Practical perspectives on machine learning interpretability for public transport demand forecasting

Gergő Galiger[1,2,*], Max von Münster[1,*], Márk Nyerges[1], Patrik Tóth[1], Péter Kovács[2]

***Abstract*—As the data generated in public transport systems is growing in complexity and quantity, data-driven artificial intelligence models are increasingly used to optimize operations, improve passenger experience, and support decision-making. However, the adoption of machine learning (ML) in this domain is impeded by the lack of interpretability of these black-box techniques. While often highly accurate, these models struggle to gain the trust of transport authorities and stakeholders due to their low transparency. This paper explores the role of interpretable ML in public transport demand forecasting based on a real-world dataset from the transport authority of Budapest. In particular, we perform a fair comparison of the inherently interpretable NeuralProphet time series model, with the post-hoc interpretation of XGBoost using SHAP on data from the metro line M4 of the city. Our results demonstrate that XGBoost outperforms NeuralProphet in terms of forecasting accuracy. Additionally, we show that the insights from the post-hoc SHAP interpretation aligns with both the seasonality components of NeuralProphet and domain knowledge from transportation experts. These results support the adequacy of using SHAP to interpret accurate models for predicting public transport demand, with implications in the broader domain of timeseries forecasting. Source code is available at: github.com/galigergergo/PTDemComp.**

***Index Terms*—Public transport modeling, passenger demand forecasting, explainable artificial intelligence, post-hoc interpretability, SHapley Additive exPlanations.**

I. Introduction

Reliable transport demand forecasting is essential for the design, operation, and evaluation of transport systems [1]. In the public transportation sector, transport planners and policy-makers depend on accurate forecasts to plan services, manage congestion, allocate resources, and design long-term infrastructure investments [2]. As urban populations grow and mobility patterns become increasingly dynamic, especially in the face of changing travel behaviors and disruptions such as pandemics or the growing shift toward home-based work, the need for forecasting methods that are both accurate and understandable has become increasingly critical [2]–[4].

Historically, transport demand models have predominantly been based on detailed behavioral and simulation-based approaches [1], [5]. These models have traditionally emphasized interpretability, giving planners control over different factors, such as service levels, prices, or socioeconomic variables in modeling demand [6]. Later, with the increasing availability of granular operational and temporal data, time series forecasting models based on data-driven ML approaches emerged as an alternative [7]. The main advantage of these tools was that they could handle complex usage patterns without requiring detailed behavioral specifications, thus leading to more accurate forecasts and less human intervention [8]. However, these models also lost their interpretability as a result of their added complexities [9]–[11].

This shift towards data-based approaches raises concerns about the transparency of these tools in the context of transport planning and service provision. The goal of this paper is to explore these concerns by comparing two different forecasting approaches in a practical public transport setting: one that is inherently interpretable and another that requires post-hoc interpretation. Specifically, we compare NeuralProphet [12], a time series forecasting model that builds on classical mathematical modeling of seasonality and events, and Extreme Gradient Boosting (XGBoost) [13], a high-performing black-box ML method whose predictions we interpret using SHAP [14]. While both methods have demonstrated success in time series forecasting in the transport domain [15], [16], their utility for transparent transport demand forecasting is less well understood.

Our analysis focuses on urban public transport in Budapest, a city of approximately 1.7 million inhabitants with a dense and multimodal transport network, which includes buses, trams, trolleybuses, metro lines, and suburban railways. Among these, the metro line M4, which began operations in 2014, plays a key role in linking major interchanges and central districts. It features fully automated operations and serves corridors with high demand potential but also varying travel patterns depending on land use and intermodal connections. Here, passenger volumes are estimated using anonymized mobile phone signaling data collected via underground cellular antennas. This service represents diverse demand structures and provides a valuable case study for comparing model performance and transparency under different transport conditions.

As the methods tested in this paper are general-purpose forecasting tools, the insights gained are expected to inform a wide range of use cases within and beyond transportation, where data patterns are characterized by strong seasonal effects. Examples include the planning of micromobility services,

[1]Gergő Galiger, Max von Münster, Márk Nyerges and Patrik Tóth are with the Department of Data Analysis and Modelling, BKK Centre for Budapest Transport, Budapest, Hungary, gergo.galiger@bkk.hu, theodor.maximilian@bkk.hu, marknyerges49@gmail.com, patrik.toth@bkk.hu

[2]Gergő Galiger and Péter Kovács are with the Department of Numerical Analysis, Eötvös Loránd University, Budapest, Hungary, galiger.gergo@inf.elte.hu, kovika@inf.elte.hu

*Gergő Galiger and Max von Münster contributed equally.

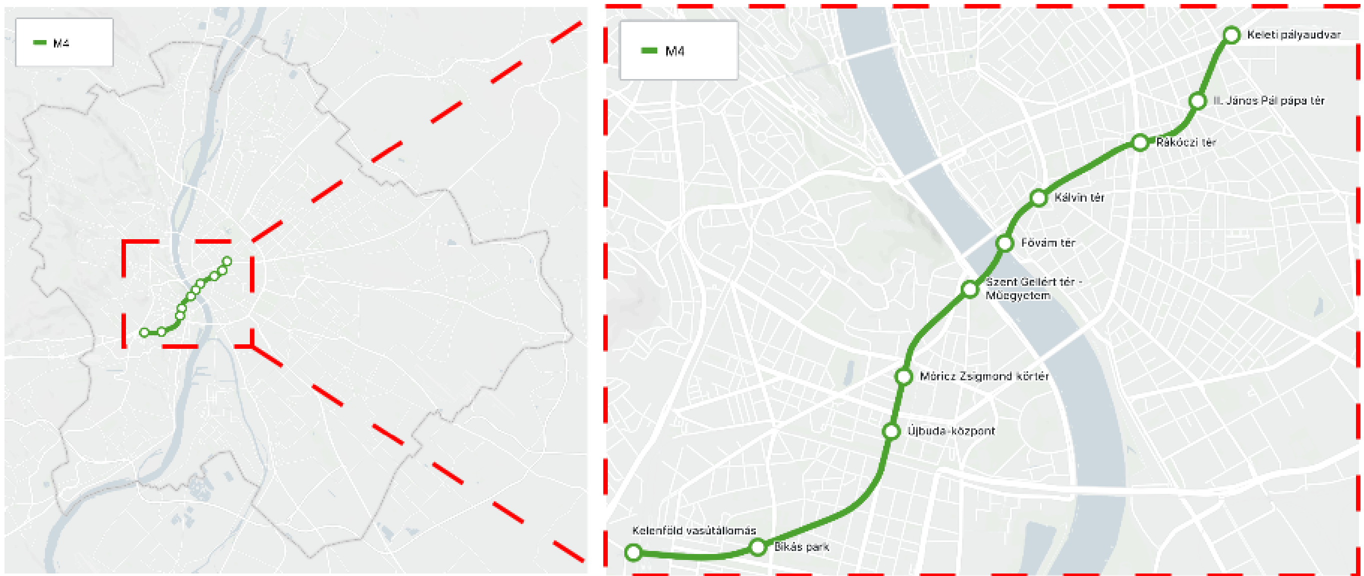

Fig. 1. Route and stations map of the metro line M4 in the city of Budapest.

traffic volume forecasting, and demand-responsive transit, as well as applications in other infrastructure sectors such as electricity load forecasting and environmental monitoring. Ultimately, our contribution is intended to guide the selection of forecasting tools that balance predictive performance with transparency, an increasingly important consideration as transport agencies seek to build trust in data-driven planning.

II. Methodology

The presented study is performed using passenger numbers from the metro line M4 of Budapest. Being a core component of the city's public transport network, this metro line is primarily used for daily commuting and is not operational during night hours. The data used in this work for public transport demand prediction contains hourly aggregated passenger numbers of the metro line M4 sourced for a one-year period in 2024. Passenger numbers are estimated using underground cellular antennas, this process being based on an algorithmic calculation prone to errors. This results in missing data points with passenger numbers only available for a total of 216 days from 5:00 to 24:00 in the final dataset used for this analysis. The route maps of the metro line M4 is shown in Figure 1, along with its respective stations serving as stops.

In the following, we present the methodology used to provide a controlled and fair comparison between the performance and interpretability of the NeuralProphet and XGBoost models on the presented dataset. As a preliminary step, we perform a series of data preprocessing steps on our dataset as detailed in the following parts of this section. Then, we train both models for public transport demand forecasting, which is then followed by two evaluation steps. Next, we compare the prediction performance of the two models with ground truth data to evaluate their ability to accurately predict transport demand. Finally, we interpret both models to gain insights on the importance of different data features on the prediction, comparing the obtained insights with expert knowledge of the ground truth data. We use SHAP to interpret the black-box XGBoost model and perform both evaluation steps on unseen data that were not included in the training step.

We define the task of public transport demand forecasting as the prediction of the hourly number of boarding passengers at a given station in a given direction. We focus on short-term hourly prediction to allow for an analysis of daily trends in the data, while only considering boarding passengers. Moreover, we limit the analysis to a single station and one direction of travel to control for spatial heterogeneity as preliminary exploratory analysis indicated that temporal trends across stations and directions were relatively consistent. The final station and direction were chosen to represent the strongest specific trends in the dataset – Kálvin tér towards Kelenföld vasútállomás, as shown in Figure 1. Additionally, this simplification ensured that observed differences in model performance could be more confidently attributed to modeling choices rather than geographic or directional variability.

To align with NeuralProphet's time-aware architecture and ensure fairness in comparison, we included the 4 most recent hourly passenger counts (lagged features) into the tabular training data for XGBoost, along with the following temporal features: hour of day, day of week, day of year, and a continuous numerical representation of time to model gradual drifts and trends. Moreover, an additional regressor referring to educational periods was also included in the training dataset of both models as preliminary exploratory analysis showed the relative importance of this period on traffic patterns of this services.

To evaluate model generalization and temporal robustness, we used a dual-level data splitting approach. First, a one-week continuous test window was reserved for final model evaluation and accuracy visualization. This period was selected to include representative hourly, weekday and weekend travel behavior. Then, the remaining dataset was split into training, validation, and testing subsets using a 7:1:2 ratio, ensuring

TABLE I
PREDICTION PERFORMANCE VALUES FOR THE ENTIRE TEST SUBSET OF THE M4 DATASET. SCORES ARE AVERAGED OVER ALL 6 HOURS OF THE PREDICTION WINDOW.

Model	NeuralProphet	XGBoost
MAE	230.2803	**91.0753**
RMSE	363.2115	**144.7094**
R-squared	0.7896	**0.9600**

that the one-week test window was included in the test set. This setup provided sufficient data for model fitting while enabling robust validation-based hyperparameter optimization and performance evaluation.

Both models were configured to produce multi-step forecasts of passenger counts over a 6-hour prediction window. The 4 lagged features were provided for each forecast window, corresponding to a standard auto-regressive setup. To allow for a comparable interpretation analysis, NeuralProphet was trained with conditional daily seasonality modeled separately for weekdays and weekends, as well as conditional weekly seasonality based on the educational periods, with the number of Fourier terms $k = 3$ in both cases. On the other hand, XGBoost was trained with a *gbtree* booster type to enforce the black-box nature of the model. Other model-specific and global parameters (e.g. learning rate) were further optimized for 50 trials using the Optuna framework. The root mean-squared error (RMSE) was minimized while training both models.

III. RESULTS AND DISCUSSION

A. Prediction performance

To compare the prediction performance of the two approaches, we evaluate both trained models on the test subset of our dataset. To provide a quantitative performance comparison, we use the mean average error (MAE), RMSE and R-squared scores for the entire 6-hour prediction window. Table I shows the results of this comparison for models on unseen test data. These results clearly show the superiority of XGBoost for accurately predicting transport demand, as compared to NeuralProphet. The differences in metric scores are particularly high in the case of the M4 metro data, which are due to the regularity of this algorithmically calculated dataset. XGBoost can fit the variance of this regular dataset particularly well, with an R-squared score very close to 1.

Figure 2 shows a qualitative comparison of the predictive performance of both models compared to ground truth values on the one-week continuous test window from the dataset. Rush-hour specific daily trends can be observed when looking at the weekdays in the M4 data, which both models are able to recognize. Similarly, the one-peak weekend-specific trends are also fitted by both models. The lower quantitative performance of NeuralProphet can also be partially explained by its inability to represent the sharp switches to zero values during the night hours and the variable afternoon peaks during weekdays, which is a result of its rigorous model-based design, specifically the smoothness of the periodic seasonality components.

B. Interpretability

In this section, we compare both models from the lens of interpretability. Our aim here is to uncover whether the inner workings of the models align with expert knowledge on the traffic patterns characteristic to the metro line M4. A further question concerns the adequacy of using SHAP for the post-hoc interpretation of XGBoost. Section III-A demonstrated the clear superiority in prediction performance of XGBoost for public transport demand forecasting. However, it is still unclear whether the insights provided by the post-hoc interpretation of this black-box model about the way a prediction is produced are comparable with the components of NeuralProphet, or with expert knowledge.

In this experiment, we leverage the specific daily patterns characteristic to our passenger demand dataset to benchmark whether NeuralProphet and XGBoost model public transport demand in accordance with the transport functions of the M4 metro service. For this, we compare the seasonality components of NeuralProphet with the XGBoost SHAP values for the corresponding temporal features. As the ground truth for the justification of the underlying expert knowledge about the expected seasonality patterns, we use the dataset-wide averages of passenger numbers. The results of this experiment are compiled in Figure 3, where we analyze the daily patterns in our dataset. Note, that the three rows in each subfigure represent different quantities, as described in the caption. Despite this, our following analysis shows that the three perspectives often align, offering complementary views of how each temporal feature affects the passenger demand forecasts of the two models.

First, we interpret both models from the perspective of the hourly patterns of M4, initially looking at Figure 3.a. On weekdays, we observe two clear peaks: a morning commute spike around 08:00, and a stronger afternoon peak around 16:00–17:00, which is typical for the Kálvin tér station, as it is located in an employment-focused zone with low residential occupancy. On weekends, the distribution flattens with a moderate rise from 10:00 to 18:00, suggesting more dispersed leisure travel.

Looking at the SHAP values for the hour feature from the XGBoost model, they closely mirror the observed counts, confirming that the model accurately captures these temporal patterns. Notably, high SHAP values during peak hours on weekdays highlight the model's sensitivity to commuter demand. Since the SHAP values are also in a large order of magnitude, the hour feature is dominant for XGBoost, i.e. the change in the passenger numbers is mostly explained by the change in the hour. Additionally, we observe that the early morning rise in weekend passenger volume (particularly at 06:00 and 07:00) is strongly reflected in the SHAP values for the hour feature. This is because, during the night, lagged passenger values are zero in this dataset, so the model relies more heavily on the hour itself to make accurate predictions. By 10:00, although passenger counts remain high, the SHAP

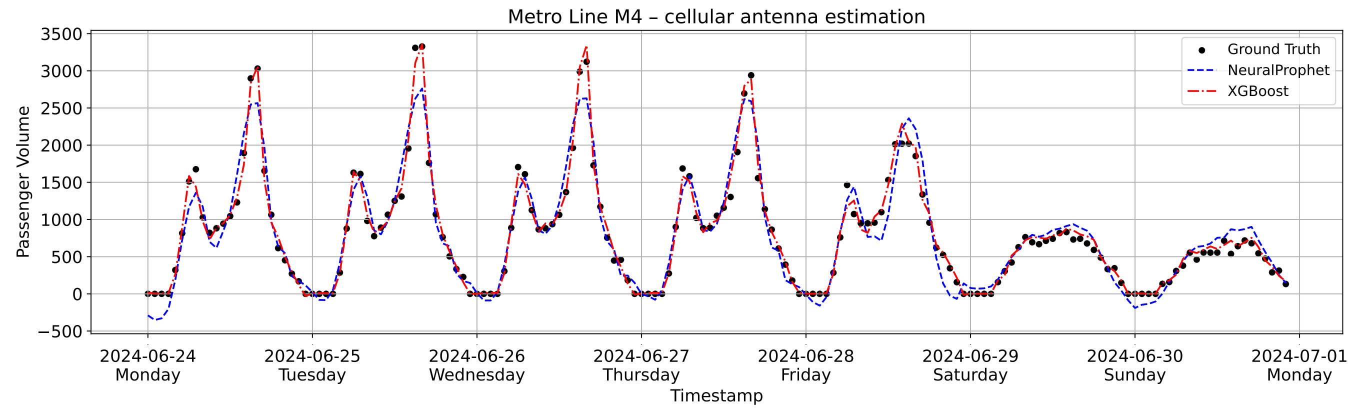

Fig. 2. Comparison of hourly prediction of both models on the one-week continuous test window from the M4 dataset. Results are shown for the first hour of the 6-hour prediction window.

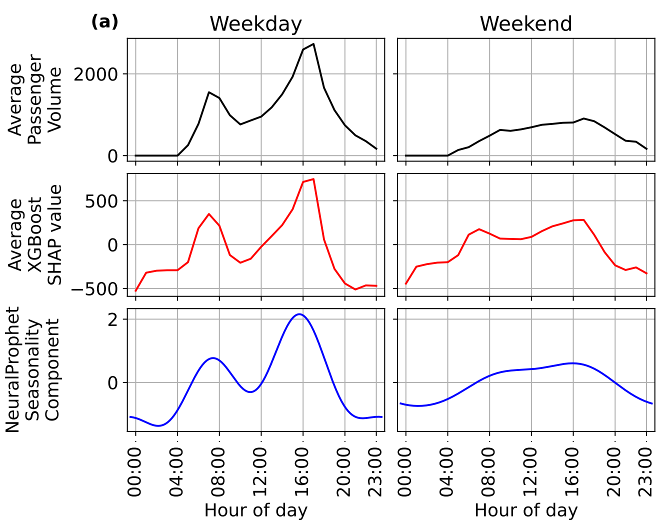

Fig. 3. Comparison of interpretable seasonality components of the two models with ground truth daily patterns over the hours of the day with separate columns for weekdays (left) and weekends (right. The rows contain the interpretable components of the two models and the dataset ground truth as follows: the top rows (black) represent the ground truth as dataset-wide average passenger numbers, the middle rows (red) show average XGBoost SHAP values for the corresponding temporal features, while the bottom rows (blue) illustrate the relevant seasonality component of NeuralProphet.

value for the hour feature drops. At this point, the model can already use the previous 4 hours of higher passenger volumes as lagged inputs, reducing the marginal contribution of the current hour. This indicates that the model shifts its reliance from calendar features to recent observed data as the day progresses.

The NeuralProphet model's periodic components also reflect strong hourly variation on weekdays and a flatter profile on weekends. This consistent alignment with SHAP values and raw averages reinforces both models' capacity to capture real-world behavior. We observe that the NeuralProphet baseline (0%) aligns with early morning (06:00) and evening (19:00) hours on weekdays. During the day, the learned seasonal component rises sharply, peaking at over +200% in the afternoon, indicating a tripling of expected demand compared to the baseline. This strong daily fluctuation is consistent with patterns seen in the raw data. However, unlike XGBoost, NeuralProphet does not reflect the same lag-based behavior. While both models use the previous 4 hours as lagged inputs, XGBoost can conditionally combine lag features with other variables, e.g. it may treat morning lags differently on weekends versus weekdays.

IV. Conclusion

Modern transportation systems increasingly rely on accurate and transparent transport demand forecasting methods for efficient service management and resource allocation. The present work demonstrates that time series modeling, particularly using the NeuralProphet and XGBoost models, is a viable framework for forecasting public transport demand. Specifically, our result show that these models can capture key temporal demand patterns such as daily cycles. However, a critical trade-off emerges here: while interpretable models provide transparency, they may struggle to capture nuanced dynamics in complex settings. In contrast, black-box models (e.g., tree ensembles or neural networks) can model intricate interactions but require post-hoc interpretation methods such as SHAP to provide insights into the model's decisions. This raises the important question of which approach to trust: an accurate black-box model with post-hoc interpretation, or a significantly worse performing but highly interpretable model.

Our findings suggest that SHAP can offer meaningful insights on black-box public transport demand forecasting models, aligning with expert knowledge from the transport domain. Additionally, our experiments on show that the interpretation of seasonality from SHAP aligns with the seasonal components of the rigorous model-based NeuralProphet approach. Both of these results support the adequacy of SHAP for interpreting black-box models for predicting public transport demand, which suggests the same implications for the broader field of timeseries forecasting. In conclusion, the focus of solving such forecasting problems should be the prediction performance, as accurate models allow for a viable post-hoc interpretation via SHAP. As a consequence, we would advocate for using XGBoost over NeuralProphet for public transport demand forecasting based on the present study.

However, our results also highlight two key advantages of model-based approaches: they provide a more nuanced interpretation and expert-based fine-tuning of models, while also ensuring a certain level of robustness. As a result, the hybrid approach of combining black-box ML components with classical mathematical modeling emerges as a promising direction for research in transport demand forecasting. The example of NeuralProphet demonstrates how applying this approach can result in inherently interpretable models with comparable performance to purely ML-based methods. As NeuralProphet only slightly deviates from the classical Prophet model, including other model-based ML techniques, such as deep unfolding [17], or variable projections [18] could further close the performance gap between these approaches.

Acknowledgments

This project was supported by the EKÖP-KDP-24 University Excellence Scholarship Program Cooperative Doctoral Program of the Ministry for Culture and Innovation of Hungary from the source of the National Research, Development and Innovation Fund. This work was supported by the University Excellence Fund of Eötvös Loránd University, Budapest, Hungary (ELTE). This project was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. The authors also thank the Márton Áron College of ELTE and the Collegium Talentum Programme of Hungary for supporting this work.

References

[1] V. A. Profillidis and G. N. Botzoris, *Modeling of transport demand: Analyzing, calculating, and forecasting transport demand*. Elsevier, 2018.

[2] R. Balcombe, R. Mackett, N. Paulley, J. Preston, J. Shires, H. Titheridge, M. Wardman, and P. White, *The demand for public transport: a practical guide*. Transportation Research Laboratory, 2004.

[3] K. Gkiotsalitis and O. Cats, "Public transport planning adaption under the covid-19 pandemic crisis: literature review of research needs and directions," *Transport Reviews*, vol. 41, no. 3, pp. 374–392, 2021.

[4] D. I. M. Gogola, D. M. D. Sitanyiová, L. Černickỳ, and M. Veterník, "New demand patterns for public transport due to demographic change," *Interreg Central Europe*, 2018.

[5] D. McFaddenĆ, "The behavioral science of transportation," *Transport policy*, vol. 14, pp. 269–274, 2007.

[6] D. McFadden and F. Reid, *Aggregate travel demand forecasting from disaggregated behavioral models*. Institute of Transportation and Traffic Engineering, University of California, 1975.

[7] R. P. Masini, M. C. Medeiros, and E. F. Mendes, "Machine learning advances for time series forecasting," *Journal of economic surveys*, vol. 37, no. 1, pp. 76–111, 2023.

[8] A. Tealab, "Time series forecasting using artificial neural networks methodologies: A systematic review," *Future Computing and Informatics Journal*, vol. 3, no. 2, pp. 334–340, 2018.

[9] A. Alwosheel, S. van Cranenburgh, and C. G. Chorus, "Why did you predict that? towards explainable artificial neural networks for travel demand analysis," *Transportation Research Part C: Emerging Technologies*, vol. 128, p. 103143, 2021.

[10] G. Galiger and Z. Bodo, "Explainable patch-level histopathology tissue type detection with bag-of-local-features models and data augmentation," *ACTA Univ. Sapientiae Inform*, vol. 15, pp. 60–80, 2023.

[11] G. Galiger, "Decision validity of heatmap-based explainable neural networks – an explanation first approach," in *Proceedings of The Workshop Biosignal*. Göttingen Research Online, 2024.

[12] O. Triebe, H. Hewamalage, P. Pilyugina, N. Laptev, C. Bergmeir, and R. Rajagopal, "Neuralprophet: Explainable forecasting at scale," *arXiv preprint arXiv:2111.15397*, 2021.

[13] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[14] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[15] K. Zaraket, H. Harb, I. Bennis, A. Jaber, and A. Abouaissa, "Hyperflophet: A neural prophet-based model for traffic flow forecasting in transportation systems," *Simulation Modelling Practice and Theory*, vol. 134, p. 102954, 2024.

[16] T. Stadler, A. Sarkar, and J. Dünnweber, "Bus demand forecasting for rural areas using xgboost and random forest algorithm," in *International Conference on Computer Information Systems and Industrial Management*. Springer, 2021, pp. 442–453.

[17] G. Galiger, N. Azadi, B. Lehner, M. Huemer, and P. Kovács, "Model-based neural networks for thermographic image reconstruction," in *2024 IEEE 3rd Conference on Information Technology and Data Science (CITDS)*. IEEE, 2024, pp. 1–6.

[18] P. Kovács, G. Bognár, C. Huber, and M. Huemer, "Vpnet: Variable projection networks," *International Journal of Neural Systems*, vol. 32, no. 01, p. 2150054, 2022.