Industrial data-driven modeling for imbalanced fault diagnosis

Kuo-Yi Lin
Department of Industrial Engineering and Management, National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan, and
Thitipong Jamrus
Department of Industrial Engineering, Research Unit on System Modeling for Industry, Khon Kaen University, Khon Kaen, Thailand

Abstract

Purpose – Motivated by recent research indicating the significant challenges posed by imbalanced datasets in industrial settings, this paper presents a novel framework for Industrial Data-driven Modeling for Imbalanced Fault Diagnosis, aiming to improve fault detection accuracy and reliability.
Design/methodology/approach – This study addressing the challenge of imbalanced datasets in predicting hard drive failures is both innovative and comprehensive. By integrating data enhancement techniques with cost-sensitive methods, the research pioneers a solution that directly targets the intrinsic issues posed by imbalanced data, a common obstacle in predictive maintenance and reliability analysis.
Findings – In real industrial environments, there is a critical demand for addressing the issue of imbalanced datasets. When faced with limited data for rare events or a heavily skewed distribution of categories, it becomes essential for models to effectively mine insights from the original imbalanced dataset. This involves employing techniques like data augmentation to generate new insights and rules, enhancing the model’s ability to accurately identify and predict failures.
Originality/value – Previous research has highlighted the complexity of diagnosing faults within imbalanced industrial datasets, often leading to suboptimal predictive accuracy. This paper bridges this gap by introducing a robust framework for Industrial Data-driven Modeling for Imbalanced Fault Diagnosis. It combines data enhancement and cost-sensitive methods to effectively manage the challenges posed by imbalanced datasets, further innovating with a bagging method to refine model optimization. The validation of the proposed approach demonstrates superior accuracy compared to existing methods, showcasing its potential to significantly improve fault diagnosis in industrial applications.

Keywords Industrial data-driven modeling, Data imbalanced, Fault diagnosis

Paper type Research paper

1. Introduction

Industrial data-driven modeling for imbalanced fault diagnosis represents a significant challenge and opportunity within the realm of industrial maintenance and reliability engineering. The natural occurrence of faults or failures in numerous industrial processes is inherently skewed. Under these circumstances, the volume of normal operational data significantly overshadows the instances of system failures or anomalies. This discrepancy introduces a distinctive hurdle for models based on data-driven fault diagnosis (Diaz and Ardalan, 2023). Such models may develop a bias towards the predominant class, thereby compromising the accurate identification of infrequent yet vital fault conditions. In particular, there are three reasons with considering the detail of fault diagnosis operations. Firstly, industrial systems are often complex and composed of multiple interconnected components. Each component might have different failure modes, contributing to the diversity and complexity of possible faults. This complexity makes it difficult to capture all fault conditions with a limited amount of failure data. Secondly, faults can manifest in various ways, even for the same component under similar conditions. This variability means that data-driven models need a broad spectrum of data to learn from, which can be challenging to
obtain for rare faults. Finally, some faults are more critical than others, posing significant risks to safety, productivity, and financial stability. Identifying and prioritizing these faults is crucial, but their rarity makes it challenging for models to learn to detect them without bias towards more frequent, less critical conditions.

To address this issue, specialized approaches and methodologies have been developed within the field of machine learning and data science. These techniques aim to enhance the sensitivity of fault diagnosis models towards the less represented fault conditions (Peng et al., 2022), thereby improving their overall effectiveness and reliability (Wu et al., 2021). Strategies such as oversampling of the minority class, undersampling of the majority class, synthetic data generation, and advanced algorithmic adjustments are commonly employed to counter the effects of data imbalance (Fan et al., 2020; Liu et al., 2021). In particular, magnetic disk drives are the least reliable component of modern computers, which offer the most cost-effective way to store large amounts of data. Fault diagnosis of hard disk is crucial in real industries, leading to enhancement of machinery reliability. As machine down, equipment fault can delay the entire production process and take much time for experts for troubleshooting. To safeguard against data loss, disk-based storage systems are equipped with redundancy measures like duplication and error correction codes. However, instead of merely reacting to disk failures, proactive strategies involve monitoring the SMART attributes of disks to identify those at higher risk of failure. By leveraging sensors and data logging technologies, engineers can amass historical data from numerous devices over specified periods. Deep learning models are then applied to this data, enabling precise fault diagnosis. This proactive approach allows for the preemptive backup of data from disks deemed at risk, significantly mitigating the potential for data loss. Research supports the effectiveness of such methods, including studies that employ recurrent neural networks for assessing disk health and predicting failures (Xu et al., 2016). Other researchers have explored the use of classification networks and regression trees for forecasting disk failures, further validating the practicality of predictive maintenance in disk-based storage systems.

Digital models, utilizing machine learning algorithms, have become crucial for simulating future states of various phenomena, aiding decision-making across multiple time horizons (Kusiak, 2020). In digital manufacturing, algorithms such as decision trees (Kusiak, 2017), clustering, and broader computational intelligence frameworks have been deployed. Nevertheless, these conventional methods sometimes falter, particularly with extreme cases, such as when labels are absent from data sets. Addressing this, Kang et al. (2018) introduced innovative methods to handle missing labels, a common issue in the digital realm.

The challenge of imbalanced datasets is inherent to the goal of enhancing stability by minimizing the failure probability. Imbalanced data issues arise when the sample size discrepancy between categories surpasses a 1:100 ratio. Traditional machine learning models often fall short in this area, as their classifiers are inclined to favor the majority class, potentially misleading users. The surge in research focused on the imbalanced learning problem is a testament to the urgency and complexity of the issue. Emerging approaches include sampling methods, cost-sensitive learning, kernel-based learning, and active learning strategies. However, each solution carries limitations: traditional under-sampling can lead to loss of information, though this is mitigated by Synthetic Minority Over-sampling Technique (SMOTE) which creates synthetic minority samples, and Informed Under-sampling which prevents the information loss seen in conventional methods. Moreover, cost-sensitive learning introduces different cost matrices to tailor to unbalanced datasets.

The research gap in this domain is the need for a cohesive framework that seamlessly integrates these various approaches while adhering to the core principles of machine learning and data science. Such a framework would address the current shortcomings by providing a robust, unified strategy for tackling imbalanced datasets, thereby enhancing the reliability and efficacy of digital models in practical applications. The proposed methods need to be
evaluated not just in isolation but as part of an interconnected system that appreciates the
nuances and intricacies of machine learning algorithms in the face of skewed data. The
development of this framework represents an exciting and necessary frontier in the field of
data science.

This study proposed an industrial data-driven modeling to addressing the challenges of
extremely imbalanced datasets, which are not adequately covered by existing literature. The
foundation of this approach is based on the real-world scenario where data centers frequently
update their storage systems with new disks to replace failed ones or to expand storage
capacity, leading to a diverse mix of disk brands and models. This diversity exacerbates the
imbalance issue, making conventional oversampling techniques potentially misleading due
to the generation of non-representative sample points and skewed classification outcomes.
Moreover, informed under-sampling in such contexts could lead to unevenly distributed
samples across different batches. To navigate these challenges, our methodology integrates
data augmentation and cost-sensitive strategies. Specifically, data augmentation is informed
by insights derived from imbalanced datasets utilizing the SMOTE, while the Cross-Entropy
loss function, enhanced with penalty factors for imbalanced categories, is employed to refine
the classification network. This dual approach not only mitigates the risk of misdiagnosis in
predicting hard disk failures but also tailors the model to accommodate the unique data
landscape of diverse disk brands and models by reconstructing the original dataset for
optimal machine learning model performance. The remainder of this study is organized as
follows. Section 2 includes the related works. The proposed method is introduced in Section 3,
and experimentally validated in Section 4. Section 5 concludes this study.

2. Related works
In the realm of industrial operations, the advent of data-driven technologies has
revolutionized the approach towards maintenance (Lu et al., 2021), reliability (Zhang et al.,
2023), and fault diagnosis (Wu et al., 2020). However, one of the significant challenges that
persist is the imbalanced nature of fault data. Typically, datasets are heavily skewed towards
normal operational data, with fault conditions being rare occurrences. This imbalance
presents a critical hurdle for machine learning models, as they tend to bias towards the
majority class, thereby undermining the detection of fault conditions. Addressing this
challenge is crucial for enhancing operational efficiency, reducing downtime, and ensuring
safety in industrial environments.

2.1 Imbalanced fault diagnosis
The importance of effective fault diagnosis cannot be overstated. Early and accurate
detection of faults can lead to timely interventions, preventing catastrophic failures, reducing
maintenance costs, and extending the lifespan of industrial equipment (Lv et al., 2020). In
sectors such as manufacturing, energy, and transportation, where equipment reliability is
paramount, the stakes are particularly high. The challenge of imbalanced fault diagnosis thus
becomes a focal point for research and development within industrial data science. Cost-
sensitive learning (Höppner et al., 2022; Xu et al., 2020) uses different cost matrices to impose
penalty factors on imbalanced categories, whose foundations and algorithms can be
naturally applied to imbalanced learning problems.

The core challenge in imbalanced fault diagnosis lies in developing models that can
accurately identify rare fault conditions despite their underrepresentation in the dataset.
Traditional machine learning algorithms, when applied directly, often perform poorly on
imbalanced datasets due to their inherent bias towards the majority class. To overcome this,
various methodologies have been proposed and developed: Aiming at the extremely
imbalanced hard disk drives, to avoid the problem of over-fitting caused by the SMOTE algorithm in the oversampling process, this study combines the SMOTE algorithm and the Cost-Sensitive algorithm to achieve the category balance for the extremely imbalanced dataset.

Data-level approaches: Techniques such as oversampling the minority class, undersampling the majority class, and generating synthetic samples are employed to balance the dataset before training the models (Zhang et al., 2020). These methods aim to provide a more balanced view of the data to the algorithms, thereby enhancing their ability to learn from both classes.

Algorithm-level approaches: These involve modifying existing algorithms or developing new ones that are inherently more robust to data imbalance. Techniques include cost-sensitive learning, where different weights are assigned to classes based on their frequency, and ensemble methods like bagging and boosting, which aggregate the predictions of multiple models to improve performance.

Hybrid approaches: Combining data-level and algorithm-level strategies can often yield superior results (Chen et al., 2021). For instance, using SMOTE to balance the dataset followed by training an ensemble of cost-sensitive models can leverage the strengths of both approaches.

Numerous industrial applications have demonstrated the effectiveness of these methodologies. For example, in the energy sector, imbalanced fault diagnosis models have been used to predict failures in wind turbines, potentially saving millions in unplanned downtime (Zhang et al., 2022). In manufacturing, similar models ensure the reliability of assembly lines, directly impacting product quality and operational efficiency.

Recent advances in the field have focused on leveraging deep learning (Souza et al., 2021; Fan et al., 2021) and artificial intelligence (Shi and Zhang, 2020) to tackle the imbalance problem (Wang et al., 2020a, b). Neural networks, with their capacity to learn complex patterns from data, have shown promise in detecting subtle indicators of faults. Furthermore, innovations in loss functions, such as focal loss, aim to make models more sensitive to the minority class by increasing the penalty for misclassifying rare events.

2.2 Hard disk fault diagnosis

The literature on hard disk fault diagnosis reflects a dynamic field of study focused on leveraging advanced data analytics to enhance hard disk drive (HDD) reliability. There are studies (Shi et al., 2021; Wang et al., 2020a, b) to develop predictive models capable of early and accurate fault detection. As research continues to evolve, the integration of emerging technologies and innovative modeling techniques should further improve the predictive accuracy and operational efficiency of HDD fault diagnosis systems, ensuring the reliability of data storage infrastructure across industries. In particular, there are three reasons with considering the detail of fault diagnosis in hard disk operations. Firstly, the quality and availability of operational and failure data can significantly impact the performance of fault diagnosis models. In many cases, data might be noisy, incomplete, or not systematically collected, which complicates the modeling process. Secondly, industrial processes can evolve over time due to changes in operation, maintenance practices, or external conditions. This evolution can alter the nature of faults and their signatures, requiring continuous adaptation of the fault diagnosis models. Finally, the implementation of fault diagnosis models in real-world settings often faces operational constraints, such as limited computational resources, the need for real-time processing, and integration with existing monitoring systems. These constraints can limit the complexity and depth of the models that can be deployed.

One of the significant challenges in HDD fault diagnosis is the imbalanced nature of the dataset, where instances of failure are much less frequent than normal operational data. This imbalance poses a problem for machine learning models, leading to a bias towards predicting
the majority class and, consequently, a high rate of false negatives for failures. Researchers have explored various techniques to address this issue, including oversampling of the minority class (Li et al., 2023; Soltanzadeh and Hashemzadeh, 2021), undersampling of the majority class (Arefeen et al., 2020), and synthetic data generation methods. Another challenge is the variability in failure modes across different HDD models and batches, which complicates the development of a universally applicable predictive model. This variability necessitates continuous model tuning and adaptation to new data, making the maintenance of predictive accuracy over time a complex task.

Industrial data-driven modeling integrates real-time monitoring and predictive analytics is a key area, enabling proactive interventions before failures occur. Additionally, the application of transfer learning and domain adaptation techniques could address the issue of model generalization across different HDD models and configurations. Utilizing these advanced technologies allows for the deployment of predictive models in closer proximity to the data source. This strategic positioning minimizes latency, thereby facilitating more prompt detection and mitigation of faults.

2.3 Summary of literature review

In this paper, an innovative approach is introduced to predict the timeframe within which a hard disk might fail, aiming to provide an early warning system for such failures. To tackle the challenge of highly imbalanced datasets, the paper employs a multifaceted strategy. This strategy includes enhancing the data, refining the model, and modifying the loss function to boost the model's accuracy and its ability to handle problems associated with long-tail distributions.

This comprehensive approach not only aims to improve the predictive performance of the model in the context of imbalanced datasets but also enhances its robustness, making it more effective in predicting hard disk failures and thereby facilitating timely interventions. The imbalance issue discussed in this study is particularly complex compared to what is typically encountered in existing research. The core of the challenge arises from the fact that in real-world settings, especially within data centers, there’s a continuous cycle of adding new hard drives. These additions are made either as replacements for those that have failed or simply to increase the storage capacity of the system. As a result, a data center might house a diverse array of hard drives, each differing in brand and model. This diversity introduces significant complexity in developing a one-size-fits-all approach for fault diagnosis.

To effectively address the challenge of diagnosing faults in varied hard drive brands and models, it’s essential to commence by refining the initial dataset. This preparation ensures that the machine learning models are equipped to interpret the data accurately. A versatile model is key in recognizing and diagnosing faults effectively across the spectrum of hard drives. The proposed strategy aligns with the fundamental principles of machine learning which necessitate data that is representative and well-prepared for the models to learn from.

In the second phase of the strategy, fine-tuning a machine learning model’s parameters becomes pivotal. This customization allows a single architecture to adapt to the idiosyncrasies of each brand and model of hard drive. This method of parameter adjustment is grounded in the core principles of data science: to derive insights from data that are both precise and applicable to the specific context. By doing so, the adaptability and accuracy of fault diagnosis in diverse storage environments are significantly improved, overcoming the challenges posed by data imbalance and variability. This approach not only demonstrates an understanding of the intricacies involved in machine learning but also showcases a tailored application of its principles to yield practical and reliable results.
3. Industrial data-driven modeling for imbalanced fault diagnosis

The framework for industrial data-driven modeling for imbalanced fault diagnosis, illustrated in Figure 1, unfolds across four pivotal stages. Initially, the study delineates the issue of hard drive failures, noting the variation in disk distributions across different manufacturers and models, with specific serial numbers identifying unique drives. The research commences by reshaping the original dataset to align with data-driven models, adjusting parameters to cater to the diverse array of disk brands and models, ensuring the model’s effectiveness across a standardized disk population.

The second phase involves rigorous data preparation, including the remediation of missing values, feature and dimension reconstruction, and the normalization of data imbalance. This phase introduces the use of SMOTE, integrated with penalty factors for skewed categories, addressing the imbalance.

Model construction constitutes the third phase, where the study systematically approaches data restructuring, class equilibrium, feature derivation, and classification, laying down the foundation for accurate fault prediction.

The final phase emphasizes the implementation of results, where a bagging technique is deployed to forecast the risk trajectory. The efficacy of this method is then meticulously evaluated through comparative analysis, demonstrating the model’s predictive accuracy and robustness in data-driven decision support.

3.1 Problem definition

The problem definition includes addressing the inherent limitations of traditional models in handling imbalanced datasets. It underscores the need for specialized approaches that enhance the model’s sensitivity to the minority class, ensuring that critical but infrequent events are accurately identified, thereby improving the reliability of fault detection in imbalanced data contexts.

However, this standard classification framework struggles with imbalanced data scenarios, where certain categories are significantly underrepresented compared to others. The issue arises when the model, despite achieving high accuracy, incorrectly classifies minority class instances (e.g. failures) as belonging to the majority class. This misclassification is particularly critical in applications where detecting rare events, such as system downtimes, is crucial.
The deep learning model excels in classification tasks, employing a foundational structure that includes a connection layer and a Soft-max layer. The connection layer enhances the model’s depth, enabling it to tackle complex classification challenges by simulating intricate patterns, while the Soft-max layer determines the probability of each category, culminating in classification via supervised learning.

Furthermore, the challenge of extremely imbalanced datasets is that they can lead to the generation of sample points that do not accurately reflect real-world conditions. This research addresses this issue by integrating data augmentation and cost-sensitive strategies to manage the imbalance. Specifically, SMOTE is utilized to augment data from underrepresented classes, and a modified Cross-Entropy loss function incorporates penalty weights for these classes to mitigate the risk of incorrect downtime predictions.

The careful selection of datasets is paramount in predicting system downtimes, aiming to capture the intricate details and patterns that often precede failures. The criteria for selection prioritize data that mirrors the true operational behaviors under which downtimes are known to occur. This encompasses an array of indicators such as system logs, error messages, performance metrics, and maintenance records. It’s crucial that the selected data not only signifies the health of the system but is also chronologically marked with timestamps to facilitate an accurate timeline analysis. The ultimate goal is to compile a dataset that represents both the standard operational data and the data leading up to downtimes, thereby enabling learning algorithms to distinguish effectively between normal operations and potential failure scenarios.

The algorithm of choice in this research is the Bagging-Based Deep Learning (BBDL) framework, selected for its effectiveness in dealing with imbalanced datasets, a common characteristic when forecasting system downtimes due to the infrequency of failure events. The BBDL method incorporates a suite of sophisticated techniques: Oversampling to enhance minority class representation, cost-sensitive adjustments to accurately weigh the importance of rare yet significant failure events, deep learning algorithms for their superior ability to decode complex data patterns, and bagging strategies to improve model stability and accuracy by aggregating multiple predictors to minimize variance and prevent overfitting. This selection is made to ensure that the chosen algorithms are adept at handling the specific challenges of predicting system downtimes.

The methodological approach of the study is intrinsically linked to its primary objective: the accurate prediction of system downtimes. Each component of the BBDL framework has been meticulously selected to counteract the specific difficulties presented by the research question, which include the lopsided nature of the data and the essential demand for prompt and precise predictions of system failures. The comprehensive nature of the BBDL framework showcases an application of machine learning and data science principles that is not just theoretically sound but also pragmatically focused on the issue at hand.

Data restructuring within this framework involves the creation of relevant features, adjustment of model dimensions, elimination of features lacking data, and imputation of missing values using average data from identical series. SMOTE is applied to generate additional minority class samples, achieving a more balanced class distribution. This preprocessing step ensures that both training and test datasets contain representative samples of minority class instances. Following the training phase, the bagging approach is employed to estimate the risk curve, systematically dividing the 49-day pre-downtime period into several intervals (or bags), thereby refining the prediction model for system downtimes.

In this research, the model is retrained using data from the last 49/n days before a hard drive failure, treating this period as indicative of impending downtime. By classifying the test dataset accordingly, the model is capable of forecasting the likelihood of a hard drive failing several days before the event actually occurs. To facilitate a clearer comprehension of this methodology, Figure 2 presents a detailed workflow diagram of the proposed approach.
The provided diagram offers a structured approach to processing imbalanced datasets for the purpose of early warning and fault diagnosis in hard drives. This approach involves several critical stages: initially, data pre-processing and dimension reconstruction are employed to refine the dataset for optimal use. Subsequently, the incorporation of work time features and oversampling with SMOTE is performed to balance the dataset. Finally, bagging and cost-sensitive classification methods are utilized, leading to the desired outcomes of early fault detection and diagnosis.

When dealing with sensitive or proprietary data within such a framework, one must navigate through a myriad of considerations. Ensuring data privacy and security is paramount, as mishandling sensitive information during pre-processing or any subsequent stage could lead to breaches of regulations like GDPR or HIPAA. Issues of data ownership and sharing also come into play, particularly when proprietary data restricts how and where the data can be utilized, potentially hindering the training and implementation of models. Moreover, when integrating specific features such as "Work Time", the potential for introducing bias necessitates judicious selection to prevent unfair or discriminatory model outcomes.

Furthermore, the utilization of proprietary data in the development of machine learning models raises substantial questions regarding intellectual property rights, compliance with legal obligations, and the adherence to industry standards. The complexity added by using such data may also obscure model transparency and explainability, posing challenges for stakeholders who depend on clear and accountable decision-making processes. To manage these concerns effectively, robust data governance frameworks are essential, ensuring that data handling and processing remain secure, compliant, and considerate of privacy and
intellectual property rights. The framework must be adaptable to accommodate these restrictions while still upholding the integrity and efficacy of the fault prediction and diagnosis process.

3.2 Data preparation
The process of data preparation encompasses several key steps, such as data preprocessing, handling missing values, adjusting features and dimensions, and data oversampling. These steps are integral to the model construction phase. Specifically, during data refactoring, the model requires preprocessing to cleanse the data, impute missing values, and adjust features and dimensions for optimal performance. To ensure that failure data is represented in both the training and testing datasets, this study employs a novel approach for dataset segmentation. Additionally, the study utilizes SMOTE to achieve class balance and employs an auto-encoder for the identification and extraction of relevant features.

(a) Data refactoring
Figure 3 illustrates the transformation process of data, highlighting the transition from raw data on the right to processed data on the left. The key steps in data refactoring involve:

1. Generating features based on date information.
2. Defining the dimensional structure of the model.
3. Discarding any feature with missing values exceeding a third of its total data count.
4. Imputing missing entries with the average values from identical serials on proximate dates.

Data refactoring serves two primary purposes in addressing this challenge. The initial phase is dedicated to translating the operational dates of hard drives into cumulative working hours. This phase encompasses three critical tasks:

1. Identifying and recording the start date of operation for each hard drive.
2. Due to the discontinuity in operation dates, a novel algorithm is required to accurately convert these dates into continuous working hours.
3. Given the vast quantity of data, it is imperative that the algorithm efficiently stores interim results to maintain the completeness of the working hours attribute.

To address the challenges outlined, this study introduces a method for iterating through the dataset to ascertain the active dates for each hard drive. During each iteration, a dictionary is utilized to record several key pieces of information: the last known active date for each drive, the current active date, the duration of activity between these two points, and the cumulative operating hours up to the present. This iterative approach is detailed in Table 1. Following each iteration, the algorithm locates the relevant data by serial number and date for the hard drive and then updates it with the time feature calculated, using the “Total” value from Table 1 for imputation. This strategy effectively navigates the issues of discontinuous dates and the preservation of intermediate data.

Additionally, this algorithm effectively addresses the variability in the number of days per month and accounts for leap years, ensuring accurate calculation of working hours. It incorporates specific algorithms to identify the month and leap year status, allowing for precise determination of working time. Thus, the computation outlined in Equation (1).
Figure 3. Framework of model
\[ \text{Total}_{\text{last}} + y^* (\text{Lastdate}_{\text{year}} - \text{Currentdate}_{\text{year}}) + m^* (\text{Lastdate}_{\text{month}} - \text{Currentdate}_{\text{month}}) + (\text{Lastdate}_{\text{day}} - \text{Currentdate}_{\text{day}}) \]

(1)

In this context, “y” represents the total number of days in the current year, while “m” indicates the total number of days within the specified month. The subsequent phase involves transforming the model number attribute into a dimension, given that the original dataset is organized by date.

(b) Class balance and feature extraction

Initially, this segment introduces a method for partitioning imbalanced datasets and employs the SMOTE to create new instances within the training dataset, thereby equalizing the distribution of classes. The essence of the SMOTE algorithm lies in its ability to closely examine minority class samples and synthetically augment the dataset with additional instances derived from these samples, as illustrated in Figure 4. The procedural steps of the algorithm are outlined as follows:

1. For each sample \( x \) of a minority class, calculate its distance to all samples in \( S_{\text{min}} \) of a minority class based on Euclidean distance, and obtain its k-nearest neighbor.

2. A sampling proportion is set according to the unbalanced proportion of samples to determine the sampling multiplier \( N \). For each small number of samples \( x \), several samples are randomly selected from its k nearest neighbors, and the selected nearest neighbor is assumed to be \( \hat{x} \).

3. For each randomly selected neighbor \( \hat{x} \), construct new samples with the original samples according \( x_{\text{new}} = x + \text{rand}(0, 1) \cdot (\hat{x} - x) \).

Given the significant imbalance in the dataset, employing a random partitioning strategy could lead to a scenario where certain categories with fewer instances are not represented in the sample set. To address this challenge, this study introduces a novel partitioning method. Initially, the dataset is bifurcated into two subsets based on the size of each category, distinguishing between those with a larger number of samples and those with fewer. Subsequently, each category is split into training and testing datasets in a consistent ratio, specifically 3:1 in this case. The concluding step involves amalgamating the training subsets from both categories into a unified training dataset, and similarly for the testing datasets.

Regarding the method for oversampling, this study employs the SMOTE for its implementation. The oversampling process is guided by Equation 2:

\[ x_{\text{new}} = x + \text{rand}(0, 1) \cdot (\hat{x} - x) \]

(2)

In this methodology, \( x \) represents a sample from a minority category, while \( \hat{x} \) denotes its nearest neighbor. The parameter in the SMOTE algorithm functions similarly to a seed for random number generation, ensuring reproducibility of results. After conducting various experiments, this study has chosen a value of 65 to optimize the sampling process.

<table>
<thead>
<tr>
<th>Each hard drive</th>
<th>Last date</th>
<th>Current date</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Last iteration</td>
<td>2019-2-12</td>
<td>2019-2-21</td>
<td>a</td>
</tr>
<tr>
<td>Current iteration</td>
<td>2019-2-21</td>
<td>2019-2-23</td>
<td>a+23-21</td>
</tr>
<tr>
<td><strong>Source(s):</strong> Authors’ own creation/work</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
For the task of feature extraction, this study employs an auto-encoder algorithm, which can be conceptualized as comprising two main components: an encoder that compresses the input data and a decoder that reconstructs it. This structure enables the auto-encoder to identify and prioritize the most significant features required to accurately reconstruct the input from the compressed representation. The learning process of the network is defined by the following mapping relationship:

\[ g(\sigma(X)) \approx X \]  

where, the encoding is \( \sigma(X) \), and the decoding is \( g(X) \). The training process is as follows: for the training samples \( X = \{x_1, x_2, x_3, \ldots, x_d\} \) (each training sample \( x_i = [x_{i1}, x_{i2}, x_{i3}, \ldots, x_{in}]^T \)), the encoder converts the input vector \( X \) into latent variables \( h = \{h_1, h_2, h_3, \ldots, h_d\} \) (each latent variable \( [h_1, h_2, h_3, \ldots, h_n]^T \)) through activation equation, which is denoted as:

\[ h = \sigma(WX + b) \]  

Where \( W \) is \( m \times n \) vector and \( b \) is \( m \) vector. Similarly, the decoding process converts the latent variable \( h \) into a reconstruction variable \( Z = \{z_1, z_2, z_3, \ldots, z_d\} \) (each reconstruction variable is \( [z_1, z_2, z_3, \ldots, z_n]^T \)), denoted as:

\[ Z = \sigma(W'h + b') \]  

where, \( W' \) is the \( n \times m \) dimension vector, \( b' \) is the \( n \) dimension vector. The network training process is to minimize the reconstruction error of \( X \) and \( Z \). As the vibration signal belongs to the real number range, the most reusable mean square error can be used:

\[ L(X, Z) = L_2(X, Z) = C(\sigma^2)||X - Z||^2 \]  

where \( C(\sigma^2) \) is a constant determined by \( \sigma^2 \).

The optimization process employs the Adam algorithm, an advanced optimization technique that dynamically adjusts the learning rate for improved model training outcomes. The Adam algorithm incorporates a method to directly integrate the first moment (mean) estimation of the gradient into its momentum calculation, enhancing its efficiency. Moreover, it features a bias correction mechanism to adjust both the first and second moment estimations that commence from zero. This adjustment addresses and mitigates the potential for high bias during the initial stages of training, a limitation observed in the RMSProp algorithm due to its absence of a correction factor for the second moment estimation. The Equation of the Adam algorithm is as follows:
\[ \begin{align*}
    m_t &= \beta_1 m_{t-1} + (1 - \beta_1) g_t, \\
    v_t &= \beta_2 v_{t-1} + (1 - \beta_2) g_t^2, \\
    \tilde{m}_t &= \frac{m_t}{1 - \beta_1^t}, \tilde{v}_t = \frac{v_t}{1 - \beta_2^t}, \\
    W_{t+1} &= W_t - \frac{\eta}{\sqrt{\tilde{v}_t + \delta}} \tilde{m}_t
\end{align*} \]

Equation (7)

where, \( W_t \) is the parameter of the model of the TTH iteration, \( g_t = \nabla J(W_t) \) is the gradient of the cost function, and \( \delta \) is a number with a small value to avoid the denominator being 0. \( m_t \) and \( v_t \) are first order momentum terms and second order momentum terms. \( \beta_1 \) and \( \beta_2 \) are the dynamic values, usually 0.9 and 0.999. \( \tilde{m}_t \) and \( \tilde{v}_t \) are their corrections. Although Adam algorithm sometimes needs to be modified from the default learning rate, it is fast and robust.

The chosen loss metric is the Mean Squared Error (MSE), detailed in equation (4). The auto-encoder aims to identify the most critical features, hypothesized to be “a” in number. When the feature count falls below 26, the auto-encoder’s loss exhibits fluctuating behavior, indicating instability. This observation suggests that if the feature count is below “a”, the auto-encoder struggles to accurately reconstruct the original data. Consequently, this approach enables the identification of a feature set that effectively represents the dataset, as determined through the auto-encoder’s processing.

3.3 Model construction

In scenarios with highly imbalanced datasets, applying a high oversampling ratio can generate sample points that deviate from real-world distributions, potentially skewing classification outcomes. To address this imbalance effectively, employing a cost-sensitive approach adjusts the penalty costs associated with misclassification, particularly favoring the minority class that is underrepresented. This strategy ensures that the classifier does not erroneously favor the majority class, thereby enhancing the overall accuracy. This research integrates the SMOTE with a cost-sensitive framework, achieved by applying varied penalty weights to different classes within the cross-entropy loss function. This combination aims to balance the dataset more accurately and improve classification performance by acknowledging the inherent value and risk of misclassifying the minority class examples.

Specifically, the core element of the cost sensitive learning method is the cost matrix. Where \( \text{cost}_{ij} \) represents the cost of predicting the ith sample as the j sample. Generally speaking, \( \text{cost}_{ij} = 0 \). If class 0 is judged as class 1, the loss caused by it is greater, then \( \text{cost}_{0j} > \text{cost}_{10} \). The greater the difference in the degree of loss, the greater the value difference, while the Soft-max function Equation is:

\[ f(z_j) = \frac{e^{z_j}}{\sum_{i=1}^{n} e^{z_i}} \]

Equation (8)

For k classification problem, let the training set be \( \{(x^1, y^1), (x^2, y^2), \ldots, (x^n, y^n)\} \), and its label be \( y^{(i)} = \{1, 2, 3, \ldots, k\} \). Given the test input \( x \), we can use the hypothesis function to estimate the probability value of each category j:

\[ p(y = j|x) \]

Equation (9)

Essentially, the role of the Soft-max layer is to calculate the likelihood of each potential outcome for a given input \( x \). To achieve this, it generates a k-dimensional vector, where the sum of the vector’s elements equals 1, with each element representing the probability of each
of the $k$ possible outcomes. The specific Equation of our classification hypothesis function is designed to reflect this, as follows:

$$h_k(x^{(i)}) = \begin{bmatrix} p(y^{(i)} = 1|x^{(i)}; \theta) \\ p(y^{(i)} = 2|x^{(i)}; \theta) \\ \vdots \\ p(y^{(i)} = k|x^{(i)}; \theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^{k} e^{\theta_j^T x^{(i)}}} \begin{bmatrix} e^{\theta_1^T x^{(i)}} \\ e^{\theta_2^T x^{(i)}} \\ \vdots \\ e^{\theta_k^T x^{(i)}} \end{bmatrix}$$ (10)

where, $\theta_1, \theta_2, \theta_3, \ldots \theta_k$ is the parameter of the model, and the purpose of $\sum_{j=1}^{k} e^{\theta_j^T x^{(i)}}$ is to normalize the probability distribution so that the sum of all probabilities is 1, $\theta$ is $\theta_1, \theta_2, \theta_3, \ldots \theta_k$ and the matrix arranged in rows:

$$\theta = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_k \end{bmatrix}$$ (11)

To set in $f\{\cdot\}$:

$$\begin{cases} f\{\text{The value is True}\} = 1 \\ f\{\text{The value is False}\} = 0 \end{cases}$$ (12)

Then the regression cost function of classification network is:

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{k} f\{y^i = j\} \log \frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{k} e^{\theta_l^T x^{(i)}}}$$ (13)

The essence of the classification process of Soft-max classification network is actually the process of regression cost function. The gradient Equation is:

$$\nabla_{\theta} J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} [x^{(i)}(f\{y^{(i)} = j\} - p(y^{(i)} = j|x^{(i)}; \theta))]$$ (14)

Since there is still no closed solution, the iterative algorithm is needed.

The deep learning model, detailed in Table 2, utilizes the Stochastic Gradient Descent (SGD) method for optimizing the classifier. Unlike traditional gradient descent, which

<table>
<thead>
<tr>
<th>Layer</th>
<th>Input</th>
<th>Output</th>
<th>Activation</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>26</td>
<td>22</td>
<td>sigmoid</td>
<td>FC</td>
</tr>
<tr>
<td>2</td>
<td>22</td>
<td>18</td>
<td>sigmoid</td>
<td>FC</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
<td>14</td>
<td>sigmoid</td>
<td>FC</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
<td>8</td>
<td>sigmoid</td>
<td>FC</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>4</td>
<td>sigmoid</td>
<td>FC</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>2</td>
<td>sigmoid</td>
<td>FC</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>2</td>
<td>/</td>
<td>Soft-Max</td>
</tr>
</tbody>
</table>

Table 2. The deep learning model

Source(s): Authors’ own creation/work
computes a gradient and updates parameters in one go, SGD selects a random sample to calculate the gradient, thereby moving towards the optimum solution incrementally. While SGD enhances computational efficiency significantly, it also introduces variability in the gradient estimation, which can introduce noise. Excessive noise may adversely affect the convergence of the SGD. To maintain neural activity, this model employs a sigmoid activation function instead of the exponential linear unit (elu) or rectified linear unit (relu) functions. Experiments conducted using both elu and relu activation functions resulted in no observable changes in loss during training, suggesting diminished neural activity with these functions. The choice of input and output sizes is determined through parameter tuning. The concluding layer of the model is a softmax layer, which facilitates the classification process.

The objective function is,

$$\text{loss}(x, \text{class}) = -\log \frac{e^{x\text{class}}}{\sum_j e^{y_j}} = -x_{\text{class}} + \log \left( \sum_j e^{y_j} \right)$$  \hspace{1cm} (15)$$

The objective function can also define as,

$$\text{loss}(x, \text{class}) = \omega_{\text{class}} \left( -x_{\text{class}} + \log \left( \sum_j e^{y_j} \right) \right)$$  \hspace{1cm} (16)$$

What the model does is to reduce this objective function in order to get better classification effect. We can use this function to coordinate the Soft-Max layer. And the training process is to guarantee the accuracy of the detection so that the objective function is chosen as the accuracy of the classification. The accuracy function is,

$$\text{loss} = -\alpha_t (1 - p_t)^\gamma \log(p_t)$$  \hspace{1cm} (17)$$

where, $\alpha_t = \begin{cases} \alpha, & \text{if} \quad p \leq \frac{b}{1-p} \text{; and } \alpha \text{ is the weight factor, } p \text{ is the probability of the softmax layer. And } \gamma \text{ is the adjustable parameter.} \\
\frac{1}{1-\alpha}, & \text{otherwise} \end{cases}$$

Through experiment, this paper adopts modified Cross-Entropy-Loss as its loss function in pre-training and adopts MSE as the training loss in order to improve the accuracy. The training epoch is 4 and the batch size adopts 10,000, whose total training data number is 259,278. The learning rate is 0.00003.

3.4 Results and implement
This study introduces an advanced n-bagging methodology to predict risk curves, as depicted in Figure 5. At its core, within the outlined oversampling classification model, lies the foundation for model training. This method innovatively alters the training data set, as shown below the figure. For instance, dividing the 49 days prior to potential failure into 7 segments or “bags”, each covering 7 days, serves as a model for this approach. These intervals are then leveraged as the basis for further retraining of the model. Through orderly analysis and categorization of the test data set, the model identifies hard drives at risk of failure and, through the bagging approach, allocates these to their appropriate segments. This technique, by evaluating the distribution of failures among the various bags, allows a predictive curve that can forecast potential downtime with significant lead time.

The methodology’s flow is outlined below, with subsequent verification through data analysis and experimentation:

(1) Identify the failing hard drive and record its serial number.
Retrieve the serial number and its associated timestamp.

Predict that the identified hard drive will fail and cease to function in the future. The period following the failure is considered the “infected window”, indicating that data from this timeframe is unreliable and should not be used.

Anticipate that the failure of the hard drive will not be abrupt but preceded by discernible indicators, allowing for the assessment of the hard drive’s risk of failure.

Segment the timeframe immediately preceding the failure into seven “bags”, each encompassing seven days, to analyze the lead-up to the failure.

Utilize common understanding that the farther a date is from the failure event, the lower the risk of the hard drive failing.

Treat the pre-failure indicators similarly to actual failures for the purpose of data oversampling.

Figure 6 illustrates the historical characteristic parameters of a hard drive nearing failure, highlighting how these parameters evolve over time, with a particular focus on the working time feature. Visual analysis reveals significant variations in features leading up to the failure, with two characteristic curves showing marked changes and one feature fluctuating intensely. These observations support the notion that failures are preceded by clear signs. The choice of seven days per “bag” aids in weekly predictive analysis. The application of the bagging methodology in conjunction with a deep learning model facilitates the nuanced classification of risk levels associated with temporal distance from the predicted failure. This study will further validate the hypothesis that the risk of hard drive failure decreases as the temporal distance from the projected failure rate increases through experimental validation.

Figure 6 appears to be a time-series plot illustrating the normalized readings of various SMART (Self-Monitoring, Analysis and Reporting Technology) attributes over time, which are commonly used for predicting hard drive failures. The x-axis represents time, likely in days, and the y-axis represents the normalized value of different SMART attributes. There are three distinct sections demarcated by dashed vertical lines, labeled as “Data Variation,” ...

Source(s): Authors’ own creation/work
“Feature Variation,” and “Downtime.” These sections possibly indicate different phases in the data monitoring process. In the “Data Variation” phase, multiple SMART attributes are tracked. The lines are relatively flat, indicating stable readings during this period, with no significant changes or anomalies detected. The “Feature Variation” phase shows a continuing trend of stability for most attributes, but a select few (notably, the orange line) begin to demonstrate an upward trend, which may signify an emerging issue with the hard drive. The “Downtime” phase is marked by a steep upward trajectory in the orange line, while other attributes remain stable. This could be indicative of a critical SMART attribute reaching a threshold that suggests imminent drive failure, leading to downtime. The red downward-pointing triangle at the end may represent the actual point of failure or shutdown. The “smart_1_normalized” through “smart_195_normalized” labels on the bottom correspond to individual SMART attributes tracked over time. Each attribute’s stability or variability could be vital for predictive models that aim to preemptively identify and address hard drive issues before a complete failure occurs.

4. Empirical study

4.1 Problem definition

This investigation examines historical hard drive data spanning from 2015 to 2023, encompassing a wide range of devices. The data originates from various sources related to hard drive performance, including internal and external sensor readings, as well as machine logs. Specific types of data analyzed include temperature measurements, acoustic signals, luminosity levels, thermal images, and records of errors and events. This comprehensive dataset provides a rich foundation for understanding the operational characteristics and failure patterns of hard drives, facilitating a detailed analysis of factors that may influence their reliability and lifespan.

The model number, assigned by manufacturers, signifies substantial differences among hard drives, such as variations in manufacturing processes. The “capacity-bytes” denotes the storage capacity of each hard drive, while the S.M.A.R.T. (Self-Monitoring, Analysis, and
Reporting Technology) number represents specific attributes critical to understanding the health and performance of the drive. This study transforms the date data into a metric of the hard drive’s operational duration, considering it alongside S.M.A.R.T. parameters as a key indicator of the device’s condition.

Additionally, the dataset encompasses approximately 120 attributes for each hard drive, yet it is characterized by a high incidence of missing values—nearly half of the data points are absent in each attribute. To address these challenges, the study employs techniques in feature engineering and methods for handling missing values.

A notable aspect of the data is the infrequency of hard drive failures, leading to an imbalance between the classes of normal and failed drives. This imbalance underscores the primary objective of the investigation: to effectively differentiate between functioning and malfunctioning hard drives based on their attributes. Although the diversity of model numbers necessitates distinct models due to their significant differences, there exists a commonality in the challenges posed by imbalanced data and the need for accurate downtime diagnosis. This shared foundation allows for a unified approach to the problem, despite the requirement for model-specific parameter adjustments. The core strategy, therefore, revolves around applying a consistent methodology to tackle the imbalance and diagnostic challenges across different hard drive models, tailoring only the parameters to suit each model’s unique characteristics.

Figure 7 presents a visual analysis combining a line graph and a detailed table view, presumably of hard drive performance data over time. The line graph at the top tracks a performance metric from Q1 of 2016 through Q4 of 2019. There is a noticeable discontinuity in
the data starting at Q1 of 2019, indicated by a dashed line and a red box highlighting the last plotted point before the gap. This suggests missing data or a transition point in the dataset. Below the graph, a table lists detailed hard drive information. Each row corresponds to a specific hard drive, as indicated by the dashed lines connecting the graph’s discontinuity and the first entry in the table. This linkage implies that the discontinuity in the graph is directly related to the data for that specific hard drive. The table columns are labeled with attributes such as “date”, “serial_number”, “model”, “capacity_bytes”, and several “smart_X_normalized” columns where X is the attribute number. These SMART attributes provide health indicators of the hard drives, which are often used for predicting failures. “Label” is another column that likely indicates the status or classification outcome for each drive at the given date. The columns with blue headers represent the various SMART attributes, while the red headers highlight critical identifiers and labels for each drive, such as the serial number and the label/classification. The annotations on the image indicate a focus on these particular aspects: discontinuity of the date, corresponding hard disk identifiers, and features of interest for analysis or predictive modeling. The red pen marks and the word “Feature” written on the side suggest that the viewer should pay particular attention to the features (SMART attributes) when analyzing the data.

4.2 Data preparation
This section outlines the methodology for parameter configuration within the Bagging-Based Deep Learning (BBDL) framework. The BBDL approach showcases a sequential application of Auto-encoder, SMOTE, and neural network models to effectively address complex issues associated with multi-featured, imbalanced datasets of hard drive information. The initial phase of BBDL involves two critical steps: data refactoring and preprocessing of data features.

Following the data refactoring process, the BBDL framework successfully generates a total of 51 distinct models, covering a variety of hard drive models such as ST12000NM0007, ST4000DM000, ST8000NM0055, and ST10000NM0086, among others. To illustrate the efficacy and advantages of the BBDL methodology, the study specifically focuses on the ST10000NM0086 model. This example serves to highlight the strengths of the BBDL approach in handling data from a specific hard drive model, providing insights into its capacity to enhance predictive accuracy and model robustness.

Subsequent to this demonstration, the study extends its analysis to additional datasets corresponding to different hard drive models. This comparative approach aims to underscore the BBDL model’s versatility and its general applicability across a broad spectrum of hard drive data. Through this comparative analysis, the research seeks to validate the BBDL model’s generalization capabilities, demonstrating its potential as a powerful tool for addressing the challenges posed by imbalanced and multifaceted datasets in the domain of hard drive failure prediction and analysis.

4.3 Model construction
This research emphasizes the significance of the working time feature within the Bagging-Based Deep Learning (BBDL) framework, thereby making it a critical input for the auto-encoder alongside the remaining 25 features. Figure 8 delineates the outcomes derived from the auto-encoder’s performance. The axes represent the number of iterations (horizontal) and the loss value (vertical), respectively. The illustration includes two graphs: the first displaying results for an auto-encoder configured with 22 outputs, and the second for one with 20 outputs. Both graphs exhibit a similar trend where the loss values initially decrease modestly during the early iterations. However, as the iterations progress, the loss values exhibit fluctuations, characterized by significant ups and downs, indicating a higher loss
value. This pattern suggests that while the auto-encoder initially learns to reconstruct the input data effectively, it encounters difficulties in maintaining a steady reduction in loss, particularly when the number of outputs is reduced. This behavior highlights the challenges in optimizing the auto-encoder’s performance, especially when it comes to balancing the complexity of the model with the need for accurate data reconstruction.

Following the data reprocessing phase, the ST1000NM0086 model is refined to encompass 26 distinct features. An intriguing observation emerges when the output dimensionality of the auto-encoder is reduced below this 26-feature threshold: the model’s loss fails to converge. This phenomenon underscores a critical insight—that diminishing the output dimensions beyond this point compromises the auto-encoder’s ability to faithfully reconstruct the original data through its decoding process. Consequently, this analysis posits that the suite of 26 features is integral, serving as a cornerstone for the model’s ability to accurately identify and process data.

This study extends its analysis by specifically examining the impact of a newly introduced feature: working time, derived from the original date data. The inclusion of this feature is hypothesized to offer a novel perspective on the data, potentially enhancing the model’s predictive capabilities. The evaluation of this feature’s effectiveness, as well as its comparative impact on the model’s performance, is meticulously documented in Table 3. This comparison aims to elucidate whether the integration of working time as a feature materially improves the model’s ability to distinguish between different operational states of the hard drive, thereby refining its predictive accuracy.
BBDL achieves accuracy at 99.99%, indicating that it correctly predicts almost all positive and negative cases. The precision is 63.63%, meaning that when it predicts a positive case, it is correct about 63.63% of the time. The recall is 100%, which means it identifies all actual positive cases. As BBDL without working time feature, its accuracy drops to 92.86%, showing a notable decrease in the overall correct predictions. The precision plummets to 0.19%, indicating that it becomes almost entirely unreliable in its positive predictions. However, the recall remains relatively high at 85.71%, meaning it still identifies a large portion of the actual positive cases, though not all as with the complete model.

The comparative analysis underscores the significance of the working time feature within the BBDL framework, highlighting its pivotal role in enhancing model performance. However, a notable observation emerges concerning the model’s precision metrics; they are relatively low. This outcome is primarily attributed to the inherent imbalances within the dataset. In scenarios where the dataset is skewed, with a predominance of one class over another, even a small number of true positive predictions (actual failures correctly identified) juxtaposed against a substantial volume of true negative instances (non-failures correctly identified) can disproportionately affect precision.

In this specific instance, the model’s precision is adversely impacted due to the rarity of true positive examples within the dataset. Despite the model correctly identifying a small number of true positive cases, the overwhelming majority of predictions are true negatives, given the dataset’s imbalanced nature. Consequently, the precision metric, which evaluates the proportion of true positive predictions out of all positive predictions (both true positives and false positives), exhibits a low value. This scenario underscores a common challenge in predictive modeling with imbalanced datasets, where achieving high precision can be particularly challenging due to the skewed distribution of classes.

4.4 Results and implement

The adjustment of the “random_state” parameter plays a crucial role in determining the outcomes of models that involve a stochastic process. This parameter is essentially a seed value provided to the random number generator used by the algorithm. It ensures the reproducibility of results by starting the generation of pseudo-random numbers from the same initial point. In the context of machine learning models, particularly those that involve random initialization of weights, selection of data points for training (such as in stochastic gradient descent), or split of data (like in cross-validation or bootstrapping), the “random_state” ensures that the results are consistent across multiple runs, given the same parameter setting.

When adjust the “random_state” value and then alter the sequence of random numbers generated by the algorithm, which in turn affects various aspects of the model training process, including the initialization of weights, selection of features for splitting in tree-based models, and the sampling of data points. This can lead to variations in the final model performance metrics such as accuracy, precision, and recall.

For instance, in Table 4, if it were to list the outcomes of a model under different “random_state” settings, then would observe variations in the results, illustrating the sensitivity of the

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>BBDL</td>
<td>99.99%</td>
<td>63.63%</td>
<td>100%</td>
</tr>
<tr>
<td>BBDL without working time feature</td>
<td>92.86%</td>
<td>0.19%</td>
<td>85.71%</td>
</tr>
</tbody>
</table>

**Table 3. Comparison of BBDL and BBDL without working time feature**
model to the initial conditions set by this parameter. It’s important for practitioners to be aware of this sensitivity, especially when comparing models or tuning hyperparameters, to ensure that the comparisons are fair and that the chosen model is robust and not overly dependent on a particular initialization or data sampling scheme.

(a) Parameters of classifier

This research incorporates pre-training techniques, utilizing Cross-Entropy-Loss as the loss function during the pre-training phase, and adjusting the training loss to (1-accuracy) * 100% in an effort to enhance model accuracy. The pre-training is conducted over 2 epochs, followed by a more extensive training period of 4 epochs. The batch size is set to 10,000, with a total of 259,278 data points used for training. The learning rate is carefully chosen at 0.00003, indicating a cautious approach to model updates to prevent overshooting the optimal solution.

The selection of an appropriate activation function is crucial as it significantly influences the model’s ability to classify data accurately. Activation functions play a key role in determining the output of neural network nodes based on the inputs they receive, and they are essential for introducing non-linearity into the model, allowing it to learn complex patterns.

Figure 9 illustrates how the loss metric decreases initially but then plateaus, indicating that further iterations do not significantly impact the loss reduction. This phenomenon suggests that the model reaches a point where adjustments in the learning rate or loss function do not materially affect its performance. One potential reason for this stagnation is that the chosen activation function might be leading to the deactivation of a large number of neurons within the network. When neurons are deactivated, they do not contribute to the network’s learning process, which can halt progress in improving the model’s classification capabilities. This condition, known as the vanishing gradient problem, can severely limit the

<table>
<thead>
<tr>
<th>Table 4. Adjustment of random_state in SMOTE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random_state</td>
</tr>
<tr>
<td>Loss</td>
</tr>
<tr>
<td>Accuracy</td>
</tr>
<tr>
<td><strong>Source(s):</strong> Authors’ own creation/work</td>
</tr>
</tbody>
</table>
network’s ability to learn from the training data, suggesting that a reassessment of the activation function might be necessary to enable further improvements in model performance.

(b) Model prediction

Figure 10 presents the risk curve forecasting the likelihood of hard drive failure, emphasizing the temporal proximity to potential downtime events. The curve assigns a value of 1 to the period encompassing the seven days immediately preceding a downtime event, signifying a markedly increased probability of hard drive failure within this timeframe. As the timeline extends further back from the imminent downtime, the risk curve demonstrates a decline in the probability of failure.

This visual representation encapsulates the model’s ability to quantify and predict the risk of hard drive failure over time. By analyzing historical data and identifying patterns that precede failures, the model can assign higher risk scores to periods closer to the failure event, reflecting an elevated likelihood of downtime. Conversely, as the distance from the predicted failure event increases, the model’s assessed risk level diminishes, illustrating a lower probability of imminent failure.

The risk curve serves as a crucial tool for preemptive measures, enabling system administrators and IT professionals to implement targeted maintenance and interventions. By accurately predicting the risk probability before actual downtime occurs, the model provides a window of opportunity for taking corrective actions, potentially averting costly downtimes and ensuring the reliability and continuity of operations. This predictive capability not only enhances the understanding of hard drive failure dynamics but also contributes significantly to improving system resilience and efficiency.

Figure 11 validates the hypothesis that the further in advance a date is from an impending hard drive breakdown, the lower the risk of that breakdown occurring. This figure portrays the outcome of predictions across each defined bag, with the left histogram showcasing the count of predicted downtimes for each bag, and the right histogram aggregating these predictions to reflect the overall risk of downtime.

**Source(s):** Authors’ own creation/work
The essence of these histograms is distilled into a risk curve, mirroring the analytical approach depicted in Figure 11. The lower curve delineates the risk associated with each individual bag, while the upper curve aggregates these risks to indicate the overall likelihood of hard drive failure over time. This dual-layered representation enables a nuanced understanding of how risk evolves as the predicted date of breakdown approaches.

The practical application of this model is significant for enterprises or factories that operate under specific risk thresholds, such as an 80% warning level. By integrating this predictive model, organizations can gain a foresighted view of potential downtimes, enabling them to enact preventative measures well in advance. This anticipatory capability is crucial for maintaining operational continuity, minimizing downtime, and optimizing maintenance schedules. Essentially, the model provides a strategic tool for operational risk management, offering a tangible means to predict and mitigate the impact of hardware failures before they culminate in disruptive downtime.

5. Discussion

The study's exploration into the efficacy of SMOTE in enhancing model performance is encapsulated in the comparison presented in Table 5. SMOTE is a widely recognized approach for addressing class imbalance within datasets by artificially generating synthetic examples of the minority class, thus aiming to provide a more balanced dataset for training machine learning models. This technique is particularly relevant in scenarios like hard drive

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>BBDL</td>
<td>99.99%</td>
<td>63.63%</td>
<td>100%</td>
</tr>
<tr>
<td>BBDL without SMOTE</td>
<td>96.80%</td>
<td>0.43%</td>
<td>75%</td>
</tr>
</tbody>
</table>

Table 5. The impact of SMOTE
failure prediction, where the event of interest (failure) is significantly rarer compared to the non-failure state, leading to a skewed distribution of classes.

Table 5 details the outcomes of models trained with and without the application of SMOTE, allowing for a direct evaluation of its impact on the model's ability to predict hard drive failures. Key performance metrics such as precision, recall, F1 score, and accuracy would be essential in this comparison, providing insight into not just the model's overall accuracy, but also its sensitivity (recall) to detecting failures and the precision in its predictions.

The utilization of SMOTE could potentially lead to improvements in these metrics, particularly in enhancing the model's recall by reducing the bias towards the majority class. This would be an important finding, as it would suggest that SMOTE effectively counters the class imbalance problem, making the model more robust in predicting rare failure events. Such an outcome would underscore the value of employing advanced data preprocessing techniques like SMOTE in predictive maintenance models, especially in industries where early detection of potential failures can significantly reduce downtime and associated costs.

SMOTE is used to handle imbalanced data by oversampling the minority class. When SMOTE is applied, the reliability of the metrics would generally be expected to increase for the minority class, as the model is given more examples to learn from. This could potentially lead to a better generalization when predicting new, unseen data. Reliability here also depends on how well the synthetic samples generated by SMOTE represent the true underlying distribution of the minority class. If the synthetic data is too similar to the original data, it could lead to overfitting, which is a reliability concern.

Without SMOTE, the model may be more biased toward the majority class, and the reliability of its performance metrics for the minority class may be lower. This is because the model has seen fewer examples of the minority class and is less likely to predict it correctly. However, the metrics without SMOTE could be considered more reliable in reflecting the model's performance on the original, unmodified dataset. It provides a clearer picture of how the model would perform in real-world conditions where the imbalance exists.

In both cases, a reliability analysis would benefit from cross-validation, where the data is split into multiple sets, and the model is trained and tested several times. This helps in assessing the stability of the model's performance. Consistent results across folds can increase confidence in the reliability of the reported metrics.

To validate the broad applicability of the Bagging-Based Deep Learning (BBDL) framework, this study conducts a comparative analysis between the ST10000NM0086 dataset and other datasets, such as ST12000NM0007 and ST4000DM000. Notably, the latter datasets are characterized by a larger overall size but possess a lower proportion of true failure instances, which inherently influences the performance metrics, specifically recall and precision. Despite these challenges, the accuracy levels across the datasets remain consistently high, indicating a robust degree of model stability.

Table 6 likely presents a detailed comparison of these performance metrics, illustrating the model's capability to maintain high accuracy across different datasets while highlighting variations in recall and precision rates. The lower recall and precision observed in datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Loss</th>
<th>Accuracy</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>ST10000NM0086</td>
<td>0.2410</td>
<td>99.99%</td>
<td>100%</td>
<td>63.63%</td>
</tr>
<tr>
<td>ST12000NM0007</td>
<td>0.2473</td>
<td>99.98%</td>
<td>97.06%</td>
<td>55.93%</td>
</tr>
<tr>
<td>ST4000DM000</td>
<td>0.2467</td>
<td>99.99%</td>
<td>98.15%</td>
<td>67.09%</td>
</tr>
<tr>
<td>ST8000NM0055</td>
<td>0.2489</td>
<td>99.97%</td>
<td>94.12%</td>
<td>38.10%</td>
</tr>
<tr>
<td>ST500LM030</td>
<td>0.2476</td>
<td>99.97%</td>
<td>100%</td>
<td>84.62%</td>
</tr>
</tbody>
</table>

Source(s): Authors’ own creation/work

Comparison between different models
with a larger size and fewer true examples underscore the impact of class imbalance on model performance. However, the consistency in accuracy across different datasets signifies the BBDL model’s strong generalization ability.

Furthermore, the analysis suggests that optimizing the parameters specific to the model’s architecture could lead to improvements in both accuracy and recall. This potential for enhancement through fine-tuning indicates that the BBDL framework is not only adaptable across various datasets but also amenable to adjustments that could further elevate its predictive performance. Such findings are encouraging, as they demonstrate the BBDL model’s versatility and effectiveness in predicting hard drive failures across a spectrum of scenarios, affirming its utility as a reliable tool for predictive maintenance in real-world applications.

The comparison highlighted in Table 7 provides a clear distinction between two training methodologies for the Bagging-Based Deep Learning (BBDL) model, with a specific focus on addressing hard drive failure prediction. The first row of the table likely presents results from a training approach that prioritizes minimizing loss, while the second row introduces a hybrid strategy that considers both loss reduction and accuracy enhancement.

This nuanced approach, which incorporates accuracy considerations into the training process, demonstrates a strategic shift towards balancing the model’s performance metrics. Although prioritizing loss minimization is a common practice aimed at improving the model’s predictive capabilities, integrating accuracy as a parallel objective introduces a more holistic perspective on model optimization.

The slight increase in loss observed when incorporating accuracy considerations suggests a trade-off, where the model may allow for a marginal elevation in error to achieve a significant improvement in correctly identifying hard drive failures. This is a critical insight for predictive maintenance applications, where the ability to accurately predict failures can offer substantial practical benefits.

The enhanced model accuracy resulting from this dual-focused training approach signifies a more effective tool for diagnosing hard drive issues, optimizing maintenance schedules, and ultimately, reducing downtime. By demonstrating that a balance between loss minimization and accuracy enhancement yields superior outcomes for hard drive failure prediction, this comparison underscores the importance of adopting multifaceted objectives in model training to address complex real-world problems effectively.

Table 8 offers a comparative analysis of the Bagging-Based Deep Learning (BBDL) model’s performance against other methodologies, specifically highlighting the effectiveness of incorporating oversampling techniques like SMOTE in improving prediction accuracy for

<table>
<thead>
<tr>
<th>Table 7. Comparison of Loss function and loss + accuracy function</th>
<th>Loss</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loss function</td>
<td>0.2504</td>
<td>99.99%</td>
</tr>
<tr>
<td>Loss + accuracy function</td>
<td>0.2510</td>
<td>96.37%</td>
</tr>
<tr>
<td>Source(s): Authors’ own creation/work</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 8. Results</th>
<th>Loss</th>
<th>Accuracy</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>BBDL</td>
<td>0.2510</td>
<td>99.99%</td>
<td>100%</td>
</tr>
<tr>
<td>SVM with oversampling</td>
<td>\</td>
<td>98%</td>
<td>\</td>
</tr>
<tr>
<td>Decision tree</td>
<td>\</td>
<td>84.195%</td>
<td>\</td>
</tr>
<tr>
<td>Regression</td>
<td>\</td>
<td>77.858%</td>
<td>\</td>
</tr>
<tr>
<td>Neural network</td>
<td>\</td>
<td>77.855%</td>
<td>\</td>
</tr>
<tr>
<td>Source(s): Authors’ own creation/work</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
hard drive failures. The table presents a clear comparison between the proposed BBDL approach, a Support Vector Machine (SVM) method without SMOTE, and a classifier without any form of oversampling.

The findings underscore a crucial insight: while the BBDL method outperforms the SVM approach, the advantage of using SMOTE becomes evident. Oversampling techniques like SMOTE are designed to address class imbalance by artificially generating new instances of the underrepresented class (in this case, hard drive failures), thereby providing a more balanced dataset for training predictive models. This balance is critical in scenarios where the occurrence of one class is significantly lower than the other, as is common in predictive maintenance for hard drive failures.

The SVM method without SMOTE, which uses a bagging oversampling approach, demonstrates inferior results compared to the BBDL model. This disparity highlights the BBDL model’s superior capacity to handle imbalanced data effectively, especially when enhanced with SMOTE. Moreover, the classifier without SMOTE shows a marginal decrease in overall accuracy but exhibits a higher precision in predicting actual failures. This suggests that while the model may correctly identify more true failures, it risks misclassifying normal operations as potential failures, reflecting a trade-off between sensitivity and specificity.

This comparison elucidates the nuanced impact of different oversampling techniques on predictive performance. The BBDL model’s superiority is attributed to its ability to better generalize from the balanced dataset provided by SMOTE, leading to more accurate predictions of hard drive failures. This is particularly valuable in practical applications where the cost of missing a potential failure (false negatives) can be significantly higher than the inconvenience of investigating a false alarm (false positives).

In summary, the analysis in Table 8 affirms the BBDL model’s effectiveness in hard drive failure prediction, underscoring the importance of appropriate data preprocessing techniques like SMOTE in enhancing model performance, especially in the presence of class imbalances. This insight is crucial for developing robust predictive maintenance systems that can reliably forecast equipment failures, enabling timely interventions to prevent downtime and loss.

6. Conclusion
This study proposed a data enhancement method based on SMOTE and Cost-Sensitive was proposed on hard drive fault diagnosis. Data information learned from imbalanced datasets was applied to data enhancement through SMOTE, where the available data is extremely imbalanced. Cross-Entropy loss function was used to train the classification network and penalty factors were imposed on imbalanced categories, which reduce misdiagnosis on hard disk downtime. A contribution of this paper is the development of a bagging algorithm designed specifically for issuing alerts ahead of hard disk failures, based on the principle that there are detectable signs before a failure occurs. The model leverages the SMOTE and a Cost-Sensitive approach to augment the minority class in the dataset, addressing the imbalance issue. Additionally, it adjusts the Cross-Entropy loss function by incorporating a weighting scheme that prioritizes the infrequent categories. By integrating focal loss into the Cross-Entropy framework, the model further hones its focus on training samples that are challenging to learn from.

Subsequent empirical studies have predominantly focused on forecasting hard disk failures and managing the incidence of false alarms, achieving a fault detection rate ranging only between 3 and 10%. Considering that nearly 90% of data is stored on hard disk drives, both users and manufacturers prioritize the reliability of these storage devices, emphasizing the need for early failure warnings to safeguard critical data. Their primary concern is the precision of fault diagnosis over the prediction of failure timing. This study aims to enhance
the detection rate of rare failures by integrating data augmentation and penalty factors, alongside introducing a bagging model designed to predict the timeframe of potential hard disk failures. This approach is intended to provide users and manufacturers with ample opportunity for data migration or backup, addressing their core concerns.

Industrial data-driven modeling for imbalanced fault diagnosis is a field of critical importance and immense challenge. The unique difficulties presented by imbalanced datasets require innovative solutions that are constantly evolving. This study contributes a systemic way to support the problem solving. Through a combination of data-level, algorithm-level, and hybrid approaches, study has been made in accurately diagnosing industrial faults. The field is set to benefit from advances in machine learning, artificial intelligence, and data collection technologies, promising even greater efficiencies and reliability in industrial operations. As we continue to navigate the complexities of imbalanced fault diagnosis, the collaborative efforts of researchers, practitioners, and industry stakeholders will be pivotal in shaping the future of this vital domain. Looking ahead, the integration of data-driven modeling for imbalanced fault diagnosis with emerging technologies such as the Internet of Things (IoT) and edge computing presents exciting opportunities. Real-time data collection and analysis could further enhance the predictive capabilities of these models, leading to even more proactive and efficient fault management strategies.

The proposed data-driven modeling approach to tackle extremely imbalanced datasets in industrial scenarios, while innovative, is not without its limitations. One key challenge stems from the reliance on SMOTE for data augmentation. While SMOTE can be effective in generating synthetic data points to balance classes, its algorithm assumes that intermediate points between minority class examples are also valid samples. In the diverse and dynamic environment of data centers with a variety of disk brands and models, this assumption might not hold true. The synthetic samples created may not adequately represent the complex, real-world failure patterns of hard disks. This can lead to a model that is artificially confident, misclassifying new or unseen data points. Additionally, the use of the Cross-Entropy loss function, despite being enhanced with penalty factors, may not completely overcome the skewed distribution of data. This function tends to push the decision boundary towards the minority class, which could result in a higher false-positive rate, potentially leading to unnecessary replacements or maintenance actions.

To address these challenges, several strategies could be implemented. Firstly, advanced data augmentation techniques that consider the underlying distribution of the data more carefully could be developed. These techniques would involve more sophisticated interpolations between data points or even the creation of entirely new data points based on the learned distribution of the minority class. Machine learning models that incorporate notions of uncertainty and confidence, such as Bayesian networks or ensembles of decision trees, could provide more reliable predictions by quantifying the uncertainty of the model in its predictions. Another approach could be to adapt the cost-sensitive learning strategy to more dynamically adjust the penalties as the model learns, perhaps through a feedback mechanism informed by the model’s performance on a validation set. Additionally, incorporating anomaly detection strategies could help identify outliers or novel patterns of disk failures that are not represented in the training set. By doing so, the model could become more robust to the actual operational conditions of the data centers, ensuring that the predictions remain accurate and useful over time.

As the domain of data-driven modeling for fault diagnosis matures, future research is poised to pivot towards crafting solutions that are more adaptable, interpretable, and holistic, fully capitalizing on the evolving capabilities of AI and machine learning technologies to bolster industrial robustness and efficacy. The trajectory of these advancements is expected to include the development of models that can dynamically adjust to new data and evolving conditions, which is critical in industrial applications where operational parameters are continually
Interpretability is another frontier, as the need for transparent AI becomes more pronounced; understanding the decision-making process of models is vital for trust and regulatory compliance. Additionally, integrating disparate data sources and types into cohesive models will likely be a key area of focus, ensuring that the rich tapestry of available data can be synthesized effectively. The establishment of standardized frameworks and performance benchmarks will play a crucial role in this progression, offering a common ground for assessing and comparing the efficacy of different fault diagnosis models. Such standards promise to propel the field forward, stimulating innovation and the refinement of methodologies in this increasingly critical area of industrial maintenance and reliability engineering.

References


IMDS


Corresponding author
Thitipong Jamrus can be contacted at: thitja@kku.ac.th

For instructions on how to order reprints of this article, please visit our website: www.emeraldgrouppublishing.com/licensing/reprints.htm Or contact us for further details: permissions@emeraldinsight.com