



Journal Homepage: [-www.journalijar.com](http://www.journalijar.com)

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI:10.21474/IJAR01/24381
DOI URL: <http://dx.doi.org/10.21474/IJAR01/24381>



RESEARCH ARTICLE

USING ARTIFICIAL INTELLIGENCE FOR THE EARLY DIAGNOSIS OF VEHICLE MALFUNCTIONS: A LEAKAGE-AWARE EVALUATION AND META-ANALYSIS

Idrees M. A. Alkandari

1. Sabah Alsalem Industrial Institute, PAAET, Kuwait.

Manuscript Info

Manuscript History

Received: 12 August 2026

Final Accepted: 14 September 2026

Published: October 2026

Key words:-

Artificial intelligence, early fault diagnosis, engine fault, Engine Fault DB, machine learning, meta-analysis, predictive maintenance, vehicle.

Abstract

The early diagnosis of the vehicle malfunctions by the artificial intelligence (AI) is a promising way to reduce the breakdowns and the maintenance cost, but the reported accuracies in the literature are usually very high and they may not reflect the real use. In this paper, six machine learning classifiers were trained on the public Engine Fault DB dataset, which contains 55,999 records of 14 engine and exhaust variables under one normal and three fault conditions. The models were evaluated by two protocols, a random five-fold split and a blocked split in which the test segments represent unseen operating conditions. In addition, a random-effects meta-analysis of nine published studies was conducted. The results showed that the four-class accuracy was about 75% under the random split, and it dropped to between 41.17% and 66.92% under the blocked split. However, the detection of the fault against the normal state was almost perfect ($AUC = 1.000$), and when the two fault types which cannot be separated were merged, the accuracy became 98.22% with all the signals and 97.34% with six on-board signals only. The meta-analysis gave a pooled reported accuracy of 96.50% (95% CI 92.63% to 98.38%) with a very high heterogeneity ($I^2 = 98.47\%$). The findings indicate that, on this dataset, the AI models can detect the presence of a fault reliably, while the isolation of the fault type under new operating conditions remains difficult and it needs a stricter validation practice.

"© 2026 by the Author(s). Published by IJAR under CC BY 4.0. Unrestricted use allowed with credit to the author."

Introduction:-

The modern vehicle contains a large number of sensors and electronic control units, and these sensors produce a continuous data about the state of the engine, the powertrain and the battery. The traditional on-board diagnostics depend on the fixed thresholds and the diagnostic trouble codes, and accordingly the fault is detected usually after it becomes severe. Therefore, the early diagnosis of the vehicle malfunctions by the artificial intelligence (AI) received a big attention in the recent years, because the AI can learn the small changes in the sensor signals before the fault becomes critical [1], [2], [3]. Several previous studies applied the machine learning and the deep learning for the vehicle fault diagnosis. As for the electric vehicles, the deep learning was used for the battery fault diagnosis with real vehicle data, and an accuracy of 97.30% was reported for the early battery failures [4], while other researches focused on the

Corresponding Author:- Idrees M. A. Alkandari

Address:- Sabah Alsalem Industrial Institute, PAAET, Kuwait.

motor faults [1], [5] and on the battery pack faults [6]. Regarding the internal combustion engine, the misfire and the combustion faults were diagnosed from the vibration, the acoustic and the crank speed signals [7], [8], [9]. Moreover, Vergara et al. [10] published the Engine Fault DB dataset, which is one of the few open datasets for the automotive engine faults, and the later studies reported accuracies near 75% on this dataset [11].

However, it should be noted that many of the previous studies reported the accuracy on a random split of the data, where the training and the test samples come from the same experimental runs. In this case, the neighboring samples in the time are very similar, and the model can recognize the operating condition instead of the fault itself, which is known as the data leakage [12]. Accordingly, the reported accuracy may be higher than the accuracy in the real vehicle, where the operating conditions are always new. Furthermore, a quantitative synthesis of the reported accuracies in this field was not found by the authors. For this reason, the current study revisits the EngineFaultDB benchmark with six classifiers and two validation protocols, namely a random split as in the previous studies and a blocked split which keeps the test segments unseen during the training. The fault detection and the fault isolation are evaluated separately, and the effect of the decision window and of the signals which are available from the OBD-II port is also examined. In addition, the reported accuracies of the recent AI vehicle diagnosis studies are synthesized by a random-effects meta-analysis, in order to place the obtained results in the context of the literature.

Materials and Methods:-

Dataset:-

The Engine Fault DB dataset [10] was collected from a C14NE spark ignition engine under the laboratory conditions, and it consists from 55,999 records and 14 variables, which are the manifold absolute pressure (MAP), the throttle position (TPS), the force, the power, the engine speed (RPM), the fuel consumption in L/h and in L/100 km, the speed, the CO, the HC, the CO₂, the O₂, the lambda and the air-fuel ratio. The label has four classes, the normal state (F0, 16,000 records), the fault type 1 (F1, 10,998), the fault type 2 (F2, 15,000) and the fault type 3 (F3, 14,001). It is worth mentioning that the records are stored in the order of the acquisition, and each class forms one continuous run.

Classifiers and Validation Protocols:-

Six classifiers were implemented with scikit-learn [13]: the logistic regression (LR), the decision tree (DT), the k-nearest neighbors (k-NN, $k = 5$), the random forest (RF, 150 trees), the histogram gradient boosting (HGB) and the multilayer perceptron (MLP, 64-32 neurons with early stopping). The standardization was applied for the LR, the k-NN and the MLP. Two validation protocols were used. In the random protocol, a stratified five-fold split was applied, as it is common in the previous studies. In the blocked protocol, each class run was divided into consecutive blocks of 500 records, and a group five-fold split was applied on the blocks, so the test blocks were never seen during the training. In other words, the blocked protocol simulates the diagnosis under the new operating conditions. The performance was measured by the accuracy and the macro-F1, and the area under the ROC curve (AUC) was used for the binary detection of the fault against the normal state.

Early Diagnosis Analyses:-

Three additional analyses were made with the HGB model under the blocked protocol. In the first analysis, a short time window of w consecutive records ($w = 1, 5, 10, 25$ and 50) was summarized by the mean, the standard deviation and the difference between the last and the first record, in order to study the trade-off between the decision time and the accuracy. In the second analysis, the model was trained with six signals only (MAP, TPS, RPM, speed, lambda and fuel consumption in L/h), which are available from a normal OBD-II port, and it was compared with the full set which needs a gas analyzer. In the third analysis, the permutation importance was computed on 6,000 test records with five repetitions.

Meta-Analysis Procedure:-

The peer-reviewed studies from 2022 to 2025 which report the accuracy of an AI model for a vehicle fault diagnosis task were collected, and nine studies were included [4], [7], [8], [9], [11], [14], [15], [16], [17]. The studies were divided into two subgroups according to the data source, the laboratory or simulation data ($k = 5$) and the field data from real vehicles ($k = 4$). For the study of Canal et al. [17], the F1-score was used because the accuracy was not reported. The accuracy was transformed to the logit scale, and since the test sizes were not reported consistently, the within-study variance was approximated by $1/[n p(1 - p)]$ with a common effective size of $n = 1000$, and $n = 200$ and $n = 5000$ were tested as a sensitivity analysis. The pooled estimate was obtained by the Der Simonian-Laird random-effects model [18], and the heterogeneity was measured by the I^2 index [19].

Results:-

Random Versus Blocked Validation:-

Fig. 1 compares the four-class accuracy of the six classifiers. Under the random split, the five non-linear models gave almost the same accuracy, between 74.69% and 75.05%, and the LR gave 58.76%. These values are in agreement with the values of 74.82% to 74.96% which were reported in [11], and this confirms the reproducibility of the pipeline. However, under the blocked split, the accuracy was reduced for all the models, and it became 66.92% for the DT, 59.32% for the k-NN, 53.54% for the RF, 52.60% for the HGB, 49.85% for the MLP and 41.17% for the LR. Moreover, the standard deviation between the folds was increased from below 0.6% to between 1% and 5%.

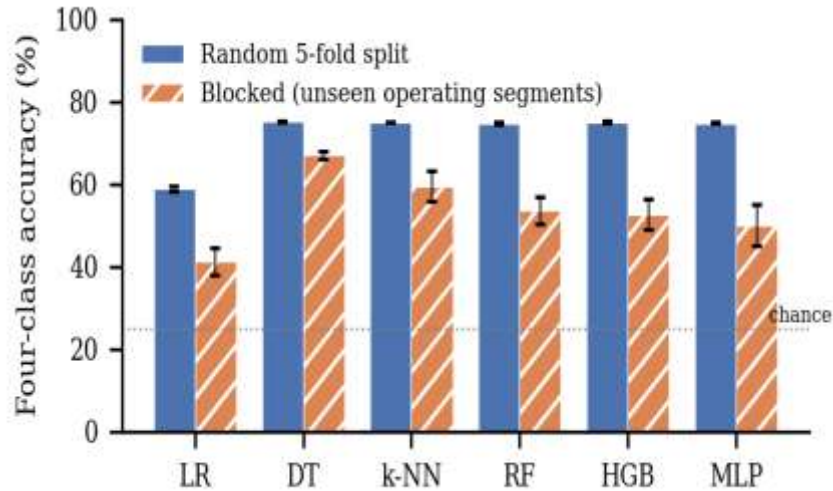


Fig. 1. Four-class accuracy of the six classifiers under the random and the blocked validation (mean $\pm$ SD over five folds).

In order to understand the reason of this drop, the confusion matrices of the HGB model are shown in Fig. 2. Under the random split, the normal state and the fault type 1 were classified perfectly, but the fault types 2 and 3 were confused at a level near the chance (55% and 48% correct). Under the blocked split, the fault type 1 was still recognized at 91%, while 87% of the fault type 2 records were predicted as the fault type 3 and 89% of the fault type 3 records were predicted as the fault type 2. In other words, the measured variables do not contain a stable information which separates the fault types 2 and 3, and the model under the random split was learning the operating point of each run.

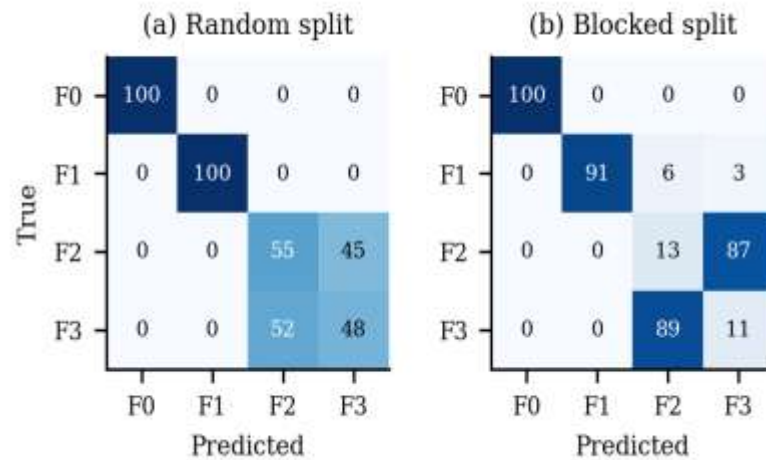


Fig. 2. Row-normalized confusion matrices (%) of the HGB model: (a) random split, (b) blocked split. F0 is the normal state.

Fault Detection and Reduced Fault Isolation:-

As it is shown in Table I, the binary detection of the fault against the normal state was almost perfect under the blocked split, with an AUC of 1.000 for all the signals and 0.9994 for the six on-board signals. When the fault types 2 and 3 were merged into one class, the three-class accuracy under the blocked split became 98.22% (macro-F1 97.33%) with all the signals and 97.34% (macro-F1 96.12%) with the on-board signals only.

Table I Hgb Performance Under The Blocked Validation

Task	Signals	Accuracy (%)	Macro-F1 (%)
4-class isolation	All 14	52.60	53.82
4-class isolation	On-board 6	52.83	53.94
3-class merged (F2+F3)	All 14	98.22	97.33
3-class merged (F2+F3)	On-board 6	97.34	96.12
Detection (AUC)	All 14	1.000	—
Detection (AUC)	On-board 6	0.9994	—

Decision Window and Feature Importance:-

Fig. 3(a) shows the effect of the window length on the macro-F1 of the four-class task. The macro-F1 was increased from 53.82% for a single record to 59.66% for a window of 50 records with all the signals, and from 53.94% to 60.77% with the on-board signals. Thus, a longer observation improves the isolation, but the gain is moderate and it comes with a longer decision time. Fig. 3(b) shows that the CO concentration was the most important variable (drop of 0.289 in the macro-F1), and it was followed by the force (0.095), the fuel consumption in L/h (0.072) and the O₂ (0.069).

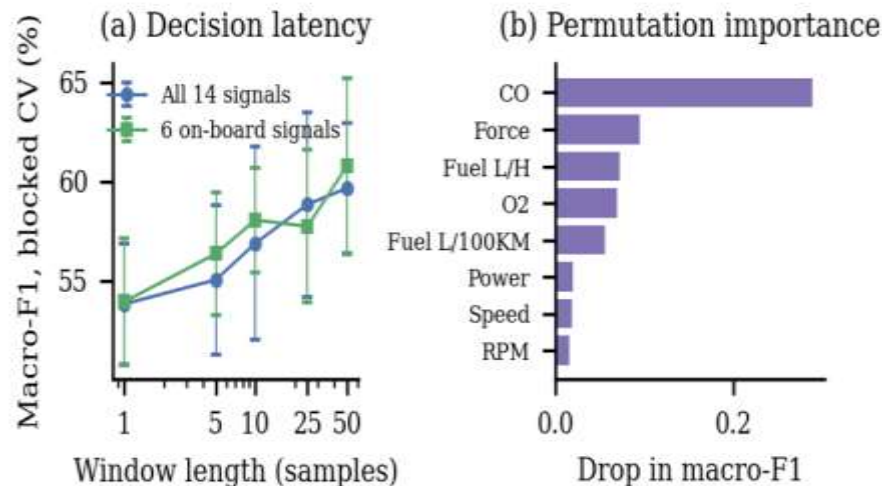


Fig. 3. (a) Macro-F1 versus the length of the decision window under the blocked validation. (b) Permutation importance of the top eight variables for the three-class model.

Meta-Analysis of the Published Studies:-

Fig. 4 presents the forest plot of the nine included studies. The reported accuracies ranged from 74.96% to 99.80%, and the pooled accuracy was 96.50% (95% CI 92.63% to 98.38%) with $I^2 = 98.47\%$ and $\tau^2 = 1.38$ on the logit scale. The laboratory and simulation subgroup gave 97.46% (89.96% to 99.39%), and the field subgroup gave 95.27% (89.22% to 98.00%), and the difference between the subgroups was not significant ($z = 0.74$, $p = 0.46$). With regard to the sensitivity analysis, the pooled accuracy was 95.78% for $n = 200$ and 96.75% for $n = 5000$, therefore the conclusion did not depend on the assumed sample size. It should be noted that the accuracies of the current study (74.88% random and 52.60% blocked, dashed lines in Fig. 4) are far below the pooled estimate.

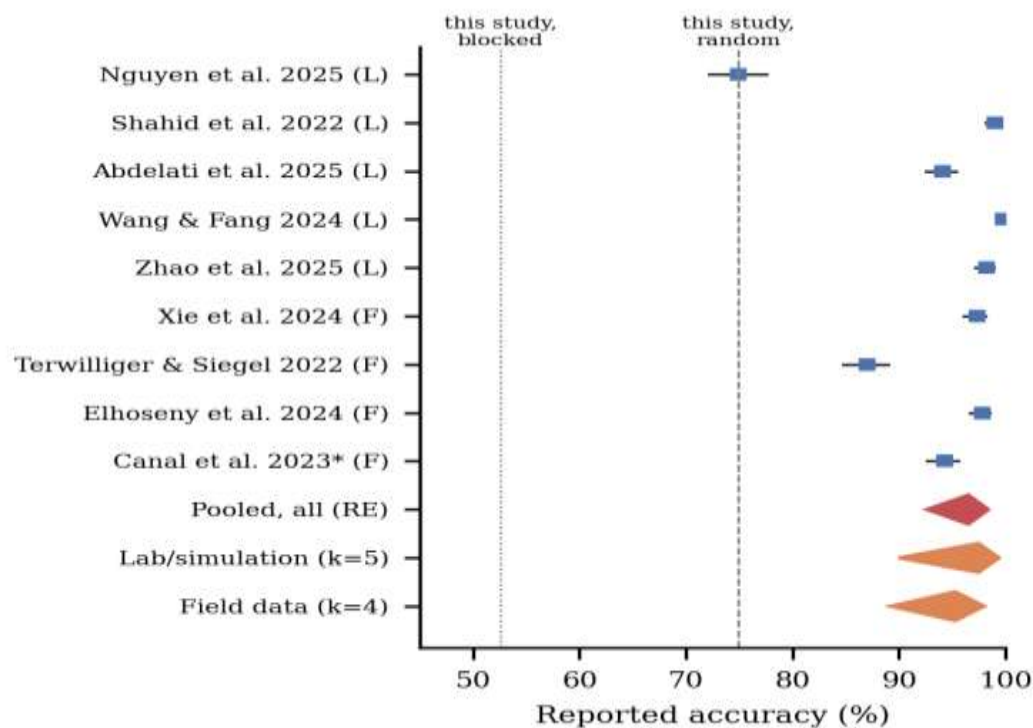


Fig. 4. Forest plot of the reported accuracies (95% CI on the logit scale, back-transformed) with the random-effects pooled estimates. (L) laboratory or simulation data, (F) field data. *F1-score used.

Discussion:-

The most important observation of this study is that the choice of the validation protocol changes the conclusion about the AI diagnosis to a large extent. Under the random split, all the non-linear models seem equivalent and the accuracy seems acceptable, while under the blocked split the accuracy dropped by up to 24.84 percentage points (for the MLP). This is in agreement with the general warning about the data leakage [12], and it explains partly why the pooled accuracy of the literature (96.50%) is much higher than the accuracy of the current study. Moreover, the simple DT was the most stable model under the blocked split, and this is consistent with the observation that the tree-based models are strong on the tabular data [20], but it also shows that the more complex model is not necessarily better in front of the new operating conditions.

On the other hand, the results are positive regarding the detection of an abnormal state. The normal state was separated from all the faults with an AUC near 1.000, even with the six signals which can be read from the OBD-II port. It should be mentioned, however, that the faults in this dataset are steady conditions and not incipient faults, therefore the word early here refers to the detection from the continuous sensor data before a trouble code is set, and this capability must be verified later on the incipient faults. Accordingly, a two-stage system is recommended, where the first stage detects any abnormal state early and warns the driver, and the second stage isolates the fault type when more data or a workshop test is available. The importance of the CO and the O₂ indicates that the mixture-related faults leave a clear signature in the exhaust, and this supports the integration of the low-cost exhaust sensors with the on-board data.

As for the meta-analysis, the very high heterogeneity ($I^2 = 98.47\%$) shows that the reported accuracies cannot be compared directly, because the studies differ in the vehicle system, the signals, the number of classes and the validation method. The absence of a significant difference between the laboratory and the field subgroups should be interpreted with caution, since the number of studies is small and the validation protocols of the field studies were not always described in detail. The limitations of this study are the use of one laboratory dataset of one engine, the unknown physical definition of the fault types 1 to 3 in the public release, and the assumed within-study variance in the meta-analysis.

Conclusion and Future Directions:-

In this paper, six AI classifiers were evaluated on the Engine Fault DB dataset under the random and the blocked validation, and a meta-analysis of nine published studies was conducted. It was found that the four-class accuracy was about 75% under the random split and dropped to between 41.17% and 66.92% under the blocked split, while the fault detection was almost perfect ($AUC = 1.000$) and the three-class accuracy reached 98.22% (97.34% with the on-board signals). The pooled reported accuracy of the literature was 96.50% with a very high heterogeneity. Several directions are suggested for the future work. The future studies should report the results under a condition-independent or a vehicle-independent split, together with the test size, in order to allow a reliable meta-analysis. Moreover, larger open datasets from the real vehicle fleets, including the electric vehicles, are needed. The domain adaptation and the transfer learning between the vehicles should be investigated to improve the isolation under the new conditions [4]. Lastly, the explainable AI methods should be integrated with the diagnosis model, so the technician can understand the decision of the model [2].

References:-

1. W. Lang, Y. Hu, C. Gong, X. Zhang, H. Xu, and J. Deng, "Artificial intelligence-based technique for fault detection and diagnosis of EV motors: A review," *IEEE Trans. Transp. Electrific.*, vol. 8, no. 1, pp. 384–406, Mar. 2022.
2. L. Cummins et al., "Explainable predictive maintenance: A survey of current methods, challenges and opportunities," *IEEE Access*, vol. 12, 2024.
3. S. Baccari, M. Haddad, H. Ghazzai, H. Touati, and M. Elhadeif, "Anomaly detection in connected and autonomous vehicles: A survey, analysis, and research challenges," *IEEE Access*, vol. 12, pp. 19250–19276, 2024.
4. L. Xie, J. Wei, X. Li, C. Chen, H. Chen, and G. Dong, "Deep learning-enabled fault diagnosis of lithium-ion batteries using real-world vehicle data with Gramian angular difference fields," *IEEE Trans. Ind. Informat.*, vol. 20, pp. 5622–5632, 2024.
5. M. A. S. Saleh and A. Ghrayeb, "Attention-enhanced AGRU framework for induction motor incipient fault diagnosis in electric vehicles," *IEEE Trans. Instrum. Meas.*, 2024, doi: 10.1109/TIM.2024.3502781.
6. L. Yao et al., "Fault identification of lithium-ion battery pack for electric vehicle based on GA optimized ELM neural network," *IEEE Access*, vol. 10, pp. 15007–15022, 2022.
7. S. M. Shahid, S. Ko, and S. Kwon, "Real-time abnormality detection and classification in diesel engine operations with convolutional neural network," *Expert Syst. Appl.*, vol. 192, Art. no. 116233, 2022.
8. M. H. Abdelati, M. Mourad, A.-H. Matar, and M. Rabie, "Enhanced engine misfire diagnosis through integration of vibration and acoustic emission signals using artificial neural networks," *J. Eng. Appl. Sci.*, vol. 72, no. 1, 2025.
9. A. M. Terwilliger and J. E. Siegel, "Improving misfire fault diagnosis with cascading architectures via acoustic vehicle characterization," *Sensors*, vol. 22, no. 20, Art. no. 7736, 2022.
10. M. Vergara, L. Ramos, N. D. Rivera-Campoverde, and F. Rivas-Echeverría, "EngineFaultDB: A novel dataset for automotive engine fault classification and baseline results," *IEEE Access*, vol. 11, pp. 126155–126171, 2023.
11. V.-T. Nguyen et al., "AI-based automotive engine diagnostics: Exploring the EngineFaultDB dataset for fault detection," *J. Sci. Technol. (Hanoi Univ. Ind.)*, 2025.
12. S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," *ACM Trans. Knowl. Discov. Data*, vol. 6, no. 4, pp. 1–21, 2012.
13. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
14. [14] J. Wang and T. Fang, "Research on fault diagnosis model of electric vehicle based on learning algorithm," in *Proc. SPIE*, vol. 13071, Art. no. 130713J, 2024.
15. Y. Zhao et al., "Enhancing battery durable operation: Multi-fault diagnosis and safety evaluation in series-connected lithium-ion battery systems," *Appl. Energy*, vol. 377, Art. no. 124632, 2025.
16. M. Elhoseny, D. D. Rao, B. D. Veerasamy, N. Alduaiji, J. Shreyas, and P. K. Shukla, "Deep learning algorithm for optimized sensor data fusion in fault diagnosis and tolerance," *Int. J. Comput. Intell. Syst.*, vol. 17, 2024.
17. R. Canal, F. K. Riffel, J. P. A. Bonomo, R. S. de Carvalho, and G. Gracioli, "Misfire detection in combustion engines using machine learning techniques," in *Proc. Braz. Symp. Comput. Syst. Eng. (SBESC)*, 2023, pp. 1–6.
18. R. DerSimonian and N. Laird, "Meta-analysis in clinical trials," *Control. Clin. Trials*, vol. 7, no. 3, pp. 177–188, 1986.
19. J. P. T. Higgins and S. G. Thompson, "Quantifying heterogeneity in a meta-analysis," *Stat. Med.*, vol. 21, no. 11, pp. 1539–1558, 2002.
20. L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?" in *Proc. NeurIPS Datasets Benchmarks Track*, 2022.