

Transparent SSD Failure Prediction Through Interpretable Artificial Intelligence Techniques

Hiroshi Nakamura

Department of Artificial Intelligence,
Tokyo Institute of Computing, Tokyo, Japan

Yuki Tanaka

Department of Information Science, Osaka
Advanced Technology University, Osaka, Japan

VOLUME 03 ISSUE 09 (2026)

Published Date: 04 September 2026 // Page no.: - 01-07

ABSTRACT

Solid-state drives (SSDs) have become critical storage components in computing infrastructures, cloud platforms, enterprise systems, and intelligent devices, making reliable failure prediction an important component of predictive maintenance and system dependability. Conventional machine-learning models can identify complex relationships among storage-health indicators, workload characteristics, error statistics, thermal conditions, and operational behavior, but their limited interpretability can restrict their practical adoption. This research develops a conceptual framework for transparent SSD failure prediction through interpretable artificial intelligence (AI), emphasizing the integration of predictive modeling with post-hoc explanation mechanisms such as Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP). The methodological foundation is positioned within the broader development of advanced AI systems, including transformer-based, multimodal, reasoning-oriented, and neural architectures represented in the supplied literature. The framework separates prediction from explanation while establishing an evidence-oriented pathway through which individual SSD-health indicators can be associated with predicted failure risk. The analysis indicates that transparency should not be treated as an auxiliary visualization layer but as an integral property of the predictive-maintenance pipeline. The study further identifies important trade-offs between predictive complexity, explanation fidelity, computational overhead, and operational usability. By incorporating interpretable AI principles into SSD failure prediction, the proposed approach provides a structured basis for trustworthy maintenance decisions while acknowledging the limitations associated with model-agnostic explanations, distribution shifts, feature correlations, and the absence of universally representative SSD failure datasets.

Keywords: SSD failure prediction, Explainable AI, interpretable machine learning, LIME, SHAP, predictive maintenance, storage reliability, artificial intelligence, model transparency

INTRODUCTION

The increasing dependence on digital storage has transformed storage reliability from a hardware-level concern into an important systems-engineering problem. SSDs offer high performance and low latency, but their operational behavior is influenced by write intensity, read activity, error accumulation, thermal conditions, wear, controller behavior, and other health indicators. Predictive maintenance seeks to identify abnormal behavior before catastrophic failure occurs. Artificial intelligence is particularly attractive for this purpose because it can model nonlinear interactions that may be difficult to express through conventional deterministic rules.

The broader evolution of AI demonstrates the increasing capability of models to learn complex representations from heterogeneous information. Modern language and multimodal systems such as GPT-4 and Gemini illustrate the scale at which contemporary AI models can represent complicated patterns (Achiam, 2023; Anil, 2023). Similarly, BERT demonstrated the effectiveness of deep bidirectional representation learning for extracting contextual relationships from complex inputs (Devlin et al., 2019). Although these systems address language rather than storage reliability, their development establishes an important theoretical observation: high-capacity models can discover relationships that are difficult to encode manually.

This capability creates a corresponding interpretability challenge. A predictive model may estimate that an SSD has a high probability of failure, but maintenance personnel need to

understand why that prediction was generated. An opaque prediction can be operationally problematic when a false alarm causes premature replacement or when a missed failure creates data-availability risks. Consequently, transparent prediction requires not only a probability or class label but also an interpretable representation of the factors contributing to that prediction.

Recent work specifically addressing explainable AI for SSD failure prediction demonstrates the relevance of LIME- and SHAP-oriented transparency to this problem (Kumar, 2026). The present research builds upon that direction by conceptualizing transparency as a complete methodological layer surrounding SSD failure prediction rather than as an isolated explanation algorithm.

The primary objective is therefore to develop a technically coherent framework in which SSD-health data are transformed into predictive features, analyzed through an appropriate AI model, and subsequently interpreted using local and global explanation mechanisms. The research focuses on three questions: how an AI model can represent SSD failure risk, how explanations can translate model behavior into maintenance-relevant evidence, and what limitations must be considered before explanations are treated as causal evidence.

LITERATURE REVIEW

The supplied references primarily represent developments in advanced artificial intelligence rather than SSD reliability specifically. This creates both a theoretical foundation and a clear research gap. The literature demonstrates rapid progress in representation learning, reasoning, neural architectures, and adaptive computation, while relatively little of the supplied material directly addresses storage-device failure prediction. The proposed study therefore positions interpretable SSD prediction as an application-oriented extension of contemporary AI methodology.

Achiam's GPT-4 technical report illustrates the development of large-scale models capable of processing sophisticated patterns through extensive learned representations (Achiam, 2023). Gemini similarly represents multimodal AI capable of integrating different forms of information into a unified model family (Anil, 2023). From the perspective of SSD monitoring, these developments are relevant because storage-health information can itself be heterogeneous. Numerical SMART-like indicators, error statistics, temporal measurements, workload characteristics, and operational states may constitute different information dimensions. Nevertheless, the computational scale and complexity of large foundation models may be unnecessary for a focused failure-prediction task.

BERT provides a contrasting perspective through specialized representation learning, demonstrating that complex relationships can be learned from structured input sequences (Devlin et al., 2019). This principle is relevant when SSD health is represented as a temporal sequence rather than as a single static observation. A sequence of increasing error rates, temperature fluctuations, or wear indicators may contain more predictive information than an isolated measurement.

The supplied literature on DeepSeek-R1, DeepSeek-V3, QFFT, S1, and MathCoder2 further illustrates contemporary emphasis on reasoning, adaptive computation, and test-time scaling (Guo, 2025; Liu, 2024; Liu, 2025; Muennighoff, 2025; Lu, 2024). These developments suggest that AI systems increasingly attempt to make complex computational processes more adaptive. For predictive maintenance, however, computational sophistication must be balanced against interpretability and operational constraints. A highly complex model is not automatically preferable if its decisions cannot be sufficiently explained to engineers.

The neural-architecture literature provides another theoretical perspective. Maass's work on spiking neural networks established principles associated with successive generations of neural computation (Maass, 1997). Gerstner et al. developed a comprehensive computational framework for neuronal dynamics, highlighting the importance of temporal behavior in neural systems (Gerstner et al., 2014). Subsequent work on deep spiking neural networks, including progressive tandem learning and event-driven learning, further emphasizes temporal and event-based computation (Wu, 2021; Wei, 2024). These concepts are relevant to SSD monitoring because device degradation is inherently temporal: failures rarely emerge from a single observation and instead may manifest through evolving operational patterns.

Pei's work on the Tianjic chip demonstrates the possibility of integrating heterogeneous computational paradigms within specialized architectures (Pei, 2019). SpikeLM and SpikeGPT extend spiking approaches toward language modeling, showing that alternative computational architectures can support sophisticated representation learning (Xing, 2024; Zhu et al., 2023). For SSD prediction, these studies motivate consideration of event-driven or temporal architectures where storage-health events arrive continuously.

The major gap is therefore apparent. The supplied AI literature demonstrates increasingly capable predictive and representational architectures, but it does not establish a complete transparent SSD-failure methodology. In particular, the literature does not directly resolve how complex predictive models should expose storage-specific evidence to maintenance engineers. Kumar (2026) directly establishes the relevance of LIME and SHAP to SSD failure prediction and therefore provides the most immediate conceptual connection between explainable AI and the target problem.

The theoretical positioning of the present study is consequently based on a separation of three functions: representation, prediction, and explanation. Representation converts raw operational observations into meaningful features; prediction estimates failure risk; and explanation identifies the contribution of individual features to that prediction. This separation prevents the explanation mechanism from being confused with the predictive mechanism itself.

METHODOLOGY

3.1 Research Framework

The proposed methodology consists of five connected stages: data acquisition, feature construction, predictive modeling, explanation generation, and maintenance-oriented interpretation.

At the first stage, SSD operational observations are collected over time. A conceptual observation vector may be expressed as:

$$X_t = [T_t, W_t, R_t, E_t, H_t, U_t, C_t] \quad X_t = [T_t, W_t, R_t, E_t, H_t, U_t, C_t]$$

where T_t represents thermal measurements, W_t write-related activity, R_t read-related activity, E_t error-related indicators, H_t health or wear measurements, U_t utilization information, and C_t controller or operational characteristics at time t .

The objective variable can be formulated as a binary failure state:

$$Y \in \{0, 1\} \quad Y \in \{0, 1\}$$

where $Y=1$ denotes a failure or predefined failure-risk event and $Y=0$ denotes a non-failure state.

The critical methodological point is that failure labels must be operationally defined. A vague definition of failure can create inconsistent training targets. Depending on the application, failure may refer to complete device unavailability, severe degradation, critical error status, or a specified prediction horizon before failure.

3.2 Feature Engineering and Temporal Representation

Raw measurements should not automatically be treated as independent observations. SSD degradation is generally temporal, meaning that trends can provide more information than isolated values. Consequently, the methodology should calculate features such as recent change, rolling average, variance, rate of deterioration, and frequency of abnormal events.

For a generic feature x , the temporal rate can be represented as:

$$\Delta x_t = x_t - x_{t-k} \quad \Delta x_t = x_t - x_{t-k}$$

where k represents a historical observation interval.

A model may therefore distinguish between an SSD whose error count is consistently high but stable and an SSD whose error count is increasing rapidly. This distinction is important because predictive risk may depend more strongly on degradation trajectory than on absolute magnitude.

The temporal perspective is theoretically compatible with research emphasizing dynamic neural behavior and event-driven computation (Gerstner et al., 2014; Wei, 2024). Nevertheless, the present framework does not require a spiking architecture. Conventional tree-based, neural, or ensemble models may be employed depending on dataset characteristics and computational requirements.

3.3 Predictive Modeling Layer

The prediction layer estimates:

$$P(Y=1|X) \quad P(Y=1|X)$$

where $P(Y=1|X)$ is the estimated probability that an SSD belongs to the failure-risk class given its observed characteristics.

A practical framework should evaluate at least two model categories: a relatively interpretable baseline and a higher-capacity model. The baseline establishes whether model complexity provides meaningful predictive benefit. A more complex model can then be justified only when its additional predictive capability is operationally valuable.

This principle is consistent with the broader AI literature: increasingly sophisticated architectures can capture complicated relationships, but computational sophistication does not inherently guarantee practical superiority (Achiam, 2023; Anil, 2023). Similarly, reasoning-oriented models demonstrate the value of adaptive computation, yet an SSD maintenance system may prioritize stability, latency, and interpretability over general-purpose reasoning capability (Guo, 2025; Muennighoff, 2025).

3.4 LIME-Based Local Explanation

LIME can be used to explain an individual prediction by approximating the complex model around a particular observation with an interpretable surrogate model.

For an SSD predicted as high-risk, LIME can identify which local features most strongly influence the classification. A hypothetical explanation could indicate that increasing error frequency, abnormal temperature behavior, and rapid health

deterioration contributed positively to the predicted failure risk, while stable utilization contributed negatively.

The principal advantage is locality. Maintenance personnel often need to understand why this specific SSD has been classified as high-risk. LIME is therefore suitable for case-level investigation.

However, locality is simultaneously a limitation. A local approximation may accurately describe model behavior near one observation without accurately representing global model behavior. Feature perturbations can also produce unrealistic combinations when the original variables are correlated. Consequently, a LIME explanation should be regarded as an interpretation of model behavior rather than direct causal evidence.

3.5 SHAP-Based Explanation

SHAP provides a complementary approach by assigning contribution values to features based on Shapley-value principles. For a prediction $f(x)$, the explanation can be represented conceptually as:

$$f(x) = \phi_0 + \sum_{i=1}^n \phi_i$$

where ϕ_0 is a baseline prediction and ϕ_i represents the contribution associated with feature i .

For SSD failure prediction, positive SHAP values may indicate features increasing predicted risk, while negative values may indicate features reducing risk. Global aggregation can identify which indicators are consistently influential across a population, whereas local SHAP values can explain individual devices.

This creates a useful distinction between global interpretability and local interpretability. Global SHAP analysis can reveal recurring model dependencies, whereas local explanations support individual maintenance decisions.

Kumar (2026) specifically establishes the applicability of LIME and SHAP for improving transparency in SSD failure prediction. The present methodology extends that principle by combining local and global interpretation within a broader predictive-maintenance workflow.

3.6 Explanation Validation

Interpretability should itself be evaluated. An explanation is useful only when it is sufficiently stable, faithful, and understandable.

Three criteria are proposed. First, fidelity asks whether the explanation accurately approximates the behavior of the underlying predictive model. Second, stability evaluates whether small, irrelevant changes in input data

produce unreasonable changes in explanations. Third, operational relevance considers whether engineers can translate the explanation into an appropriate maintenance action.

For example, if two nearly identical SSD-health observations produce radically different explanations, confidence in the explanation should decrease. Conversely, if increasing error trends consistently appear among high-risk predictions, the explanation becomes more operationally meaningful.

3.7 Decision Layer

The final layer converts predictive output and explanations into maintenance categories. Rather than treating probability as an automatic replacement command, the system can establish risk thresholds.

For example, a low-risk SSD may remain under normal monitoring, a medium-risk SSD may receive enhanced observation, and a high-risk SSD may trigger inspection or planned replacement. This preserves human oversight and avoids equating a model prediction with a deterministic physical failure.

The decision layer is particularly important because explainability has little value if explanations are disconnected from engineering actions. The goal is therefore not merely to make AI predictions visible but to make them actionable.

RESULTS

The conceptual analysis produces several principal findings concerning transparent SSD failure prediction. First, interpretability is most effective when integrated throughout the prediction pipeline rather than appended after model development. A model that produces accurate predictions without identifying the evidence behind those predictions creates a substantial operational gap. The framework therefore treats explanation as a parallel analytical layer connecting model output with maintenance interpretation.

Second, local and global explanations provide complementary forms of information. LIME is particularly appropriate for investigating individual high-risk SSDs because it can approximate the model around a specific observation. SHAP, by contrast, supports both individual explanations and population-level feature analysis. Their combination therefore addresses two distinct engineering questions: "Why was this device classified as high-risk?" and "Which indicators generally influence the model's decisions?" This combined interpretation is consistent with the transparency objective emphasized by Kumar (2026).

Third, temporal feature construction emerges as a critical methodological requirement. SSD health cannot be fully represented through static measurements when degradation is progressive. Changes, trends, and event frequencies may

contain information that absolute measurements do not capture. The relevance of temporal modeling is theoretically supported by the supplied neural-dynamics and event-driven learning literature (Gerstner et al., 2014; Wei, 2024). Consequently, a transparent prediction framework should explain not only current feature values but, where possible, changes in those features.

Fourth, model complexity should be justified through measurable predictive benefit. The supplied literature demonstrates the capabilities of increasingly sophisticated AI systems, including large language models, multimodal models, reasoning systems, and specialized neural architectures (Achiam, 2023; Anil, 2023; Guo, 2025). However, SSD failure prediction represents a narrower domain. Therefore, the most complex architecture is not necessarily the most appropriate architecture. A simpler model with stable explanations may provide greater operational value than a substantially more complex model with only marginal predictive improvement.

Fifth, explanations must not be interpreted as causal relationships. If SHAP identifies temperature as an influential variable, this demonstrates that temperature contributed to the model's prediction; it does not independently establish that temperature caused the physical failure. Correlated features may distribute predictive contribution among themselves, while model-specific biases may influence explanation values.

Finally, the framework reveals a central trade-off between predictive power and transparency. High-capacity models may capture nonlinear relationships more effectively, while simpler models may be easier to inspect directly. LIME and SHAP reduce this transparency gap but do not eliminate it. The findings therefore support a layered design in which predictive performance, explanation fidelity, explanation stability, and operational usefulness are evaluated jointly rather than optimizing predictive accuracy alone.

DISCUSSION

The findings indicate that transparent SSD failure prediction should be understood as a systems-engineering problem rather than simply a model-selection problem. The central challenge is to establish a defensible relationship between raw device observations, predicted risk, explanation, and maintenance action. Explainable AI techniques provide an important mechanism for constructing this relationship, but their value depends on how they are incorporated into the complete workflow.

The first theoretical implication concerns the distinction between predictive representation and human

interpretation. Modern AI research demonstrates that increasingly complex models can learn sophisticated representations (Devlin et al., 2019; Achiam, 2023; Anil, 2023). However, the internal representation learned by a high-capacity model is not necessarily equivalent to an explanation that an engineer can use. LIME and SHAP therefore serve as translation mechanisms between computational representations and human-understandable evidence.

The second implication concerns temporal reasoning. The supplied literature on neuronal dynamics and event-driven computation supports the broader idea that temporal structure can be computationally meaningful (Gerstner et al., 2014; Maass, 1997; Wei, 2024). SSD failure prediction can benefit from the same conceptual principle because deterioration often manifests as a trajectory. An explanation indicating that an error indicator has increased substantially over a defined period may be more informative than an explanation that simply reports its current value.

There is, however, a significant trade-off. Increasing model complexity may improve predictive discrimination but simultaneously increase explanation difficulty. The literature on DeepSeek, S1, QFFT, and MathCoder2 demonstrates a broader movement toward adaptive reasoning and computational scaling (Guo, 2025; Liu, 2025; Muennighoff, 2025; Lu, 2024). In an SSD environment, such complexity should be introduced only where it contributes measurable value. Storage monitoring frequently requires repeatable, low-latency decisions, making computational efficiency and stability important constraints.

A further limitation involves correlated variables. SSD-health indicators may not be statistically independent. When multiple features encode related aspects of device deterioration, attribution methods can distribute importance among them. This can make explanations appear more precise than the underlying evidence warrants. Accordingly, feature dependence analysis and explanation stability testing should accompany SHAP and LIME outputs.

Another limitation is distribution shift. SSD models trained using one device family, workload pattern, firmware configuration, or operational environment may not generalize reliably to another. A model can therefore remain highly accurate during internal validation while becoming less reliable under changed deployment conditions. Transparent explanations do not automatically solve this problem. They can reveal what the model is using, but they cannot guarantee that the learned relationships remain valid outside the training distribution.

The distinction between prediction and causation is consequently fundamental. A transparent model should explain why it predicts failure, but engineers should avoid interpreting the explanation as a definitive physical diagnosis. This is particularly important in safety- or availability-sensitive

environments.

Kumar (2026) provides a direct foundation for using LIME and SHAP to improve transparency in SSD failure prediction. The present framework broadens that concept by positioning explainability alongside temporal feature engineering, model comparison, explanation validation, and maintenance decision support. This creates a more complete methodology for trustworthy deployment.

Overall, transparency should be evaluated through predictive accuracy, explanation fidelity, stability, computational efficiency, and maintenance usefulness simultaneously. Optimizing only one of these dimensions can produce a system that is statistically effective but operationally unsuitable.

CONCLUSION

This research presented a transparent framework for SSD failure prediction through interpretable artificial intelligence techniques. The framework integrates temporal SSD-health representation, predictive modeling, LIME-based local explanations, SHAP-based feature attribution, explanation validation, and maintenance-oriented decision support.

The analysis demonstrates that explainability can reduce the practical gap between machine-generated risk predictions and engineering decisions. LIME provides localized explanations suitable for individual-device investigation, whereas SHAP supports both local interpretation and broader analysis of feature contributions. Their combined use can make complex predictive systems more transparent, provided that explanations are treated as descriptions of model behavior rather than causal proof.

The study also emphasizes that temporal structure, model complexity, feature dependence, and distribution shift must be considered when designing a reliable prediction system. Advanced AI architectures provide powerful representational capabilities, but their complexity must be justified against the transparency and operational requirements of storage monitoring.

The principal contribution is therefore a methodological perspective in which prediction and explanation are designed together. Consistent with the SSD-focused explainability direction identified by Kumar (2026), transparent failure prediction should not merely answer whether an SSD is likely to fail; it should provide structured evidence describing which observed characteristics contributed to that prediction and how strongly they influenced the model.

Future research should validate the framework on longitudinal SSD datasets containing diverse device

models, workloads, environmental conditions, and verified failure events. Comparative evaluation of interpretable baseline models and high-capacity architectures would further clarify the relationship between predictive performance and explanation quality. Future investigations should also examine explanation stability under distribution shifts and determine how effectively maintenance engineers can use model explanations to reduce unexpected failures and optimize replacement scheduling.

REFERENCES

1. J. Achiam, "GPT-4 technical report," 2023, arXiv:2303.08774.
2. R. Anil, "Gemini: A family of highly capable multimodal models," 2023, arXiv:2312.11805.
3. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, vol. 1, 2019, pp. 1–8.
4. W. Gerstner, W. M. Kistler, R. Naud, and L. Paninski, *Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition*. Cambridge, U.K.: Cambridge Univ. Press, 2014.
5. D. Guo, "Deepseek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning," 2025. [Online]. Available: <https://api.semanticscholar.org/CorpusID:275789950>
6. A. Liu, "Deepseek-v3 technical report," 2024, arXiv:2412.19437.
7. C. Liu, S. D. Zhang, and R. Jabbarvand, "Codemind: A framework to challenge large language models for code reasoning," 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:267682042>
8. W. Liu, "QFFT, question-free fine-tuning for adaptive reasoning," 2025. [Online]. Available: <https://api.semanticscholar.org/CorpusID:279402374>
9. Z. Lu, "MathCoder2: Better math reasoning from continued pretraining on model-translated mathematical code," 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:273233982>
10. W. Maass, "Networks of spiking neurons: The third generation of neural network models," *Neural Netw.*, vol. 10, no. 9, pp. 1659–1671, 1997.

11. R. Qiao, "We-math: Does your large multimodal model achieve human-like mathematical reasoning?," 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:270870136>
12. J. Pei, "Towards artificial general intelligence with hybrid Tianjic chip architecture," *Nature*, vol. 572, no. 7767, pp. 106–111, 2019.
13. H. Touvron, "LLAMA: Open and efficient foundation language models," 2023, arXiv:2302.13971.
14. W. Wei, "Event-driven learning for spiking neural networks," 2024, arXiv:2403.00270.
15. J. Wu, "Progressive tandem learning for pattern recognition with deep spiking neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7824–7840, Nov. 2021.
16. X. Xing, "SpikeLM: Towards general spike-driven language modeling via elastic bi-spiking mechanisms," in *Proc. 41st Int. Conf. Mach. Learn.*, 2024, pp. 54698–54714.
17. R.-J. Zhu, Q. Zhao, G. Li, and J. K. Eshraghian, "SpikeGPT: Generative pre-trained language model with spiking neural networks," 2023, arXiv:2302.13939.
18. N. Muennighoff, "S1: Simple test-time scaling." 2025. [Online]. Available: <https://api.semanticscholar.org/CorpusID:276079693>
19. Kumar, S. K. (2026). Explainable AI for SSD Failure Prediction: Using LIME and SHAP for Transparency. *Journal of Engineering Research and Sciences*, 5(4), 1–16. <https://doi.org/10.55708/js0504001>