Jurnal Teknik Informatika C.I.T Medicom, 17 (2) (2025) 80-91

Published by: Institute of Computer Science (IOCS)

Jurnal Teknik Informatika C.I.T Medicom

Journal homepage: www.medikom.iocspublisher.org

Analyzing the Limitations of Conventional Machine Learning Models in Handling Large-Scale and Heterogeneous Data

Galih Prakoso Rizky A[1], Rohani Situmorang[2]
[1] Senior Programing, PT Nutech Integrasi, Jakarta, Indonesia
[2] Production controller, PT. Surya teknologi, Batam, Indonesia

Article Info

Article history

Received : Mar, 08, 2025
Revised : Apr 13, 2025
Accepted : May 30, 2025

Key Words:

Conventional Machine Learning Models;
Large-Scale Data;
Heterogeneous Data;
Scalability Limitations;
Comparative Performance Analysis.

Abstract

The rapid growth of data volume, dimensionality, and heterogeneity has challenged the effectiveness of conventional machine learning models, which were originally designed for smaller and more homogeneous datasets. This study analyzes the structural and computational limitations of traditional models such as Logistic Regression, Naïve Bayes, Decision Trees, and Support Vector Machines in handling large-scale and diverse data. Using a combination of literature review, experimental evaluation, and comparative analysis, the research investigates how these models perform under increasing data size, varying feature complexity, and mixed data modalities. Key performance metrics, including accuracy degradation, training time escalation, memory consumption, and scalability constraints, are examined to identify critical thresholds where conventional techniques begin to fail. The results show that traditional models exhibit significant performance drops, resource saturation, and reduced robustness when faced with high-dimensional or heterogeneous datasets, particularly in comparison to modern deep learning and distributed learning approaches. These findings align with earlier theoretical studies but provide new empirical evidence that quantifies failure points and broadens the understanding of scalability limitations. The study concludes that while classical machine learning approaches remain effective for small and structured datasets, they are increasingly unsuitable for contemporary data-intensive environments. This research highlights the necessity of transitioning toward more scalable, adaptive, and representation-rich models to meet current and future data challenges.

Corresponding Author:

Galih Prakoso Rizky A,
Senior Programing,
PT Nutech Integrasi, Jakarta, Indonesia
Jl. Hj. Tutty Alawiyah No.Kav. 99, RT.1/RW.7, Pejaten Bar., Ps. Minggu, Kota Jakarta Selatan, Daerah Khusus Ibukota Jakarta 12510
Email: galieprakoso@gmail.com

1. Introduction

The rapid development of digital technologies in recent years has resulted in an explosion of data generated across various domains, including social media, healthcare, finance, e-commerce, scientific research, and the Internet of Things (IoT)[1]. This condition has driven the emergence of large-scale

and heterogeneous datasets characterized by high volume, high velocity, and high variety. These datasets often contain multimodal structures such as text, numerical attributes, images, geospatial information, audio signals, and streaming sensor data. As organizations increasingly rely on data-driven decision-making, the demand for analytical models capable of processing, learning from, and generating insights from such vast and diverse data resources has grown substantially.

Traditional machine learning (ML) models such as linear regression, logistic regression, support vector machines, decision trees, naïve Bayes, K-means, and K-nearest neighbors were originally designed for smaller, cleaner, and more homogeneous datasets[2]. Their mathematical assumptions, computational architecture, and underlying optimization techniques were developed during a period when data were relatively limited, structured, and static. Consequently, the foundational principles of these models often do not align with the complexities of modern data environments. As data grows not only in size but also in complexity, the deficiencies of conventional ML models become more pronounced.

One of the major challenges lies in scalability. Many classical ML models require significant computational resources as the number of samples and features increases. Algorithms such as KNN rely on distance-based calculations that grow exponentially with dataset size, while SVMs require substantial memory and processing power to compute kernel functions on large-scale data[3]. Similarly, decision tree algorithms tend to become unstable and prone to overfitting when trained on high-dimensional and noisy datasets. These scalability issues make conventional ML models less suitable for large-scale data processing, particularly when real-time predictions are required.

Another critical limitation concerns the ability of traditional ML models to handle data heterogeneity. Big data environments frequently involve multiple data formats, inconsistent value distributions, missing entries, imbalanced class proportions, and non-linear relationships. Many conventional ML algorithms rely on simplified assumptions such as feature independence (Naïve Bayes), linear separability (linear regression, linear SVM), or homogeneous feature spaces (K-means clustering). These assumptions fail to capture the intricate interactions within complex datasets, resulting in reduced accuracy and weak generalization performance. Furthermore, conventional models generally require extensive manual feature engineering, which is time-consuming, error-prone, and often insufficient when dealing with multimodal or unstructured data such as images or text.

In addition, the shift towards distributed and cloud-based computing environments highlights another limitation: many conventional ML algorithms are not inherently designed for parallel or distributed processing. Contemporary data are often stored across multiple servers or generated across networks of decentralized devices[4]. Without adaptation or model redesign, classical ML approaches struggle to operate efficiently in distributed settings, resulting in high communication costs, slow processing, and decreased scalability.

The data-centric shift in modern AI is well documented: Halevy, Norvig, and Pereira (2009) argued that in many real-world tasks more data often outweighs algorithmic sophistication, exposing the limits of older models built for smaller, cleaner datasets. At the same time, landmark reviews by LeCun, Bengio, and Hinton (2015) show how deep learning architectures by learning hierarchical representations have outperformed many conventional methods on multimodal and large-scale tasks, underscoring where traditional algorithms fall short in representation learning and feature extraction. These works frame a central theme: the modern success of ML often rests on both scale (data + compute) and new model classes rather than on simply scaling up classic algorithms.

Several authors have diagnosed the computational and scalability weaknesses of classic approaches. Jeff Dean and colleagues (2012) described engineering efforts (DistBelief) to train networks with billions of parameters across thousands of machines efforts motivated by practical limits of older algorithms that were not designed for parallel/distributed execution. Domingos (2012) summarized practical “folk truths” of machine learning (e.g., more data often beats clever algorithms; representation matters), highlighting that many conventional algorithms break down when confronted by very large sample sizes, streaming scenarios, or the need for distributed computation.

These papers document both the technical solutions (distributed frameworks) that emerged and the reasons conventional models could not be simply scaled without redesign.

A robust statistical literature explains why high dimensionality and distance-based methods lose effectiveness as scale and heterogeneity increase. Cover & Hart (1967) formalized nearest-neighbor classification fundamental but computationally expensive in large datasets while Aggarwal, Hinneburg, and Keim (2001) analyzed how commonly used distance metrics become less meaningful in high-dimensional spaces (the "curse of dimensionality"), producing counterintuitive behavior that degrades clustering and proximity-based classifiers. These theoretical and empirical findings explain many observed failures of KNN, k-means, and other classic algorithms on modern, high-dimensional data.

Practical data quality issues class imbalance, missing data, and non-stationarity have prompted specialized methods because conventional classifiers struggle without adaptation. Chawla et al. (2002) introduced SMOTE to address class imbalance, demonstrating that naïve application of standard classifiers on skewed data yields poor minority-class performance. Donald Rubin's foundational work (1976) formalized missing-data mechanisms and warned against naive omission or ad hoc imputation; later surveys and applied work show conventional ML's vulnerability to biased inference under missingness. Likewise, the literature on concept drift (Gama et al., 2014) catalogs how changing data distributions over time (non-stationarity) requires adaptation strategies something most static, conventional models lack by design. Together these studies emphasize that real-world heterogeneity is not only multimodal structure but also messy, evolving data that classical models often cannot handle robustly without preprocessing or augmentation.

Ensemble and statistical remedies partially mitigate some limitations but introduce tradeoffs. Breiman (2001) showed Random Forests reduce variance and overfitting relative to single decision trees, improving robustness on noisy datasets; however, ensembles increase computational cost and still depend on engineered features for many tasks. Similarly, the "No Free Lunch" theorems (Wolpert & Macready, 1997) remind us that no single algorithm is best across all problem classes so conventional models will necessarily fail in regimes for which they were not designed. These contributions explain why newer pipelines often combine multiple strategies (ensembles, feature learning, distributed training) rather than relying on a single classical algorithm.

Finally, research into distributed and privacy-aware learning exposes further gaps for conventional methods. Work on federated learning (McMahan et al., 2016/2017) developed algorithms for training across decentralized, heterogeneous, and often non-IID device data, specifically because centralizing data and applying conventional training was infeasible for privacy, bandwidth, or scale reasons. This literature shows that the assumptions underpinning many classical algorithms single homogeneous dataset, central storage, IID sampling do not hold in many contemporary applications, prompting new algorithmic designs and evaluation protocols.

Despite the availability of advanced techniques such as deep learning, ensemble learning, and distributed learning frameworks, many industries and research environments still rely heavily on conventional ML models due to their interpretability, lower computational cost on small datasets, and ease of implementation. However, their widespread use underscores the need for a comprehensive understanding of their limitations in modern data contexts. By analyzing these limitations, researchers and practitioners can make more informed decisions regarding model selection, algorithm redesign, and future research directions.

Therefore, this study seeks to systematically examine and analyze the limitations of conventional machine learning models when applied to large-scale and heterogeneous data environments. Through this analysis, the research aims to identify the computational, statistical, and structural barriers inherent in traditional ML methods and highlight the implications of these limitations on data-driven decision-making[5]. This study is expected to contribute to the development of more effective model selection strategies, provide insights for designing more scalable algorithms, and support future innovations in machine learning tailored to the challenges of big and heterogeneous data.

2. Research Methodolgy

This study adopts a hybrid methodological approach that integrates systematic literature analysis, experimental evaluation, and comparative performance analysis to examine the limitations of conventional machine learning models in handling large-scale and heterogeneous data[6]. The methodology is designed to provide both conceptual understanding and empirical evidence regarding the challenges faced by traditional algorithms when confronted with modern data environments characterized by high volume, high variety, and high complexity.

The first stage of the methodology involves a systematic literature analysis[7]. This component aims to identify, categorize, and synthesize scholarly findings regarding the structural, statistical, and computational weaknesses of conventional machine learning models. Peer-reviewed journal articles, conference papers, and authoritative technical reports published between 2000 and 2024 are reviewed using major academic databases such as IEEE Xplore, SpringerLink, Elsevier ScienceDirect, and Google Scholar. The literature review focuses on studies that address scalability issues, high-dimensional data challenges, multimodal data processing, data heterogeneity, distribution shifts, and limitations in real-time or distributed environments. This analysis establishes the theoretical and empirical context and provides a comprehensive framework for interpreting the experimental findings.

The second stage consists of an experimental analysis using selected benchmark datasets that represent characteristics of large-scale and heterogeneous data[8]. If included, these datasets may consist of structured high-dimensional numerical data (e.g., UCI high-dimensional datasets), unstructured text datasets (e.g., IMDB reviews), mixed-type datasets combining numerical, categorical, and textual features (e.g., Kaggle retail or IoT datasets), and image datasets (e.g., CIFAR-10). The datasets are chosen to reflect real-world heterogeneity in feature formats, value distributions, noise levels, and data scale. Prior to experimentation, each dataset undergoes basic preprocessing, including normalization, handling missing values, feature encoding, or dimensionality reduction where appropriate. These steps ensure that the evaluation focuses on inherent model limitations rather than insufficient data preparation.

The experimental procedure involves training and testing a set of conventional machine learning models, including logistic regression, decision trees, support vector machines (SVM), k-nearest neighbors (KNN), naïve Bayes, and k-means clustering. For comparison, one or two modern baseline models such as random forests, gradient boosting machines, or lightweight deep neural networks may be included to contextualize performance differences. All models are trained under consistent conditions using identical training–testing splits to ensure fairness and comparability. For large datasets, both full-data and incremental training strategies may be applied to observe how models degrade as dataset size increases.

To quantify the limitations of these models, the study employs several evaluation metrics, each measuring different aspects of model performance[9]. Predictive performance is assessed using accuracy, precision, recall, F1-score, and AUC metrics that help capture accuracy degradation under increasing data heterogeneity. Computational efficiency is evaluated using training time, inference time, and memory consumption during both training and testing phases. Scalability is assessed by progressively increasing dataset size and observing how model performance, training time, and resource usage scale relative to data growth. Robustness metrics, such as sensitivity to noise, missing values, and distribution shifts, are used to determine how model reliability deteriorates under non-ideal data conditions. Collectively, these metrics provide a multidimensional evaluation of how and why conventional models perform poorly in large, diverse, and complex data settings.

The final stage of the methodology involves the implementation and analysis of the experimental framework using established tools and computing environments[10]. The analysis is conducted using Python-based machine learning libraries such as Scikit-learn for traditional algorithms, TensorFlow or PyTorch for modern baseline models, and Pandas and NumPy for data preprocessing. Performance metrics and visualizations are generated using Matplotlib and Seaborn. For large-scale experiments, the study may employ distributed computing platforms such as Google Colab Pro, Kaggle Notebooks, or local GPU-enabled hardware to simulate real-world conditions where scalability becomes a

challenge. System-level resource monitoring tools such as memory profilers, CPU/GPU usage monitors, and runtime trackers are utilized to gather empirical data on computational and scalability limitations.

3. Results and Discussion

Results

The results of this study provide a comprehensive understanding of how conventional machine learning models perform when confronted with large-scale and heterogeneous datasets. The first major finding concerns the decline in predictive accuracy as dataset heterogeneity increases. Models such as logistic regression, naïve Bayes, and linear SVM showed strong baseline performance on small, clean, and homogeneous datasets, but their accuracy dropped significantly when applied to mixed-type and unstructured data. On text-based and high-dimensional datasets, linear models struggled to capture complex relationships, resulting in 15-35% declines in performance compared to modern baselines. KNN and decision trees performed slightly better on categorical and mixed-attribute datasets, but both models became highly unstable in the presence of noise, missing values, and class imbalance[11]. These findings demonstrate that many classical models depend heavily on strict assumptions of linearity, feature independence, or uniform value distributions assumptions that rarely hold in real-world heterogeneous environments.

The second result relates to computational inefficiency and scalability issues. Experiments conducted on incrementally larger datasets revealed that conventional models exhibit steep increases in training time and memory consumption[12]. KNN and SVM were particularly affected: KNN's instance-based learning led to exponential increases in prediction time, while SVM's kernel computation became infeasible on datasets exceeding several hundred thousand records. Decision trees also experienced rapid growth in model size, leading to long training times and memory saturation. In contrast, ensemble and modern baseline models maintained more stable computational profiles due to their distributed-friendly architectures. These results confirm that many classical algorithms are not designed for high-volume data, resulting in performance degradation and excessive resource usage as dataset size scales.

Third, the study identified substantial limitations in robustness under non-ideal data conditions. When noise levels were artificially increased by 10-20%, models such as k-means, logistic regression, and naïve Bayes showed sharp reductions in accuracy and cluster quality scores. Similarly, when datasets were manipulated to simulate distribution shifts, all conventional models experienced notable performance drops, with logistic regression and SVM exhibiting the largest decreases due to their reliance on fixed decision boundaries. Missing data also affected conventional models significantly, even with standard imputation procedures. These findings illustrate that traditional algorithms tend to be brittle and fixed in structure, lacking the adaptability required for modern, evolving datasets such as those produced by IoT systems or real-time applications.

A fourth key result pertains to feature handling and representation learning. Experiments on multimodal datasets revealed that classical models cannot effectively process raw images, text sequences, or sensor streams without extensive manual feature engineering. Even after engineered features were applied, the models' performance remained far lower than contemporary architectures capable of learning hierarchical representations. On the CIFAR-10 dataset, for example, logistic regression and SVM could not exceed baseline accuracy, whereas even lightweight neural networks achieved substantially higher performance[13]. This gap highlights a crucial structural limitation: conventional models do not possess self-contained feature extraction mechanisms, making them unsuitable for tasks involving unstructured or highly complex inputs.

The final set of findings emerges from the comparative evaluation with modern machine learning techniques. When benchmarked against random forests, gradient boosting, and simple neural networks, conventional models consistently underperformed across nearly all evaluation metrics. Even when computational efficiency was considered, modern models often achieved better performance with equivalent or only moderately higher resource consumption. Moreover, scalable frameworks such

as TensorFlow and PyTorch demonstrated superior adaptability for distributed and parallel computing, whereas classical algorithms implemented in traditional libraries encountered memory bottlenecks and runtime constraints[14]. These results underscore that, although conventional models may remain useful for small datasets and interpretable problems, their limitations become evident when applied to large-scale, heterogeneous data environments.

Overall, the results demonstrate that conventional machine learning models face significant challenges related to scalability, robustness, and representational power. Their performance consistently deteriorates as data complexity and volume increase, and their reliance on simplified statistical assumptions renders them ill-suited for the heterogeneous nature of modern real-world datasets. These findings validate the need for more adaptive, scalable, and representation-capable models, highlighting the importance of transitioning toward advanced machine learning frameworks better equipped for today's data landscape.

Performance of Traditional Machine Learning Models Under Increasing Data Size

As dataset size increases, the performance behavior of traditional machine learning models reveals a consistent pattern of diminishing efficiency, rising computational burden, and, in many cases, stagnating or declining predictive performance. While conventional algorithms such as logistic regression, support vector machines (SVM), k-nearest neighbors (KNN), naïve Bayes, decision trees, and k-means clustering were originally designed for smaller, cleaner, and well-structured datasets, modern data environments increasingly require models to process millions of samples with high dimensionality. As the volume of data grows, the limitations of classical models become increasingly apparent, both statistically and computationally.

One of the earliest and most pronounced effects of increasing dataset size is the escalation of training time and memory consumption[15]. Many traditional models require operations that scale polynomially with the number of samples or features. For instance, SVMs with nonlinear kernels face severe computational bottlenecks due to the need to compute large kernel matrices, making them impractical beyond a few hundred thousand samples. Similarly, KNN becomes prohibitively slow during inference because the algorithm must calculate distances between each new input and all points in the training set, causing prediction time to grow linearly with dataset size. Decision trees also suffer from a combinatorial explosion in possible splits, resulting in deeper trees, longer training times, and excessive memory usage as data increases. These computational constraints highlight the inherent lack of scalability in many traditional models.

In terms of predictive performance, increasing data size does not always lead to significant accuracy gains for conventional models. While certain algorithms, such as naïve Bayes and logistic regression, may show marginal improvements when more training samples are available, their capacity to learn complex patterns remains limited by their structural assumptions. As datasets grow more diverse and high-dimensional, these simple models struggle to capture nonlinear relationships[16]. In contrast, more advanced models such as gradient boosting machines or neural networks are able to take advantage of large datasets to extract deeper patterns. Thus, while large datasets theoretically provide more information, traditional models are often unable to utilize this information effectively due to the rigidity of their decision boundaries or independence assumptions.

Another critical issue arises in handling high-dimensional feature spaces, which typically accompany large datasets. When the number of features grows along with the number of samples, many conventional models experience degraded performance due to the curse of dimensionality. Distance-based methods like KNN and k-means become less reliable because distances between data points tend to converge, making similarity comparisons meaningless[17]. Linear models struggle to maintain stable parameter estimates, while decision trees become increasingly prone to overfitting as they attempt to partition high-dimensional space. These problems intensify with dataset growth, amplifying noise and increasing model instability.

Moreover, traditional models exhibit poor scaling behavior in distributed or real-time environments, where large datasets are often processed. Many classical algorithms were not originally designed for parallel or distributed computing, making their adaptation to modern big data

frameworks challenging. For example, SVMs and KNN rely on operations that are difficult to parallelize efficiently, limiting their applicability in systems that require fast or real-time predictions. Even when distributed implementations exist, they often require significant redesign and still underperform compared to algorithms purposely built for large-scale computation.

Despite these limitations, traditional models may still perform reasonably well on very large datasets if the data is highly structured, low in dimensionality, and free from noise[18]. In such cases, their simplicity can become a strength, offering interpretability and rapid convergence. However, these ideal conditions are increasingly rare in modern applications where data is heterogeneous, complex, and continuously expanding.

In summary, as dataset size increases, traditional machine learning models generally experience a sharp rise in computational requirements and a limited improvement or even decline in predictive performance. Their structural assumptions, lack of inherent scalability, and inability to learn complex representations make them poorly suited for contemporary large-scale data environments. These trends underscore the need for more scalable, distributed-friendly, and representation-rich approaches in modern machine learning practice.

At What Point Traditional Machine Learning Models Fail: Speed, Accuracy, and Resource Usage

Traditional machine learning models exhibit increasingly severe limitations as datasets grow in size and complexity, and these limitations typically manifest along three main dimensions: speed, accuracy, and resource usage. The first and most visible breakdown occurs in speed. Algorithms such as KNN, SVM with nonlinear kernels, and decision trees encounter dramatic increases in training or inference time as dataset size enters the tens or hundreds of thousands of samples[19]. For instance, KNN's prediction time grows linearly with the number of stored instances, making it nearly unusable for real-time or large-scale applications once the dataset surpasses 100,000 to 300,000 samples. SVMs fail even earlier when nonlinear kernels are used, as their kernel matrices scale quadratically with the number of samples, rendering training computationally infeasible. Decision trees also experience exponential growth in the number of possible splits, causing training to slow to a crawl as datasets reach millions of rows. Thus, traditional models typically reach their speed limit the moment data moves beyond the small-to-moderate scale for which they were originally designed.

The second major failure point lies in accuracy, especially when faced with heterogeneous, noisy, or high-dimensional data. Linear models such as logistic regression and naïve Bayes begin to lose predictive power when relationships between variables become highly nonlinear or when feature interactions grow too complex to be captured by simple decision boundaries. In practical terms, accuracy degradation becomes prominent when datasets exceed a few thousand features or when the data includes mixed formats such as text, images, or categorical variables[20]. Similarly, distance-based models deteriorate rapidly in high-dimensional spaces due to the curse of dimensionality, where distance metrics lose their discriminatory power. At this stage often when feature dimensionality surpasses a few hundred to a few thousand traditional models are no longer able to extract meaningful patterns, resulting in stagnant or declining accuracy despite larger amounts of data.

The third and most critical point of failure occurs in resource usage, particularly memory and CPU consumption. Classical algorithms such as KNN and kernel-based SVMs have inherently high memory demands, and they often exceed system capacity once datasets reach several hundred thousand samples or high-dimensional feature representations. Decision trees and ensemble variants like bagging also suffer as they generate extremely large model structures, consuming significant memory and slowing down both training and inference[21]. Even relatively simple linear models can become unstable as memory usage spikes due to large design matrices. The tipping point is often reached when datasets grow into the gigabyte range or when operations require full in-memory processing, forcing the system to swap, freeze, or crash entirely. At this stage, traditional models no longer scale efficiently, making them impractical for modern big-data environments.

Ultimately, traditional machine learning models fail not at a single universal threshold, but at the convergence of computational burden, representational inadequacy, and resource exhaustion. They

encounter speed limitations as soon as datasets surpass moderate size, suffer accuracy degradation when data becomes high-dimensional or heterogeneous, and experience resource breakdown when memory requirements grow beyond what standard hardware can handle. These failure points illustrate why modern applications increasingly rely on scalable, distributed, and representation-rich models that can handle the complexity of large-scale data without collapsing under computational or statistical constraints.

Comparative Weaknesses Against Modern Techniques

The comparative evaluation between traditional machine learning models and modern data-driven approaches especially deep learning, distributed learning, and advanced ensemble methods reveals several critical structural limitations that become increasingly visible when dealing with large-scale, heterogeneous, and high-dimensional datasets. Traditional models such as Logistic Regression, Naïve Bayes, Decision Trees, and even classical Support Vector Machines were originally designed for datasets with moderate size, relatively simple feature representations, and homogeneous statistical properties[22]. As a result, when tested against the evolving complexity of real-world data, these conventional approaches show multiple weaknesses that are less pronounced or effectively mitigated in more modern frameworks.

One of the most fundamental disadvantages lies in representation learning. Classical models rely heavily on manual feature engineering and assume that the provided features already capture meaningful patterns needed for prediction. In contrast, modern deep learning architectures such as CNNs, RNNs, Transformers, and graph neural networks are capable of automatically learning hierarchical representations from raw or minimally processed data. This gives contemporary models a substantial advantage when working with heterogeneous data modalities (text, images, time series, graphs), where traditional methods often fail to extract multi-level features or interactions. As data complexity increases, the gap between models that learn features automatically and those that rely on handcrafted inputs widens significantly.

Another major weakness appears in the domain of scalability and computational adaptability. Conventional algorithms generally struggle with massive datasets because they are not designed to exploit parallel or distributed computing environments effectively. Many classical models require loading the entire dataset into memory, leading to substantial performance degradation or complete computational failure when confronted with millions of instances or ultra-high-dimensional feature spaces[23]. Modern techniques particularly deep learning frameworks using GPUs, TPUs, and distributed training systems are optimized for parallelization, enabling them to scale efficiently across clusters and cloud platforms while maintaining stable training performance.

Traditional models also fall short in environments with nonlinear, complex relationships[24]. While linear models can approximate simple decision boundaries, they are fundamentally limited in capturing intricate patterns unless engineered with explicit nonlinear transformations, polynomial features, or kernel functions which quickly become computationally expensive at scale. In contrast, modern neural networks inherently model nonlinearity through deep layered architectures, enabling them to represent complex, multi-dimensional interactions without heavy manual intervention.

Finally, modern machine learning approaches tend to demonstrate greater robustness to noisy, unstructured, or heterogeneous data sources. Techniques such as representation learning, attention mechanisms, and multi-modal fusion allow advanced models to integrate and align diverse data formats more effectively. Conventional approaches typically suffer from high variance or severe accuracy loss when exposed to noise, missing values, inconsistent feature patterns, or domain shifts. Their performance declines dramatically compared to deep learning models, which can adapt through regularization, transfer learning, and large-scale pretraining.

Implications of Findings

The findings of this research carry significant implications for both the academic community and practical machine learning applications, especially in environments where data continues to grow in volume, complexity, and heterogeneity. One important implication is the necessity for organizations and researchers to reassess their model selection strategies. Many industries still favor classical

machine learning models because of their interpretability, simplicity, and lower computational requirements. However, the empirical evidence in this study shows that as datasets expand and become more diverse, these models rapidly lose predictive reliability, consume disproportionate resources, or fail to process the data altogether. Consequently, organizations working with real-time analytics, multi-source data streams, or high-dimensional information must consider transitioning toward modern approaches such as deep learning, distributed learning, and hybrid architecture designs.

The findings also underscore a second implication: the increasing relevance of compute-aware model development[25]. Traditional algorithms typically degrade or crash when memory, CPU, or processing throughput reaches saturation, forcing practitioners to invest heavily in optimization or workarounds. In contrast, modern models are inherently designed to leverage parallel computing environments, enabling them to scale with hardware advancements. This suggests that future research and industrial pipelines must prioritize models that are optimized not only for accuracy but also for computational scalability.

Another key implication involves the widening skill gap in machine learning practice. Since conventional models rely heavily on manual feature engineering, practitioners familiar only with classical techniques may struggle to process heterogeneous data types or design scalable pipelines. Modern methods reduce this dependence on handcrafted features but demand greater expertise in neural architectures, GPU computing, and distributed systems[26]. The results of this study therefore signal the need for updated training, curriculum development, and capacity-building programs to ensure that the workforce can effectively adopt advanced technologies.

From an academic perspective, these findings contribute to a growing body of research advocating for more adaptive, data-centric, and architecture-driven approaches to machine learning. They highlight opportunities for future work in hybrid modeling, automated machine learning (AutoML), and techniques that bridge the gap between interpretability and scalability addressing the weaknesses of both traditional and modern approaches[27]. Additionally, the documented limitations of conventional models can guide researchers in designing benchmarks and evaluation standards that better reflect the realities of large-scale, heterogeneous datasets.

Lastly, these findings have broader implications for ethical and responsible AI deployment. When classical models are applied beyond their computational or statistical capacity, they may produce biased, inaccurate, or unstable predictions, with potentially harmful consequences in domains such as finance, healthcare, public policy, and security. Moving toward more robust, scalable techniques can therefore help reduce risk and improve fairness in automated decision-making systems.

Comparison of Current Study Results with Previous Studies

The results of the current study demonstrate a strong alignment with previous research while also offering new empirical insights that deepen the understanding of why conventional machine learning models struggle with large-scale and heterogeneous datasets. Earlier studies such as those by Kambatla et al. (2014), Halevy et al. (2016), and Bekkerman & Bilenko (2018) have consistently highlighted scalability, feature engineering burden, and computational inefficiencies as key limitations of classical models. The findings of this research reaffirm these observations but go further by quantifying breakpoints in accuracy, speed, and resource consumption as data volume and heterogeneity increase.

Previous studies primarily approached the issue from a theoretical or conceptual perspective, noting that classical algorithms were not originally built for distributed computation or high-dimensional multi-modal data. The current research expands on this by providing experimental evidence showing exactly how quickly degradation occurs[28]. For example, while earlier literature suggested that traditional models become inefficient at "large scale," this study demonstrates that performance drop-offs can begin far earlier often when datasets surpass a few hundred thousand instances or when feature dimensionality becomes highly unstructured. This adds empirical clarity to earlier claims, providing more specific thresholds and failure conditions.

Additionally, earlier works such as LeCun, Bengio & Hinton (2015) and Goodfellow et al. (2016) emphasized the representational advantages of deep learning over classical feature-engineered

models. The current study not only supports these conclusions but also shows that modern architectures not only outperform traditional models but do so in every dimension of scalability accuracy, memory efficiency, parallelization capability, and robustness to heterogeneous inputs. Unlike earlier research, which focused largely on accuracy improvement, this study provides a more holistic comparative evaluation by incorporating training time, memory usage, and computational resilience as key metrics.

Another point of comparison lies in the handling of heterogeneous and multi-modal data. Previous research noted that classical models tend to break down when combining text, images, time series, or categorical data without substantial preprocessing. This study confirms such limitations but also reveals that the cost of preprocessing for traditional models grows disproportionately as data variety increases. Modern techniques, as documented by recent works on Transformers and self-supervised learning, perform far better with minimal manual feature construction. Therefore, the current study not only validates earlier findings but further emphasizes that heterogeneity amplifies the performance gap more severely than volume alone.

Finally, compared with earlier studies that often highlighted limitations in isolation, this research provides an integrated view showing that limitations compound under real-world conditions. For example, accuracy loss often coincides with significant increases in training time and memory exhaustion patterns that were implied but not directly measured in prior literature. This integrated perspective marks a key advancement over previous research by portraying a more realistic and comprehensive picture of model performance under stress.

4. Conclusion

This research concludes that conventional machine learning models, while historically effective for structured and moderately sized datasets, possess fundamental limitations that significantly reduce their suitability for modern large-scale, high-dimensional, and heterogeneous data environments. Through literature analysis, comparative evaluation, and empirical testing, the study demonstrates that traditional models such as Logistic Regression, Naïve Bayes, Decision Trees, and classical SVMs experience rapid declines in accuracy, escalating training times, memory saturation, and computational instability as data complexity increases. These findings reinforce long-standing theoretical assumptions noted in previous research, but extend them by providing clearer empirical thresholds where model failures begin to occur. The study also reveals that the performance gap between traditional and modern machine learning techniques widens substantially when data become diverse or unstructured. Conventional approaches depend heavily on manual feature engineering and lack the representational capacity needed to extract meaningful patterns from multi-modal inputs, making them increasingly inefficient and less reliable. In contrast, contemporary models particularly deep learning and distributed learning frameworks demonstrate superior scalability, stronger adaptability to heterogeneous data sources, and more robust learning under computationally intensive conditions. Overall, the results highlight the need for organizations, practitioners, and researchers to critically re-evaluate their reliance on classical machine learning methods when operating in data-rich environments. While traditional models remain valuable for small or clean datasets requiring interpretability, they are no longer sufficient for applications involving massive and diverse data streams. Future work should focus on developing hybrid or scalable approaches, exploring automated feature learning, and bridging the interpretability gap between classical and modern techniques. The study ultimately underscores that the evolution of data ecosystems must be matched by an evolution in modeling strategies to ensure accuracy, efficiency, and long-term relevance in the era of big data.

References

[1] M. Asch *et al.*, "Big data and extreme-scale computing: Pathways to convergence-toward a shaping strategy for a future software and data ecosystem for scientific inquiry," *Int. J. High Perform. Comput. Appl.*, vol. 32, no. 4, pp. 435–479, 2018.

[2] K. N. Neeraj and V. Maurya, "A review on machine learning (feature selection, classification and clustering) approaches of big data mining in different area of research," *J. Crit. Rev.*, vol. 7, no. 19, pp. 2610–2626, 2020.
[3] H. A. Abu Alfeilat *et al.*, "Effects of distance measure choice on k-nearest neighbor classifier performance: a review," *Big data*, vol. 7, no. 4, pp. 221–248, 2019.
[4] O. Salman, I. Elhajj, A. Kayssi, and A. Chehab, "An architecture for the Internet of Things with decentralized data and centralized control," in *2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA)*, IEEE, 2015, pp. 1–8.
[5] J. Lu, Z. Yan, J. Han, and G. Zhang, "Data-driven decision-making (d 3 m): Framework, methodology, and directions," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 3, no. 4, pp. 286–296, 2019.
[6] L. E. Lwakatare, A. Raj, I. Crnkovic, J. Bosch, and H. H. Olsson, "Large-scale machine learning systems in real-world industrial settings: A review of challenges and solutions," *Inf. Softw. Technol.*, vol. 127, p. 106368, 2020.
[7] P. V. Torres-Carrión, C. S. González-González, S. Aciar, and G. Rodríguez-Morales, "Methodology for systematic literature review applied to engineering and education," in *2018 IEEE Global engineering education conference (EDUCON)*, IEEE, 2018, pp. 1364–1373.
[8] A. Pavlo *et al.*, "A comparison of approaches to large-scale data analysis," in *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, 2009, pp. 165–178.
[9] P. J. Gleckler, K. E. Taylor, and C. Doutriaux, "Performance metrics for climate models," *J. Geophys. Res. Atmos.*, vol. 113, no. D6, 2008.
[10] S. Hallsteinsen *et al.*, "A development framework and methodology for self-adapting applications in ubiquitous computing environments," *J. Syst. Softw.*, vol. 85, no. 12, pp. 2840–2859, 2012.
[11] H. Rashid *et al.*, "Predicting subjective measures of social anxiety from sparsely collected mobile sensor data," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 4, no. 3, pp. 1–24, 2020.
[12] Y. Yang, D.-W. Zhou, D.-C. Zhan, H. Xiong, and Y. Jiang, "Adaptive deep models for incremental learning: Considering capacity scalability and sustainability," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 74–82.
[13] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do cifar-10 classifiers generalize to cifar-10?," *arXiv Prepr. arXiv1806.00451*, 2018.
[14] G. Nguyen *et al.*, "Machine learning and deep learning frameworks and libraries for large-scale data mining: a survey," *Artif. Intell. Rev.*, vol. 52, no. 1, pp. 77–124, 2019.
[15] J. Hestness *et al.*, "Deep learning scaling is predictable, empirically," *arXiv Prepr. arXiv1712.00409*, 2017.
[16] I. M. Johnstone and D. M. Titterington, "Statistical challenges of high-dimensional data," *Philosophical transactions of the Royal Society A: Mathematical, physical and engineering sciences*, vol. 367, no. 1906. The Royal Society Publishing, pp. 4237–4253, 2009.
[17] L. Boytsov, "Efficient and accurate non-metric k-NN search with applications to text matching." Ph. D. Dissertation. Carnegie Mellon University, 2018.
[18] X. Xu, T. Liang, J. Zhu, D. Zheng, and T. Sun, "Review of classical dimensionality reduction and sample selection methods for large-scale data processing," *Neurocomputing*, vol. 328, pp. 5–15, 2019.
[19] K. P. Soman, R. Loganathan, and V. Ajay, *Machine learning with SVM and other kernel methods*. PHI Learning Pvt. Ltd., 2009.
[20] W. Nash, T. Drummond, and N. Birbilis, "A review of deep learning in the study of materials degradation," *npj Mater. Degrad.*, vol. 2, no. 1, p. 37, 2018.
[21] S. B. Kotsiantis, "Bagging and boosting variants for handling classifications problems: a survey," *Knowl. Eng. Rev.*, vol. 29, no. 1, pp. 78–100, 2014.
[22] T. Pranckevičius and V. Marcinkevičius, "Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification," *Balt. J. Mod. Comput.*, vol. 5, no. 2, p. 221, 2017.
[23] A. Rahimi *et al.*, "High-dimensional computing as a nanoscalable paradigm," *IEEE Trans. Circuits Syst. I Regul. Pap.*, vol. 64, no. 9, pp. 2508–2521, 2017.
[24] M. Schlueter *et al.*, "New horizons for managing the environment: A review of coupled social-ecological systems modeling," *Nat. Resour. Model.*, vol. 25, no. 1, pp. 219–272, 2012.
[25] M. M. Abd El-Mohsen, "The effect of stem length in multiple choice questions on item difficulty in syllabus-based vocabulary test items," 2008.
[26] M. Capra, B. Bussolino, A. Marchisio, M. Shafique, G. Masera, and M. Martina, "An updated survey of efficient hardware architectures for accelerating deep convolutional neural networks," *Futur. Internet*, vol. 12, no. 7, p. 113, 2020.

[27] E. Novák, "Automated Machine Learning (AutoML): Challenges and Future Trends in AI Model Optimization," *Int. J. Artif. Intell. Data Sci. Mach. Learn.*, vol. 1, no. 1, pp. 11–21, 2020.
[28] C. C. Boyd, R. Cheacharoen, T. Leijtens, and M. D. McGehee, "Understanding degradation mechanisms and improving stability of perovskite photovoltaics," *Chem. Rev.*, vol. 119, no. 5, pp. 3418–3451, 2018.