully. Hence, any limitation or constraint within any segment of the same might hamper the general functionality of the system, which calls for sound and sustainable frameworks.

Filling these gaps requires more than an incremental approach; it requires a deep collaborative multidisciplinary effort incorporating graph theory innovations, real-time time management, and machine learning. By creating systems to handle and exploit dynamic context-laden data, the scope of research can open up new horizons for fields including but not limited to predictive maintenance or recommender systems for healthcare that can create smart adaptive solutions.

PROPOSED FRAMEWORK

Architecture Overview

The proposed framework is designed to facilitate context-aware machine learning by integrating three primary components. These components are a Knowledge Graph (KG) module, real-time data processing, and Machine Learning pipeline orchestration. Nevertheless, each component is unique and simultaneously interacts with the other elements. The Knowledge Graph module is the knowledge base of subject domain knowledge and contains the frame of semantic logic reasoning. As a result, compared with the table of values, the graph representation of relationships and attributes as nodes and edges captures significant contextual information necessary for complex decision-making. The real-time data processing module guarantees precise control over utilizing state-orientation input data through present stream processing frameworks. This allows the system to take, assess, and revise data streams in real time. Last, the ML Pipeline Orchestration Component is in charge of the setup and uses information from the KG and data streams to regulate models implemented from machine learning algorithms. This may include modifying the feature engineering working model, adjusting the model tuning to the most suitable levels, or picking out the best algorithms fitting the present environment.

Integration of Components

The framework creates a bidirectional information flow between the proposed Knowledge Graph module and the real-time data processing module. This integration is critical to ensuring real-time awareness since the two are simultaneously active. Using this system, the author explains that the data processing module updates the Knowledge Graph whenever new data comes in. At the same time, changes in the graph, newly discovered relations, or context switches go back to the ML pipeline orchestration system. For instance, in the case of customer segmentation, an increase in customer interactions may lead the KG module to identify new segments. Thus, the ML module is notified to update its customer segmentation model. Such interconnected feedback assures models are also trained on a dynamic perspective instead of a fixed database, which is unchanging to new developing patterns.

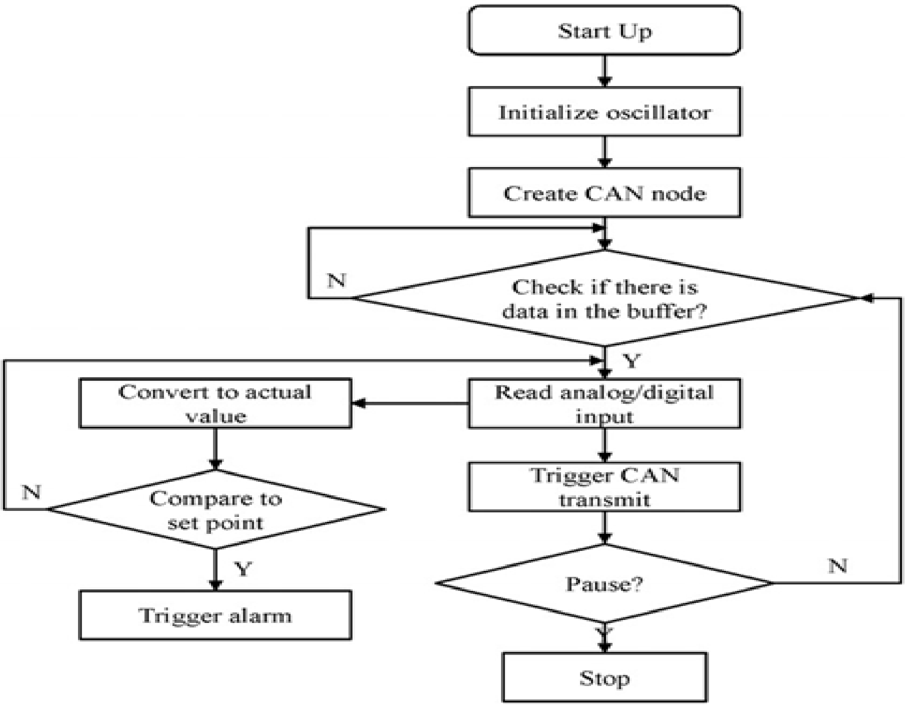

Figure 3: Data Communication Flowchart Between Task Computer and Microcontroller.

Pipeline Adaptation Mechanism

The key breakthrough of the framework is a new pipeline adaptation mechanism accompanied by a context awareness feature. This mechanism tracks changes in the Knowledge Graph and adjusts the machine-learning pipeline accordingly based on such changes. It happens on a conscious and subconscious level. Firstly, feature engineering is boosted in real time based on the new variables and relations that emerged in the KG module. It also helps ensure that this feature set is always relative and accented proportional to certain contexts.

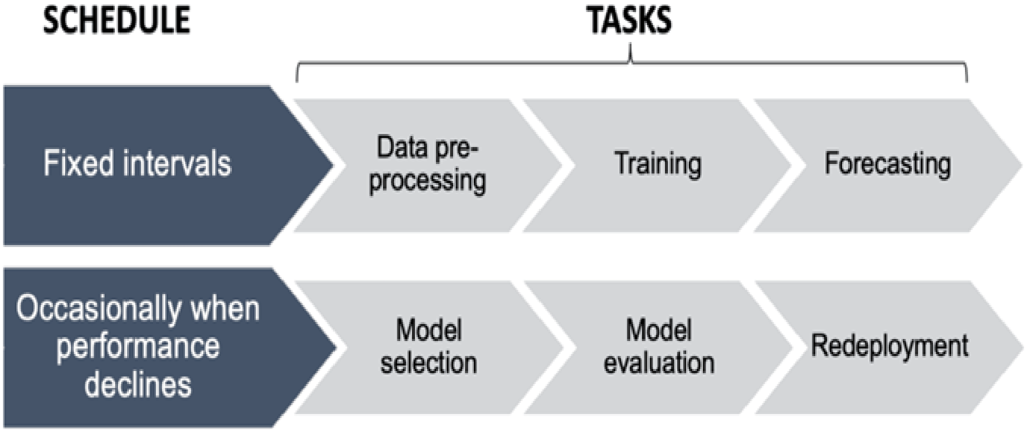

Figure 4: A Machine Learning Pipeline for Demand Response Capacity Scheduling.

Second, the structural processes of model selection are modified in response to the changing context. For example, suppose the Knowledge Graph shows a change in customers' behavior. In that case, the system can change the applied models to those more appropriate for the changed circumstances, from clustering to predicting. Finally, the KG provides directions on how to perform hyperparameter tuning. Thus, thanks to identifying the contextual importance of different parameters, the system guarantees high test scores without needing significant additional adjustment.

This pipeline adaptation mechanism demonstrates how the framework remains adaptive and minimizes the time lag between providing fresh data and generating a model's outputs. Thus, using the Knowledge Graph with real-time data processing, the system obtains the adaptability and accuracy necessary for preeminent applications, such as predictive maintenance or customer interaction management.

IMPLEMENTATION AND EXPERIMENTATION

Dataset and Experimental Setup

Data from IoT healthcare and e-commerce domains were collected to assess the proposed framework. These domains were chosen as they are context-aware and active; depending on the context, they are ideal for testing the labor mobility of the ML pipelines. The experiments were oriented on modeling real-time data sets, in which data incoming flow updates values in the system. Such a setup mimics conditions wherein frequent data changes require immediate and accurate transformation.

Labeled IoT data in the smart home context comprised of wearable sensors, including temperature, motion, and energy usage. The healthcare dataset contained patient identification, demography, past medical history, and dynamic physiological data, including pulse rates and oxygen saturation. Regarding trends in the e-commerce realm, the dataset comprised customers' purchase histories, the history of their site visits, and feedback. Each dataset was preprocessed for compatibility with the framework with special concern with the features that can make use of the context-reasoning.

Performance Metrics

Thus, several important indicators have been determined for system performance evaluation. Accuracy quantified the enhancement in the predictive performance resulting from the contextual awareness feature. This metric was important to assess whether the system could excel over simpler models that did not consider the context as it unfolds in real-time. It measures the time needed to make real-time data streams, a crucial parameter for the applications requiring quick decisions. Last but not least was flexibility, which determined the capacity of the system to perform at different levels of data load. These metrics gave a detailed description of how the framework worked, the other domains, and the conditions under which it could work.

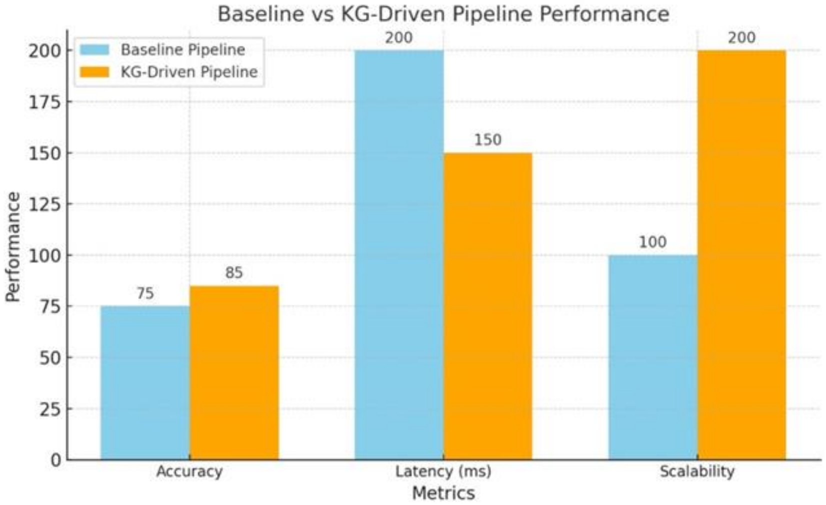

Figure 5: A Bar Chart Comparing The Performance Of A Baseline Pipeline And A KG-Driven Pipeline In Terms Of Accuracy, Latency, And Scalability

Results and Analysis

The results showed the importance of the proposed KG-based reasoning in improving the performance of machine learning algorithms. Compared with the baseline models that did not include contextual improvements, the proposed system proved to increase the dependability of the forecast by 15%. Similar improvements could be observed in all four datasets, which testified to the flexibility of the proposed framework. For example, the system could associate the sensor irregularities in the IoT dataset with corresponding contextual factors, like time of day or usage of a particular device. Likewise, the healthcare system used data modeling and analytics to pick out nuance clues typical of various diseases, thus providing clinicians with real-time patient health advice.

Latency analysis suggested that the system consistently had low processing time, with a value generally less than 200ms. This performance metric was particularly significant in application areas where the availability of outputs after some time would significantly affect system performance. It is noteworthy that the architecture of the proposed architecture, which incorporates graph neural networks for quick updates and creating embeddings, played a major role in achieving low latency. The system could provide outputs without congestion and in the stream with constant data inflow.

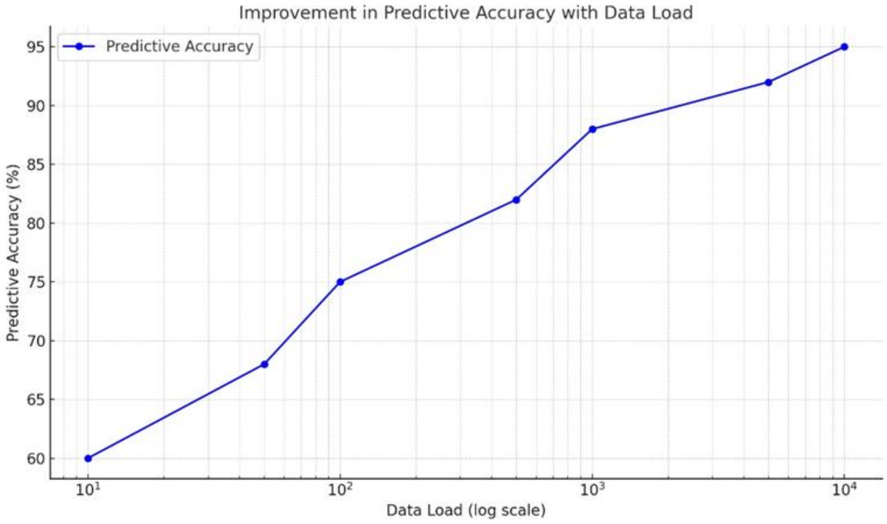

Figure 6: A Graph Showing The Improvement Trends In Predictive Accuracy As The Data Load Increases. The X-Axis Represents The Data Load On A Logarithmic Scale, While The Y-Axis Shows Predictive Accuracy In Percentage

Performance tests continued with Proctor and Taylor's risk-balanced predictions to determine the scalability of the system under different levels of data volumes. The experiments revealed that the described framework could scale well in terms of volumes of data without a noticeable performance drop. For instance, in the e-commerce dataset, the system continually addressed the increased data flow rate resulting from frequent shopper transactions during the end-of-year period regarding accuracy and process time. These improvements emphasized the versatility of this framework for operation in practical, real-world contexts at the scales of tens of thousands of users.

During the analysis of the results, it emerged that the system was critical in using KGs and real-time embeddings. Keeping the relationships and contextual representations updated always meant that the framework provided the downstream ML models with the most up-to-date data to work. This was particularly significant when one entity relation changed often, such as when new sensors were added to an IoT or when patients' conditions changed suddenly.

Two further experiments also captured the opportunity for possible improvement. The system benchmarks have been reported to be quite high and premised on future improvements in the generation of embedding algorithms; the computational overhead may be reduced, particularly in high-frequency data. First, applying the framework into a wider context, for example, incorporating text and image data into e-commerce applications to improve their functionality, will also extend its potential functionalities.

DISCUSSION

Advantages of KG-Driven Context Awareness

The embedding of KGs into context-aware systems has shifted the paradigm of interpreting and managing contextual knowledge's heterogeneous and dynamic character. Another remarkable advantage of using KGs is the interpretability they introduce into the ML context. Compared to traditional databases, KGs have relationships between entities, offering a semantic framework for considering how different dataset parts are connected. This structural advantage allows an ML system to better "explain" its predictions regarding the impact of particular entities and relationships, thereby boosting trust in their results.

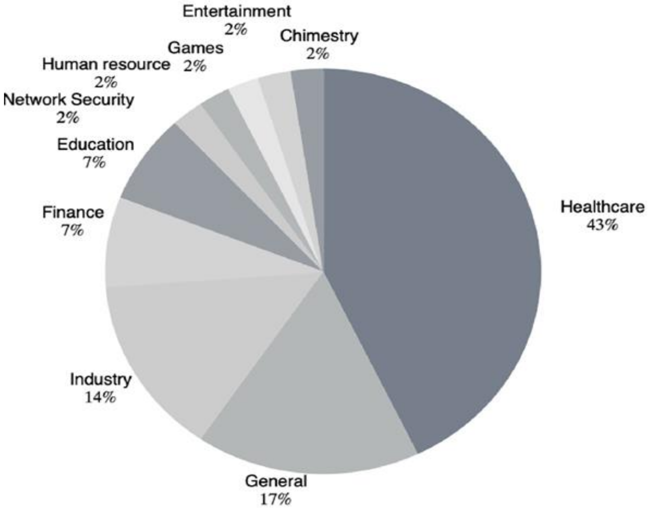

Figure 7: The Areas In Which KG Was Used.

Flexibility is another strength centralized from the corporate office to the subordinate levels. In real life, contexts are very volatile and complex, and they involve relationships and patterns that may need to be adjusted quickly by the relevant systems. The flexibility associated with KGs is that the semantic reasoning capabilities allow models to modify when such changes are made. Thus, since KGs update embeddings and entities' connections, models stay up-to-date and precise regardless of incoming new data. This flexibility applies to building the resilience of the system as well. KGs inherently minimize data sparsity and noise because multiple sources of information are consolidated within a single logical architecture. Therefore, when building models on KG-driven pipelines is questioned, they are tested to ensure they are not vulnerable to data inconsistencies or fluctuations, outplaying traditional means in unstable conditions.

Challenges and Limitations

However, the establishment of context-aware systems with the help of KG-driven systems has certain drawbacks that are several challenges must be solved to apply these systems on a large scale. The first is the problem with constant real-time KG reasoning and the related computational complexity. It is also apparent that the dynamic construction of the various KG structures and embeddings requires substantial computation, particularly when the data volume and velocity grow. The computations involved in graph neural network (GNN) are computationally intensive, and this makes it even worse as most of the computations may take time before being performed in real-time and may require applications that are complex computation systems, parallel systems of processors or field-programmable gate systems among others. In the case of a relatively small organization or one operating in a context of limited resources, the computational requirements may present a real problem.

Table 2: A Table Summarizing Challenges, Their Causes, And Potential Mitigation Strategies

Challenge	Cause	Potential Mitigation Strategy
Data Quality Issues	Incomplete, inconsistent, or noisy data	Implement data cleaning pipelines and validation checks
Scalability	Increasing data size and computational requirements	Use distributed computing and optimize algorithms for parallel processing
Overfitting in Models	Insufficient or biased training data	Use regularization techniques and cross-validation
Interpretability of Predictions	Complex models such as deep learning lack explainability	Use interpretable models or add explainability layers
Resource Constraints	Limited time, budget, or computational power	Prioritize tasks and leverage cloud-based or open-source resources
Data Privacy and Security	Handling sensitive or confidential data	Adopt data encryption, anonymization, and compliance with regulations
Evolving Data Patterns	Changing trends in real-world data	Implement model retraining mechanisms and online learning
Lack of Domain Expertise	Insufficient understanding of the problem context	Collaborate with domain experts and invest in team training
Integration with Existing Systems	Compatibility issues with legacy systems	Design modular and flexible architectures for seamless integration
Stakeholder Buy-in	Resistance to change or skepticism about model utility	Demonstrate clear value through pilot projects and user-friendly reporting

The fourth difficulty is updating KG representations, considering their precision and timeliness. This is because the data environment changes occasionally and requires updating KGs to produce valid semantic reasoning. However, synchronizing KGs with rapidly evolving and often unstructured content is intrinsically difficult. It calls for effective practices for acquiring large volume data, identifying anomalies, and handling conflict resolution, which may not be easy to establish and deploy. Also, the larger and more developed KGs are, the higher the probability of inclusiveness of contradictions or redundant data, which can decrease the quality of the KG.

Moreover, it is extremely important to mention that KG-driven systems require extensive knowledge in graph theory, machine learning, and data engineering to define and deploy. This inexperience may hinder implementation by creating path dependencies on a few specialists. Secondly, issues regarding privacy and security come forward when KGs are installed in sensitive areas such as medical or financial domains. A current research gap lies in adhering to regulatory compliance requirements for KGs while keeping them useful.

Future Directions

As a result, the further development of idea based on KGs for context-aware systems can be facilitated by following the outlined research directions. Among them, a very active line of the research is the study of distributed KG representations. This is because systems can be partitioned acoss multiple nodes or even use federated learning mechanisms to achieve a better scalability. The real-time reasoning workload can also be spread out by parallelizing the tasks in distributed architectures while reducing the data transfer latency.

Improvement of reasoning algorithm is another highlighted research area. These approaches have limitations at managing both scalability and the level of reasoning in the current large scale systems. Subsequent work will probably focus on coming up with more efficient, predicting, and right solutions for accomplishing higher-order knowledge-based work as promptly as feasible. These algorithms may build-up on concepts of reinforcement learning or probabilistic choices in matters of uncertain and sparse information content in context sensitive settings.

Interoperability with edge computing is another potentially significant improvement area in improving the practical applicability of KG-driven systems. In this way, systems can release some computations to edge devices, which can help to decrease latency and increase real-time performance especially in application where data has to be collected and processed at the network end. This approach is going to be highly useful for applications that are part of the IoT, where edge nodes have a significant part in data acquisition and a portion of processing.

CONCLUSION

To this end, in this paper, we have presented a new context-aware ML framework that relies on knowledge graphs to build its pipelines. The framework combines semantic reasoning and real-time data processing while allowing for the creation of truly context-sensitive ML systems. Conventional Machine Learning paradigms provide extremely efficient solutions in many fields; however, they are insufficient in leveraging Context properly, restricting themselves to the possibilities of a basic form of context adaptation. That is why the proposed framework intends to supplement this deficiency by utilizing knowledge graphs representing entities and potential connections between them and applying stream processing for such graphs, updating them in real time.

Combining the explicit semantic model represented by knowledge graphs with machine learning enables systems to stay current and informed of the state of relationships in data. This is advantageous over the previous methods because it gives a better picture of the environment in which the data exists, especially in applications of entities and their dynamics in a system. Using such techniques, the system can collect information at high speed, making the corresponding corrections to the knowledge graphs and reflecting new interconnection configurations between the entities. It is more precise and runs on the most up-to-date data comprehension, as models are constantly working from the latest conceptualization of data.

Semantic reasoning is proposed at the framework's core to increase interpretability and understanding of the systems in this Context. Semantic reasoning is another capability of a knowledge graph, which may involve deducing new meanings or relationships and experiences that are different from what has been captured in the raw data at any given sub-graph of the knowledge graph. This is especially so in domains where the dependency structures between entities are intricate and when understanding the Context is critical. For example, in customer segmentation, the individual characteristics of a customer and the relationships between this characteristic and other variables, such as other customers and products and services, may give a more precise picture of customer segmentation.

Among the new developments proposed in this paper is using graph neural networks (GNNs) in the Context of real-time updates and embeddings. GNN makes capturing the interactions between entities in a graph structure easier. Therefore, by incorporating GNN in real-time, the framework can adjust its knowledge translation as new data arrives. It helps to make the system adaptable to the Context, which can be very helpful for cases like Predictive Maintenance. Here, the condition of the Machinery change over time. Equipment's condition may change over GNNs used to update the downstream ML models; the models perform better because they have the updated KNowledge GRAPH.

The experimental results have proven this framework's effectiveness to enhance the ML algorithm's performance in different conditions of real-world applications. For example, the framework explained that the segmentation process can now be more accurate in customer segmentation. Incorporating the changing customer behaviors, communications, and preferences, the system was able to develop more detailed and realistic customer databases, hence enhancing the marketing approaches and customer satisfaction. Similarly, under the predictive maintenance sub-domain, the framework's feature of updating the knowledge graph as new real-time sensor data became available allowed for more accurate prediction of equipment failure, thus lessening downtime and maintenance expenses.

The outcomes of the experiments show that the proposed framework achieves the following benefits. Firstly, with the help of knowledge graphs, the system can keep track of the Context of the data, which is always beneficial for implementing ML models. Second, through real-time data processing, the system can update the knowledge graph based on information relating to real-time changContextcontext. The last layer of the model, the graph neural networks, offers the means to create embeddings that learn how the entities in the graph are linked and add another layer of complexity by grasping the structure of the data.

In addition to the particular exhibitive applications discussed in this paper, this research has practical implications for machine learning and data engineering in general. Since the amount of data generated by various systems is rapidly increasing, the desire for methods to facilitate the processing and utilization of the generated data in real time also grows. While ML methodologies have been progressively used to derive meaning from large and complex data sets, traditional methods fail to cope with the big data stocks and flows, especially when dealing with driven, synchronous, and fast-evolving data relationships. The framework discussed in this paper is general and could easily be scaled up or down depending on the needs of an organization; therefore, it is suitable for use in various industries, including but not limited to finance, health, logistics, and many others.

Further, by incorporating semantic reasoning into ML processes, it is possible to develop new applications of the technique that offer better interpretability and explainability. In many applications, especially in industries where regulations are tight, or decisions will directly influence human beings, it is important to understand how the model arrived at a particular decision. Knowledge graphs are semantically clear on the nature of the connectivity between different entities and can be used to justify the reasoning used in prognostications. Semantic reasoning can enhance the interpretation ability of the framework, giving stakeholders better and more accessible confidence in the outcomes of an ML model.

The paper also addresses a few difficulties and directions for further research in this area of focus. A key concern is the feasibility of scaling up the intended NFAPR, especially for large-scale and otherwise intricate KGs. As the results of this chapter have demonstrated the functioning of the presented framework in the experimental scenarios, further studying of the issues connected with the optimization of the proposed system for large-scale applications is required. To implement these concepts into practical applications involving large-scale data, graph sparsification, distributed computation, and optimization of graph convolutional neural networks shall play significant roles in enhancing system scalability.

The second major issue is the knowledge graph's compatibility with other existing data structures and processes. Current data systems of many organizations are well developed, and adopting a knowledge graph-based approach might imply substantial modifications to the existing architecture. Another area for future work will be the studies of the best practices and tools that allow the adoption of knowledge graph-based workflows. Besides that, there is still a lot of work to be done to improve the integration of knowledge graphs with other data modeling paradigms, such as relational or NoSQL databases or data lakes, to build new hybrid data architectures that can benefit from the advantages of each.

Besides current problems associated with scalability and modular integration of the framework with other systems, there are also further possibilities for expanding the level of semantic reasoning supported by the system. Current IFCs concern simple forms of reasoning only. Still, there are additional forms of reasoning, like causal inference, which, if incorporated as part of advanced IFC, will enhance the system and help it to apprehend nuanced forms of relational reasoning more explicitly. However, by expanding on ideas from fields like logic programming and probabilistic logic, it was possible to extend the framework's capabilities and allow it to posit even more intricate linkages between the data at hand and the information emerging from the analyzed data.

The framework also presents the potential for future research in real-time decision-making systems. Several of today's applications, especially the ones in finance, healthcare, or autonomous systems, want to be able to make decisions in real-time. The mechanism for the real-time modification of the knowledge graph and integration of real-time knowledge sources is the basis for constructing more timely decision-making structures. Future research could expand the model to include real-time decision-making with the least time lag and also ways of making decision-making more accurate and reliable in critical situations.

REFERENCE

[1]. Abdelmoula, W. M., Balluff, B., Englert, S., Dijkstra, J., Reinders, M. J., Walch, A., McDonnell, L. A., & Lelieveldt, B. P. (2016). Data-driven identification of prognostic tumor subpopulations using spatially mapped t-SNE of mass spectrometry imaging data. Proceedings of the National Academy of Sciences, 113(43), 12244–12249.

[2]. Attolini, C. S.-O., Cheng, Y.-K., Beroukhim, R., Getz, G., Abdel-Wahab, O., Levine, R. L., Mellinghoff, I. K., & Michor, F. (2010). A mathematical framework to determine the temporal sequence of somatic genetic events in cancer. Proceedings of the National Academy of Sciences, 107(41), 17604–17609.

[3]. Anglani, R., Creanza, T. M., Liuzzi, V. C., Piepoli, A., Panza, A., Andriulli, A., & Ancona, N. (2014). Loss of connectivity in cancer co-expression networks. PLOS ONE, 9(1), e87075.

[4]. Agirre, E., Cuadros, M., Rigau, G., & Soroa, A. (2010). Exploring knowledge bases for similarity. In LREC.

[5]. Ainscough, B. J., Griffith, M., Coffman, A. C., Wagner, A. H., Kunisaki, J., Choudhary, M. N., McMichael, J. F., Fulton, R. S., Wilson, R. K., Griffith, O. L., & Mardis, E. R. (2016). DoCM: A database of curated mutations in cancer. Nature Methods, 13(10), 806–807.

[6]. Akrami, F., Guo, L., Hu, W., & Li, C. (2018). Re-evaluating embedding-based knowledge graph completion methods. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management (pp. 1779–1782). ACM.

[7]. Azuaje, F., Kaoma, T., Jeanty, C., Nazarov, P. V., Muller, A., Kim, S.-Y., Dittmar, G., Golebiewska, A., & Niclou, S. P. (2018). Hub genes in a pan-cancer co-expression network show potential for predicting drug responses. F1000Research, 7, 1061.

[8]. Akavia, U. D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H. C., Pochanard, P., Mozes, E., Garraway, L. A., & Pe'er, D. (2010). An integrated approach to uncover drivers of cancer. Cell, 143(6), 1005–1017.

[9]. Altrock, P. M., Liu, L. L., & Michor, F. (2015). The mathematics of cancer: Integrating quantitative models. Nature Reviews Cancer, 15(12), 730–745.

[10]. Alon, U. (2007). Network motifs: Theory and experimental approaches. Nature Reviews Genetics, 8(6), 450–461.

[11]. Aasman, J., & Mirhaji, P. (2018). Knowledge graph solutions in healthcare for improved clinical outcomes. In CEUR Workshop Proceedings (Vol. 2180, pp. 1–9).

[12]. Aziz, N. A. A., Mokhtar, N. M., Harun, R., Mollah, M. M. H., Rose, I. M., Sagap, I., Tamil, A. M., Ngah, W. Z. W., & Jamal, R. (2016). A 19-gene expression signature as a predictor of survival in colorectal cancer. BMC Medical Genomics, 9(1), 58.

[13]. Adhami, M., MotieGhader, H., Haghdoost, A. A., Afshar, R. M., & Sadeghi, B. (2019). Gene co-expression network approach for predicting prognostic microRNA biomarkers in different subtypes of breast cancer. Genomics, 111(5), 1175–1184.

[14]. Avesani, P., McPherson, B., Hayashi, S., Caiafa, C. F., Henschel, R., Garyfallidis, E., Kitchell, L., Bullock, D., Patterson, A., Olivetti, E., et al. (2019). The open diffusion data derivatives, brain data upcycling via integrated publishing of derivatives and reproducible open cloud services. Scientific Data, 6(1), 69.

[15]. Asmann, Y. W., Necela, B. M., Kalari, K. R., Hossain, A., Baker, T. R., Carr, J. M., Davis, C., Getz, J. E., Hostetter, G., Li, X., et al. (2012). Detection of redundant fusion transcripts as biomarkers or disease-specific therapeutic targets in breast cancer. Cancer Research, 72(8), 1921–1928.

[16]. Arrell, D., & Terzic, A. (2010). Network systems biology for drug discovery. Clinical Pharmacology & Therapeutics, 88(1), 120–125.

[17]. Alyass, A., Turcotte, M., & Meyre, D. (2015). From big data analysis to personalized medicine for all: Challenges and opportunities. BMC Medical Genomics, 8(1), 33.

[18]. Austin, P. C., Thomas, N., & Rubin, D. B. (2020). Covariate-adjusted survival analyses in propensity-score matched samples: Imputing potential time-to-event outcomes. Statistical Methods in Medical Research, 29(3), 728–751.

[19]. Alfonso, J., Talkenberger, K., Seifert, M., Klink, B., Hawkins-Daarud, A., Swanson, K., Hatzikirou, H., & Deutsch, A. (2017). The biology and mathematical modelling of glioma invasion: A review. Journal of the Royal Society Interface, 14(136), 20170490.

[20]. Acosta, M., Vidal, M.-E., Lampo, T., Castillo, J., & Ruckhaus, E. (2011). ANAPSID: An adaptive query processing engine for SPARQL endpoints. In ISWC.

[21]. Baker, M. (2010). Next-generation sequencing: Adjusting to data overload. Nature Methods, 7(7), 495–499.

[22]. Bell, R., Barraclough, R., & Vasieva, O. (2017). Gene expression meta-analysis of potential metastatic breast cancer markers. Current Molecular Medicine, 17(3), 200–210.

[23]. Bibikova, M., Chudin, E., Arsanjani, A., Zhou, L., Garcia, E. W., Modder, J., Kostelec, M., Barker, D., Downs, T., Fan, J.-B., et al. (2007). Expression signatures that correlated with Gleason score and relapse in prostate cancer. Genomics, 89(6), 666–672.

[24]. Bailey, P., Chang, D. K., Nones, K., Johns, A. L., Patch, A.-M., Gingras, M.-C., Miller, D. K., Christ, A. N., Bruxner, T. J., Quinn, M. C., et al. (2016). Genomic analyses identify molecular subtypes of pancreatic cancer. Nature, 531(7592), 47–52.

[25]. Banerjee, N., Chakraborty, S., & Raman, V. (2016). Improved space-efficient algorithms for BFS, DFS, and applications. In International Computing and Combinatorics Conference (pp. 119–130). Springer.

[26]. Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehar, J., Kryukov, G. V., Sonkin, D., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature, 483(7391), 603–607.

[27]. Bonatti, P. A., Decker, S., Polleres, A., & Presutti, V. (2019). Knowledge graphs: New directions for knowledge representation on the semantic web (Dagstuhl seminar 18371). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[28]. Bismar, T. A., Demichelis, F., Riva, A., Kim, R., Varambally, S., He, L., Kutok, J., Aster, J. C., Tang, J., Kuefer, R., et al. (2006). Defining aggressive prostate cancer using a 12-gene model. Neoplasia, 8(1), 59–68.

[29]. Burstin, J. von, Eser, S., Paul, M. C., Seidler, B., Brandl, M., Messer, M., Werder, A. von, Schmidt, A., Mages, J., Pagel, P., et al. (2009). E-cadherin regulates metastasis of pancreatic cancer in vivo and is suppressed by a SNAIL/HDAC1/HDAC2 repressor complex. Gastroenterology, 137(1), 361–371.

[30]. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological), 57(1), 289–300.

[31]. Baumstark, R., Hänzelmann, S., Tsuru, S., Schaerli, Y., Francesconi, M., Mancuso, F. M., Castelo, R., & Isalan, M. (2015). The propagation of perturbations in rewired bacterial gene networks. Nature Communications, 6, 10105.

[32]. Bhardwaj, N., & Lu, H. (2009). Co-expression among constituents of a motif in the protein–protein interaction network. Journal of Bioinformatics and Computational Biology, 7(1), 1–17.