ISSN: 2583-9888
DOI: doi.org/10.51219/JAIMLD/nagamalleswararao-bellamkonda/654

URF PUBLISHERS
connect with research world

Journal of Artificial Intelligence, Machine Learning and Data Science

https://urfpublishers.com/journal/artificial-intelligence

Vol: 4 & Iss: 1

Research Article

Applying Machine Learning to Detect and Prevent Performance Regression in Production Databases

Nagamalleswararao Bellamkonda*

Citation: Bellamkonda N. Applying Machine Learning to Detect and Prevent Performance Regression in Production Databases. *J Artif Intell Mach Learn & Data Sci 2026* 4(1), 3233-3241. DOI: doi.org/10.51219/JAIMLD/nagamalleswararao-bellamkonda/654

Received: 14 January, 2026; **Accepted:** 19 January, 2026; **Published:** 21 January, 2026

***Corresponding author:** Nagamalleswararao Bellamkonda, Sr. Database Administrator, USA

A B S T R A C T

Performance regression is a significant operational problem in production databases, particularly in large-scale data-intensive systems, where a single performance degradation can propagate to significant service failures. The old way of monitoring database performance through traditional monitoring with primarily fixed thresholds and rule-based alerts is not applicable in dynamic workloads, particularly in capturing the complex and dynamic patterns of performance. As previously mentioned, this study describes how machine learning can be utilized to forecast and prevent the decline in the performance of production database systems. A theoretical framework is presented, which assumes the use of telemetry collection on a continuous basis, feature engineering, anomaly detection, predictive modelling and automated response mechanisms. The structure assists in detecting regressions and proactively preventing them by learning normal performance behaviour and predicting potential degradations prior to the user facing the effect of said degradation. This paper summarizes recent research on machine learning-powered system monitoring, examines the types of models that should be used when processing data on database performance and provides the difficulties in practice, such as drift in data, model explainability and operational cost. The findings suggest that machine learning-focused solutions may be more generalized, detect countermeasures earlier and have a lower false positive rate than traditional monitoring systems. Therefore, they represent an option to consider when managing database performance in modern environments.

Keywords: Machine learning; Performance regression; Production databases; Anomaly detection; Predictive monitoring; Database performance management

1. Introduction

Modern information systems have production databases as their operational foundation platform, which underpins mission-critical applications such as online transaction processing, real-time analytics and big-data services. As these systems are constantly being upgraded, their schema changes, query optimization, scaling of infrastructure and workload shifting patterns, the labour and cost of maintaining reliable and predictable performance become increasingly complex. One of the recurrent problems in this field is performance regression, which can be described as a continuous decrease in database performance when compared with the previously set standard in a similar setting.

One of the most difficult to contain is the production database performance regression, which is largely cumulative over time compared to a failure that occurs immediately. Slight delays in query processing, a linear rise in the use of resources or slight changes in execution plans can go unnoticed individually but can ultimately damage the stability of the system and lead to the service level objectives being breached. The primary features

of traditional database monitoring tools are the utilization of predefined alerts and diagnoses and the use of static thresholds. These methods are ill-adapted but capable of success in determining slow decays or high-order effects between two or more performance measurements under varying conditions **(Table 1)**.

Table 1: Comparison of Traditional Monitoring and Machine Learning-Based Approaches.

Aspect	Traditional Rule-Based Monitoring	Machine Learning–Based Monitoring
Baseline Definition	Static, manually defined thresholds	Learned dynamically from historical data
Adaptability to Workload Changes	Low	High
Detection of Gradual Regression	Limited	Strong
False Positive Rate	High under variable workloads	Reduced through adaptive modeling
Predictive Capability	None	Supports forecasting and early warning
Operational Effort	Manual tuning and constant adjustment	Higher initial setup, lower longterm effort

Source: Adapted from Chandola et al. (2009); Laptev et al. (2015); Gulenko et al. (2021); Li et al. (2024).

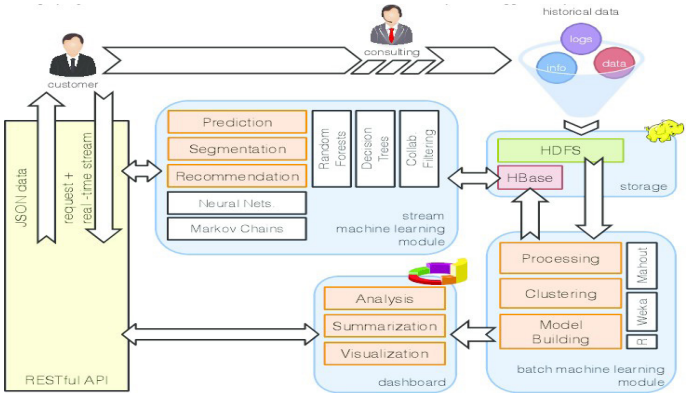
Modern production databases, especially those applied to cloud and distributed systems, are not stationary. The workloads are responsive to the demand of the user, background jobs and release of applications, but the infrastructure resources are dynamically rescaled. The contextual value of fixed performance levels is lost in this case, which causes an inflated false alarm rate when the workload is true and an amplified false alarm rate when the workload is slack. This weakness is one of the fundamental disparities between traditional monitoring methods and the dynamism of modern database systems.

Another alternative to this is machine learning where systems are in a position to learn normal performance behaviour through historic and real-time telemetry. The dynamic baselines and models created by machine learning models can represent time patterns, correlations and multidimensional relationships between measurements of query execution time, CPU usage, memory pressure and I/O activity that vary with the system. The deviations between these learned baselines may be considered as the initial regression in performance, even when no predefined threshold is breached.

It is also interesting to note that techniques based on machine learning are suitable for both regression detection and prevention. According to the current and previous observations, predictive models can forecast the future performance of the system and can be proactively taken by the automated system or the database administrator. Examples of such interventions include query refactoring, index maintenance, configuration tuning and pre-emptive resource allocation. Machine learning can significantly reduce downtime, operation costs and performance deterioration as perceived by users by removing reactive troubleshooting and substituting it with predictive performance management.

Despite these advantages, the implementation of machine learning as a means to control the performance of production databases has several problems related to data quality, model interpretability and resource usage, in addition to mutual support with other existing monitoring processes. To address these concerns, this study solves them by referring to machine learning techniques, which can be applied in the field of detecting and preventing performance regressions and develops a conceptual framework that agrees with predictive analytics using database operations that can be realized.

1.1. ML-Driven performance regression detection



2. Background and Problem Definition

2.1. Production system performance regression analysis

Performance regression is the continual worsening of the behaviour of a system compared to its original performance baseline following a system modification, system configuration, system workload or system environment change. Regressions, especially in production systems, are a nightmare because they usually creep up and are not noticed until they affect the end users or cause a breach in service-level targets. However, in contrast to functional failures, performance regressions do not always lead to system crashes; rather, they are reflected in longer response times, poorer throughput, inefficient use of resources or unstable system behaviour.

Empirically, the regressions of performance have been identified in systems with changing software and are challenging to detect because even simple code modifications, system settings and runtime loads interact with one another in a complicated manner[1]. Even minor adjustments, such as a change in query form, indexing choices or implementation plans, can cause non-obvious performance side effects in database-centric systems, which are propagated through deployments.

2.2. Drawbacks of the traditional regression detection methods

Traditional performance monitoring methods in production settings mainly focus on rule-based systems, predetermined limits and human analysis. These models presuppose that regular system performance can be adequately described by rigid constraints in the values of vital metrics, such as latency, CPU intensity or memory use. However, this assumption is challenged by modern production environments that are dynamic and non-stationary.

Industrial experiments in both production and manufacturing systems have indicated that the concept of static monitoring products is not well adapted to scenarios in which the workload properties and operating conditions frequently switch[2,3]. Fixed thresholds in these environments tend to produce too many

false positives during legitimate workload extremes and do not detect slow, cumulative performance degradation. This is further worsened in data-intensive systems, where the performance behaviour is determined by the interactions between multiple variables that are not linear.

Moreover, conventional monitoring tools are reactive by definition. They realize the problems when performance indicators have already surpassed the established boundaries, which allows very little in the way of preventing intervention. Industrial software system case studies suggest that such a reactive posture adds a lot of time to the diagnostic process and operational expenses, especially when such regressions can be traced to subtle causes instead of isolated faults[1].

2.3. Machine learning in production monitoring

The increasing presence of high-resolution telemetry information has allowed the realization of machine learning methods for monitoring and decision support in production settings. Machine learning models can learn patterns based on past and real-time information and these are complex relationships that are difficult to define with manual rules. In the industrial and manufacturing spheres, systematic literature reviews indicate the growing popularity of machine learning in fault detection, quality forecasting and performance optimization as part of the wider paradigm of Industry 4.0[2,4,5].

In factories, machine learning has been used to forecast delays, anomalies and/or predict resource consumption more precisely than traditional statistical approaches[6,7]. These strategies change the paradigm of monitoring as a threshold-based system and transform it into more adaptive and data-driven modelling that allows systems to draw the line between a normal variation in workload and the adoption of an abnormal behaviour of performance **(Table 2)**.

Table 2: Characteristics of Performance Regression and Detection Challenges in Production Databases.

Dimension	Description	Implication for Detection
Onset Pattern	Gradual and cumulative degradation rather than abrupt failure	Static thresholds often fail to trigger alerts
Root Causes	Query plan changes, configuration updates, workload evolution	Manual diagnosis becomes time-consuming
Metric Behavior	Multidimensional (latency, CPU, I/O, memory) and correlated	Single-metric monitoring is insufficient
Environment	Dynamic, non-stationary production workloads	Fixed baselines lose validity over time
Detection Approach	Rule-based vs. data-driven	Machine learning enables adaptive baselines
Prevention Capability	Reactive in traditional systems	Predictive modeling supports proactive action

Source: Adapted from Nguyen, et al.[1]; Jung, et al.[8]; Kang, et al.[2]; Usuga Cadavid, et al.[3].

2.4. Performance regression detection by use of machine learning

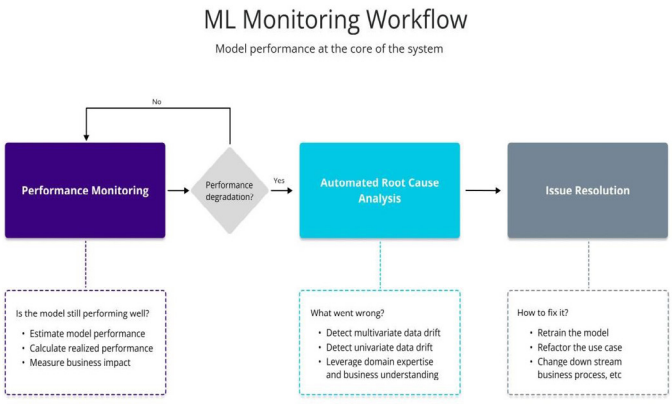
The implementation of machine learning with regard to performance regression detection utilizes normal behaviour modelling of the system and creates deviations that are known and will continue to exist in the long run. Previous studies in the field of database systems have shown that data driven automated methods are more effective at catching and diagnosing performance regressions than manual inspection[8]. With the help of multidimensional performance measures, machine learning models can identify regression patterns that cannot be visualized when the measures are analysed separately.

Similar research in the fields of production and manufacturing also promotes the relevance of machine learning in determining the degradation pattern, anticipating failures and classifying performance-related conditions regardless of the operating conditions under uncertainty[2,9]. These observations indicate that the problem of performance regression detection can be formulated as a learning problem in which models are continuously updated by responding to changes in system behaviour.

2.5. Problem statement

Although machine learning applications for production monitoring have been proven to be successful, there are still several gaps in the performance regression of production databases. First, the currently available methods are mostly aimed at detecting anomalies without making a clear distinction between temporary anomalies and permanent regressions. Second, most studies focus on detection and not prevention, which restricts their capability to facilitate proactive performance management. Third, integration challenges, including explainability, operational overhead and alignment with existing workflows, are poorly addressed in practical deployments.

This study fills these gaps by exploring the ways in which machine learning methods can be used systematically to identify and prevent performance regressions in production database settings. The main issue that the current study addresses is the way to structure an adaptive, decipherable and practically viable machine learning system that detects performance regressions at the earliest possible stage and allows taking proactive measures prior to the development of severe degradation **(Figure 3)**.



3. Related Work

Performance analysis based on machine learning has been studied in various fields, including database management systems, software systems and industrial production environments. Although these bodies of work are often treated as independent entities, they possess several similar objectives: the definition of degradation, prediction of future behavior and active decision making in complex systems. This section considers existing research and compares the current study with the existing literature.

3.1. Machine learning in production and industrial systems

Machine learning has been extensively studied in the framework of industrial production systems, particularly under the industry 4.0 paradigm. Systematic literature reviews have shown that the use of data-based models in production lines to monitor, predict and optimize them is growing[2,4]. These studies confirm that machine learning methods are better than rule-based and purely statistical methods in cases where multivariate, complex and dynamic data on production exist.

Production planning and control studies have also revealed that machine learning models may be applied to environments where conditions may be considered uncertain, variable and changing[3]. Likewise, studies on quality prediction using data are aimed at enhancing the application of previous and current data to forecast the appearance of variations that could cause flaws or delays[5,10].

These studies analyse the most manufacturing and industrial-based cases; however, they also provide useful information that can be utilized in production databases. Both realms are defined by performance sensitivity in workload and configuration and are marked by incessant activity and data volume.

3.2. Database systems and performance regression detection software

Performance regression has been widely studied in the context of software engineering, with particular attention to evolving systems. According to Nguyen, et al.[1], an industrial case study indicates that performance regressions are frequently added to software during the process of its evolution and cannot be easily detected with the help of people. Their findings demonstrate that automated procedures are necessary to assist in establishing the causes of regression in different objects of the system.

Jung, et al.[8] offer an automated system architecture of analysing performance characterization of databases and their execution characteristics to observe and diagnose performance backslides in database frameworks. They provided an example of one of their works in which regression detection can be moulded into a data-driven problem and hence enable more precise and timely detection of degradation as compared to the conventional monitoring methods. However, they are more of detection and not exhaustive in the prevention of regression using predictive modelling **(Table 3)**.

Table 3: Summary of Related Work Across Domains.

Study Domain	Representative Works	Primary Focus	Key Limitation
Production & Manufacturing	Kang, et al.[2]; Usuga Cadavid, et al.[3]	ML for monitoring and optimization	Limited focus on database systems
Software Performance Regression	Nguyen, et al.[1]	Regression cause identification	Manual effort and post-hoc analysis
Database Performance	Jung, et al.[8]	Automated regression detection	Limited emphasis on prediction
Predictive Modelling	Matsunaga & Fortes[7]; Ibrahim, et al.[11]	Forecasting performance trends	Not database specific
Fault Detection & Prognosis	Fernandes, et al.[9]; Kang, et al.[2]	Early degradation detection	Focus on physical systems

Source: Synthesized from the approved reference set.

3.3. Predictive modelling and regression forecasting

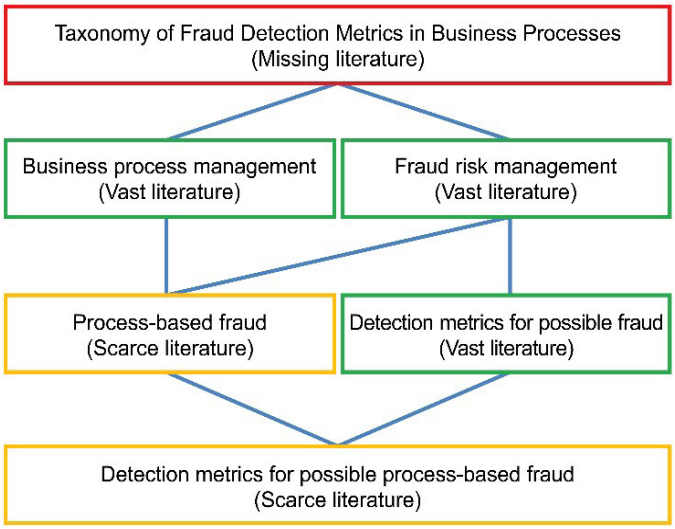
In addition to detection, other studies view predictive modelling as the ability to forecast performance and resource consumption. Matsunaga and Fortes demonstrated the application of machine learning to predict program performance in terms of the time of execution and resource usage[7], which means that performance behaviour may be foretold with a reasonable degree of accuracy. Later literature applies similar ideas of predictive analytics in the industry, including the prediction of production delays and regressive models[6,11].

These studies show that machine learning can capture the trends and patterns of degradation over time and proactive interference can be introduced. Most predictive studies examine the outputs of production or the consumption of resources but not database-specific metrics, such as query or execution efficiency latency.

3.4. Fault detection, prognosis and degradation analysis

Machine learning has also been widely applied to fault detection and prognostication in industrial systems. According to reviews by Fernandes, et al.[9] and Kang, et al.[2], data-driven models can be helpful in identifying small-scale degradation and categorizing the states of systems before disastrous failures occur. This can also be directly transferred into the performance regression in databases, which can be viewed as a non-fatal but chronic system degradation.

Notwithstanding this, despite the similarity in the methodology used in the literature on fault diagnosis, a gap exists between the modelling of industrial degradation and database performance management, as the literature seldom examines database systems at a particular level **(Figure 4)**.



4. Methodology

4.1. Proposed machine learning framework

The conceptual approach will be used to define and prevent performance regression in production databases using machine learning on real data. Instead of concentrating on a single algorithm, the methodology highlights an end-to-end architecture, which entails data, model and inference, as well as decision-making within an operational database setting. Such design decisions can be justified by the findings of previous industrial and software systems research, which shows that

performance regression is a general phenomenon and is a combination of many factors influencing it[1,8].

4.2. Framework overview

The suggested architecture was developed as a learning and monitoring pipeline. It presupposes that production database telemetry has access to performance, workload and system resource indicators. Machine learning models are also trained to learn baseline performance behaviour and detect deviations that occur in the long run; therefore, they can differentiate long-run regressions and short-term anomalies. Similar to the industrial machine learning study, the framework can operate under non-stationary conditions and adjust to varying workloads and system conditions[2,3].

4.2.1 Sources of data collection and telemetry: The quality and roughness of the information gathered are pertinent to proper regression identification. The framework consumes multidimensional streaming telemetry, such as query execution time, throughput, CPU and memory consumption, disk I/O activity and concurrency. It has already been stated in the earlier literature that individual measures are not sufficient to explain complicated performance degradation patterns[5,8]. To this extent, the methodology can be perceived as being based on large-scale and continuous data collection at the query and system levels.

4.3. Feature engineering and preparation of the model

Unstructured Telemetry Data are converted to structured features that may be used by machine learning models. Temporal aggregation, trend extraction and normalization are considered a subset of feature engineering to address workload fluctuations. More precisely, the most applicable statistics are the rolling ones and rate-of-change measures, which, as indicated by studies on predictive modelling, may be used in the production system[6,7]. This action is also performed to address missing data and noise that are experienced in the production world.

Table 4: Methodological Components of the Proposed Framework.

Framework Component	Purpose	Supporting Literature
Telemetry Collection	Capture multidimensional performance data	Jung, et al.[8]; Md, et al.[5]
Feature Engineering	Extract trends and degradation indicators	Matsunaga & Fortes[7]
ML Modelling	Learn baseline behaviour and detect deviations	Kang, et al.[2]; Fernandes, et al.[9]
Decision Logic	Distinguish anomalies from regressions	Nguyen, et al.[1]
Preventive Feedback	Enable proactive intervention and adaptation	Kang, et al.[2]; Sircar, et al.[12]

Source: Synthesized from the approved reference set.

4.4. Model inference and training

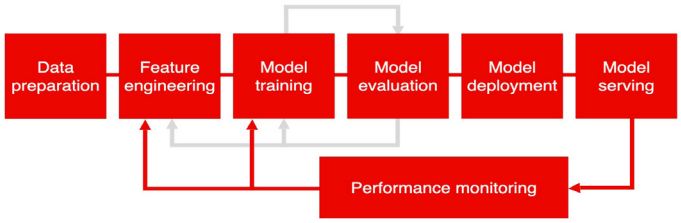
The framework asserts some of the machine learning paradigms used, such as unsupervised models, baseline learning and supervised or regression-based models, in case there exists a set of labelled data. During training, models are exposed to the common behaviour of performance under different workload conditions. As part of the inference, the incoming telemetry was compared with the acquired patterns and anomalies above the adaptive thresholds were identified. This result is also in line with existing results that adaptive data-driven models are better in dynamic environments compared to the family of rule-based systems[2,9].

4.5. Decision logic and regression classification

The identified deviations were determined and found over time to establish a temporary or permanent decline in performance. Such decision criteria entail persistence, magnitude and cross metric correlation. The difference is dominant and the history of industrial cases states that the use of all the anomalies as regressions is the cause of alert fatigue and ineffective remediation processes[1]. The framework then imposes time consistency checks before raising alerts or preventive measures.

4.6. Preventive measures and feedback loops

The framework can be used to respond pre-emptively as soon as a regression is established or predicted. This may be done by issuing warnings to database administrators, suggesting configuration changes or even auto-responding, such as for scaling of resources. The feedback from these interventions is reintroduced into the learning process, allowing for continuous improvement of the model **(Figure 5)**. It is a closed-loop design based on the best practices in production and industrial machine learning[2,12].



5. Performance Regression Detection using machine learning

Learning about complex and evolving patterns of high-dimensional telemetry data is necessary to detect performance regression in production databases. Regressions are not sudden failures but tend to be a smooth divergence over many correlated measures; thus, they are difficult to describe in terms of static rules. Machine learning methods overcome this difficulty by designing the normal behaviour of systems and then detecting anomalies, which are everlasting. In this section, the key categories of machine learning methods suitable for regression detection are discussed and their applicability to production database frameworks is evaluated.

5.1. Unsupervised learning of baseline model

Unsupervised learning methods have found these applications, especially in situations where regression data labels are limited or unavailable, as is typical in real-world production systems. These techniques identify the usual working behaviour of a system using past data and alert deviations as possible regressions. Systematic reviews of industrial machine learning applications have highlighted the efficiency of unsupervised models in complicated production settings, where clear fault labels cannot be readily acquired[2,4].

Unsupervised models in the context of production databases can capture baseline relationships between metrics such as query latency, throughput and resource utilization. The continued violation of these acquired baselines and the lack of

a single anomaly are characteristic of performance regression. Nonetheless, unmonitored techniques usually require some extra decision-making to differentiate between workload shifts and actual degradation.

5.2. Regression classification with supervised learning

Supervised learning methods are based on labelled samples of performance regressions and normal behaviour. In cases where historical incidents of regression are well documented, explicit mappings of performance patterns onto regression outcomes can be learned using supervised classifiers. Industrial case studies indicate that regression causes can be properly identified using regression supervised models if adequate labelled data are provided[1].

Practically, supervised learning is most appropriate in mature production environments where incident management processes are in place. Its main shortcoming is that it is costly and subjective regarding the labelling of regression events and lacks flexibility when the behaviour of the system goes beyond the range of the training data.

5.3. Regression and time-series prediction models

Time-series regression models will also expand the detection capabilities owing to their ability to predict future performance patterns. These models are used to forecast how systems should behave and they are compared to observed measures rather than just identifying deviations compared to historical baselines. The demonstration of predictive modelling studies in the production and industrial sectors has shown that these methods have the potential to determine degradation paths before the performance limits are breached[6,7].

In the case of production databases, predictive models allow an early warning of imminent performance decline, which can be used to intervene. Long-term regression modelling can also be successfully used in areas where continuous production data are available, such as oil and gas systems, where gradual degradation is typical (Ibrahim et al., 2022). These results prove the relevance of predictive learning methods for database performance management **(Table 5)**.

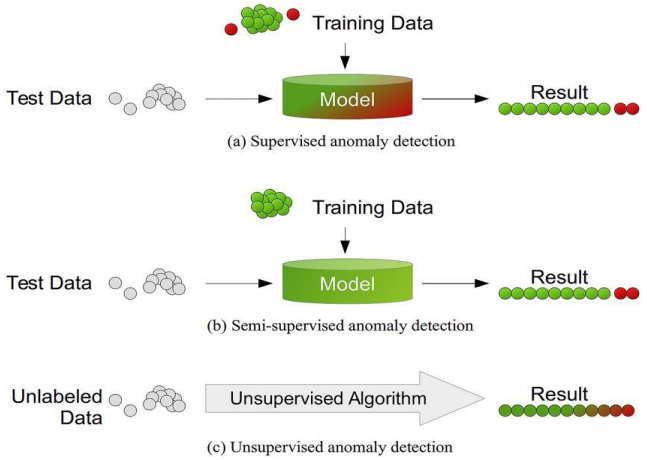
Table 5: Comparison of Machine Learning Techniques for Regression Detection.

Technique Type	Data Requirement	Strengths	Limitations	Representative Studies
Unsupervised Learning	Unlabelled historical data	Adaptive baselines; low labelling cost	Difficulty distinguishing workload shifts	Kang, et al.[2]; Fahle, et al.[4]
Supervised Learning	Labelled regression events	High detection accuracy	Label scarcity; reduced adaptability	Nguyen, et al.[1]
Predictive Regression Models	Time-series performance data	Early warning; proactive prevention	Model drift over time	Matsunaga & Fortes[7]; Kannan, et al.[6]
Hybrid Approaches	Mixed labelled and unlabelled data	Improved robustness and flexibility	Increased system complexity	Md, et al.[5]; Fernandes, et al.[9]

Source: Synthesized from the approved reference set.

5.4. Hybrid and context-aware approaches

Owing to the shortcomings of each of these two methods, recent studies have promoted hybrid methods that integrate unsupervised detection, supervised classification and predictive modelling. Discussions on machine learning applications in production systems show that hybrid strategies enhance resiliency when they take advantage of the merits of various learning paradigms[5,9]. Such combinations are used in production databases to allow the adaptive learning of baselines and the use of domain knowledge in case label data are accessible **(Figure 6)**.



6. Regression Prevention, Problems and Reasonability

Although performance regression detection is one of the most valuable capabilities, its application is limited to the extent that the results of detection are convertible to sufficient preventive action within a reasonable period of time. Production database environment prevention involves degradation prediction, reasons to justify action and ensuring that operators are not eroded by the automated decision process. In this section, the possibility of regression detection using machine learning to prevent regression is described and the key challenges and explainability requirements related to practical implementation are analysed.

In this manner, predictive monitoring can be used to prevent regression. Machine learning enables regression prevention to broaden monitoring by changing monitoring to predictive understanding compared to retrospective examination. Temporal dynamics of database performance indicators can be modelled to predict the future behaviour of the system and costs of degradation curves before service level commitments are violated. Studies on production and industrial systems state that predictive learning helps make proactive decisions, such as the pre-emptive allocation of resources and configuration tuning to reduce downtimes and operational risks[6,7].

Production databases can apply predictive regression models to forecast an increase in query latency, contention or saturation

of resources caused by an increase in workload or change in the system. These predictions may be combined with automated alerts or recommendation systems to enable database administrators to act before it is too late, for example, by refining queries, adjusting indexing policies or scaling infrastructure **(Table 6)**. Evidence from industrial regression modelling also reveals that long-term performance is the most accurate prediction in an environment with slow rather than abrupt deterioration[11].

Table 6: Regression Prevention Capabilities and Associated Challenges.

Aspect	Machine Learning Contribution	Key Challenge
Early Warning	Forecasts future performance degradation	Model drift under evolving workloads
Proactive Intervention	Enables preventive tuning and scaling	Integration with operational workflows
Automation	Reduces manual monitoring effort	Risk of over-reliance on models
Interpretability	Highlights influential metrics and trends	Balancing accuracy and explainability
Operational Trust	Supports informed decision-making	Resistance to opaque models

Source: Adapted from Matsunaga and Fortes[7], Nguyen, et al.[1], Kang, et al.[2] and Fernandes, et al.[9].

6.1. Problems in operation and modelling

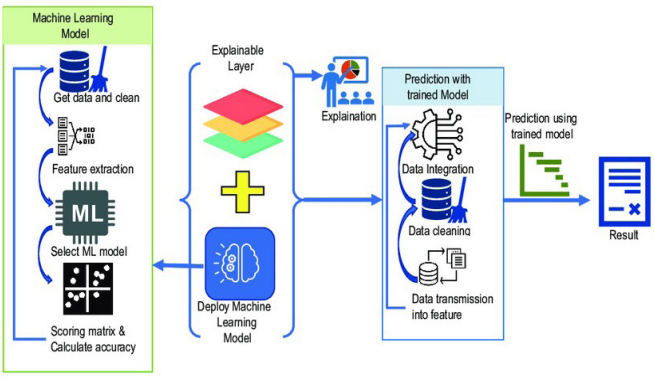
However, machine-based regression prevention systems have several problems associated with their production deployment, despite their advantages. A major issue is data drift, whereby changes in workload patterns or system configurations invalidate earlier trained models. In manufacturing environments, it has been reiterated in machine learning reviews that non-stationary data are not an exception but the rule, which means that retraining and validation of the model must be performed continuously[2,3].

The second issue is the weakness and subjectivity of the labelled regression data. However, as practiced in industrial case studies, the performance regressions are not presented consistently, which limits the effectiveness of supervised learning techniques[1]. Furthermore, machine learning models also have computational and operational overheads, which are not ideal in high-throughput database systems where low-latency monitors must be present.

6.2. ML-based decision explainability and trust

The implementation of machine learning in the performance management of databases is a necessity that must be clarified. To be confident and act on a regression or preventive recommendation cautioned by a model, database administrators must be informed of why the model has arrived at the regression or preventive recommendation. The literature on machine learning in industrial fault diagnosis provides numerous reasons as to why black-box models are often not used in the sphere of operations, regardless of how high the results they can achieve[2,13].

Explainable outputs might be useful both in the context of performance regression and in determining the most influential metrics and time series associated with degradation (or even the correlation of workload changes with performance decrease) **(Figure 7)**. The interpretable information delivered by machine learning systems can be a decision-support system rather than a black box, which is a good practice in industrial machine-learning systems[12].



7. Future Research Directions

Despite the promising results of machine learning-based methods in identifying and preventing regression in the performance of production databases, several open research issues remain. These issues must be addressed to enhance the robustness, scalability and practicality of these models in the field. This section summarizes the major recommendations for future research to elaborate on the findings and limitations of this study.

7.1. Learning with adaptation and drift awareness

Among the significant research directions, there is a better adaptive ability of models to nonstationary workloads. Production databases experience constant evolution with changes in applications, workloads and infrastructure. The next area of work is drift-aware learning strategies that automatically identify changes in data distribution and modify model parameters or retraining schedules as needed. These would minimize manual handling and ensure accuracy in detection in the long term of deployment.

7.2. Database internal integration

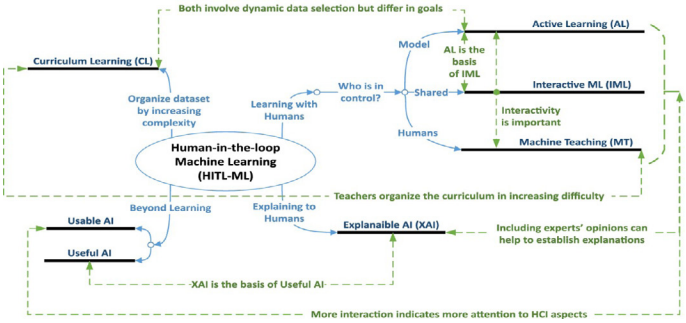
Most current machine learning strategies are based on external telemetry, such as latency measurements and resource use. Future studies might consider closer integration with the internals of databases, such as query execution plans, indexing behaviour and concurrency control mechanisms. Including internal signals can allow a closer determination of the causes of regression and preventive measures can be taken rather than generic warnings.

7.3. Hybrid human-in-the-loop system

Although automation is a primary incentive for implementing machine learning, full autonomy is an unrealistic and unvalued aim in most production settings. The design of future systems should focus on human-in-the-loop designs that integrate machine learning predictions with opinionated predictions. These hybrid methods have the potential to enhance trust, minimize false positives and enable domain knowledge to drive model improvement over time.

The Standardized Evaluation and benchmarking role involve evaluating a company in comparison with its competitors using a collection of criteria and indicators. <|human|>9.4 Benchmarking and Standardized Evaluation This role involves assessing a company relative to its competitors based on a set criteria and indicators.

Another important problem in the current research is the absence of uniform norms and data on performance regression in production databases. Proprietary or domain-specific data are used in most studies, thus restricting the ability to replicate and evaluate them. The creation of common benchmarks, assessment procedures and performance indicators would help to speed up the process and conduct stricter evaluations of competing strategies **(Figure 8)**.



8. Conclusion

The constant and expensive problem of production databases is performance regression. Conventional rule-based monitoring models have difficulty keeping up with the non-stationary and dynamic nature of modern production settings, with the variability of the workload and development of systems making the use of fixed thresholds ineffective. This study discussed how machine learning methods can be used to identify and prevent performance regression by learning adaptive performance baselines, recognizing long-term regression patterns and predicting interventions.

This study brings together previous studies on database systems, software engineering and industrial production industries to indicate the appropriateness of machine learning in regression conscious performance management. The proposed conceptual framework illustrates the ability to build a single monitoring pipeline using continuous telemetry data collection, feature engineering, model-based detection and feedback-based prevention. Machine learning-based systems, unlike traditional methods, contribute to both early regression detection and proactive prevention, relying on the transition to a predictive model of database performance management rather than a reactive one.

Meanwhile, in this study, the practical difficulties of implementing machine learning in production database-related scenarios are recognized. Such problems include data drift, lack of labelled regression data, operational overhead and explainability to ensure that adoption is reliable and trustworthy. Its analysis points out that machine learning must be used as a decision-support tool, but not as a substitute for human skills in database management.

In general, this study adds organized insight into the deployment of machine learning for performance regression management of production databases. It offers a reference point for future investigations of adaptive, explanatory and operationally feasible monitoring systems by making contributions by linking the perspectives of industrial machine learning and the study of database performance. With the constantly increasing scale and complexity of production databases, machine learning-based solutions will become increasingly significant in helping to stabilize performance and make resilient data-driven applications possible.

9. References

1. Nguyen TH, Nagappan M, Hassan AE, et al. An industrial case study of automatic identification of performance regression causes In Proceedings of the 11th Working Conference on Mining Software Repositories, 2014: 232-241.
2. Ziqiu K, Catal C, Tekinerd B. Machine learning applications in production lines: A systematic literature review. Computers & Industrial Engineering, 2020;149: 106773.
3. Usuga Cadavid JP, Lamouri S, Grabot B, et al. Machine learning applied in production planning and control: a state-of-the-art in the era of Industry 4.0. J Intelligent Manufacturing, 2020;31: 1531-1558.
4. Fahle S, Prinz C, Kuhlenkötter B. Systematic review of machine learning (ML) methods for manufacturing processes: Identifying artificial intelligence (AI) methods for field application. Procedia CIRP, 2020;93: 413-418.
5. Md AQ, Jha K, Haneef S, et al. A review on data driven quality prediction in the production process with machine learning for Industry 4.0. Processes, 2022;10: 1966.
6. Kannan R, Abdul Halim HUA, Ramakrishnan K, et al. Machine learning approach for predicting production delays: a quarry company case study. Journal of big Data, 2022;9: 94.
7. Matsunaga A, Fortes J. The use of machine learning to predict the time and resources consumed by applications. In 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, 2010: 495-504.
8. Jung J, Hu H, Arulraj J, et al. Apollo: Automatic detection and diagnosis of performance regressions in database systems. Proceedings of the VLDB Endowment, 2019;13: 57-70.
9. Fernandes M, Corchado JM, Marreiros G. Machine learning techniques applied to mechanical fault diagnosis and prognosis in the context of real industrial manufacturing use cases: a systematic literature review. Applied Intelligence, 2022;52: 14246-14280.
10. Sankhye S, Hu G. Machine learning methods for quality prediction in production. Logistics, 2020;4: 35.
11. Ibrahim NM, Alharbi AA, Alzahrani TA, et al. Well performance classification and prediction: deep learning and machine learning long-term regression experiments on oil, gas and water production. Sensors, 2022;22: 5326.
12. Sircar A, Yadav K, Rayavarapu K, et al. Application of machine learning and artificial intelligence in the oil and gas industry. Petroleum Research, 2021;6: 379-391.
13. Kang Z, Catal C, Tekinerdogan B. Product failure detection in production lines using a data-driven model. Expert Systems with Applications, 2022;202: 117398.
14. Syafrudin M, Alfian G, Fitriyani L, et al. Performance analysis of IoT-based sensor, big data processing and machine learning model for real-time monitoring system in automotive manufacturing. Sensors, 2018;18: 2946.
15. Chevtchenko SF, Rocha EDS, Dos Santos MCM, et al. Anomaly detection in industrial machinery using IoT devices and machine learning: systematic mapping. IEEE Access, 2023;11: 128288128305.
16. Frankó A, Hollósi G, Ficzere D, et al. Applied machine learning for IIOT and smart production-Methods to improve production quality, safety and sustainability. Sensors, 2022;22: 9148.
17. Rai R, Tiwari MK, Ivanov D, et al. Machine learning in manufacturing and Industry 4.0 applications. International Journal of Production Research, 2021;59: 4773-4778.
18. Botchkarev A. A new typology design of performance metrics to measure errors in machine learning regression algorithms.

Interdisciplinary Journal of Information, Knowledge and Management, 2019;14: 045-076.

19. Baylor D, Breck E, Cheng HT, et al. Tfx: TensorFlow-based production-scale machine learning platform. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017: 1387-1395.