Using Explainable Artificial Intelligence (X-AI) To Address The "Black-Box" Dilemma in Autonomous Driving

Yibai Li

University of Nottingham, Nottingham, NG7 2RD, England, UK

* Corresponding Author Email: ssyyl55@nottingham.ac.uk

Abstract. Deep learning-based autonomous vehicles (AVs) demonstrate significant potential in reducing traffic accident rates and enhancing transportation efficiency. However, the "black-box" nature of artificial intelligence systems such as deep neural networks (DNNs) has raised widespread concerns regarding the explainability, transparency, and safety of their decision-making processes. This paper focuses on the application of explainable artificial intelligence (XAI) in the autonomous driving domain as an effective approach to address current technological challenges. The paper first reviews two fundamental architectures of autonomous driving systems and provides a basic overview of X-AI technologies, explaining their significance within these systems. It then systematically outlines the classification framework of X-AI and its critical importance in high-risk domains. Through analysis of multiple cutting-edge research frameworks—including the SafeX framework for modular architectures, the XAI integration framework for end-to-end systems, and the XAI-ADS system for cybersecurity— This paper delves into how X-AI enhances the safety, regulatory compliance, and user trust of autonomous driving systems. Finally, it outlines future development directions for X-AI in autonomous driving technology and proposes research recommendations for building explainable, safe, and responsible autonomous driving systems.

Keywords: Artificial intelligence, autonomous driving, explainable AI, "Black-Box".

1. Introduction

Autonomous driving has garnered significant attention across numerous technological fields, particularly technologies based on deep learning (DL) neural networks. Research indicates that by around 2030, sales of AVs are projected to surge to 30% of global automotive sales [1]. The non-profit Society of Automotive Engineers (SAE) developed a widely adopted classification system for autonomous driving levels [2]. Currently, most AVs remain at Level 2. Advancing to higher levels demands more sophisticated engineering, complex algorithms, and precise implementation—areas where deep learning has emerged as a key solution to current technical challenges.

However, despite the evident potential and benefits of neural network-based autonomous vehicles in daily life, society remains deeply concerned about their functional safety. This concern primarily stems from the "unexplainability" inherent in current artificial intelligence and deep learning technologies. Humans struggle to comprehend the unique decision-making criteria employed by computers. This results in situations where, if autonomous driving makes an erroneous judgment that causes danger, humans have no way of understanding why the computer made the wrong decision or how to correct it. Currently, according to a survey by the American Automobile Association, as many as 68% of drivers in the United States do not trust autonomous driving technology and remain wary of AVs [3].

For such technical challenges, explainable artificial intelligence (X-AI) represents a highly promising approach. This technology enables humans to understand the inner workings of deep learning systems and gain insights into their outcomes—a capability that is particularly crucial in the autonomous driving domain, where safety and decision-making are paramount concerns. This paper first summarizes current autonomous driving architectures alongside XAI technologies and their classifications, highlighting XAI's critical importance to the field. It then conducts a comprehensive review of XAI applications in autonomous driving through extensive analysis of

recent research, covering diverse technical frameworks and underlying principles. Finally, it outlines the technology's future prospects and directions for further research.

2. Technical Overview

2.1. Architecture of the Autonomous Driving System

Currently, there are two mainstream architectures for autonomous driving systems: modular and end-to-end. The modular pipeline divides the autonomous driving system into four interconnected modules: perception, localization, planning, and control. These modules respectively perform the functions of acquiring data, constructing environmental maps, generating navigation paths and motion decisions, and sending commands to execute motion based on those decisions. Due to recent technological breakthroughs in deep learning and computer vision algorithms, coupled with enhanced performance of various high-end sensor devices, end-to-end autonomous driving systems have gradually become the primary focus for researchers. This architecture directly uses raw sensor data as visual input to generate control commands. Unlike the modular architecture, it directly produces driving behavior by packaging the four major modules into a composite machine learning task. Comparing the two architectures, the modular architecture is easier to develop, explain, and debug. However, it may suffer from higher latency, and errors can accumulate across layers due to sequential processing. The end-to-end architecture, conversely, is simpler and more efficient, avoiding information loss in intermediate layers. However, the end-to-end architecture relies heavily on “deep learning,” which poses significant challenges in understanding the system's reasoning and decision-making processes. Addressing this challenge is the critical role of XAI [4,5].

2.2. X-AI

The “black-box” issue of AI algorithms is gaining increasing attention. Humans are becoming more concerned with how computers arrive at their outputs rather than merely requiring the results themselves. This involves a crucial property known as “explainability,” and artificial intelligence possessing this quality is termed “explainable artificial intelligence,” or “X-AI.”

Explainable Artificial Intelligence (X-AI) is defined as “artificial intelligence technology capable of expressing its logic to human users, identifying its strengths and weaknesses, and imparting knowledge about its potential future performance” [6]. This technology will help humans deepen their understanding of AI's logic and thought processes.

2.3. Classification of X-AI

Figure 1 illustrates the classification system of XAI [7].

The manner in which information is represented within AI models categorizes X-AI into three types: symbolic systems, sub-symbolic systems, and mixed systems combining both. Among these, constructing technical algorithmic systems using the sub-symbolic method presents the greatest difficulty.

The second classification method relates to the stage at which XAI generates explanations during the decision-making process. Post-hoc XAI systems operate after a decision has been made, making them broadly applicable to any AI method. Ante-hoc XAI systems, while more precise, are somewhat limited because they generate explanations directly from the information used to make the decision.

The third classification method categorizes XAI by its syntactic and semantic forms. Under this framework, four primary types of XAI emerge. Surrogate systems compress complex wholes into simpler, more explainable models—effectively “substituting” them. Example systems explain certain aspects of the decision-making process based on input examples. Important systems primarily explain the influence level of input features on decisions as specified by the user. Inherent systems are the simplest; for explainable preprocessing methods, the explanation pattern can be determined based on the system's inherent structure.

Classifying XAI systems based on their scope is also an effective approach, categorizing them into local, cohort, and global systems. Local systems provide explanations applicable only to a given input instance. Cohort systems offer explanations applicable to a group of instances. Global systems, meanwhile, provide explanations applicable to all input instances.

The final classification method involves categorization by “medium.” This refers to the manner in which the system's explanation is conveyed to relevant nodes—such as through visual charts, images, textual communication, or tactile feedback. For system design, this is a critical factor, as it directly impacts the effectiveness and comprehensibility of explanation.

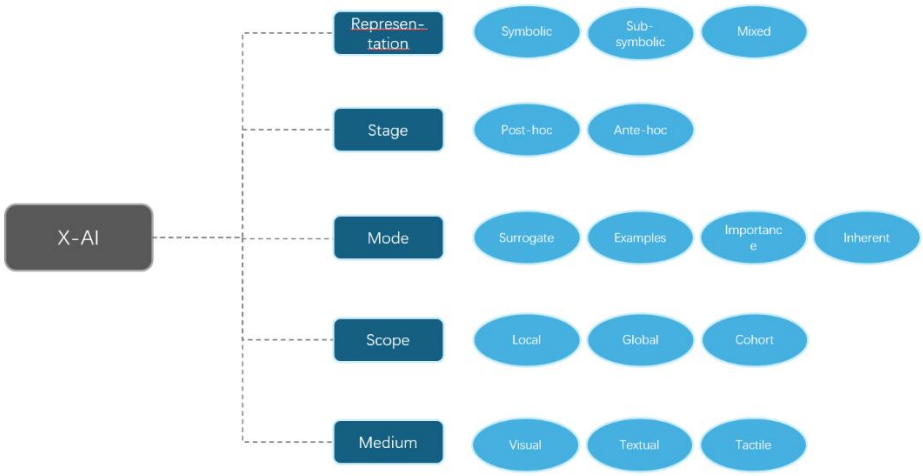


Fig. 1 Classification of X-AI [8]

3. X-AI Based Autonomous Driving System Framework

3.1. SafeX

Kuznietsov et al. proposed a conceptual framework named SafeX, which can be applied to develop safer and more explainable autonomous driving systems [8]. It integrates the user interface with safety and explainability within the system, serving as a framework suitable for modular autonomous driving systems. The structure of this framework is illustrated in Figure 2. At its core lies the Explainable Monitoring System (EMS), which facilitates the connection between the user interface and the autonomous vehicle. The system has two primary functions: First, EMS extracts necessary information from the autonomous vehicle based on user queries and generates user-understandable explanations. Second, it provides a monitor for each AD module (Perception Monitor, Planning Monitor, and Control Monitor), delivering safety feedback to the module's outputs.

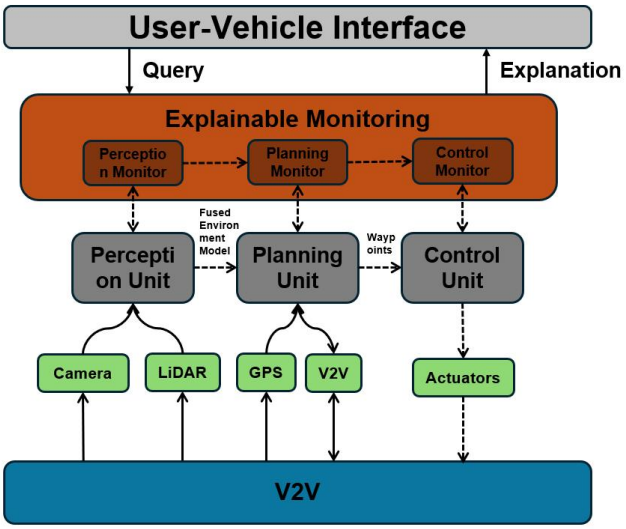


Fig. 2 SafeX Framework Architecture Diagram [8]

Figure 2 illustrates the architecture of the SafeX framework. It provides explanation and monitoring information for each AD module and transmits it to the EMS above. The three units on the left—Perception Unit, Planning Unit, and Control Unit—are all black-box modules based on deep learning technology.

Regarding this framework, Kuznietsov et al. proposed two variants of SafeX.

The first variant is that, for the perception unit, cameras can generate thermal images to highlight pixels relevant to the black-box model's predictions for explanation. The perception monitor can then validate the object detector's predictions for pedestrians. For the planning unit, the motion planning algorithm takes the detector's environmental observations as input, ultimately outputting the AV's projected trajectory. For the control unit, a DNN-based controller can be employed. This controller can be combined with cross-comparable clustering to serve as a tool for explaining control signals.

The second variant trains the perception module as a concept bottleneck model. By autonomously extracting semantic concepts across diverse environments, the algorithm gains enhanced explainability. For the planning module, Monte Carlo methods can generate search trees that provide explanations for planning decisions. Regarding the control module, DNN controllers with explainable projection mechanisms can be employed to facilitate auxiliary explanations.

3.2. E2E-XAI Integration Framework

Atakishiyev et al. propose a novel, comprehensive framework capable of providing full explanations for end-to-end autonomous driving systems [4]. This framework integrates approaches for developing XAI, learning end-to-end systems, and safety components to inform regulatory systems and processes. It divides end-to-end autonomous driving systems into three distinct components: an end-to-end control component, the safety-regulatory compliance component (srC), and an explanation component.

3.2.1 End-to-end control component

The function of a control system is to map the perceived environment onto the vehicle's behavior. If the control system is a total function capable of mapping every instance of the environment onto corresponding actions, it is termed an end-to-end control system (eeC).

The most prevalent learning paradigms for end-to-end autonomous driving are reinforcement learning (RL) and imitation learning (IL). Additionally, differentiable learning has recently emerged as an end-to-end driving architecture. While prioritizing the planning component, this learning pipeline optimizes multiple modules across the entire driving architecture. Overall, the end-to-end

learning pipeline employs a single deep neural network as a unified task, mapping sensor models of the world to real-time control commands for the autonomous vehicle.

3.2.2 Safety-regulatory compliance component

The srC acts on behalf of regulatory authorities, with one of its primary roles being to validate the safety of any combination of eeC components with AV actions. This component performs comprehensive compliance testing on eeC components to ensure the vehicle's public safety meets ISO 26262 standards. According to the standard, two fundamental approaches exist to confirm compliance: “simulation” and “verification.” When establishing regulatory compliance through simulation, the basic approach involves selecting a set of autonomous actions for simulation and then evaluating whether the results are satisfactory. In this method, the accuracy of the simulator determines the quality of compliance. However, this compliance process can be extremely costly and prone to security vulnerabilities. The fundamental approach to establishing regulatory compliance through “verification” involves validating AVs based on offline simulations using actual sensor suites and learning software stacks on real roads to pass safety inspections for regulatory compliance. Additionally, another function of this component is to defend against cyberattacks. It is anticipated that “automation” represents the future development goal for SRC. The complexity of this component lies in ensuring its testing methodology meets specific legal standards. For instance, regulators may require the eeC to achieve a minimum pass rate of 90% across N safety tests conducted.

3.2.3 Explanation component

This section forms the core of the framework's “explainability,” enabling it to provide human users with explanations for autonomous driving decisions consistent with the eeC and srC components. This component must be capable of explaining how it makes driving decisions during operation and communicating this information to users through both in-driving and post-driving analysis. In this section, Atakishiyev also engages in a more detailed discussion primarily focused on the temporal granularity of explanations.

Providing timely feedback greatly assists occupants in responding to emergencies. According to research by Koo et al., conveying explanations before an event occurs proves more beneficial than providing them afterward [9]. Pre-action explanations enable occupants to anticipate the current state and proactively monitor the AV's subsequent actions. Should the AV make a potentially hazardous decision, occupants can promptly take control. The timing of this handover request—specifically, the transition time between “autonomous” and “manual” driving—is another critical parameter. Research by Huang et al. indicates that excessively short transition times cause users stress due to insufficient reaction time and severely impair transition performance. Users demonstrate better performance when the transition time is ample (exceeding 10 seconds). During sudden emergencies, the advance time for transition control may be only a few seconds. However, this time can be appropriately extended during non-emergency periods [10-12]. At the same time, to provide users with a smoother, more relaxed driving experience, the system should ideally only offer explanatory information when necessary. If the vehicle provides occupants with an excessive amount of information during travel, it may lead to undue psychological stress for users, resulting in a poor driving experience. The method of delivering explanations involves human-machine interface (HMI) design. In this regard, meeting diverse user preferences while creating personalized designs tailored to varying cognitive abilities (such as color blindness or hearing impairments), presents a significant challenge for automotive manufacturers.

3.3. XAI-ADS

Due to the limited resilience of vehicular ad hoc networks (VANETs) in autonomous vehicles against external attacks, artificial intelligence models are required for anomaly detection within AV VANETs. However, such detection methods also lack explainability. Therefore, NAZAT et al.

propose a novel end-to-end XAI autonomous driving framework to enhance anomaly detection while enabling visualization and explainability—the "XAI-ADS" framework [13].

3.3.1 Framework Structure

The main components and processes of this framework are shown in Figure 3.

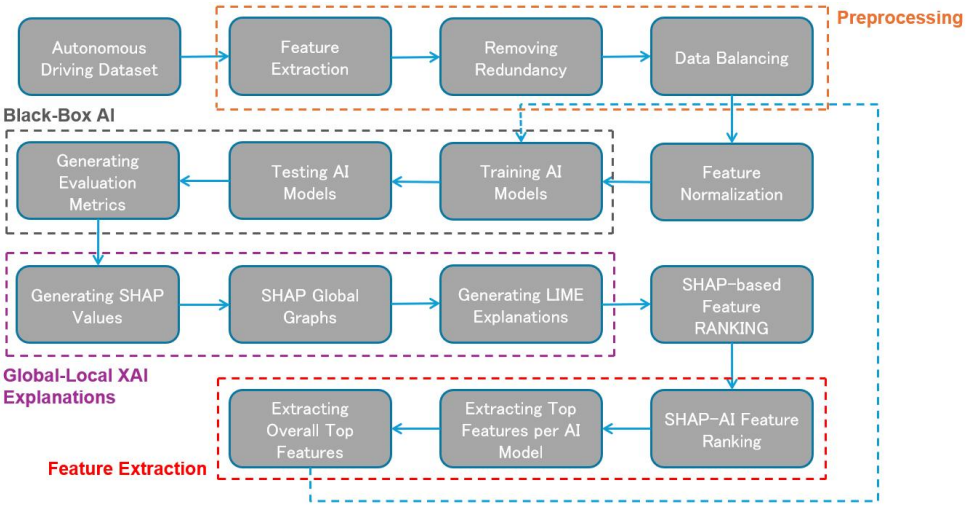


Fig. 3 Schematic Diagram of the XAI-ADS Framework Structure [13]

First, this framework assumes a decentralized adversary model where attackers reside within the VANET. Attackers can communicate with each benign AV in the network to compromise them. To effectively defend against these adversarial models, sensors are installed on all AVs. These sensors collect data during communication with other AVs and feed it into AI models to determine whether other AVs are benign or anomalous. The framework uses the VeReMi dataset (based on misbehavior in VANETs) and the Sensor dataset (based on monitoring ten different sensors on AVs) to classify AV anomalies [14]. The system selects fundamental features from each dataset to construct AI models. Redundancy is eliminated in VeReMi by removing duplicate rows and rows containing empty features, while no redundancy was found in the Sensor dataset. Dataset content is categorized into benign ("0") and anomalous ("1") classes. The classified VeReMi dataset exhibits a distribution where normal data far outweighs abnormal data. Therefore, random undersampling is required to remove data of no value to the model. The Sensor dataset, being largely balanced, does not require this step. Concurrently, standard scalar feature scaling is applied to normalize the datasets. This critical preprocessing step ensures no single feature dominates the analysis.

Finally, 70% of the data was used to train anomaly detection AI models, with hyperparameters optimized for each model to achieve optimal results. The models were then evaluated on the remaining 30% data subset using accuracy (Acc), precision (Prec), recall (Rec), F1 score (F1), and area under the ROC curve (AucRoc) as metrics [15].

3.3.2 Framework Explainability

The AI models constructed above are all considered "black-boxes". Therefore, these models require XAI steps for explanation. Explanations are categorized into "global explanations" and "local explanations."

For global explanations, the framework utilizes SHAP global summary plots to illustrate the relative importance of each feature considered by the model when determining whether an AV is abnormal. Simultaneously, SHAP swarm plots are employed to test the correlation between significant feature values and model decisions. For local explanations, the framework employs the more accessible Local interpretable model- agnostic explanations (LIME). It creates simpler, locally substitute models for each data instance that mimic its prediction. By evaluating the results of these substitute models, LIME reveals the decision-making process of the original model.

Additionally, the team proposed a novel feature selection method based on XAI to enhance feature reliability.

3.4. Summary and Discussion

This study primarily reviews three recently prominent XAI autonomous driving frameworks, each representing distinct application pathways for XAI across different architectural and functional domains within autonomous driving. These frameworks possess unique characteristics and advantages, serving specific architectural and application scenarios, and even exhibiting complementary relationships in certain aspects.

Kuznietsov et al. proposed the SafeX framework. This framework is applicable to modular autonomous driving architectures, embedding explainability capabilities into submodules such as perception, planning, and control. It employs EMS to implement safety monitoring and user feedback for each module's output, representing a solution that emphasizes intra-module transparency and operational controllability. It is suitable for system environments with clear structures and well-defined responsibilities. However, for real-time decision-making systems like AVs, the framework's related work places excessive emphasis on the form and content of explanations while neglecting temporal aspects of explanation. To address this shortcoming, Atakishiyev et al. proposed the E2E-XAI integrated framework. From a holistic systems perspective, this framework tightly integrates control logic, compliance verification, and human-machine interaction, emphasizing explainability throughout the entire decision-making process. It is better suited for real-time interaction and accountability requirements regarding driving behavior and safety responsibility in future highly automated and intelligent scenarios. The team also conducted a detailed discussion on the issue of temporal granularity. However, a significant limitation of this research is the lack of detailed elaboration on how to integrate various XAI techniques into this framework. Compared to the SafeX framework, the practicality of this framework remains to be examined. In contrast, the XAI-ADS framework proposed by NATAZ et al. focuses more on protecting autonomous driving systems against security threats in VANET communication environments. This framework primarily employs AI models for anomaly detection while integrating tools like SHAP and LIME to enhance both local and global explainability of black-box models, thereby improving the transparency and auditability of AI models in cybersecurity tasks. It is better suited for deployment in V2X systems where attacks are frequent and communication environments are complex.

4. Conclusion

This paper systematically reviews the application pathways and technical practices of XAI in the field of autonomous driving, addressing the increasingly prominent issue of AI “black-box”. By examining current mainstream autonomous driving architectures and XAI technology classifications, it identifies the explainability challenges faced by existing autonomous driving systems and the technical requirements for XAI. Furthermore, this paper conducts a detailed analysis of three representative X-AI autonomous driving frameworks—SafeX, E2E-XAI, and XAI-ADS, demonstrating how X-AI technologies effectively achieve explainability and enhance societal trust at different levels. Overall, against the backdrop of stricter policy regulations and heightened user safety demands, the technical bottlenecks surrounding the “black-box” issue in AI have long plagued the autonomous driving sector. Currently, XAI, as a key solution to this problem, has gradually transitioned from theoretical research to practical application. Multiple XAI autonomous driving frameworks are maturing steadily, demonstrating significant growth potential. In the future, the development of technologies like large language models (LLM) systems will create an increasingly urgent demand for XAI.

However, current XAI research and its integration with autonomous driving still face shortcomings in terms of feasibility, explainability completeness, and real-world experimental cases.

Looking ahead, the development of X-AI in the autonomous driving field can be pursued from multiple angles. First, teams in this field should explore as many explanation methods as possible for various sensors, not just cameras. Simultaneously, research should focus on applying XAI to different integrated architectures, particularly the integration of XAI across diverse sensors. Secondly, X-AI algorithms also require further optimization. In this regard, “generative methods” may represent a promising future direction for XAI research in autonomous driving. These methods hold the potential to self-explain their decisions and may alleviate the long-tail distribution problem in critical scenarios. Simultaneously, the advancement of this technology necessitates fostering interdisciplinary collaboration between XAI, autonomous driving, and other fields, such as ethics and psychology. It will help researchers better understand the requirements of relevant stakeholders, ensuring that the system's explainability genuinely meets people's needs and expectations.

In summary, building explainable, safe, and trustworthy autonomous driving systems is a core step toward the gradual societal adoption of self-driving vehicles, and X-AI will play an irreplaceable strategic role in this developmental journey.

References

[1] Zhang Q, Zada M, Khan S, et al. Exploring the role of tourist pro-environmental behavior in autonomous vehicle adoption: A TPB and PLS-SEM approach. Sustainability, 2024, 16(20): 9021.

[2] International S. SAE Standards News: J3016 automated-driving graphic update. 2025.

[3] Moye B. AAA: Fear of self-driving cars on the rise. AAA Newsroom, 2023.

[4] Atakishiyev S, Salameh M, Yao H, et al. Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions. IEEE Access, 2024.

[5] Zhao J, Zhao W, Deng B, et al. Autonomous driving system: A comprehensive survey. Expert Systems with Applications, 2024, 242: 122836.

[6] Chamola V, Hassija V, Sulthana A R, et al. A review of trustworthy and explainable artificial intelligence (XAI). IEEE Access, 2023, 11: 78994–79015.

[7] Speith T. A review of taxonomies of explainable artificial intelligence (XAI) methods. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022: 2239–2250.

[8] Kuznietsov A, Gyevnar B, Wang C, et al. Explainable AI for safe and trustworthy autonomous driving: A systematic review. IEEE Transactions on Intelligent Transportation Systems, 2024.

[9] Koo J, Kwac J, Ju W, et al. Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. International Journal on Interactive Design and Manufacturing, 2015, 9(4): 269–275.

[10] Huang G, Pitts B J. Takeover requests for automated driving: The effects of signal direction, lead time, and modality on takeover performance. Accident Analysis & Prevention, 2022, 165: 106534.

[11] Mok B, Johns M, Lee K J, et al. Emergency, automation off: Unstructured transition timing for distracted drivers of automated vehicles. In: 2015 IEEE 18th International Conference on Intelligent Transportation Systems, 2015: 2458–2464.

[12] Wan J, Wu C. The effects of lead time of take-over request and nondriving tasks on taking-over control of automated vehicles. IEEE Transactions on Human-Machine Systems, 2018, 48(6): 582–591.

[13] Nazat S, Li L, Abdallah M. XAI-ADS: An explainable artificial intelligence framework for enhancing anomaly detection in autonomous driving systems. IEEE Access, 2024, 12: 48583–48607.

[14] Van Der Heijden R W, Lukaseder T, Kargl F. Veremi: A dataset for comparable evaluation of misbehavior detection in VANETs. In: International Conference on Security and Privacy in Communication Systems, 2018: 318–337.

[15] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 2011, 12: 2825–2830.