Secure Systems Design and Technology Development. Veliko Tarnovo, Bulgaria
Proceedings of the International Scientific Conference, 2026. Volume III, 279-283

Methods for Data Arhiving and Recovery Using Artificial Intelligence

Elena Shtereva
Naval Academy "Nikola Y. Vaptsarov"
Varna, Bulgaria
73Vasil Drumev St.
eshtereva@naval-acad.bg
ORCID: 0009-0008-1799-7106

***Abstract* - In the face of digital renewal, data is one of the most valuable stockpiles for any organization. They not only facilitate operational activity, but also play an important role in strategic decision-making. This data can include everything from clinical trials and client databases, to reports and internal communications. Through data processing, companies can identify weak links in their operations and offer solutions to improve them.**

Incorporating artificial intelligence into data archiving processes represents a big step in information management. Methods using artificial intelligence are automated, intelligently indexed, and offer search, which significantly increases the efficiency and accuracy of archival systems.

The report presents methods that compare the evolution and improvement of data archiving and recovery processes through artificial intelligence, while addressing potential challenges and limitations. An assessment of the risk of data loss in artificial intelligence systems was made using the Weibull model to predict failures. A Bayesian distribution flowchart has been developed, demonstrating the efficiency of a data backup system using implemented AI.

Keywords - Artificial Intelligence (AI), Data Backup and Recovery, Optimization.

I. INTRODUCTION

In a time of accelerated development of information technologies and constantly increasing volumes of data, maintaining their reliable protection and storage becomes a top priority. Data backup plays a key role in ensuring its integrity and the ability to recover information in case of incidents. Therefore, the choice of appropriate methods for data backup is essential for the effective functioning of information systems.

Differential backups accumulate large backups over time because it copies all the changes from the last full backup. Recovery is slow and complicated because it requires a sequential restore from multiple backups.

The same increase in disk space used is also seen with incremental backups, which require multiple small backups and make backups complex. With incremental backups, recovery is faster, but over time, the backups become larger. Damage or loss of intermediate backups makes data recovery incomplete or impossible. Often, there is a lack of reliable verification of the integrity of backups.

Conventional backup methods have been proven over time, but with advances in technology, big data, and fast recovery requirements, they are becoming less and less effective. Traditional methods are difficult to adapt to cloud environments, dynamic infrastructures, and big data. Limited flexibility and scalability make it difficult to quickly restore individual files or versions. This necessitates the introduction of innovative approaches to ensure better optimization, reliability, and flexibility.

AI-assisted data backup changes classic approaches by adding intelligence, automation, and adaptability to backup and backup systems. Backup focuses on important data, which optimizes resources and speeds up the process.

II. APPROACHES TO DATA BACKUP AND RECOVERY USING ARTIFICIAL INTELLIGENCE.

A. Dynamic archive management with integrated AI

Intelligent data classification and prioritization using artificial intelligence analysis data and determines which files or groups of data are critical, frequently used, or strategically important. Machine learning can recognize commonly used and important files that need to be backed up periodically.

The result of using artificial intelligence in data backup is increased efficiency in optimizing resources, automating processes, higher reliability in automatic detection and error correction. Fast data recovery with intelligent selection of versions, priorities and security through early

Online ISSN 3044-7771
https://doi.org/10.68302/std2026.vol3.82

threat detection with quick responses.[11] Automatic file classification and prioritization uses machine learning, with models such as Random Forest, Support Vector Machine (SVM) or neural networks to analyse the metadata and contents of the files. [16] Critical documents are automatically marked and a more frequent schedule for their backup is set, and for unimportant or old files, archiving is done less frequently. Time series and forecasting models, Long Short-Term Memory (LSTM) or neural networks, analyse system load and data usage to predict when backups will take place without disrupting users. Backups are scheduled at night or during periods of low activity, but dynamically adapt to changes in workload. [5] To protect the archive from anomalies, algorithms, such as Isolation Forest or Autoencoders, are used to detect unusual patterns in the archive files that trigger a recovery process from an older archive. AI analysis file types and content to choose the most efficient method of compression and deduplication. [1] With built-in AI modules - Cohesity and Rubrik, as well as Z standard, LZMA, or Brotli algorithms with machine learning, the system dynamically selects compression with a higher text ratio and faster video speed, which reduces time and space used. When restoring via a chatbot using Natural Language Processing (NLP), models such as GPT or BERT are applied, which allow you to restore the last created archive and automatically start the process. When using a GPT model, it is not necessary to determine the location of the archive in advance. To implement such solutions, tools such as the OpenAI GPT API, Dialogflow, and Microsoft Bot Framework are used.

B. Backup methods and AI data recovery systems

A study of archiving methods using artificial intelligence has been carried out, which allow efficient backup through intelligent compression, storage optimization and improved accuracy in digitization of archives. The results of the study are summarized in Table 1, presenting the potential problems and limitations associated with the implementation of such technologies.

TABLE 1 METHODS OF ARCHIVING USING ARTIFICIAL INTELLIGENCE.

Methods and research	***Indicator***	***Results***
Study of data compression with AI	High compression	~90% compression using AI [10],[15]
Perceptron-parallelized data compression	Indicative framework with three-layer perceptrons (machine learning) and parallelization within blocks	Up to 17.78% improvement in speed compression/ratio over SZ2.1 PW_REL in experiments [7]
3D and multimedia archives	Collaborative quantization with common codes and conditional decoding from photos	Compresses 3D scenes to the kilobyte range with better compression and adaptability for updates [9]
Model data trimming	Server Merge and Separation of Local Models for Parameter Change and Communication	Achieves very high compression rates (>99.7%) with minimal quality loss (<1.18%) in IoT/FL scenarios [2]
Optimizing Forest Data	Predictive accuracy	R^2 = 0.9965 (train), R^2 = 0.9520 (test) for cloud storage performance optimization.[14]
Media digitization of an archive	Digitization accuracy	Improved from 92.3% output level to 95.7% after setup. [3]

The statistical indicator R^2 in Table 1 is applied to measure how well a model explains the variation in the observed data.

The AI analyses the version of the files using contextual data and the most appropriate version for recovery. An example of this is documents with multiple revisions, of which the AI suggests the most relevant version according to user demand or criticality. The AI detects abnormal activity in the archives, which is an indication of malware. It automatically triggers a restore from the last healthy backup and prevents data loss. Examples of AI software bundles that detect threats in data backup and recovery are listed in Table 2.

TABLE 2. EXAMPLES OF AI BUNDLES.

Bundle	***Applications of artificial intelligence***
Rubrik	Automate backup and recovery, analyze risks and priorities with machine learning [12]
Cohesity	Intelligent deduplication and cybersecurity with artificial intelligence [4]
Veeam (с AI модули)	Analytics and automation of backup processes, anomaly detection [13]
IBM Spectrum Protect	Artificial Intelligence for Data Management and Backup Optimization in Hybrid Clouds
Commvault	Manage backups and backups using AI

From the data presented in Tables 1 and 2, the advantages of using artificial intelligence over standard data archiving methods can be summarized:

- The reliability and integrity of archived data is increased by using automatic restoration or the creation of new backups without human intervention.
- Artificial intelligence optimizes deduplication processes by intelligently recognizing duplicate or similar data. Compression adapts according to data type, achieving better results [10].
- Integration of chatbots that, through natural language, allow users to request the restoration of specific files or versions.

C. Monitoring data in systems with integrated AI

When an incident occurs in a backup system, the AI consolidates hundreds of alerts, speeding up analysis by 96%. Behavioral profiling in Extended Detection and Response reduces false alerts by 89%, and automated AI scenarios master ransomware 18 times faster than manual response. As a result, the average time to detect a problem or incident and the average response and recovery time after an incident are reduced from hours to minutes. [17],[20],[22],[23]

AI systems predict when is the best time to backup. A calculation based on Bayes' theorem (1) is applied in combination with time series and anomaly analysis. [18]

The Bays formula has the form (1):

$$\boldsymbol{P}\left(\frac{\boldsymbol{C}}{\boldsymbol{D}}\right) = \frac{\mathbf{P(D/C) * P(C)}}{\boldsymbol{P(D)}}, \tag{1}$$

where:

$\mathbf{P}(\boldsymbol{C}/\mathbf{D})$ (posterior probability), is the probability that hypothesis C is correct after observing data D (in case of anomalies or risk of loss);

$\mathbf{P}(\boldsymbol{D}/\mathbf{C})$ (likelihood), is the probability of occurrence of the observed data D if there are anomalies or risk of loss, in hypothesis C;

P(C) (prior probability) is a preliminary probability of hypothesis C before presenting the data (based on the history of the frequency of failures and the risk of loss);

P(D) (marginal probability) is the total probability for the D data;

All of these quantities are dimensionless and take values between 0 and 1 (or between 0% and 100% if expressed as a percentage).

The risk of data loss is assessed based on usage patterns and potential failures. The initial probability P(C) changes as new information enters the archive system, allowing dynamic management of archives to minimize the impact of system failures.[19],[21] The prediction of failures in it is based on the Weibull distribution model (2), which uses statistics on time to failure to estimate the probability and frequency of future failures. [22]:

$$\boldsymbol{f(t)} = \frac{\boldsymbol{\beta}}{\boldsymbol{\eta}}\left(\frac{\boldsymbol{t}}{\boldsymbol{\eta}}\right)^{\boldsymbol{\beta-1}} \boldsymbol{e}^{-(\boldsymbol{t}/\boldsymbol{\eta})^{\boldsymbol{\beta}}}, \tag{2}$$

where:

t is the time to failure

η is scale parameter (average time to failure)

β is shape parameter (determines the shape of failures, accidental failures).

D. Results and discussion of the research done

Using the Weibull model helps administrators predict future hardware failure and save information in time. No time is wasted on too frequent backups of robust systems, and efforts are focused where the statistical model predicts a problem. When a predefined threshold is reached, the system automatically initiates a backup of critical data, migrates it to more reliable media, and takes risky devices out of service. This avoids unnecessary backups of stable systems and concentrates resources on components with the highest probability of failure.

Figure 1 shows a flowchart of an algorithm linking the logic and communication of archiving processes to the use of Weibull's model. The algorithm follows the steps:

1. The administrator configures the backup threshold.
2. The monitoring system sends data to the Weibull model.
3. The Weibull model returns a probability of failure.
4. If the risk is high→ Backup Manager starts a backup.
5. Storage devices receive the command to transfer and decommission.

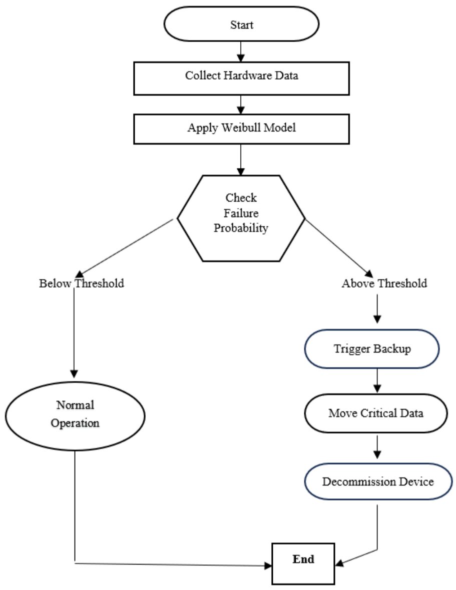

Fig. 1. Flowchart of an archiving algorithm according to the Weibull model

Table 3 presents calculations with a Bases classification of data backup with a traditional method and with AI deployed over a period of 1 year. The efficiency of the backup system can be evaluated through indicators such as backup frequency, resource utilization, recovery time and resistance to data loss. Within a one-year period, there are significant differences between traditional reactive systems and intelligent proactive solutions based on predictive models.

TABLE 3. COMPARISON OF THE PERFORMANCE OF BACKUP SYSTEMS OVER A PERIOD OF 1 YEAR

Indicator	***Traditional backup***	***AI - Based Backup***
A priori probability P(C)	0.7	0.3
Probability of correct classification P(D/C)	0.1	0.8
Accuracy of classification P(C/D),conditional probability	0.226	0.774
Paramet η (Weibull)	8 years	10 years
Parameter β	1.5	1.5
System reliability after 1 year R(t)	0.957	0.976
Performance Index E	0.216	0.756

To calculate the reliability of a backup system for 1 year, Weibull's formula was used:

$$\boldsymbol{R(t) = e^{-(t/\eta)^{\beta}}} \tag{3}$$

where:
t - time of use of the backup system;
β - increasing probability of failure of the backup system;
η - average time to failure of the backup system.
For the calculation of **R(t)** according to formula (3) the parameters are given the following values: **t** = 1 year; average time to failure for traditional systems **η** = 8 years, average time to failure for artificial intelligence systems
η = 10 years; $\boldsymbol{\beta}$ = 1.5
The calculated value of the reliability of the traditional backup system for 1 year is **R (1)** =0.957, and for an AI system – **R (1)** =0.976. The AI system is 1.9% more reliable than the traditional one.
The calculation of the efficiency index E, which combines the accuracy of the model with the temporal reliability of a backup system, is performed according to formula (4). P(C|D) is a conditional probability reflecting the accuracy of the classification.

$$\boldsymbol{E = P(C|D)xR(t)} \tag{4}$$

where:
$P(C_1)$ = 0.7
$P(C_2)$ = 0.3
P(D|C_1) = 0.1
P(D|C_2) = 0.8
P(D) = 0.8×0.3+0.1×0.7 -> P(D) = 0.31

P(C_1|D) $= \frac{0.1X0.7}{0.31}$ -> P(C_1|D) = 0.226

P(C_2|D) $= \frac{0.8X0.3}{0.31}$ -> P(C_1|D) = 0.774

For the traditional archiving system, the value of E=0.226x0.957=0.216, and for the archiving system with integrated artificial intelligence: E=0.774x0.976=0.756. The results obtained show that the archiving system with integrated AI technology is 52% more efficient than the traditional archiving method.

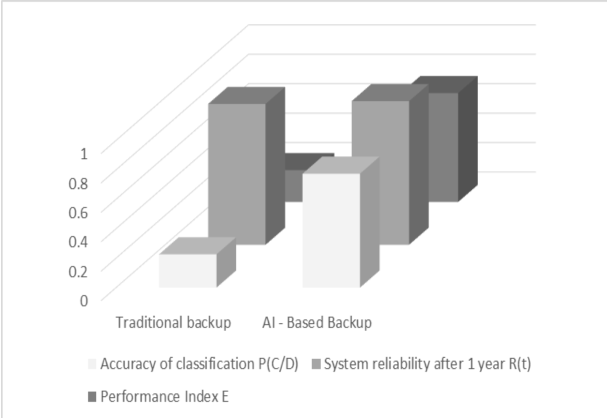

Fig. 2. Result of system research within 1 year

From the diagram in Figure 2, it can be seen that an archiving system with integrated AI is more efficient than a system using archiving with traditional methods. While both systems manage to reliably preserve data after a year of operation, the traditional system has a serious problem with its proper classification. An archiving system with integrated AI solves this problem, providing both high reliability and high accuracy of classification, which logically leads to its high efficiency index.

III. CONCLUSION

Traditional backup methods, although proven over time, turn out to be insufficiently effective in the conditions of dynamic infrastructures, cloud environments and the need for quick access to specific data. Their limited flexibility and scalability require the search and implementation of modern solutions. In this sense, the use of artificial intelligence has established itself as a key factor for the modernization of data archiving processes. The research on predicting failures in archiving systems demonstrates the potential of the proposed approach for improving reliability and supporting data management processes. By assessing the probability of failure, the backup system allows taking preventive actions such as archiving critical information, migrating to more secure media, and decommissioning risky devices. Fault forecasting helps in a more rational use of resources and reduces unnecessary data backup operations. The result is a balance between efficiency and reliability of the infrastructure. Future research in this area can be aimed at studying an archiving system with the integration of machine learning methods to improve the accuracy of Weibull model predictions through real-time analysis. It is also possible to implement adaptive algorithms that dynamically change the threshold values according to the load conditions and the backup period. Monitoring of the backup system with risk visualization will help for automated decision-making through artificial intelligence.

Integration with cloud infrastructure will allow scalable backups, as well as faster recovery in case of incidents.

IV. ACKNOWLEDGMENTS

This report was financed by the Ministry of Education and Science under the National Science Program "Security and Defense", carried out in implementation of the National Strategy for the Development of Scientific Research 2017-2030 and adopted by Decision of the Council of Ministers No. 731 of October 21, 2021. The material reflects only the author's opinion and the Ministry of Education and Science is not responsible for the content.

REFERENCES

[1] J. S. Caicedo-Vivas and W. Alfonso-Morales, “Short-term load forecasting using an LSTM neural network for a grid operator,” Energies, vol. 16, no. 23, p. 7878, 2023, doi: 10.3390/en16237878.

[2] G. Colavizza et al., “Archives and AI: An overview of current debates and future perspectives,” Journal on Computing and Cultural Heritage, vol. 15, no. 1, pp. 1–15, 2022, doi: 10.1145/3479010.

[3] A. L. Cushing and G. Osti, “‘So how do we balance all of these needs?’: How the concept of AI technology impacts digital archival expertise,” Journal of Documentation, vol. 79, no. 7, pp. 12–29, 2023, doi: 10.1108/JD-08-2022-0170.

[4] A. U. Desai, “A review of the applications of machine learning in cybersecurity and its challenges,” Journal of Digital Security and Forensics, vol. 1, no. 1, 2024, doi: 10.29121/digisecforensics.v1.i1.2024.17.

[5] L. Han et al., “Power load forecast based on CS-LSTM neural network,” Mathematics, vol. 12, no. 9, p. 1402, 2024, doi: 10.3390/math12091402.

[6] J. Kim et al., “Key backup and recovery for resilient DID environment,” Mathematics, vol. 12, no. 6, p. 830, 2024, doi: 10.3390/math12060830.

[7] N. B. Malvi et al., “Sustainable data storage solutions using AI,” in Proceedings of a Scientific Conference on Data Technologies, 2025, pp. 321–340.

[8] McKinsey Global Institute, The State of AI in 2019. McKinsey & Company, 2019.

[9] T. Mehra, “AI-driven approach to advancing backup strategies and optimizing storage solutions,” International Journal of Scientific Research in Engineering and Management, vol. 8, no. 12, pp. 1–7, 2024, doi: 10.55041/IJSREM39778.

[10] T. Mehra, “Building a resilient backup infrastructure: Combining data redundancy, encryption, and RBAC for maximum protection,” International Journal of Scientific Research in Engineering and Management, vol. 9, no. 3, pp. 1–9, 2025, doi: 10.55041/IJSREM41975.

[11] E. E. S. Murphy and J. D. Mehr, “Intelligent tiers of backup data,” U.S. Patent EP2422274B1, May 6, 2020. [Online]. Available: https://data.epo.org/publication-server/rest/v1.2/publication-dates/20200506/patents/EP2422274NWB1/document.html

[12] P. Novak et al., “Multistage malware detection method for backup systems,” Technologies, vol. 12, no. 2, p. 23, 2024, doi: 10.3390/technologies12020023.

[13] L. Peng et al., “SecDedup: Secure data deduplication with dynamic auditing in the cloud,” Information Sciences, vol. 644, p. 119279, 2023, doi: 10.1016/j.ins.2023.119279.

[14] N. Saad, N. Elmqaddem, and E. Vasseur, “Artificial intelligence at the service of digital archiving: Optimization of practices and ethical challenges,” 2022.

[15] A. Sghir et al., “Advanced strategies for big data resource and storage optimization: An AI perspective,” International Journal of Advanced Computer Science and Applications, vol. 16, no. 8, 2025, doi: 10.14569/IJACSA.2025.0160896.

[16] D. Shailesh et al., “Systems and methods for prioritizing and detecting file datasets based on metadata,” U.S. Patent US11487825B1, 2022. [Online]. Available: https://patentimages.storage.googleapis.com/76/24/01/d500122b6890c1/US11487825.pdf

[17] A. Chagovec, T. Bakardjieva, A. Ivanova, F. Sapundzhi, V. Spasova, and A. Ivanova, “AI-driven threat detection and automated incident response for securing cloud workloads,” in Proceedings of the 6th International Electronic Conference on Applied Sciences, MDPI, Basel, Switzerland, Dec. 9–11, 2025. [Online]. Available: https://sciforum.net/paper/view/27875

[18] AI-Driven Autonomous Backup Systems for Proactive Disaster Recovery in Cloud Environments,” Budapest Journal of Computational Intelligence and Robotics, vol. 6, pp. 1–9, 2022

[19] Karthick R., “Advanced Threat Detection in Backup Systems: AI, Encryption, and RBAC,” *ESP Journal of Engineering Technology Advancements*, ICECT 2024.

[20] Microsoft, “SCI XDR Ebook: Supercharge your security operations with XDR,” Microsoft Security, 2022.

[21] R. Thomas and S. A. Fathima, “AI-Based Backup Security Solutions: Detecting Anomalous Activities and Strengthening Encryption,” *International Journal of Multidisciplinary on Science and Management*, ICETETI 2024, pp. 174–179.

[22] StoneFly Inc., “Threat Detection, Monitoring, and Ransomware Protection with AI-Based XDR and SIEM Platform,” *PRWeb*, May 2025.

[23] A. Author et al., “ZADIG: A Novel Extended Detection and Response System,” *IEEE Xplore*, 2025.