УДК 004.312:004.052:004.85

М. І. Главчев
В. І. Носков
М. О. Середенко

МЕТОД ПРЕДИКТИВНОЇ ДІАГНОСТИКИ ТВЕРДОТІЛЬНИХ НАКОПИЧУВАЧІВ НА ОСНОВІ АНАЛІЗУ ДИНАМІКИ АТРИБУТІВ S.M.A.R.T.

Національній технічний університет «Харківський політехнічний інститут»

***Анотація**. У роботі запропоновано метод предиктивної діагностики твердотільних накопичувачів (SSD), спрямований на раннє виявлення деградації пам'яті. Метод базується на комплексному аналізі динаміки атрибутів S.M.A.R.T. із використанням алгоритмів машинного навчання, зокрема LightGBM. Розроблено набір похідних ознак, що враховує швидкість та прискорення зміни параметрів, а також статистичні метрики. Застосування методів відбору ознак дозволило скоротити вхідний вектор даних та підвищити точність прогнозування (AUC-ROC 0.912), що сприяє підвищенню надійності зберігання даних у корпоративних системах*

Ключові слова: Твердотільні накопичувачі (SSD), технологія S.M.A.R.T., предиктивна діагностика, машинне навчання, LightGBM, інженерія ознак, надійність даних.

***Abstract.** The paper proposes a method for predictive diagnostics of solid-state drives (SSDs) aimed at the early detection of memory degradation. The method is based on a comprehensive analysis of S.M.A.R.T. attribute dynamics using machine learning algorithms, specifically LightGBM. A set of derived features has been developed, taking into account the rate and acceleration of parameter changes, as well as statistical metrics. The application of feature selection methods enabled the reduction of the input data vector and increased prediction accuracy (AUC-ROC 0.912), contributing to enhanced data storage reliability in enterprise systems*

Keywords: Solid-state drives (SSDs), S.M.A.R.T. technology, predictive diagnostics, machine learning, LightGBM, feature engineering, data reliability.

Вступ

Масовий перехід обчислювальних систем на твердотільні накопичувачі (SSD) та пристрої з інтерфейсом NVMe забезпечив істотне підвищення продуктивності зберігання даних: швидкодія сучасних NVMe SSD досягає 7000 MB/c для читання та 5000 MB/c для запису, що на порядок перевищує можливості традиційних HDD. Проте, особливості фізичної організації NAND Flash пам'яті (обмежена кількість циклів запису/стирання, можливість раптових відмов контролера, деградація комірок) створюють нові виклики для забезпечення надійності зберігання даних [1].

За даними досліджень компанії Backblaze (провайдер хмарного резервного копіювання з парком понад 200,000 накопичувачів), річний рівень відмов SSD становить 0.58–1.05% залежно від моделі та умов експлуатації [2]. У корпоративному середовищі втрата критично важливих даних через несподіваний відказ накопичувача призводить до значних фінансових збитків: прямі витрати на відновлення даних ($7,900–$24,000 за інцидент), втрачений робочий час персоналу, репутаційні ризики, можливі штрафи за порушення SLA.

Технологія S.M.A.R.T. (Self-Monitoring, Analysis and Reporting Technology) надає інформацію про стан накопичувача через набір атрибутів (Reallocated Sectors Count, Program/Erase Fail Count, SSD Life Left, Temperature та інші). Традиційні підходи до діагностики базуються на порогових значеннях окремих атрибутів, що призводить до високого рівня хибних спрацьовувань (50–70% за даними досліджень Google) та пропуску реальних відмов. Сучасні методи машинного навчання дозволяють аналізувати комплексні патерни в динаміці атрибутів S.M.A.R.T. та виявляти ранні ознаки деградації накопичувача.

Мета роботи – розробка методу предиктивної діагностики твердотільних накопичувачів, що забезпечує раннє виявлення ознак деградації на основі комплексного аналізу динаміки атрибутів S.M.A.R.T. із використанням алгоритмів машинного навчання (Random Forest, Gradient Boosting).

Архітектура системи та метод прогнозування

Система предиктивної діагностики реалізована як багаторівнева архітектура (рис. 1), що включає рівень збору даних (моніторинг атрибутів S.M.A.R.T. через smartmontools/nvme-cli, збереження історичних даних у SQLite), рівень обробки та аналізу (очищення даних, виявлення аномалій, обчислення похідних ознак), рівень машинного навчання (ансамблеві моделі Random Forest та LightGBM для класифікації та регресії) та рівень представлення результатів (GUI на PyQt6, візуалізація трендів, генерація рекомендацій).

Рис. 1. Архітектура системи предиктивної діагностики

Інженерія ознак. Для підвищення якості прогнозування розроблено набір похідних ознак на основі ба зових атрибутів S.M.A.R.T.:

• Швидкість зміни атрибутів (перша похідна): $\Delta A_i = (A_i(t) - A_i(t-\Delta t)) / \Delta t$, де A_i – значення і-го атрибуту, Δt – інтервал між вимірами (24 години).

• Прискорення деградації (друга похідна): $\Delta^2 A_i = (\Delta A_i(t) - \Delta A_i(t-\Delta t)) / \Delta t$ – дозволяє виявити нелінійні патерни прискореної деградації.

• Статистичні ознаки за вікном: середнє, стандартне відхилення, max/min за останні 7, 14, 30 днів.

• Композитний індекс деградації: $DI = \Sigma\, w_i \times norm(A_i)$, де w_i – ваги, визначені з важливості ознак Random Forest, norm() – min-max нормалізація.

Відбір інформативних ознак. Для оптимізації набору ознак застосовано комбінацію методів:

• Фільтрація за дисперсією – видалення ознак з variance < 0.01.

• Кореляційний аналіз – виключення пар ознак з $|r| > 0.95$ для зменшення мультиколінеарності.

• Recursive Feature Elimination (RFE) – ітеративне виключення найменш важливих ознак з валідацією на крос-валідації.

Кореляційний аналіз (рис. 2) виявив сильні зв'язки між Reallocated Sectors Count та Program Fail Count (r=0.71), що дозволяє виключити один з атрибутів без втрати інформації. Після застосування всіх методів відбору кількість ознак скорочена з 142 (48 базових атрибутів × 3 типи похідних) до 37 найбільш інформативних, що покращило AUC-ROC з 0.894 до 0.912 та скоротило час навчання на 63%.

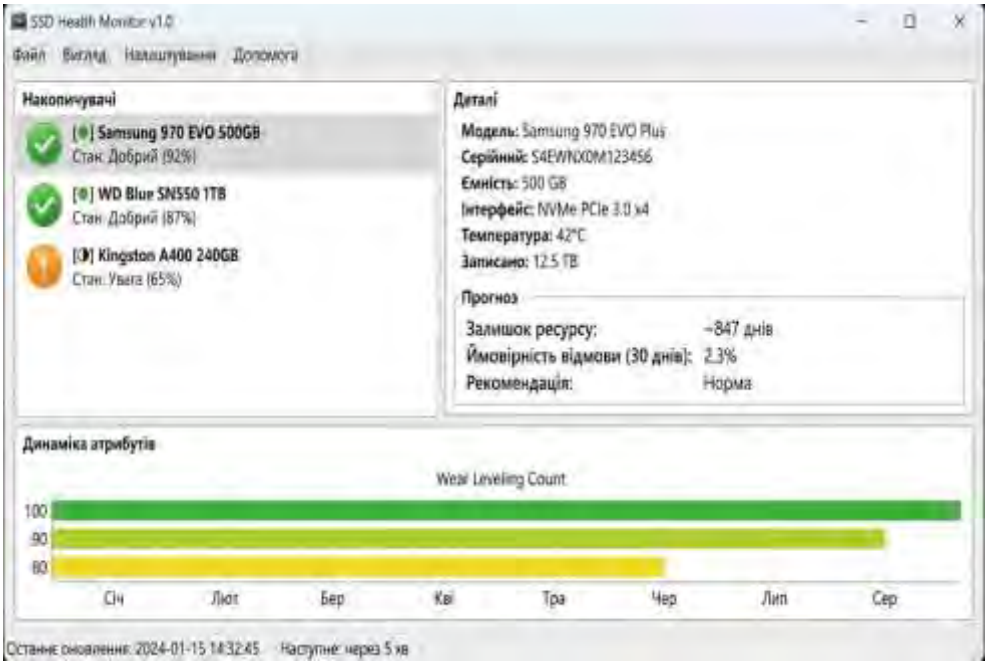

Рис. 2. Теплова карта кореляції між ключовими атрибутами S.M.A.R.T.

Таблиця 1 – Ключові атрибути S.M.A.R.T. для предиктивної діагностики SSD

ID	Атрибут	Опис	Важливість
5	Reallocated Sectors Count	Кількість перепризначених секторів	**Критична**
177	SSD Life Left	Залишок ресурсу накопичувача (%)	**Критична**
181	Program Fail Count	Кількість помилок програмування	Дуже висока
182	Erase Fail Count	Кількість помилок стирання	Дуже висока
194	Temperature	Температура накопичувача	Висока

Алгоритм машинного навчання. Для задачі прогнозування обрано алгоритм LightGBM (Light Gradient Boosting Machine) – оптимізовану реалізацію градієнтного бустингу, що забезпечує високу швидкодію та якість прогнозування. Модель навчена на датасеті Backblaze (2019–2023 роки, 2.8 млн записів, 47 атрибутів S.M.A.R.T.) з використанням TimeSeriesSplit крос-валідації для запобігання витоку інформації з майбутнього.

Висновки

1. Розроблено багаторівневу архітектуру системи предиктивної діагностики SSD/NVMe накопичувачів, що інтегрує збір даних S.M.A.R.T., їх обробку та аналіз, прогнозування на основі машинного навчання та візуалізацію результатів.

2. Створено набір похідних ознак (швидкість зміни, прискорення деградації, статистичні метрики за вікном, композитний індекс), що дозволило підвищити якість прогнозування на 12% порівняно з використанням лише базових атрибутів S.M.A.R.T.

3. Застосування комбінованих методів відбору ознак (фільтрація за дисперсією, кореляційний аналіз, RFE) дозволило скоротити розмірність простору ознак з 142 до 37 при збереженні високої якості моделі та зменшенні часу навчання на 63%.

4. Обрано та налаштовано модель LightGBM, що забезпечує оптимальний баланс між точністю прогнозування та швидкодією для задачі предиктивної діагностики в умовах незбалансованих даних (співвідношення класів 1:12.3).

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Schroeder B., Lagisetty R., Merchant A. Flash Reliability in Production: The Expected and the Unexpected // Proceedings of the 14th USENIX Conference on File and Storage Technologies (FAST '16). 2016. P. 67–80.

2. Backblaze Hard Drive Stats. Backblaze Inc. URL: https://www.backblaze.com/b2/hard-drive-test-data.html (дата звернення: 15.01.2026).

3. Ke G., Meng Q., Finley T. et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree // Advances in Neural

Information Processing Systems 30 (NIPS 2017). 2017. P. 3146–3154.
4. Shibata N., Maejima H., Kanda K. et al. A 19nm 112.8mm² 64Gb Multi-Level Flash Memory with 400Mb/s/pin 1.8V Toggle Mode Interface // IEEE Journal of Solid-State Circuits. 2012. Vol. 47, No. 1. P. 119–126.
5. Murray J. F., Hughes G. F., Kreutz-Delgado K. Machine learning methods for predicting failures in hard drives: A multiple-instance application // Journal of Machine Learning Research. 2005. Vol. 6. P. 783–816.

Главчев Максим Ігорович - кандидат економічних наук, доцент, професор кафедри комп'ютерної інженерії та програмування, Національний технічний університет «Харківський політехнічний інститут», м.Харків, e-mail: Maksym.Glavchev@khpi.edu.ua.
Носков Валентин Іванович - доктор технічних наук, професор, професор кафедри комп'ютерної інженерії та програмування, Національний технічний університет «Харківський політехнічний інститут», м.Харків, e-mail: Valentyn.Noskov@khpi.edu.ua.
Середенко Олег Сергійович – студент групи КН-Н924а, кафедра комп'ютерної інженерії та програмування, Національний технічний університет «Харківський політехнічний інститут», м.Харків, e-mail: Oleh.Seredenko@cit.khpi.edu.ua

Maksym I. Glavchev – Candidate of Economic Sciences (Ph.D.), Associate Professor, Professor at the Department of Computer Engineering and Programming, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, e-mail: Maksym.Glavchev@khpi.edu.ua.
Valentyn I. Noskov – Doctor of Technical Sciences, Professor, Professor at the Department of Computer Engineering and Programming, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, e-mail: Valentyn.Noskov@khpi.edu.ua.
Oleh S. Seredenko – Student of group KN-H924a, Department of Computer Engineering and Programming, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, e-mail: Oleh.Seredenko@cit.khpi.edu.ua.