Enhancing Ski Safety: Utilizing XAI to Explain LSTM-Based Velocity Prediction from IMU Data

Daniel Homm
Departamento de Informática
Universidad Técnica Federico Santa María
Valparíso, Chile
daniel.homm@usm.cl

I. Introduction

Skiing is a famous sport all over the world. However, one should not ignore the risk of injuries caused by skiing [1]. For example, there were a total of 24,340 injuries encountered over nine years at two ski resorts in California [2]. Furthermore, Knee injuries are the most common in Germany at 23% of all injuries for male skiers and 44% for female skiers [3]. One cause for this kind of injury is an incorrect retention value of the binding on skis [1]. This value is responsible for the correct minimal force to release the ski boots from the binding. However, the proper settings highly depend on the current conditions [1]. This means that depending on a skier's current actions, various forces act on the ski binding, some of which should trigger a release while others should not. Therefore, correct ski binding adjustments are essential for an optimal skiing experience and significantly prevent serious injuries. V. Senner et al. propose using a mechatronic ski binding to better analyze and adapt the threshold given different situations [1], [4], [5]. Such a ski binding can use distinctive features, including receiving direct feedback from various sensors to adapt the trigger threshold according to the skier's behavior [1]. One crucial factor of such an adaptation is the velocity of a skier [1], [6]. With the help of Global Navigation Satellite Systems (GNSS), the current speed can be calculated with high accuracy [7]. However, GNSS has different drawbacks, such as the dependency on a connection to a satellite and the cost of such systems [6]. As one well-suited alternative, one can use Inertial-Measurement-Units (IMUs). Reliable IMUs are just like GNSS, independent of other systems. However, compared to satellite systems, IMUs are more affordable. Furthermore, IMUs are often used to predict various human activities [8], [9]. Thus, such a unit is likely to predict other essential features concerning a skier's behavior as well. Therefore, costs related to needed sensors can be kept to a minimum.

In the context of velocity estimation with IMU recordings, P. Carqueville proposed using a long short-term memory (LSTM) model. The data used to train, validate, and test the model included four days of measurements with over two hours of skiing and over five hours of sensor recordings in total. Throughout this time, one 3D accelerometer, one 3D gyroscope, and a temperature measurement unit for the gyroscope recorded the input features. The recording rate of all measurements was 200Hz. The total amount of data was then split into 60% training, 16% validation, and 24% test data [6]. The original data preprocessing took place by first normalizing train, validation, and test split with the mean and standard deviation of the training and validation data to keep the test data unseen. Furthermore, the time sequences of each split were divided into samples with 250 time steps to limit the temporal context of the trained LSTM cells. The trained LSTM achieved good results with a mean-squared error (MSE) smaller than 3.9km/h for one lift ride and the following ski run [6]. To better understand the model's predictions, Fig. 1 visualizes its results for the test data. Nevertheless, adaptations for safety components in sports need to be based on accessible parts. These allow developers to acknowledge possibilities and limits, as well as to be able to reason about failure scenarios. To be able to interpret complex Machine Learning models, in recent years, the topic of eXplainable Artificial Intelligence (XAI) has become more popular.

The following work is structured as follows. Section II gives an overview of related studies to this research. Subsequently, Section III presents a short introduction to the theoretical foundations for the study. In Section IV, a summary of the used procedure follows. Sections V and VI divide the experiments and their results into the explanations of the model and, second, proposed adaptations for the model. Finally, the Section VII provides a summary conclusion.

II. Related Work

Deep-learning methods are nowadays a standard approach for forecasting time series variables [10]. In general, recurrent neural networks such as LSTM models improved the possibilities of analyzing time series data [11]. This is why multiple authors use LSTMs in several topics apart from sports, such as medicine [12], stock market predictions [13], and different types of velocity predictions [14], [15]. More specifically, machine learning approaches for velocity predictions with IMU-related input features have shown promising results [16], [17]. Furthermore, Feigl et al. show in [16] that recurrent neural networks outperform state-of-the-art predictions for human velocity and orientation predictions. However, one reason for the improved performance of the used models is

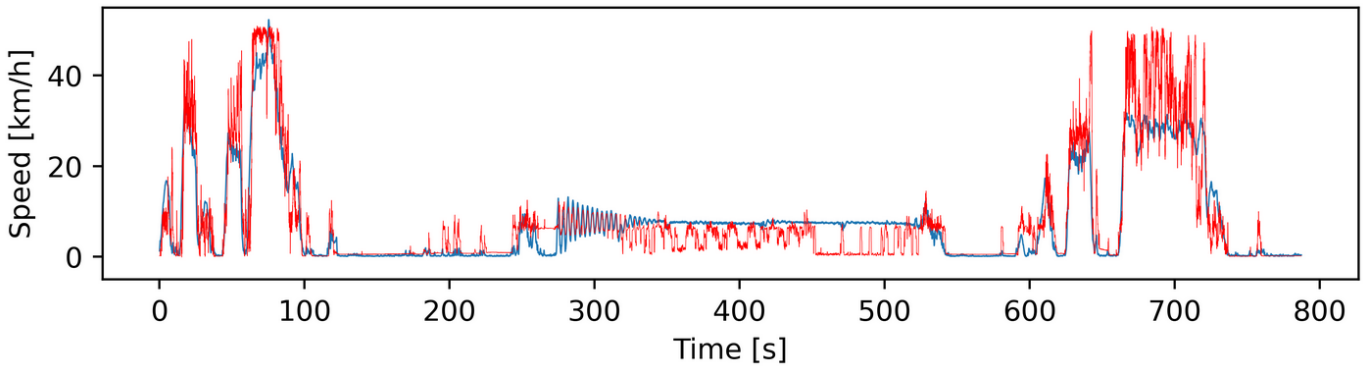

Fig. 1: LSTM prediction results of the original model for the complete test set compared to label velocities. The blue curve holds the real velocities measured by the GNSS sensor, and the red curve is the predicted velocity of the LSTM model [6].

their complexity. Models that are too complex for humans to understand are often called black-box models. Because one cannot understand the process of predictions, a lack of trust lasts [18]. In addition, possible unknown limitations can remain hidden. For this reason, different methods like SHAP [19], LIME [20], and saliency methods [21], [22] were developed to better understand how predictions of a model are derived. Regarding time series analysis, multiple methods like saliency methods and SHAP explanations can help reveal some aspects of network build for time series data. Another general explanation approach is DeepLift [23]. DeepLift is a method for explaining deep neural networks using different mathematical approximations to decrease the run time on large models. For SHAP explanations, approximations like Deep SHAP [19] and TimeSHAP [24] show promising results for explaining recurrent neural networks. TimeSHAP is not only able to capture feature importance but also the contribution of individual events in sequences [24]. In addition, TimeSHAP provides possibilities for local and global explanations of models based on time series data like recurrent neural networks.

III. Explanations via Deep SHAP

SHAP uses the Shapley values' theoretical foundation to explain each input feature's contribution in a machine-learning model. SHAP calculates SHAP values using Formula 1.

$$\phi(i) = \sum_{S \in F \setminus \{i\}} \frac{(|S|!(|F| - |S| - 1)!}{|F|!} \phi_i \qquad (1)$$

In Formula 1, F denotes the power set over all possible features. The variable i refers then to one specific feature, and $\phi_i = [f_{S \cup \{i\}}(X_S \cup \{i\}) - f_S(X_S)]$ is the calculated difference between the prediction that considers the feature i and a prediction without the feature i. Thus, Formula 1 computes the sum of all weighted differences in the outcome predictions. On the one hand, this foundation makes SHAP theoretically robust, resulting in good performance for various types of models [25]. On the other hand, Formula 1 means that one must consider the contribution of a feature in every possible coalition. Thus, contributions for 2^i coalitions, where i is the number of features, must be checked. To tackle this issue, the authors of [19] describe the combination of the DeepLift method [23] and the idea of SHAP [19] to determine the importance of features. DeepLift belongs to the category of gradient-based explanation methods. However, the concept of DeepLift is to avoid expensive derivatives and instead calculate differences to specific reference values and propagate these differences through the network [23]. The authors of [23] have proven that DeepLift is indeed an additive feature attribution method. Therefore, in Deep SHAP, the SHAP values can be approximated using DeepLift with the reference value defined as the expected output of the model $E[x]$. Furthermore, Deep SHAP uses calculated SHAP values for different network components to approximate SHAP values for each feature [19].

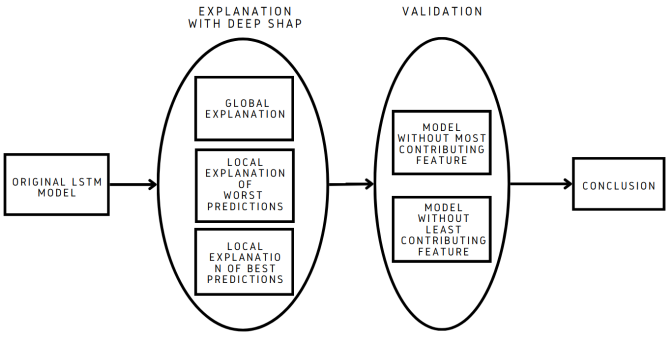

Fig. 2: Overview of the methods used in the present study.

IV. Method

A. Data

I used the original data recordings from [6] to further investigate the proposed model. In addition, I applied the same preprocessing procedure, including normalizing and splitting into time series as described in Section I, chosen by [6]. Thus, I conducted the experiments with the original train and test data.

B. Explanation Method

I investigated how the proposed model from [6] calculates its predictions using the Deep SHAP [19] described in Section III. To explain the model, I considered random subsets of the test predictions and averaged their results to approximate global feature importance. Furthermore, I interpreted subsets of the best and worst predictions separately. To validate the explanations, I trained different model adaptations and evaluated their predictions. Fig. 2 is an overview of the methods used in the present study.

V. Using Deep SHAP to explain LSTM for Velocity Predictions of Skiers

The LSTM model I explain in this research consists of two LSTM layers to enable a good capture of dependencies to past values followed by two dense layers. The features that define the input for the LSTM model are ACC-X, ACC-Y, ACC-Z to capture the information of the accelerometer, and GYRO-X, GYRO-Y, GYRO-Z, as well as GYRO-TEMP for the gyroscope. Here, the X, Y, and Z labels signify the measurements of each corresponding axis of both sensors, and GYRO-TEMP represents the temperature curve of the gyroscope. The test set for the model consists of 616 batches, with each 256 time series. As a first step, an approximation of the global behavior of the model can give a general idea of how a model produces its predictions. A widely used approach to compute such an explanation with locally calculated SHAP values is calculating the average over multiple local explanations. This is why I generated local explanations for 30 randomly chosen batches. Fig. 4 reveals that the influence of each feature rapidly diminishes over time. Therefore, the time steps after t-80 hold more than 75% of the total contribution of the two most prominent features, GYRO-TEMP and ACC-Z, over the complete time frame.

Afterward, I averaged the resulting SHAP values for each of the seven features and 250 lag values separately. The results shown in Fig. 3 show that, on average, the temperature of the gyroscope influences the outcome prediction the most. The averaged SHAP values of the temperature sum up to +1.35 over the complete window size. In second and third place are the z-axis of the accelerometer with +1.11 and the x-axis of the gyroscope with +0.92. The analysis of individual predictions in the test data reveals a consecutive sequence of 51 predictions that show errors above 35 km/h. This sequence also includes the most significant error of 40 km/h. For this worst prediction, Fig. 5 shows the given feature importance for each lag value. The analysis of Fig. 5 still reveals the most significant contributions for one lag value to be at t-1. However, compared to the average prediction, lag values that are further in the past have a higher influence on the worst prediction.

Additionally, as visible in Fig. 6, for the worst prediction, the x-axis of the gyroscope holds the highest contribution. Furthermore, apart from the y- and z-axes of the accelerometer, all other axes of both sensors play a more significant role in the prediction than the temperature. By further analyzing the sequence of bad predictions, the total contributions of all included forecasts show similar rankings. In total, the model predicted 52 outcomes with an error above 35 km/h in the test data. The analysis of the misprediction outside the consecutive sequence, shown in Fig. 7, results in the z-axis of the gyroscope being the most contributing feature for this prediction. The temperature of the gyroscope follows at second. While the influence of the z-axis of the gyroscope plays a significant role in all mispredictions, the temperature's contribution differs widely for the sequential and the individual mispredictions. Furthermore, Fig. 8 reveals that the distribution of the contributions over the past value for the worst prediction outside of the consecutive sequence does not show high contributions for values further in the past. Thus, it is more similar to the global approximation than the contribution for the worst prediction.

In contrast, about 50% of all test predictions imply errors below. Furthermore, around 48% of these predicted are related to labels above 5km/h. 0.1km/h. In the following parts of this investigation, the reference to predictions above or below a threshold of 5km/h always refers to the label of the given prediction fulfilling this statement.

Fig. 9 and 10 show the average results of 30 randomly chosen good predictions above 5km/h. It is worth mentioning that Fig. 9 also reveals a strong influence for earlier values between t-210 and t-170. Moreover, Fig. 10 shows that for these predictions, the x-axis of the gyroscope contributes the most. The z-axis of the accelerometer and the temperature of the gyroscope make the second and third-highest contributions, respectively. In contrast, results for 20 random samples of the best predictions below 5km/h vary vastly from the contributions for predictions of higher velocity. In Fig. 11, a small peak at around t-185 is visible. Nevertheless, Fig. 8 also shows that the more recent inputs between t-1 and t-50 contribute the most to these predictions. Even though the top three ranked features in Fig. 12 are the same as in Fig. 10, the order has changed. For the predictions below the threshold, the temperature of the gyroscope plays a much more significant role than for those above.

VI. Model Adaptations

Training adaptions of the original model according to different findings in explanations can help to verify the results. I chose one model that did not contain the feature that contributed the most globally to take a closer look at such an adaptation. Therefore, I trained a model with the same procedure as the original but excluded the temperature feature. Fig. 13 shows that an offset from velocities close to zero without including the temperature becomes visible. This increases the average prediction error for predictions below 5km/h from 1.44km/h to 2.69km/h. However, it is also noticeable that a decrease in the error for velocities above 5km/h arises. Removing the temperature decreased the average error for predictions above 5km/h from 4.81km/h to 3.62km/h.

If the goal is to change the model structure, explanations can also provide valuable insights to safely reduce the input size of

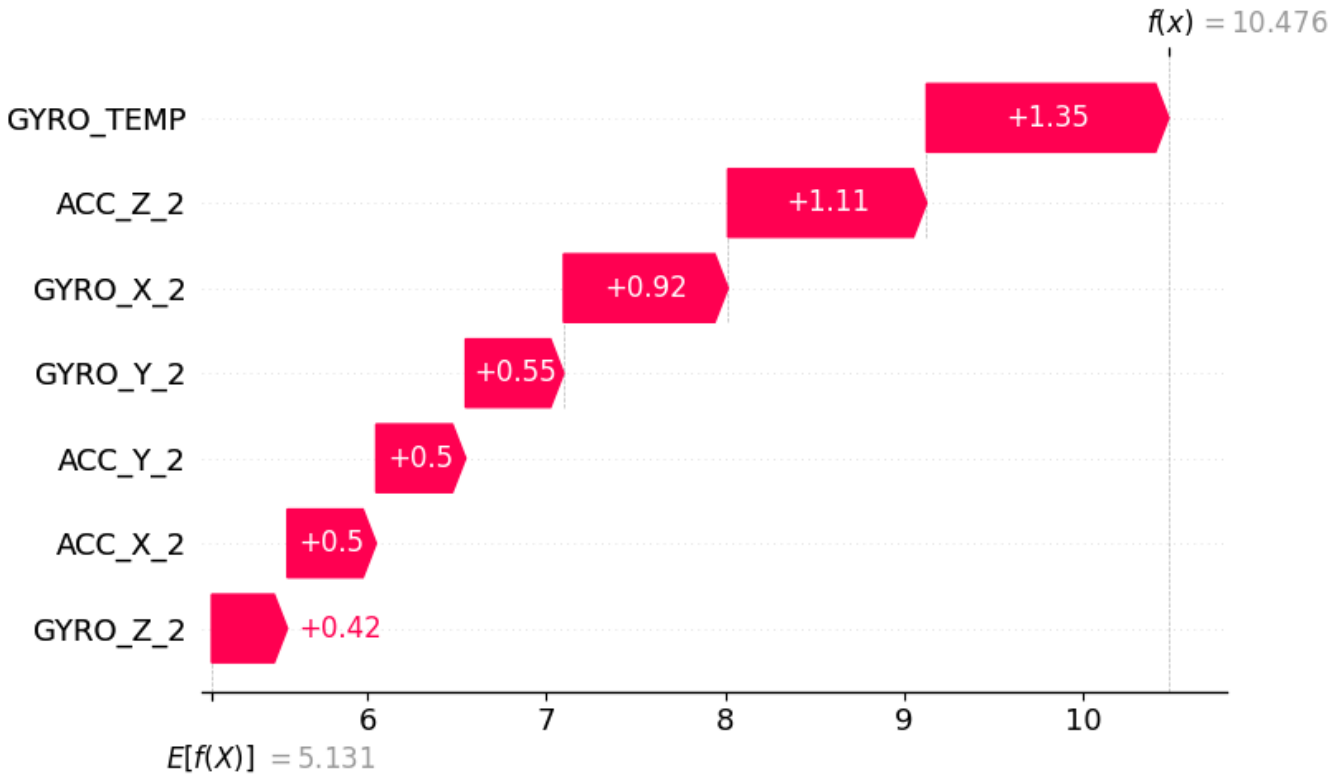

Fig. 3: SHAP waterfall legacy for the mean feature importances of all lag values for 30 randomly chosen batches for each feature.

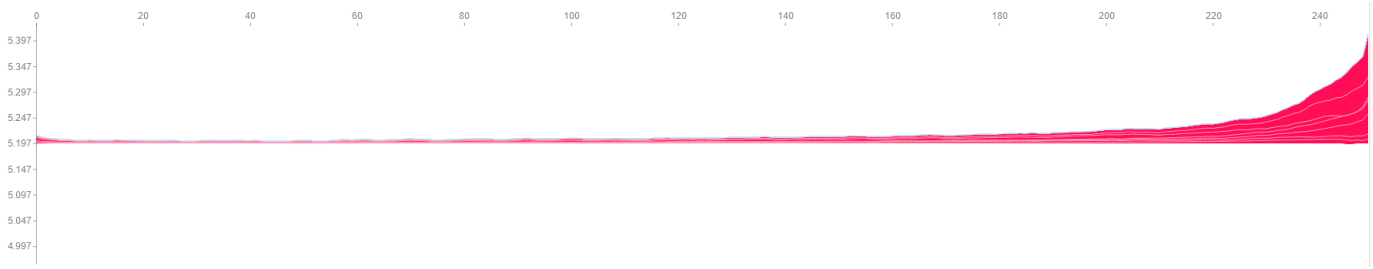

Fig. 4: SHAP force plot, indicating the average global contribution for each lag value t-w. The x-axis is labeled with the time difference t-w. The y-axis describes the expectation value as a reference point and the contributions of the features to differences to the reference to value.

a model. Thus, removing features with low contribution should result in only slight changes in corresponding predictions. To validate this, I chose the least significant feature for the best predictions above 5km/h. Fig. 10 shows that this is the y-axis of the gyroscope. By removing this feature, the accuracy of the predictions above 5km/h only increases marginally. However, including predictions below the threshold, for which the contribution is higher than for those above 5km/h, the average prediction error was increased by 0.16km/h.

VII. Conclusion

Deep SHAP can be used to explain local and global feature importance for LSTM models. It is possible to use predictions of model adaptations to validate the resulting explanations. For the original model, Deep SHAP reveals that the temperature of the gyroscope holds the highest contribution for the test set. However, explanations of different local predictions have shown that the global explanation is mainly similar to predictions below 5km/h. The explanations differ vastly for predictions above the 5km/h threshold. Because of this, predictions above 5km/h with minor errors have higher correlations with the x-axis of the gyroscope and the z-axis of the accelerometer. The result is that even though the temperature significantly contributes to predictions on the total test set, removing this feature implies a decrease of the mean absolute error by

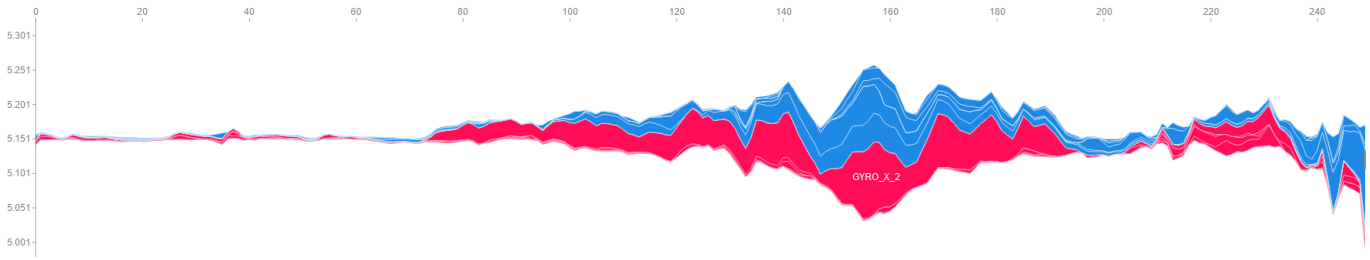

Fig. 5: SHAP force plot, indicating the SHAP values over all lag values for the worst prediction of the test data. The X-Axis is labeled with t-w. The y-axis describes the expectation value as a reference point and the contributions of the features to differences to the reference to value.

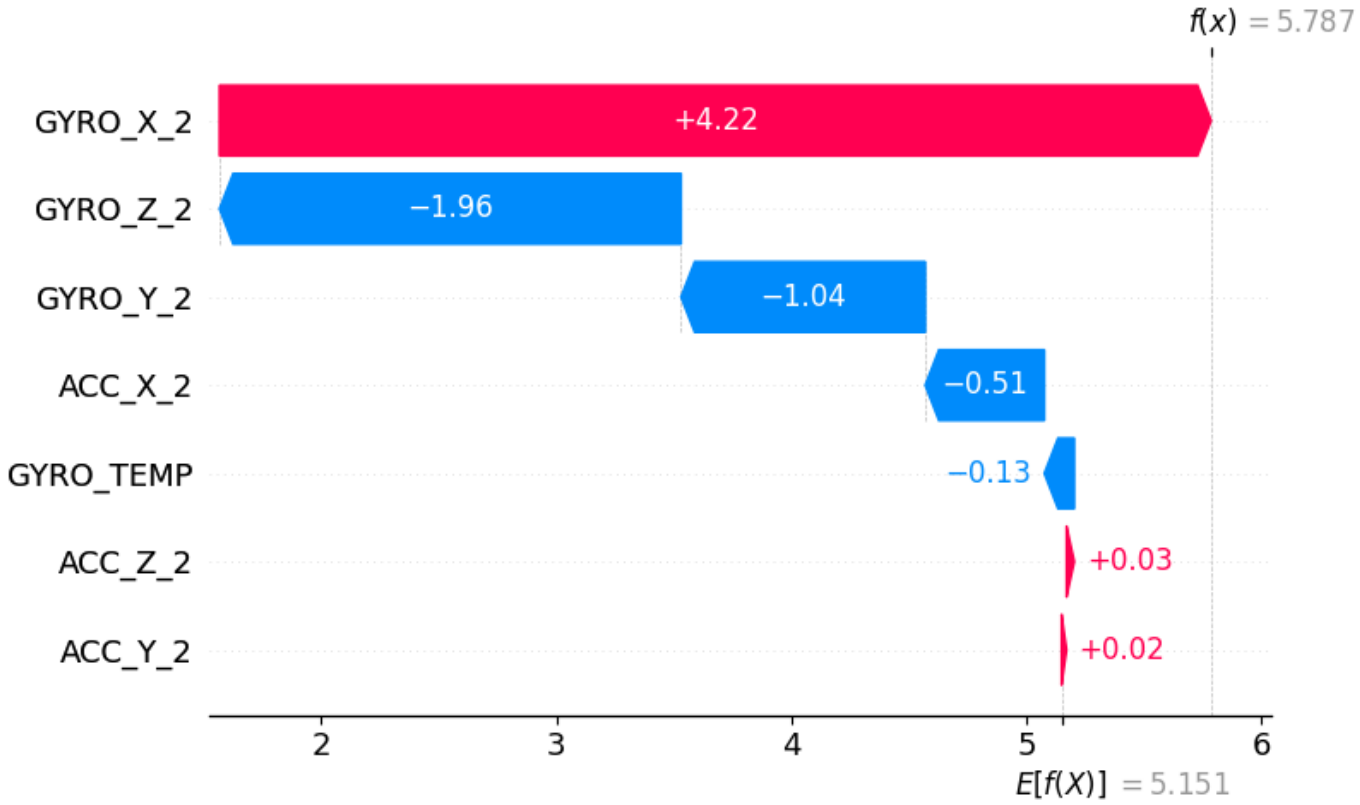

Fig. 6: SHAP waterfall legacy for the summed feature importances of all lag values for the highest misprediction.

0.09km/h. Contrary to this, removing the least significant features for good predictions denotes a total error increase. To sum up, it is essential to clarify that explanations that reveal contributions, like Deep SHAP, only indicate how much a model weighs a feature for its predictions. In general, Deep SHAP provides valuable information about time series models, introducing possibilities to understand and improve current models. However, one must distinguish this information from a feature that positively impacts a model's predictions in general.

References

[1] A. Hermann and V. Senner, *Knee injury prevention in alpine skiing. A technological paradigm shift towards a mechatronic ski binding*, vol. 24. Elsevier BV, Oct. 2021.

[2] T. M. Davidson and A. T. Laliotis, "Alpine skiing injuries. a nine-year study," *West. J. Med.*, vol. 164, pp. 310–314, Apr. 1996.

[3] ARAG and deutscher Skiverband, *Prozentuale Verteilung von Verletzungen alpiner Skifahrer in der Saison 2016/2017 auf Körperregionen*. 2017.

[4] V. Senner, F. I. Michel, S. Lehner, and O. Brügger, "Technical possibilities for optimising the ski-binding-boot functional unit to reduce knee injuries in recreational alpine skiing," *Sports Engineering*, vol. 16, pp. 211–228, Oct. 2013.

[5] V. Senner, S. Lehner, M. Nusser, and F. I. Michel, *Skiausrüstung und Knieverletzungen beim alpinen Skifahren im Freizeitsport: Eine Expertise zum gegenwärtigen Stand der Technik und deren Entwicklungspotenzial*, vol. Nr. 69 of *bfu-Report*. bfu, 2014.

[6] P. Carqueville, "Determination of skiing speed by means of imu-data and machine learning." ISSS-SITEMSH 2022 conference, Apr. 2022.

[7] A. Wägli, "Trajectory determination and analysis in sports by satellite and inertial navigation," 2008.

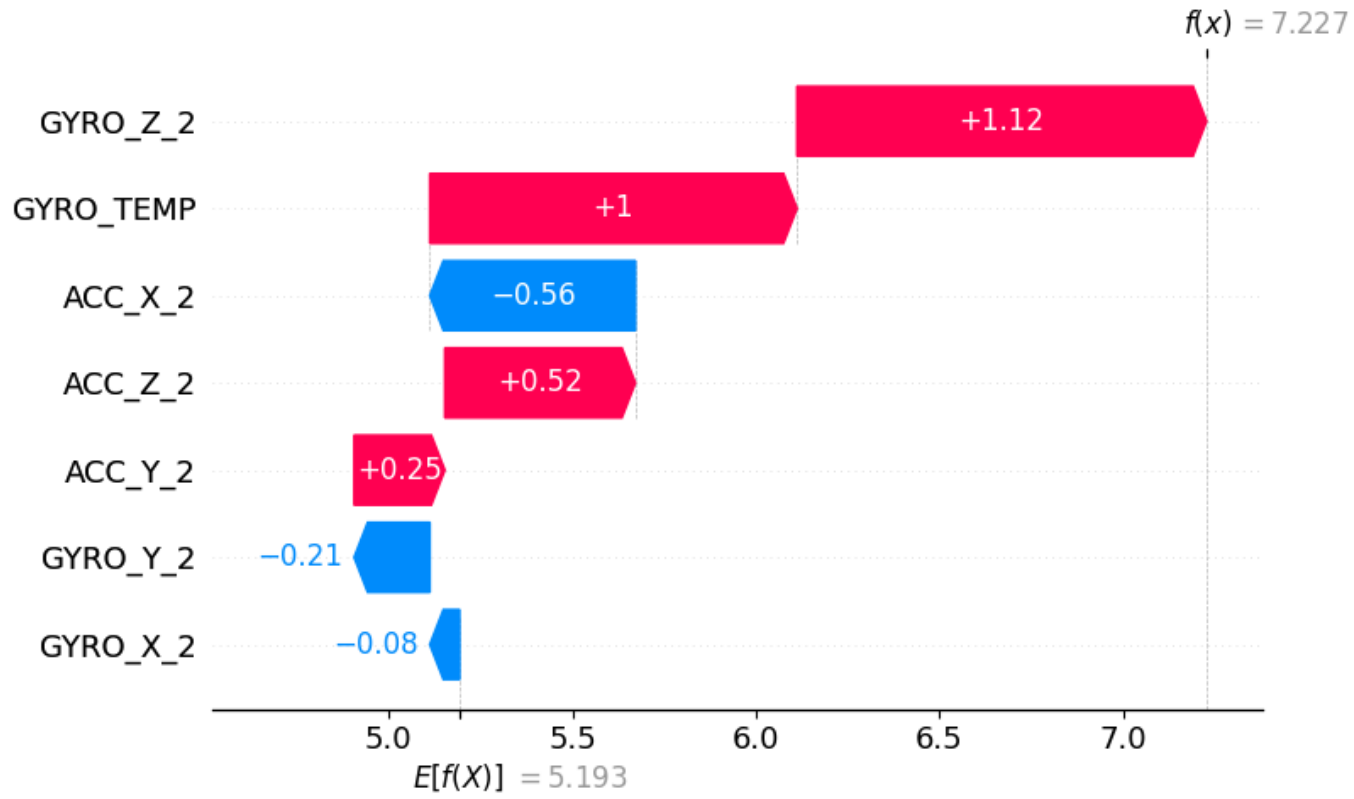

Fig. 7: SHAP waterfall legacy for the summed feature importances of all lag values for the second largest misprediction located in a different batch as the highest misprediction.

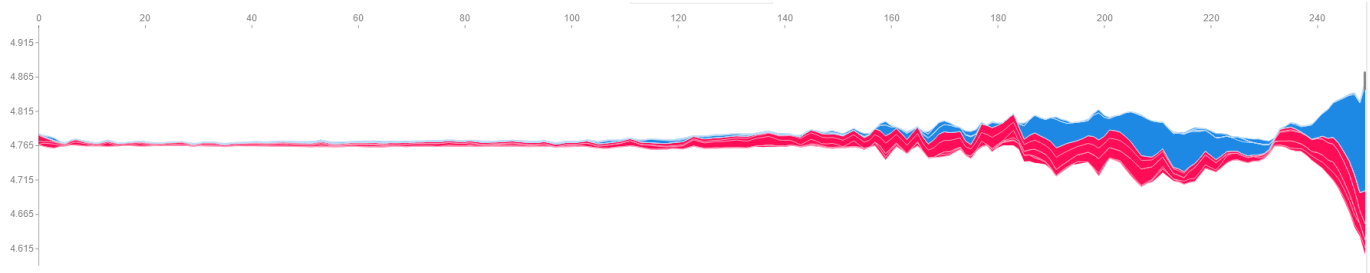

Fig. 8: SHAP force plot, indicating the SHAP values over all lag values for the worst prediction of the test data outside of the consecutive sequence of bad predictions. The x-axis is labeled with t-w. The y-axis describes the expectation value as a reference point and the contributions of the features to differences to the reference to value.

[8] F. Young, R. Mason, C. Wall, R. Morris, S. Stuart, and A. Godfrey, "Examination of a foot mounted IMU-based methodology for a running gait assessment," *Frontiers in Sports and Active Living*, vol. 4, Sept. 2022.

[9] S. Seenath and M. Dharmaraj, "Conformer-based human activity recognition using inertial measurement units," *Sensors*, vol. 23, p. 7357, Aug. 2023.

[10] S. Siami-Namini, N. Tavakoli, and A. S. Namin, "The performance of LSTM and BiLSTM in forecasting time series," in *2019 IEEE International Conference on Big Data (Big Data)*, IEEE, Dec. 2019.

[11] S. Siami-Namini, N. Tavakoli, and A. S. Namin, "A comparison of ARIMA and LSTM in forecasting time series," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, Dec. 2018.

[12] A. Barman, "Time series analysis and forecasting of covid-19 cases using lstm and arima models," 2020.

[13] A. Yadav, C. K. Jha, and A. Sharan, "Optimizing LSTM for time series prediction in indian stock market," *Procedia Computer Science*, vol. 167, pp. 2091–2100, 2020.

[14] I. I. Zulfa, D. C. R. Novitasari, F. Setiawan, A. Fanani, and M. Hafiyusholeh, "Prediction of sea surface current velocity and direction using LSTM," *IJEIS (Indonesian Journal of Electronics and Instrumentation Systems)*, vol. 11, p. 93, Apr. 2021.

[15] J. Wang and L. Li, "Traffic flow velocity prediction based on real data LSTM model," in *SAE Technical Paper Series*, SAE International, Dec. 2021.

[16] T. Feigl, S. Kram, P. Woller, R. H. Siddiqui, M. Philippsen, and C. Mutschler, "Rnn-aided human velocity estimation from a single imu," *Sensors*, vol. 20, p. 3656, June 2020.

[17] R. van den Tillaar, S. Bhandurge, and T. Stewart, "Can machine learning with imus be used to detect different throws and estimate ball velocity in team handball?," *Sensors*, vol. 21, p. 2288, Mar. 2021.

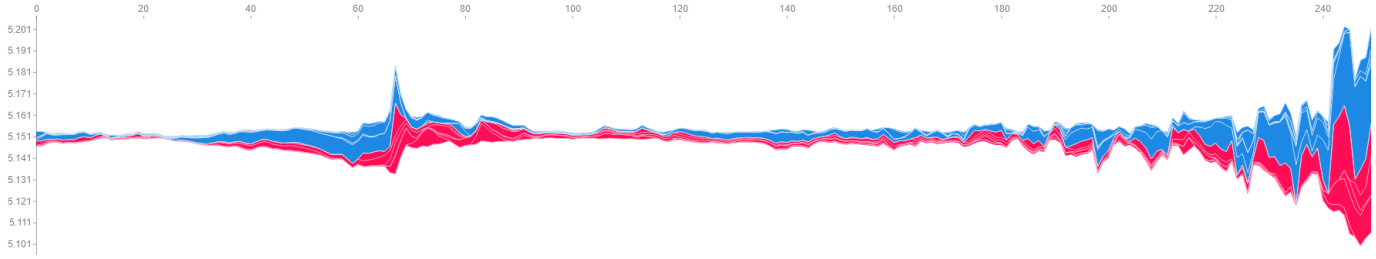

Fig. 9: SHAP force plot for the average of 30 randomly sampled good predictions above 5km/h. It shows the average influence of features for each lag value. The x-axis is labeled with t-w. The y-axis describes the expectation value as a reference point and the contributions of the features to differences to the reference to value.

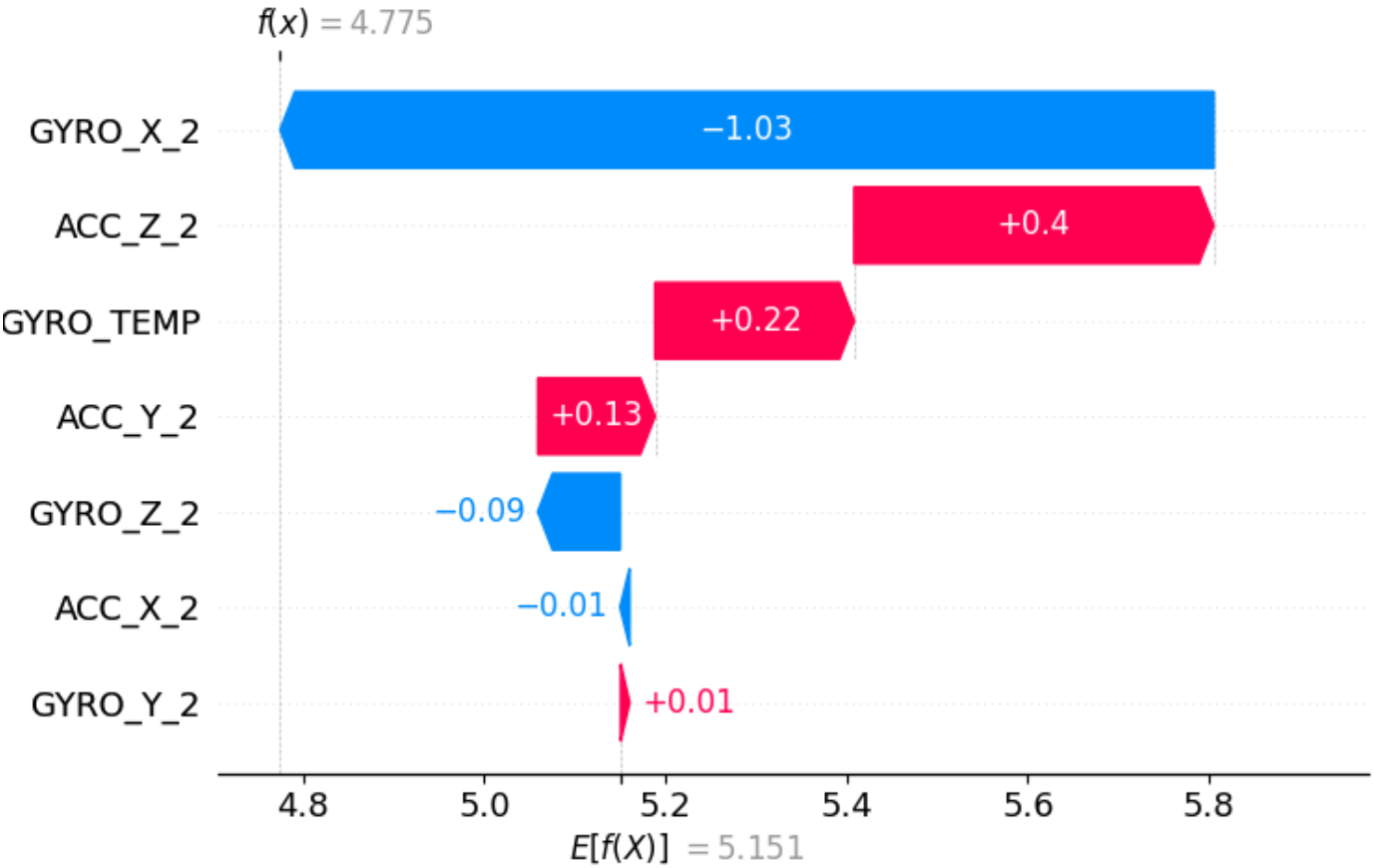

Fig. 10: SHAP waterfall legacy for the average of 30 randomly sampled good predictions above 5km/h. It indicates the total summed, average influence over all lag values for each feature.

[18] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, and U. R. Acharya, "Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)," *Computer Methods and Programs in Biomedicine*, vol. 226, p. 107161, Nov. 2022.

[19] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017.

[20] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," 2016.

[21] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," 2016.

[22] A. Kapishnikov, T. Bolukbasi, F. Viégas, and M. Terry, "Xrai: Better attributions through regions," 2019.

[23] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," ICML'17, p. 3145–3153, JMLR.org, 2017.

[24] J. a. Bento, P. Saleiro, A. F. Cruz, M. A. Figueiredo, and P. Bizarro, "Timeshap: Explaining recurrent models through sequence perturbations," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery Data Mining*, KDD '21, (New York, NY, USA), p. 2565–2573, Association for Computing Machinery, 2021.

[25] U. Schlegel, H. Arnout, M. El-Assady, D. Oelke, and D. A. Keim, "Towards a rigorous evaluation of XAI methods on time series," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, IEEE, Oct. 2019.

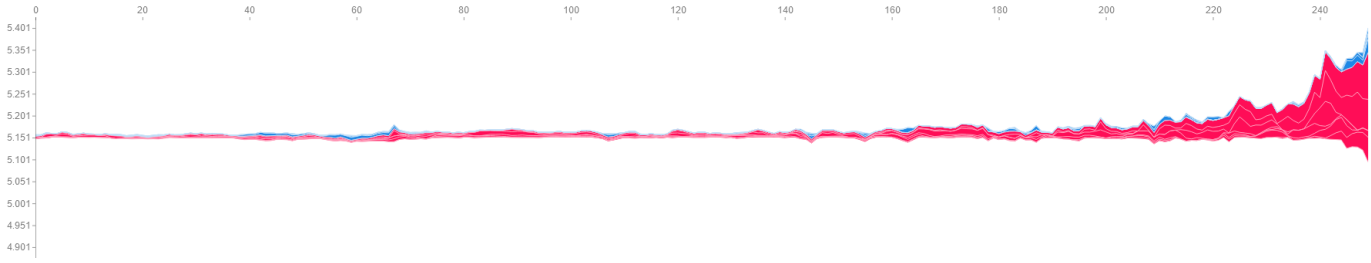

Fig. 11: SHAP force plot for the average of 20 randomly sampled good predictions below 5km/h. It shows the average influence of features for each lag value. The y-axis describes the expectation value as a reference point and the contributions of the features to differences to the reference to value.

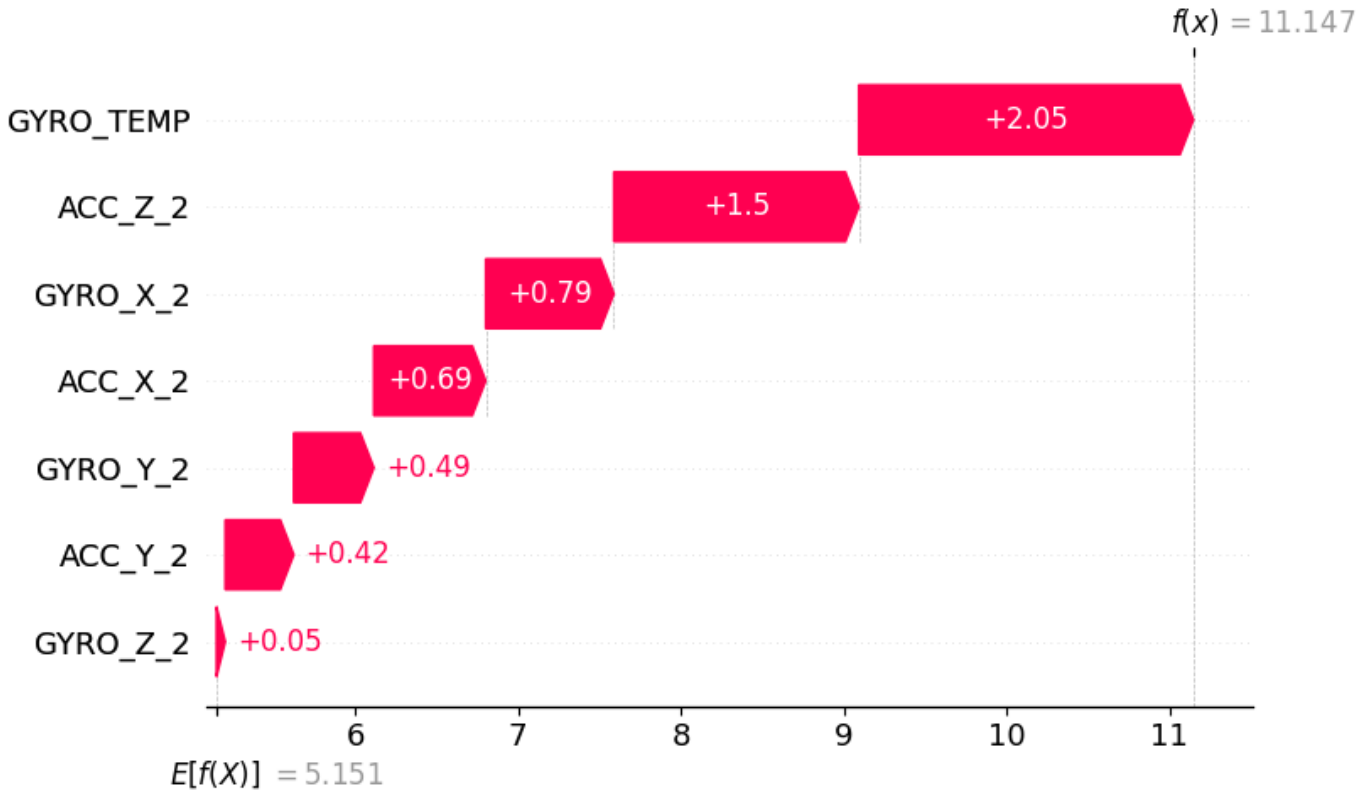

Fig. 12: SHAP waterfall legacy for the average of 20 randomly sampled good predictions below 5km/h. It indicates the total summed, average influence over all lag values for each feature.

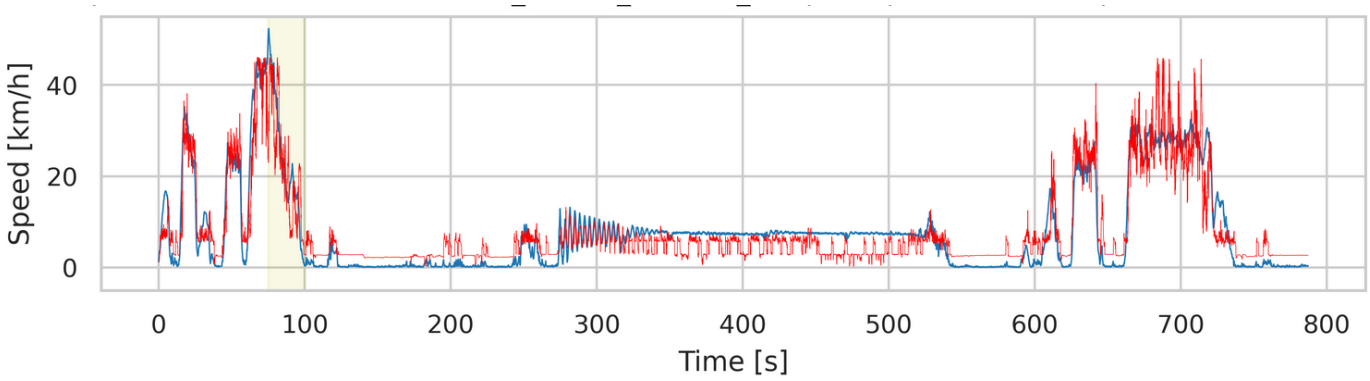

Fig. 13: LSTM prediction results in comparison to label velocities for model adaptation without gyroscope temperature as a feature. The blue curve holds the real velocities, measured by the GNSS sensor, and the red curve is the predicted velocity of the LSTM model