Data-Driven Model Categorization: Advancing Physical Systems Analysis Through Graph Neural Networks

Andrija Grbavac[1], Martin Angerbauer[1], Michael Grill[1] and André Casal Kulzer[2]

[1]*Research Institute for Automotive Engineering and Powertrain Systems Stuttgart (FKFS), University of Stuttgart, Pfaffenwaldring 12, Stuttgart, Germany*

[2]*Institute of Automotive Engineering (IFS), University of Stuttgart, Stuttgart, Germany*

Keywords: Graph Neural Networks, Physical System Models, Application of AI.

Abstract: Efficiently categorizing physical system models is crucial for data science applications in scientific and engineering realms, facilitating insightful analysis, control, and optimization. While current methods, often relying on Convolutional Neural Networks (CNNs), effectively handle spatial dependencies in image data, they struggle with intricate relationships inherent in physical system models. Our research introduces a novel approach employing Graph Neural Networks (GNNs) to enhance categorization. GNNs excel in modeling complex relational structures, making them apt for analyzing interconnected components within physical systems represented as graphs. Leveraging GNNs, our methodology treats entities as system components and edges as their arrangements, effectively learning and exploiting inherent dependencies and interactions. The proposed GNN-based approach outperforms CNN-based methods across a dataset of 55 physical system models, eliminating limitations observed in CNN approaches. The results underscore GNNs' ability to discern subtle interdependencies and capture non-local patterns, enhancing the accuracy and robustness of model categorization in a data science framework. This research contributes to advancing model categorization, emphasizing the application of data science for understanding and controlling complex physical systems. The innovative use of GNNs opens new avenues for revolutionizing the categorization of intricate physical system models in scientific and engineering domains.

1 INTRODUCTION

In the dynamic landscape of automotive engineering, the design of powertrains has evolved exponentially, marked by the integration of numerous sophisticated concepts to enhance efficiency, performance, and sustainability (Mirzadeh Phirouzabadi et al., 2020). The complexity of modern automotive powertrains has led to many combinations of interconnected subsystems, creating an urgent need for robust modeling techniques to help engineers understand and optimize these complicated systems. Among the various modeling approaches, physical system models have emerged as indispensable tools, providing a comprehensive representation of the underlying dynamics and interactions within physical systems (Wellstead, 1979).

These models capture the nuanced relationships between components and their dynamic behaviors, offering engineers valuable insights for analysis, control, and optimization. As the automotive industry navigates the realms of hybridization, electrification, and advanced control strategies (Boulanger et al., 2011; Conway et al., 2021; Cook et al., 2002), accurate and efficient mastery of these models has become paramount. Throughout this process, various concepts and, therefore, physical system models are examined, leading to an increasing number of models in the database. This paper delves into the significance of categorizing physical system models within the context of automotive powertrain design, emphasizing its pivotal role in addressing the growing intricacies of modern vehicle propulsion systems.

Recognizing the limitations of conventional methodologies, such as Convolutional Neural Networks (CNNs) (Grbavac et al., 2023), in effectively analyzing physical system models, our research propounds an innovative approach. We advocate for the application of Graph Neural Networks (GNNs), which excels in describing, modeling, and analyzing complex relational structures (Hamilton, 2020), to augment the categorization and analysis of these

Grbavac, A., Angerbauer, M., Grill, M. and Kulzer, A.
Data-Driven Model Categorization: Advancing Physical Systems Analysis Through Graph Neural Networks.
DOI: 10.5220/0012747100003756

In *Proceedings of the 13th International Conference on Data Science, Technology and Applications (DATA 2024)*, pages 277-284
ISBN: 978-989-758-707-8; ISSN: 2184-285X

dynamic and interconnected systems. This novel perspective extends the current understanding of physical system models and positions GNNs as promising tools for unraveling the complexities inherent in automotive powertrain designs. Through this exploration, we aim to contribute to the ongoing discourse on advancing engineering methodologies, fostering a deeper understanding of the intricate dynamics shaping the future of automotive propulsion systems.

The objective of this paper is to address the challenges and explore novel methods in categorizing physical system models, a critical task with implications across various domains. In Chapter 2, we delve into the background methods of model categorization, emphasizing the significance of this process and highlighting existing challenges faced in accurately categorizing diverse models. We introduce the concept of graph representation as a promising approach for modeling physical systems, setting the stage for our subsequent exploration. Chapter 3 presents our novel methodology, leveraging GNNs for model categorization. We provide an overview of our approach, detailing how GNNs offer a unique and effective strategy for this task. In Chapter 4, we undertake a comprehensive performance evaluation, beginning with a description of our dataset and the network parameters used. We then present our results, analyzing trends in test accuracy achieved through our proposed methodology. Finally, in Chapter 5, we discuss the outcomes and implications of our study. Our approach not only outperforms traditional CNN methods but also exhibits increased capacity in handling diverse categories of models. Moreover, our method shows promise for enabling node-level analysis and demonstrates improved time performance compared to existing techniques. Through this structured investigation, we aim to advance the field of model categorization and contribute to more effective approaches for understanding and analyzing complex physical systems.

2 BACKGROUND

2.1 Significance of Categorizing Models

Categorizing physical system models is imperative for harnessing the wealth of data stored within structured, hierarchical formats like Extensible Markup Language (XML) as in this paper. These structures contain information about the architecture of the models by linking parts together. The parts contain physical properties and, therefore, specific behavior in the system (Wellstead, 1979). This holds significant importance in unlocking the wealth of knowledge stored within databases, enabling engineers to leverage past experiences and insights effectively. In engineering and simulation domains, where the landscape is rich with diverse models and simulations, the ability to categorize and retrieve relevant models efficiently is paramount. However, this process often proves to be time-consuming and labor-intensive, hindering engineers from fully capitalizing on the wealth of knowledge accumulated from past projects.

One of the challenges engineers face is the arduous task of searching for similar simulation models amidst vast databases. Without a systematic categorization framework in place, engineers must sift through an extensive array of models, consuming valuable time and resources. Engineers on an expert level benefit in general from their long term experience, whereas beginners rely on input to work efficiently and learn the physical systems fast.As significant fluctuations occur and the demand for new career entrants increases in response to employee shortages (Akyazi et al., 2020), the imperative for implementing such tools intensifies. The lack of efficiency not only impedes productivity but also limits the ability to derive meaningful insights and solutions from past projects. As a result, valuable knowledge and experience remain underutilized, preventing organizations from maximizing the return on investment in simulation and modeling efforts.

Regarding the machine learning approach, data scientists encounter the challenge of dealing with small numbers of physical system models that contain specific concepts or architectures. Unlike other machine learning applications where training sets are often abundant, the availability of models containing particular architectures or configurations may be limited in engineering domains. In some cases, data scientists may only have access to one or two models that encompass a specific architecture, posing significant challenges for training and validation purposes.

The scarcity of relevant training data compounds the difficulty of developing accurate and robust categorization models, as traditional machine learning algorithms as proposed in (Grbavac et al., 2023) may struggle to generalize effectively with limited samples (Raudys et al., 1991).

In summary, the significance of categorizing models lies in its ability to streamline knowledge discovery, enhance productivity, and facilitate informed decision-making in engineering and simulation domains. By overcoming the challenges associated with database searchability and data scarcity, engineers can harness the full potential of model data, driving innovation and advancement in their respective fields.

2.2 Existing Challenges in Model Categorization

Despite the abundance of information encoded in XML structures, much of it remains untapped, representing a reservoir of valuable insights waiting to be unlocked. Efficiently categorizing models based on this data becomes crucial for comprehensive model analysis and leveraging metadata for various applications, such as search engines and predictive modeling.

Presently, there exists a prevalent approach that utilizes CNNs to categorize these models (Grbavac et al., 2023). However, this method encounters limitations when confronted with models that exhibit multiple branches. If parts of a model, that are characteristic for a specific category, are spread over multiple branches, the pre-processing of the approach using CNNs, by design, struggles to capture the relationships within a model. This leads to incomplete categorization and compromised analysis.

One alternative approach proposes enhancing the CNN methodology by dissecting each pathway within the model, from its inception to termination nodes. By scrutinizing individual paths and calculating cross-correlations along each path, it becomes possible to identify overlaps and therefore branches within the model. However, this approach introduces computational challenges, notably the time-consuming nature of cross-correlation calculations, which exhibit a computational complexity of $O(n*m)$, where n represents the number of different paths and m denotes their average length (Hale, 2006). Despite its potential, this method's efficiency is hindered by its computational demands, necessitating further exploration of alternative techniques for scalable model categorization.

2.3 Introduction to Graph Representation of Physical System Models

As introduced in (Wellstead, 1979), physical system models are intricate constructs that abstract real-world systems into interconnected components, such as mechanical, thermal, fluid, magnetic, electrical elements, or equations of state, mathematical operations and physical properties. Each component, or object, within these models possesses distinct properties and behaviors, representing various physical, chemical, and mathematical elements.

Conceptually, these models can be viewed as graphs, where nodes represent individual components, and edges denote the connections or interactions between them. As properties flow between interconnected components, the graph evolves dynamically, reflecting the propagation of information and the exchange of attributes within the system. Through this interconnected network of nodes and edges, physical system models encapsulate the complex relationships and dependencies inherent in real-world systems.

By representing physical system models as graphs, we can leverage graph-based methodologies to analyze and categorize these models effectively. Graphs provide a natural abstraction for capturing the structural and functional aspects of physical systems, enabling us to explore their intricate interconnections and emergent behaviors in a systematic manner (Hamilton, 2020).

3 THE NOVEL APPROACH

3.1 Graph Neural Networks in Model Categorization

GNNs offer a powerful framework for analyzing and categorizing graph-structured data, making them well-suited for addressing the complexities inherent in physical system models. Unlike traditional machine learning approaches that operate on fixed-dimensional data representations, GNNs can directly operate on graph-structured data, preserving the spatial and relational information encoded within the graph (Hamilton, 2020).

GNNs excel in learning representations of nodes within a graph by aggregating information from their neighboring nodes iteratively. Through a process known as message passing, GNNs propagate information through the graph, updating node representations based on both local and global context. This enables GNNs to capture the hierarchical structure and dependencies present in physical system models, allowing them to discern intricate patterns and relationships that may be overlooked by conventional methods. A more detailled overview can be found in (Hamilton, 2020).

By leveraging the rich structural information embedded within graphs, the novel approach examines if GNNs can categorize physical system models based on their inherent properties and interconnections, and facilitate more accurate and robust categorization outcomes.

3.2 Methodology Overview

The proposed methodology for categorizing physical system models can be divided in three key steps: data extraction and pre-processing, model building, and training and validation:

1. **Extraction of Data from XML Files and Labelling.**
   The simulation model files often store much information, including parameters, setups, or object definitions. This information has to be filtered. The extraction process in this methodology involves parsing the XML files to retrieve parts with their specific IDs and types and the links between them. A part is an element such as pipes, compressors, actuators, or valves as depicted in Figure 1. These extracted components and the links between them serve as the basis for constructing the graph representation of the physical system model. On top, each model is labeled with its corresponding multiclass categories. Pairing these two leads to a fundamental dataset to train the neural network.

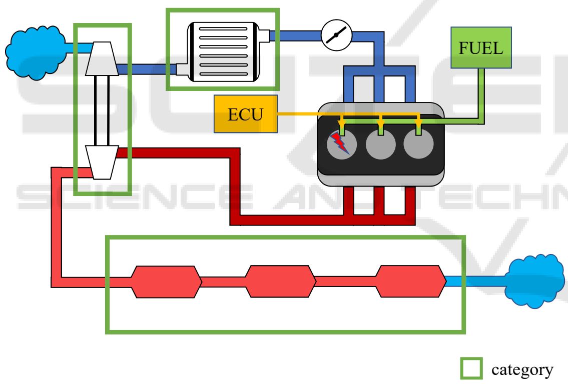

Figure 1: Example of a physical system model of a combustion engine including its periphery. Parts are connected with each other via links and form a graph of physical entities. The categories of the model (framed in green) contain a subgraph of the model (Grbavac et al., 2023).

2. **Building a Graph Representation.**
   Using the extracted parts as nodes and the links as connections, a graph representation of the physical system model is constructed. Each node in the graph corresponds to a specific part of the model, while the edges represent the connections between these parts. The nodes can have various features, with template IDs being utilized in this methodology to represent object types or behaviors inherent in the simulation model.
3. **Building a GNN Architecture.**
   As shown in Figure 2 the GNN architecture (Gilmer et al., 2017; Chollet et al., 2024) consists of several key components designed to effectively extract features from the graph representation of the physical system model and perform categorization:
   (a) **Message Passing.**
   Message passing is performed with a specified number of steps, involving information exchange between nodes within a given n-hop-neighborhood. This iterative process allows nodes to aggregate information from their neighboring nodes, capturing local and global context within the graph structure.
   (b) **Readout.**
   Following message passing, a readout mechanism is applied to generate graph-level representations. This process involves partitioning the graph into subgraphs. These then follow padding, multi-head attention with a specified number of attention heads, projection using sequential dense layers, and two-layer normalization, followed by average pooling to obtain a consolidated representation of the graph.
   (c) **Dense Layer for Categorization.**
   A dense layer is introduced to further refine the learned features and prepare them for categorization. This layer facilitates the extraction of higher-level features that are relevant for distinguishing between different categories of physical system models.
   (d) **Dense Output for Categories.**
   Finally, a dense output layer with softmax activation is employed to produce probability distributions over predefined categories. This layer maps the learned features to the respective categories, enabling the model to predict the most likely category for a given physical system model.
4. **Model Evaluation and Validation.**
   (a) **Training and Validation.**
   The constructed GNN architecture is trained using a dataset comprising labeled physical system models. During training, the model learns to map input graph representations to their corresponding categories through iterative optimization of model parameters.
   (b) **Cross-Validation.**
   To assess the robustness and generalization capability of the categorization model, K-Fold cross-validation is employed. This involves partitioning the dataset into training sets mul-

tiple times, iteratively training and evaluating the model on different subsets of the data.

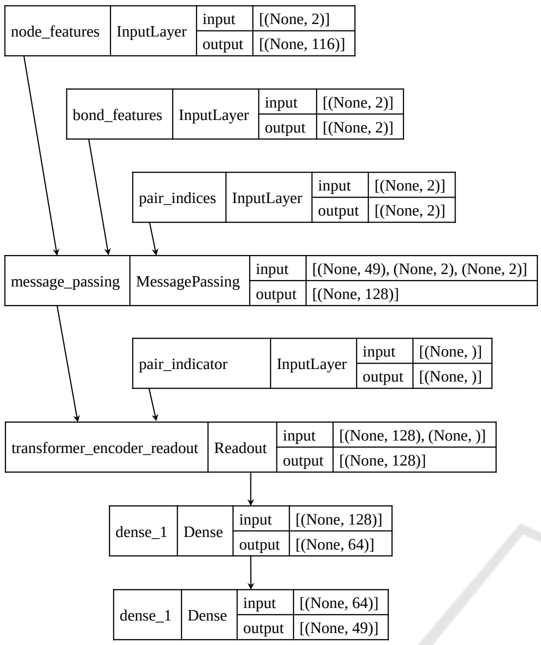

Figure 2: Scheme of the network architecture used in this study.

4 PERFORMANCE EVALUATION

4.1 Dataset Description

The dataset used in this study primarily consists of system models of automotive powertrain systems, with an overrepresented amount of internal combustion engines. This emphasis on combustion engines stems from their historical dominance in the automotive industry, resulting in a larger availability of simulation models for such powertrains compared to alternative propulsion systems.

The labeling of the dataset focuses on subsystems that are characteristic for powertrain concepts. One example are turbochargers, which serve as distinguishing factors between naturally aspirated engines and turbocharged engines. The latter category can be further subdivided into single- and two-staged turbocharged engines, reflecting variations in turbocharging configurations commonly found in automotive applications.

Furthermore, the labels in the dataset exhibit variability in size and variance, leading to differences in complexity and detail across different categories. This variation reflects the diverse nature of powertrain configurations and allows an evaluation of the categorization performance across a spectrum of complexities.

The dataset comprises 55 GT-POWER simulation models (Gamma Technologies, 2024), each containing approximately 50 to 300 objects representing various components and subsystems of automotive powertrains. These models capture different aspects of powertrain design and operation, ranging from basic configurations to more intricate systems. Based on these concepts the models are labelled by 49 categories, ranging from 1 object up to subgraphs of 143 objects.

4.2 Network Parameters

The experimental setup involves exploring a range of hyperparameters to optimize the performance of the categorization model. The following parameters are varied to assess their impact on model performance:

- **Message Passing Steps.** The number of message steps during message passing ranges from 1 to 16, allowing for the exploration of different levels of information propagation through the graph.
- **Number of Attention Heads.** The number of attention heads in the transformer encoder readout ranges from 1 to 15, enabling the model to capture varying degrees of attention and focus on different aspects of the graph representation.
- **Number of Dense Units.** The number of units in the dense layer for categorization ranges from 32 to 1024, influencing the complexity and expressiveness of the learned feature representations.

Additionally, to ensure robust evaluation and validation of the categorization model, the dataset is split into 10 folds using a K-Fold cross-validation strategy. A binary cross-entropy loss function is chosen and optimized by an Adaptive Momentum (Adam) optimizer as it acchieves great results with graph classification (Gilmer et al., 2017).

4.3 Results

The categorization model demonstrates excellent performance overall, achieving an accuracy of over 90% on the test dataset. This high level of accuracy highlights the model's ability to effectively categorize physical system models, even when faced with complex categories and variations in model characteristics.

Figure 3 shows the average test accuracy for all 49 classes. These are shown for the parameter combination with the best overall results. Regarding the categorization ability for each class, the best results are achieved with 4 message steps, 10 attention heads, and 32 dense units. This configuration achieves very good accuracy for each class, with the exception of class no. 4 (Single Staged Turbocharged Engine), which exhibits slightly lower accuracy. However, all other classes achieve at least 83% accuracy, demonstrating the model's strong categorization capabilities across diverse categories.

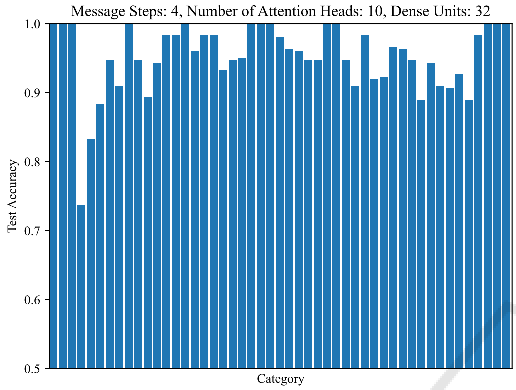

Figure 3: Average accuracy of the 49 categories after a training with 4 message steps, 10 attention heads and 32 dense units over 10 folds. With an overall accuracy of 98% this parameter combination forms the best overall results.

Trends of the Test Accuracy.

Several notable trends emerge from the experimental results, providing insights into the impact of hyperparameters on categorization performance:

- When using 32 neurons in the categorizing dense layer, the median of the test accuracy remains consistently high regardless of the number of message steps and attention heads (see Figure 4a-c). However, as shown in (see Figure 4d-f and 4g-i), with an increasing number of neurons in the dense layer, the median tends to decrease as the number of message steps increases.
- As shown in see Figure 4a-c, the number of attention heads shows a small influence on the mean accuracy when the network has 32 dense units in the categorization layer, with hardly better results observed with a higher number of attention heads. However, this trend is not consistent when using a higher number of dense units, indicating a more complex relationship between these parameters (see Figure 4g-i).
- Overall, the number of attention heads has no significant influence on accuracy, suggesting that other factors play a more dominant role in determining categorization performance.
- However, the variance of accuracy over the k folds increases with a higher number of dense units in the categorizing layers (Figure 4a, d and g) and a higher number of message steps Figure 4c, f and i), indicating increased sensitivity to variations in these hyperparameters.

5 DISCUSSION

The results obtained from the experimental evaluation of the novel GNN based approach for categorizing physical system models reveal significant improvements over the traditional CNN approach (Grbavac et al., 2023). Several key points highlight the superiority of the GNNs approach and suggest potential areas for further enhancement and analysis.

5.1 Superior Performance over CNN Approach

The performance of the novel GNN-based approach significantly surpasses that of the CNN approach. The CNN approach, which extracts paths from start to end nodes and categorizes them individually, is inherently limited in its ability to handle categories that span multiple branches (see Figure 5). In contrast, the GNN approach demonstrates superior adaptability and robustness in categorizing models with complex structural relationships that extend across multiple branches. This highlights the inherent advantages of the GNN approach in handling graph data with intricate interconnections.

5.2 Increased Capacity for Handling Categories

The GNN approach exhibits a much higher capacity for handling categories compared to the CNN approach. While the categories in the CNNs paper are limited to a maximum of 15 objects, the novel approach extends this limitation significantly, accommodating categories consisting of up to 143 objects. This expansion not only increases the breadth of categories that can be considered but also enhances the complexity of these categories, allowing for more nuanced and detailed categorization of physical system models.

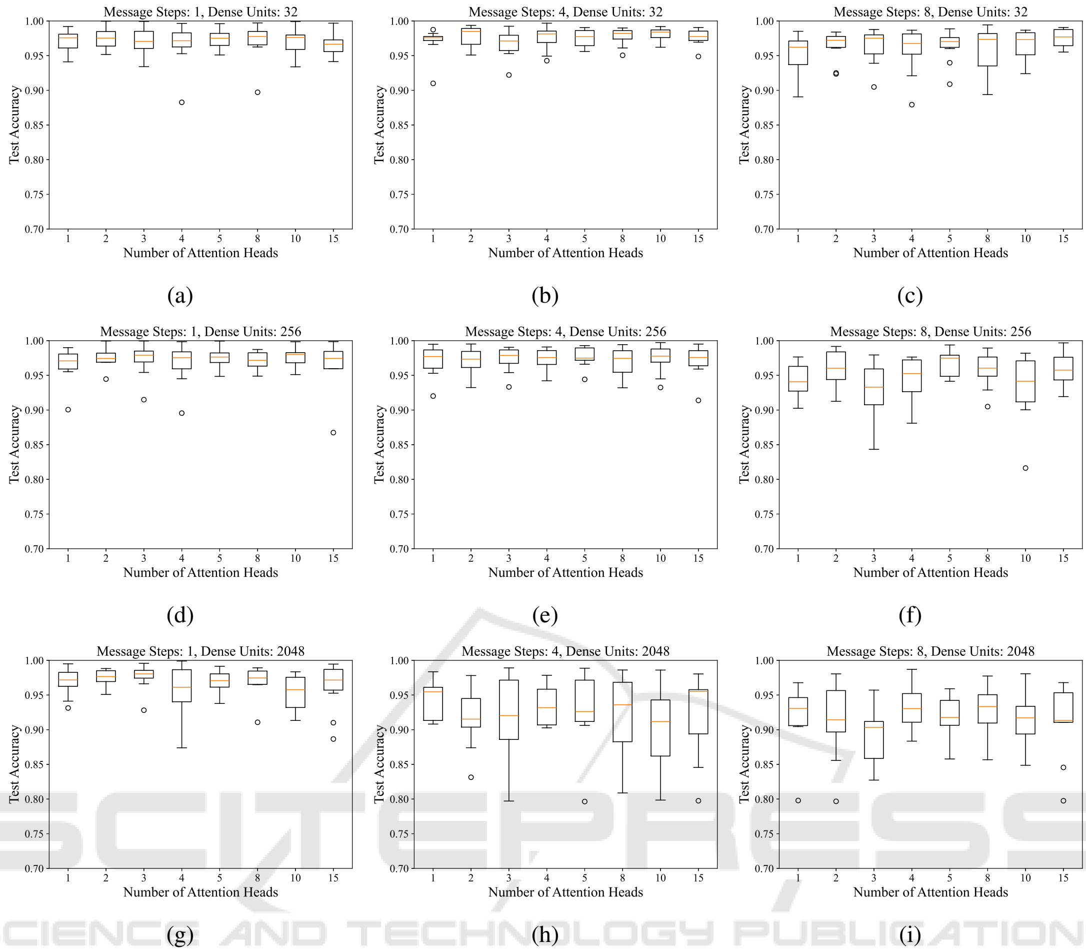

Figure 4: Investigation of parameter variations on the test accuracy of the GNN model. The boxplots represent the K-Fold distribution for different parameter combinations. The box plot highlights the median, quartiles, and outliers, providing insights into the variability and central tendencies of the test accuracy.

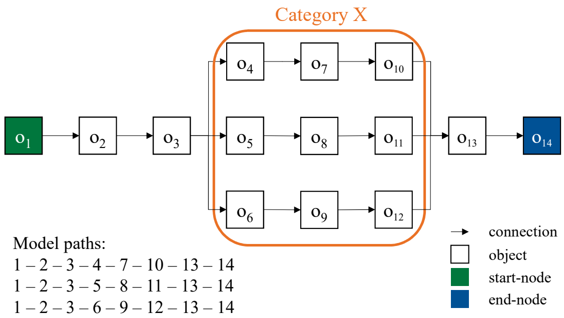

Figure 5: Example of a branched graph and the extraction of the model paths based on the CNN approach. To categorize all elements of the category X, the paths have to be categorized based on a characteristic length in each path. Otherwise an intersection has to be calculated via cross-entropy (Grbavac et al., 2023).

5.3 Potential for Node-Level Analysis

To further analyze the differences in accuracies resulting from different parameter combinations, the novel approach can be extended to enable node-level analysis. By examining embeddings on a node level and comparing them with node-level labeling, the performance of the GNN can be assessed in greater detail. This extension would provide insights into how effectively the GNN model captures and utilizes information at the level of individual nodes, shedding light on its performance beyond the graph level.

5.4 Improved Time Performance

In addition to its superior categorization performance, the novel GNN approach also outperforms the CNN approach in terms of time efficiency. The CNN approach's preprocessing phase is significantly more time-consuming due to its large recursive search through the graph and the training process involving a substantial number of extracted paths. In contrast, the GNN approach achieves comparable categorization accuracy with approximately ten times faster processing times, making it a more practical and efficient solution for real-world applications.

6 CONCLUSIONS

Tthe present study has demonstrates the effectiveness of a novel GNN based approach for categorizing physical system models, particularly focusing on automotive powertrain systems. Through rigorous experimentation and analysis, several key findings have emerged, highlighting the significant advancements achieved by the proposed methodology.

Firstly, the GNN approach shows superior performance compared to traditional CNN methods, particularly in handling complex graph structures with branched pathways. While the CNN approach struggles with categorizing models that span multiple branches, the GNN approach, leveraging message passing, exhibits remarkable adaptability and robustness in capturing the intricate interconnections within the graph.

Furthermore, the GNN approach indicates an increased capacity for handling a wider range of categories, including those with higher complexity and variance. By extending the limitations imposed by previous CNN-based methods, the novel approach enables more refined categorization of physical system models, thereby enhancing the depth of analysis and insights derived from the categorization process.

Moreover, the potential for node-level analysis presents exciting opportunities for further refinement and optimization of the GNN approach. By examining embeddings at the node-level it can provide valuable insights into its effectiveness in capturing information at a granular level.

Finally, the improved time efficiency of the GNN approach gives a significant practical advantage over traditional CNN methods. With processing times approximately ten times faster than CNN-based approaches, the GNN approach offers a more efficient and scalable solution for real-world applications.

In summary, the findings from this study underscore the promising potential of GNN-based methodologies for advancing the field of system model categorization. As we continue to refine and optimize these approaches, we can expect further advancements in our ability to analyze and understand complex physical systems, ultimately driving innovation and progress in engineering and simulation research.

REFERENCES

Akyazi, T., Alvarez, I., Alberdi, E., Oyarbide-Zubillaga, A., Goti, A., and Bayon, F. (2020). Skills needs of the civil engineering sector in the european union countries: Current situation and future trends. *Applied Sciences*, 10(20):7226.

Boulanger, A. G., Chu, A. C., Maxx, S., and Waltz, D. L. (2011). Vehicle electrification: Status and issues. *Proceedings of the IEEE*, 99(6):1116–1138.

Chollet, F. et al. (2024). Keras. https://keras.io.

Conway, G., Joshi, A., Leach, F., García, A., and Senecal, P. K. (2021). A review of current and future powertrain technologies and trends in 2020. *Transportation Engineering*, 5:100080.

Cook, J., Sun, J., and Grizzle, J. (2002). Opportunities in automotive powertrain control applications. In *Proceedings of the International Conference on Control Applications*, volume 1, pages xlii–xlli vol.1.

Gamma Technologies (2024). GT-SUITE.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). Neural message passing for quantum chemistry.

Grbavac, A., Angerbauer, M., Grill, M., Itzen, D., Milojevic, S., Hagenbucher, T., and Kulzer, A. (2023). Categorizing simulation models using convolutional neural networks. Technical report, SAE Technical Paper.

Hale, D. (2006). An efficient method for computing local cross-correlations of multi-dimensional signals. *CWP Report*, 656:282.

Hamilton, W. L. (2020). *Graph representation learning*. Morgan & Claypool Publishers.

Mirzadeh Phirouzabadi, A., Savage, D., Blackmore, K., and Juniper, J. (2020). The evolution of dynamic interactions between the knowledge development of powertrain systems. *Transport Policy*, 93:1–16.

Raudys, S. J., Jain, A. K., et al. (1991). Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on pattern analysis and machine intelligence*, 13(3):252–264.

Wellstead, P. E. (1979). *Introduction to physical system modelling*, volume 4. Academic Press London.

ACRONYMS

Adam Adaptive Momentum.

CNN Convolutional Neural Network.

GNN Graph Neural Network.

XML Extensible Markup Language.