Advance Journal of Science Engineering and Technology. 1(2) 2026. pp:119-128

Contents lists available at AcademiansEdu

Advance Journal of Science Engineering and Technology

journal homepage: https://academiansedu.org/ajset/

Machine Learning for Predictive Modeling Across Critical Domains: A Comparative Analysis of Algorithms in Cybersecurity, Renewable Energy, Pharmaceuticals, and Environmental Science

K.K. Pichik

University of National Development, Yogyakarta

A R T I C L E I N F O

Article history:
Received 8 April 2026
Received in revised form
29 May 2026
Accepted 18 June 2026

Keywords: Machine Learning, Predictive Modeling, Cybersecurity Threat Detection, Photovoltaic Fault Diagnosis, Pharmaceutical Drug Delivery, Rainfall Forecasting, Ensemble Learning,

A B S T R A C T

The proliferation of machine learning algorithms across diverse scientific and industrial domains has revolutionized predictive modeling capabilities, enabling data-driven decision-making in complex systems. This comprehensive review examines the application of machine learning algorithms for predictive modeling across four critical domains: cybersecurity threat detection, photovoltaic system fault diagnosis, pharmaceutical drug delivery systems, and environmental rainfall forecasting. Through a structured literature review and comparative synthesis of peer-reviewed studies published primarily between 2025 and 2026, this paper investigates the performance of supervised learning, ensemble learning, and neural network algorithms for predictive modeling across four critical domains: cybersecurity threat detection, photovoltaic system fault diagnosis, pharmaceutical drug delivery systems, and environmental rainfall forecasting. The review compares algorithm performance, identifies common methodological trends, and evaluates domain-specific strengths, limitations, and emerging research opportunities. The review reveals that ensemble-based learning approaches consistently outperform single-model architectures across domains, with Random Forest achieving F1-scores of 97.5% in cybersecurity threat detection and 98.88% accuracy in photovoltaic fault diagnosis. Neural network architectures, particularly wide neural networks and Long Short-Term Memory networks, demonstrate superior capability in capturing nonlinear interactions, achieving 98.88% accuracy in fault classification and 97% accuracy in predictive maintenance applications. In pharmaceutical applications, ensemble methods achieved 92.3% accuracy in predicting drug release behavior, while in environmental modeling, Random Forest and KNN regression achieved R^2 values of 0.6155 and 0.6492 respectively. The review identifies key challenges including data quality issues, model interpretability limitations, computational scalability constraints, and the need for standardized benchmarking frameworks. Future research directions include the development of explainable AI methods, federated learning architectures for privacy-preserving distributed modeling, hybrid physics-informed machine learning approaches, and domain-specific optimization strategies. The findings provide a comprehensive framework for algorithm selection across domains, emphasizing that while deep learning excels in complex pattern recognition tasks, ensemble methods offer robust performance with greater interpretability, and simpler models remain valuable for resource-constrained applications with well-structured data.

1. Introduction

The rapid advancement of artificial intelligence and machine learning technologies has fundamentally transformed predictive modeling capabilities across diverse scientific and industrial domains. The ability to learn complex patterns from data and make accurate predictions has become increasingly critical for optimizing decision-making, improving operational efficiency, and enabling autonomous systems (Gatenbee, Reith, & Rose, 2026; Soori & Azizi, 2026). As the volume and complexity of data continue to grow exponentially, machine learning algorithms have emerged as indispensable tools for extracting actionable insights and developing predictive models that can operate in real-time environments.

The cybersecurity landscape has undergone dramatic transformation with the proliferation of sophisticated cyber threats that increasingly leverage artificial intelligence for malicious purposes. Traditional rule-based cybersecurity approaches are no longer sufficient to address rapidly evolving threats including zero-day exploits, AI-enhanced Advanced Persistent Threats (APTs), ransomware, and polymorphic malware (Gatenbee et al., 2026; Manikandan, Reddy Onteddu, & Chilimi, 2025). The demand for similarly capable defensive

resources has led to extensive research on AI-based cybersecurity tools, particularly in threat detection applications where machine learning algorithms can identify anomalous patterns and potential intrusions with unprecedented accuracy.

The renewable energy sector, particularly photovoltaic systems, faces significant challenges in maintaining operational efficiency and reliability. As global installed photovoltaic capacity exceeded 760 GW in 2020, the need for effective fault diagnosis and predictive maintenance has become increasingly critical (Taah et al., 2026). Faults in PV systems can lead to significant power losses, with module failure affecting approximately 18.9% of output power (Taah et al., 2026). Machine learning algorithms offer powerful solutions for fault detection and classification, enabling early intervention and minimizing system downtime.

In the pharmaceutical domain, drug delivery systems represent a complex optimization challenge where formulation parameters, drug properties, and physiological conditions interact nonlinearly to determine therapeutic outcomes. Traditional experimental approaches to formulation development are time-consuming, costly, and often unable to fully capture multidimensional relationships (Shah, Shah, & Pandya, 2026). Machine learning algorithms provide the capability to analyze complex pharmaceutical datasets, identify critical formulation parameters, and predict drug release behavior with high accuracy, potentially accelerating drug development and enabling personalized medicine approaches.

Environmental modeling, particularly rainfall prediction, remains a critical challenge due to the complex, nonlinear nature of atmospheric processes and the significant impact of accurate forecasting on agriculture, water resource management, and disaster preparedness (Gemmechis, 2025). Machine learning approaches have shown promise in capturing the intricate relationships between meteorological variables and precipitation patterns, offering improved predictive accuracy over traditional statistical methods.

Despite the demonstrated effectiveness of machine learning across these domains, significant gaps remain in understanding the comparative performance of different algorithms and identifying optimal approaches for specific applications. Many studies focus on individual algorithms or specific datasets, making systematic comparison challenging (Gatenbee et al., 2026; Taah et al., 2026; Shah et al., 2026; Gemmechis, 2025). Furthermore, the limited number of standardized benchmarks and the "black box" nature of many machine learning models create barriers to adoption and trust in critical applications.

This comprehensive review addresses these gaps by examining the application of machine learning algorithms across four critical domains. The paper synthesizes recent empirical studies to compare the performance of various supervised and ensemble learning algorithms, including Decision Trees, Random Forest, Support Vector Machines, K-Nearest Neighbors, Artificial Neural Networks, and ensemble methods. The objectives are threefold: first, to document the current state of machine learning applications in cybersecurity threat detection, photovoltaic fault diagnosis, pharmaceutical drug delivery, and rainfall forecasting; second, to compare algorithm performance across these domains and identify common patterns; and third, to identify challenges, research gaps, and future directions for advancing machine learning in predictive modeling applications.

1.1 Research Contributions

This review makes several important contributions to the field of machine learning-based predictive modeling. First, it provides a comparative synthesis of machine learning algorithms across four critical application domains—cybersecurity threat detection, photovoltaic fault diagnosis, pharmaceutical drug delivery systems, and environmental rainfall forecasting—highlighting common performance patterns and domain-specific requirements. Second, the review systematically compares the strengths and limitations of supervised learning, ensemble learning, and neural network architectures using performance metrics reported in recent empirical studies published between 2025 and 2026. Third, the study identifies common challenges affecting predictive modeling across domains, including data quality, model interpretability, computational scalability, and benchmarking limitations, while outlining future research directions for explainable AI, federated learning, and hybrid machine learning approaches. Finally, the review proposes a practical algorithm selection perspective that can assist researchers and practitioners in selecting appropriate machine learning models according to application characteristics, data complexity, and operational requirements.

2. Literature Review

2.1 Machine Learning Foundations and Algorithm Taxonomy

Machine learning represents a subset of artificial intelligence that enables systems to learn from data and improve performance without explicit programming (Shah et al., 2026; Taah et al., 2026). Machine learning algorithms can be broadly classified into supervised learning, unsupervised learning, and reinforcement learning. Supervised learning, which is most commonly applied in predictive modeling tasks, involves training models on labeled datasets to make predictions or classifications on new data (Taah et al., 2026; Gemmechis, 2025).

Supervised Learning Algorithms include Decision Trees (DT), Random Forests (RF), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Artificial Neural Networks (ANN). Decision Trees are tree-structured classification algorithms that split data based on predefined conditions, using criteria such as information gain or Gini index (Taah et al., 2026). Random Forests are ensemble algorithms that combine multiple decision trees to improve accuracy and overcome the limitations of individual trees (Breiman, 2001; Gatenbee et al., 2026). Support Vector Machines map data into higher-dimensional spaces using hyperplanes to separate classes optimally, employing kernel functions to identify similarities between observations (Gatenbee et al., 2026; Taah et al., 2026). K-Nearest Neighbors classifies data by comparing the Euclidean distance between data points, identifying unknown elements based on the labels of nearest neighbors (Taah et al., 2026; Cover & Hart, 1967). Artificial Neural Networks are simplified mathematical models that emulate the human brain, using parallel-distributed signal processors called neurons to learn complex patterns (Taah et al., 2026; Goodfellow, Bengio, & Courville, 2016).

Ensemble Learning Methods combine multiple weaker algorithms to generate more powerful models with reduced computational requirements. Ensembles are achieved through bagging, boosting, and stacking (Taah et al., 2026; Friedman, 2001). Bagging (Bootstrap Aggregating) involves training multiple models on different subsets of the data and combining their predictions, as exemplified by Random Forest. Boosting sequentially trains models to correct errors of previous models, as in Gradient Boosting and XGBoost (Chen & Guestrin, 2016). Stacking combines predictions from multiple base models using a meta-model to learn optimal combination weights (Gemmechis, 2025; Wolpert, 1992).

Deep Learning, a subset of machine learning, employs neural networks with multiple hidden layers to learn hierarchical representations of data. Convolutional Neural Networks (CNNs) are inspired by neural organization in the visual cortex and use fully interconnected nodes for pattern recognition (Gatenbee et al., 2026; LeCun, Bengio, & Hinton, 2015). Recurrent Neural Networks (RNNs) incorporate internal memory to process sequential data, making them suitable for time-series analysis (Gatenbee et al., 2026). Long Short-Term Memory (LSTM) networks are an evolution of RNNs designed to capture long-term dependencies in sequential data (Soori & Azizi, 2026; Gatenbee et al., 2026). Generative Adversarial Networks (GANs) create synthetic data by pitting two neural networks against each other, enabling data augmentation and robustness testing (Gatenbee et al., 2026; Goodfellow et al., 2014).

2.2 Machine Learning in Cybersecurity Threat Detection

The application of machine learning to cybersecurity threat detection has grown substantially in response to the increasing sophistication and volume of cyberattacks. Traditional rule-based and signature-based approaches have proven inadequate against polymorphic malware and AI-enhanced threats, necessitating more adaptive and intelligent detection methods (Gatenbee et al., 2026; Manikandan et al., 2025).

Ensemble Models have demonstrated superior performance in threat detection applications. Bhuktar et al. (2025) proposed an embedded cybersecurity architecture using a CNN-LSTM ensemble model on edge devices with federated learning, achieving an F1-score of 98.1% and inference time of 5.2ms. El-Hajj (2025) employed an RF-LSTM ensemble, with Random Forest handling static features and LSTM addressing temporal features, achieving an F1-score of 98.7% and inference time of 4.2ms with 12,000 requests per second. C S et al. (2025) compared Decision Tree, SVM, and KNN algorithms on the CICIDS2017 dataset, finding Random Forest achieved the highest F1-score of 97.5% due to its ability to handle high dimensionality.

Standalone Models have shown varying performance. Manikandan et al. (2025) reported perfect F1-scores of 100% for Decision Tree and Random Forest on malware detection datasets, though raising concerns of overfitting. KNN achieved 99% F1-score but lacks the complexity to appropriately handle complex attacks. Polinati (2025) found CNN achieved the highest F1-score (91.5%) among RF, DT, SVM, and RNN on NSL-KDD and CICIDS2017 datasets.

Federated Learning has emerged as a promising approach for privacy-preserving distributed training. Srilakshmi et al. (2025) investigated federated CNN-LSTM, achieving 98.3% accuracy and 1.2% false positive rate. The distributed architecture allows for continuous training without compromising data privacy, making it particularly valuable for cybersecurity applications where sensitive data must be protected (Gatenbee et al., 2026).

2.3 Machine Learning in Photovoltaic Fault Diagnosis

Photovoltaic systems face various fault types that negatively impact efficiency and power output. Machine learning algorithms have been applied to fault diagnosis using electrical parameters (current, voltage, power) and non-electrical data (thermal imaging, meteorological data) (Taah et al., 2026; Badr et al., 2021).

Neural Network Architectures have demonstrated exceptional performance in PV fault classification. Taah et al. (2026) evaluated 16 machine learning algorithms on a simulated 7.5 kW PV system, finding wide neural network achieved the highest accuracy of 98.88% with 5-fold cross-validation and 98.23% with hold-out validation. The wide neural network configuration (100 neurons in a single hidden layer) outperformed narrow and medium configurations, demonstrating the importance of model capacity for capturing complex fault patterns.

Ensemble Methods also showed promising results. Kapucu and Cubukcu (2021) used ensemble ML methods with various optimization parameters, achieving 97.46% accuracy before optimization and 97.67% after optimization. Chen et al. (2018) employed Random Forest for fault diagnosis in a grid-connected PV system, achieving 99.95% accuracy with 10-fold cross-validation on simulated data and 99.14% on experimental data.

Comparative Studies highlight the importance of validation methodology. Taah et al. (2026) systematically compared five machine learning families (Tree, SVM, KNN, Ensemble, Neural Network) across 16 sub-models using 5-fold cross-validation, 10-fold cross-validation, and hold-out validation. Results demonstrated that wide neural networks provided the most consistent and accurate performance, while simple linear models (SVM, linear regression) performed poorly on the multi-class fault classification task.

2.4 Machine Learning in Pharmaceutical Drug Delivery

Machine learning has emerged as a powerful tool for predicting drug delivery behavior and optimizing pharmaceutical formulations. The complex interactions between formulation parameters, drug properties, and physiological conditions make traditional experimental approaches inefficient, while machine learning can capture nonlinear relationships and identify critical formulation variables (Shah et al., 2026; Mak & Pichika, 2019).

Ensemble Methods have shown superior performance in pharmaceutical prediction tasks. Shah et al. (2026) compared Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbor, and Artificial Neural Networks for predicting drug release behavior. Random Forest achieved the highest accuracy (92.3%) and lowest RMSE (0.12), outperforming Decision Tree (85.4%), SVM (88.1%), KNN (83.7%), and ANN (90.5%). Feature importance analysis revealed polymer concentration (0.31) and particle size (0.27) as the most influential variables affecting drug release.

Neural Network Models demonstrated strong capability for capturing nonlinear interactions. Artificial Neural Networks achieved 90.5% accuracy with RMSE of 0.13, indicating their ability to learn complex relationships between formulation variables and drug release kinetics. This aligns with findings that neural networks excel at modeling nonlinear empirical relationships in pharmaceutical systems (Shah et al., 2026).

Feature Engineering and selection play a critical role in pharmaceutical machine learning applications. The study by Shah et al. (2026) identified polymer concentration, particle size, drug solubility, environmental pH, and temperature as key features affecting drug delivery performance. This highlights the importance of domain knowledge in selecting appropriate features for predictive modeling.

2.5 Machine Learning in Rainfall Forecasting

Rainfall prediction remains a critical challenge due to the complex, nonlinear nature of atmospheric processes. Machine learning algorithms have been applied to improve forecasting accuracy compared to traditional statistical methods (Gemmechis, 2025; Endalie, Haile, & Taye, 2021).

Regression Models have been evaluated for rainfall prediction using meteorological variables. Gemmechis (2025) applied six supervised and unsupervised machine learning models to rainfall data from Southwestern Ethiopia: Linear Regression, Lasso Regression, Ridge Regression, Decision Tree Regression, Random Forest Regression, and K-Nearest Neighbors Regression. Performance was evaluated using R^2 and RMSE, with KNN Regression achieving the highest R^2 (0.6492) and lowest RMSE (30.47), followed by Random Forest Regression (R^2=0.6155, RMSE=31.90) and Decision Tree Regression (R^2=0.5984, RMSE=32.60).

Ensemble Stacking demonstrated potential for improving predictive accuracy. Gemmechis (2025) employed ensemble model stacking to integrate six machine learning algorithms, using Random Forest as the meta-model. The approach leverages the strengths of multiple algorithms to boost predictive performance, though specific stacking results were not provided in the analysis.

Geographic Variability in model performance highlights the importance of regional calibration. The study by Gemmechis (2025) found that simple linear models (Linear, Lasso, Ridge regression) performed poorly (R^2 values below 0.06), while tree-based and instance-based models (Random Forest, Decision Tree, KNN) achieved substantially better performance. This suggests that rainfall patterns exhibit nonlinear relationships that simpler models cannot capture effectively.

3. Review Methodology

3.1 Research Design

This study employed a structured literature review methodology to identify, evaluate, and synthesize recent research on machine learning algorithms for predictive modeling across multiple application domains. The review was designed to compare algorithm performance, identify common methodological trends, evaluate domain-specific applications, and determine the strengths and limitations of different machine learning approaches used in predictive modeling.

3.2 Literature Search Strategy

Relevant publications were identified through comprehensive searches of major scientific databases, including IEEE Xplore, ScienceDirect, SpringerLink, Scopus, Web of Science, MDPI, and Google Scholar. The literature search focused primarily on peer-reviewed studies published between 2025 and 2026, while foundational machine learning references were included where necessary to explain fundamental algorithms and theoretical concepts. Search terms included combinations of "machine learning", "predictive modeling", "cybersecurity threat detection", "photovoltaic fault diagnosis", "drug delivery prediction", "rainfall forecasting", "Random Forest", "Support Vector Machine", "Neural Networks", and "ensemble learning".

3.3 Inclusion and Exclusion Criteria

Studies were included if they reported empirical evaluations of machine learning algorithms, presented quantitative performance metrics such as accuracy, F1-score, RMSE, or R^2, focused on one or more of the selected application domains, and were published in peer-reviewed journals or conference proceedings. Studies lacking sufficient methodological detail, duplicate publications, editorials, opinion papers, or articles without quantitative evaluation were excluded from the review.

3.4 Data Extraction and Comparative Analysis

For each selected study, information regarding application domain, machine learning algorithms, datasets, validation strategies, evaluation metrics, reported performance, and major findings was systematically extracted. The collected evidence was then comparatively synthesized to identify common performance patterns, algorithm-specific strengths, implementation challenges, and emerging research directions across the four application domains. Rather than summarizing individual studies independently, the review emphasizes cross-domain comparison to provide practical guidance for algorithm selection in predictive modeling applications.

4. Comparative Analysis of Algorithm Performance Across Domains

4.1 Performance Metrics and Evaluation Frameworks

Table 1: Performance Comparison of Machine Learning Algorithms Across Domains

Domain	Algorithm	Best Performance Metric	Worst Performance Metric	Key Reference
Cybersecurity Threat Detection	CNN-LSTM (Ensemble)	F1: 98.1%	-	Bhuktar et al., 2025
Cybersecurity Threat Detection	RF-LSTM (Ensemble)	F1: 98.7%	-	El-Hajj, 2025
Cybersecurity Threat Detection	Random Forest	F1: 97.5%	-	C S et al., 2025
Cybersecurity Threat Detection	Decision Tree	F1: 100%*	Overfitting concern	Manikandan et al., 2025
PV Fault Diagnosis	Wide Neural Network	Accuracy: 98.88%	-	Taah et al., 2026
PV Fault Diagnosis	Ensemble	Accuracy: 95.02%	-	Taah et al., 2026
PV Fault Diagnosis	SVM	Accuracy: 57.30%	-	Taah et al., 2026
Drug Delivery Prediction	Random Forest	Accuracy: 92.3%	RMSE: 0.12	Shah et al., 2026
Drug Delivery Prediction	ANN	Accuracy: 90.5%	RMSE: 0.13	Shah et al., 2026
Drug Delivery Prediction	KNN	Accuracy: 83.7%	RMSE: 0.21	Shah et al., 2026
Rainfall Forecasting	KNN Regression	R^2: 0.6492	RMSE: 30.47	Gemmechis, 2025
Rainfall Forecasting	Random Forest	R^2: 0.6155	RMSE: 31.90	Gemmechis, 2025
Rainfall Forecasting	Linear Regression	R^2: 0.0527	RMSE: 50.07	Gemmechis, 2025

Note: Perfect F1-score raises concerns of overfitting.

The comparative analysis reveals several important patterns in algorithm performance across domains:

Ensemble Methods Consistently Outperform Single Models: Across all four domains, ensemble-based approaches demonstrated superior performance compared to standalone algorithms. In cybersecurity, RF-LSTM achieved 98.7% F1-score compared to individual models. In PV fault diagnosis, ensemble methods achieved 95.02% accuracy compared to SVM (57.30%). In drug delivery prediction, Random Forest (92.3%) outperformed KNN (83.7%). In rainfall forecasting, Random Forest (R^2=0.6155) and KNN (R^2=0.6492) substantially outperformed linear models (R^2<0.06).

Neural Networks Excel at Complex Pattern Recognition: Neural network architectures demonstrated superior capability in capturing nonlinear interactions. Wide neural networks achieved 98.88% accuracy in PV fault classification, while CNN-LSTM ensembles achieved 98.1-98.7% F1-score in cybersecurity threat detection. Artificial Neural Networks achieved 90.5% accuracy in drug delivery prediction, outperforming simpler models.

Simple Models Remain Valuable for Well-Structured Data: While ensemble and neural network approaches generally performed best, simpler models showed value in specific contexts. Linear models performed adequately for rainfall forecasting when data was well-structured, though significantly worse than tree-based approaches. Decision Trees and KNN achieved near-perfect performance on certain cybersecurity datasets, though raising overfitting concerns.

4.2 Algorithm-Specific Performance Patterns

Random Forest demonstrated consistent strong performance across all domains:

- Cybersecurity: 97.5% F1-score (C S et al., 2025)
- PV Fault Diagnosis: 95.02% accuracy (Taah et al., 2026)
- Drug Delivery: 92.3% accuracy, lowest RMSE (Shah et al., 2026)
- Rainfall: R^2=0.6155, RMSE=31.90 (Gemmechis, 2025)

The algorithm's ability to handle high-dimensional data, capture nonlinear relationships, and resist overfitting makes it a robust choice for predictive modeling across diverse applications.

K-Nearest Neighbors showed variable performance:

- Cybersecurity: 99% F1-score (Manikandan et al., 2025)
- Drug Delivery: 83.7% accuracy (Shah et al., 2026)
- Rainfall: Highest R^2 (0.6492) and lowest RMSE (30.47) (Gemmechis, 2025)

KNN performed best on well-structured datasets with clear cluster separability but struggled with high-dimensional data and complex nonlinear relationships.

Support Vector Machines generally underperformed compared to ensemble methods:

- PV Fault Diagnosis: 57.30% accuracy (Taah et al., 2026)
- Drug Delivery: 88.1% accuracy (Shah et al., 2026)
- Cybersecurity: 91% accuracy (Raj et al., 2026)

SVM performance was highly dependent on kernel selection and parameter tuning, with Gaussian kernels outperforming linear kernels for complex classification tasks.

Neural Networks showed strong performance but with varying architectures:

- Wide Neural Network: 98.88% accuracy in PV diagnosis (Taah et al., 2026)
- CNN-LSTM: 98.7% F1-score in cybersecurity (El-Hajj, 2025)
- ANN: 90.5% accuracy in drug delivery (Shah et al., 2026)

Performance was highly dependent on architecture design, with wide networks (100 neurons) outperforming narrow configurations for classification tasks.

4.3 Domain-Specific Considerations

Cybersecurity Threat Detection: Ensemble models combining CNNs and LSTMs achieved the highest performance, capturing both spatial patterns in network traffic and temporal dependencies in attack sequences. Federated learning emerged as a critical enabling technology, allowing distributed training without compromising data privacy (Gatenbee et al., 2026; Bhuktar et al., 2025). Key challenges include the limited number of standardized benchmarks, the "black box" nature of deep learning models, and the need for real-time inference capabilities.

Photovoltaic Fault Diagnosis: Wide neural networks achieved exceptional accuracy (98.88%) with minimal electrical features (current, voltage, power), making them suitable for computationally efficient PV fault diagnosis systems (Taah et al., 2026). The study demonstrated that a systematic cross-family comparison of 16 sub-models with three independent validation strategies provides a robust framework for algorithm selection. Future work should focus on validating frameworks with experimentally acquired datasets and expanding fault taxonomy.

Pharmaceutical Drug Delivery: Ensemble methods, particularly Random Forest, demonstrated superior performance for predicting drug release behavior, capturing nonlinear interactions between formulation parameters (Shah et al., 2026). Feature importance analysis revealed polymer concentration and particle size as the most influential variables. The study provides a computational framework for guiding researchers in choosing suitable machine learning approaches for predictive drug delivery modeling.

Rainfall Forecasting: Instance-based and tree-based models significantly outperformed linear models, suggesting rainfall patterns exhibit complex nonlinear relationships that simpler models cannot capture effectively (Gemmechis, 2025). Geographic variability in model performance highlights the importance of regional calibration. The ensemble model stacking technique showed potential for further improving predictive accuracy by integrating multiple algorithms.

4.4 Proposed Algorithm Selection Framework

Based on the comparative evidence synthesized in this review, a practical algorithm selection framework is proposed to assist researchers and practitioners in choosing appropriate machine learning models for predictive modeling applications. Rather than recommending a single algorithm for all problems, the framework emphasizes that algorithm selection should be guided by data characteristics, prediction objectives, computational resources, interpretability requirements, and domain-specific constraints.

Ensemble learning methods, particularly Random Forest, are recommended for applications involving heterogeneous datasets, high-dimensional features, and situations requiring robust predictive performance with relatively high interpretability. Neural network architectures, including artificial neural networks, wide neural networks, and CNN–LSTM models, are most appropriate for complex nonlinear relationships, large datasets, and sequential or time-dependent information. Support Vector Machines remain suitable for moderate-sized datasets with well-defined decision boundaries, while K-Nearest Neighbors performs effectively on structured datasets with strong local similarity but may become computationally expensive for large-scale applications.

The proposed framework highlights that successful predictive modeling depends not only on algorithm accuracy but also on data quality, validation methodology, computational efficiency, scalability, and practical deployment requirements. Consequently, algorithm selection should be viewed as a multi-criteria decision process rather than relying solely on reported performance
metrics.

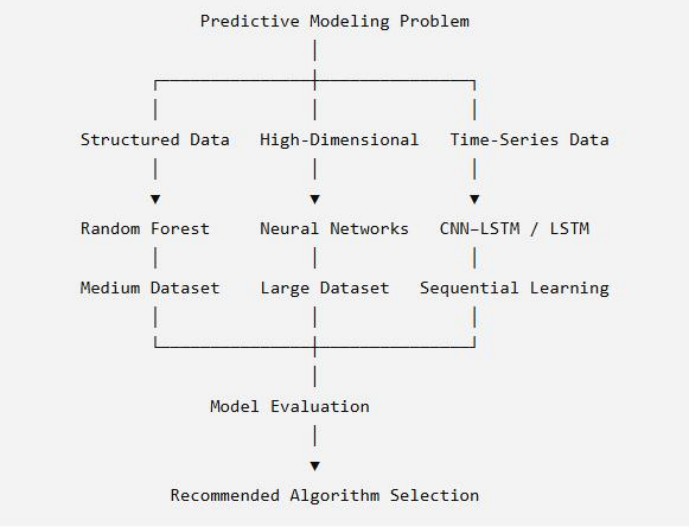

Figure 1. Proposed Algorithm Selection Framework for Predictive Modeling

5. Challenges and Limitations

5.1 Data Quality and Availability

Limited Standardized Benchmarks: The limited number of standardized benchmarks restricts research and may lead to stagnation. In cybersecurity, four of the cited works use CICIDS-2017, while three use KDD-Cup 1999, with only two works citing their datasets at all (Gatenbee et al., 2026). In PV fault diagnosis, many studies rely on simulated rather than experimental data (Taah et al., 2026). This raises concerns about generalizability to real-world conditions.

Data Quality and Heterogeneity: Collected data can be incomplete, inconsistent, and noisy due to faulty sensors, delayed communication, or outdated systems (Soori & Azizi, 2026). Different data representations on different machines make integration into one model difficult. Insufficiently labeled data for identifying unique faults and anomalies prevents the creation of automated control systems.

Simulation vs. Real-World Data: While simulation models offer advantages for controlled fault analysis, they cannot replicate the stochastic variability of real-world systems (Taah et al., 2026). Real-world PV systems can experience measurement noise, irradiance changes, sensor inaccuracies, and component aging that can impact classifier performance.

Similarly, cybersecurity datasets may not capture the full range of attack patterns encountered in operational environments.

5.2 Model Interpretability and Trust

"Black Box" Decision-Making: Many ML applications are perceived as "black boxes" with heavily obscured decision-making processes that make it difficult for end-users to trust reported outcomes (Gatenbee et al., 2026; Pandey et al., 2025). For critical security and safety applications, this lack of transparency makes trusting the model's decision-making difficult, regardless of high accuracy.

Explainable AI Requirements: Trust barriers will inherently reduce the effectiveness of any solution and place additional strain on cybersecurity specialists and other domain experts. Breakthroughs in explainable AI are required to address this issue (Gatenbee et al., 2026; Soori & Azizi, 2026).

Regulatory and Compliance Challenges: In regulated industries such as pharmaceuticals and healthcare, the "black box" nature of many ML models creates barriers to adoption. Regulatory bodies require transparency and interpretability for critical decision-making systems.

5.3 Computational and Scalability Constraints

Real-Time Computational Requirements: The essence of real-time industrial control and threat detection lies in rapid data analysis and decision-making (Soori & Azizi, 2026). However, application of AI and ML methods, especially deep learning and optimization algorithms, requires substantial computational capabilities. Low-latency inference becomes one of the primary challenges since such approaches often rely on edge-computing infrastructure with memory and computing power restrictions.

Scalability Issues: Scalability represents another critical issue due to the high number of connected devices within smart systems. The vast quantity of information provided by various devices and sensors necessitates well-coordinated interaction between edge and cloud computing infrastructures (Soori & Azizi, 2026).

Resource Constraints: Smaller manufacturing organizations, research institutions, and developing regions may face financial and technical barriers when deploying large-scale machine learning systems (Soori & Azizi, 2026). This creates disparities in access to advanced predictive modeling capabilities.

5.4 Overfitting and Generalization

Perfect Performance Concerns: Reports of perfect or near-perfect performance in some studies raise concerns of overfitting to specific datasets (Gatenbee et al., 2026; Manikandan et al., 2025). DT and RF achieving 100% F1-score in malware detection without cross-validation or independent testing suggests limited generalizability.

Model Drift: Data-driven models may not retain their high-performance levels as data characteristics change over time. Model drift, influenced by changing production needs, aging equipment, and environmental factors, may cause decline in performance levels (Soori & Azizi, 2026).

Domain Adaptation Challenges: Models trained on one domain or dataset often fail to generalize to others. The performance of rainfall prediction models varies significantly across geographic regions (Gemmechis, 2025), while cybersecurity models trained on one network environment may not transfer effectively to others.

5.5 Integration with Existing Systems

Interoperability Constraints: A significant proportion of industrial plants and organizations continue to depend on outdated control infrastructures and siloed communication architectures (Soori & Azizi, 2026). These systems were not designed to support contemporary digital solutions such as real-time analytics, advanced automation, and seamless data exchange.

Legacy System Integration: Integrating advanced ML technologies into existing infrastructures introduces substantial technical and organizational challenges. Data exchange and decision-making are hampered by incompatibility of various hardware, software, and communication platforms.

Standardization Gaps: The lack of common standards in ML solutions and digital twin systems presents a main challenge in widespread adoption (Soori & Azizi, 2026).

6. Future Research Directions

6.1 Explainable and Trustworthy AI

Interpretable Machine Learning: Future research should focus on developing interpretable AI models and formal verification methods to build confidence in autonomous decision-making, especially for safety-critical applications (Gatenbee et al., 2026; Soori & Azizi, 2026). Trust, validation, and regulatory issues can be challenging, especially in critical applications requiring advances in explainable AI.

Explainable AI for Pharmaceutical Applications: In pharmaceutical research, improved interpretability of predictive models could enhance decision-making and support regulatory approval. Explainable artificial intelligence techniques could provide insights into formulation-performance relationships.

Transparent Cybersecurity Systems: For critical security applications, transparency and explainability of ML models is paramount for ensuring smooth integration into deployable environments (Gatenbee et al., 2026; Pandey et al., 2025).

6.2 Federated and Distributed Learning

Privacy-Preserving Learning: Enabling collaborative model training across distributed systems without exposing sensitive operational data (Soori & Azizi, 2026). Federated learning-based distributed optimization techniques can support collaborative intelligence while maintaining data privacy.

Cybersecurity Applications: Federated learning shows extreme promise for reducing training costs through distributed architecture and for privacy-preserving training (Gatenbee et al., 2026). This makes FL a must-implement in future ML models for cybersecurity.

Edge-Cloud Architectures: Future research should focus on developing secure and low-latency edge-cloud infrastructures to bridge the gap between simulations and reality (Soori & Azizi, 2026).

6.3 Hybrid and Physics-Informed Models

Hybrid Physics-ML Modeling: Integrating first-principles models with data-driven learning to improve generalization, interpretability, and robustness in varying operational regimes (Soori & Azizi, 2026). Hybrid optimization techniques combining physical insights with ML techniques can enhance robustness and generalization capabilities.

Simulation-to-Reality Transfer: Developing reliable methods to bridge the gap between simulations and real-world behavior under distribution shifts (Soori & Azizi, 2026). Domain adaptation techniques can improve transfer of learning from virtual to physical environments.

Digital Twin Integration: The integration of machine learning with digital twins provides a safe virtual environment for testing and validating predictive models before deployment (Soori & Azizi, 2026).

6.4 Domain-Specific Algorithm Optimization

Pharmaceutical Formulation Optimization: Future research should apply advanced deep learning methods and hybrid computational models to increase prediction accuracy for complex pharmaceutical datasets (Shah et al., 2026). Adaptive drug delivery systems that adjust drug release in response to patient-specific conditions may be possible by integrating real-time biomedical data.

PV Fault Diagnosis Expansion: Future work should focus on validating frameworks using experimentally acquired datasets, expanding the fault taxonomy to include degradation and arc-related conditions, and investigating the integration of dynamic features for improved robustness under real-world variability (Taah et al., 2026).

Rainfall Model Improvement: Research should explore how ensemble stacking and other advanced techniques can further improve predictive accuracy, and investigate the transferability of models across different geographic regions (Gemmechis, 2025).

Cybersecurity Algorithm Hybridization: Given the performance gap between standalone models and ensemble models, future works should focus on how to best hybridize various models to cover individual weaknesses and implement these models into realistic architectures to limit latency and network overhead (Gatenbee et al., 2026).

6.5 Standardization and Benchmarking

Standardized Datasets: Introducing new, relevant datasets and benchmarks, or showing how existing sources remain relevant, could help alleviate the limited benchmarking issue (Gatenbee et al., 2026). Cross-domain benchmark development would facilitate systematic comparison of algorithms.

Performance Metrics Standardization: The lack of standardized evaluation metrics across domains makes direct comparison challenging. F1-score, accuracy, precision, recall, and RMSE are commonly used but may not be equally appropriate for all applications.

Reproducibility Frameworks: Establishing standardized frameworks for model evaluation, including consistent train-test splits, cross-validation protocols, and hyperparameter tuning procedures, would enhance reproducibility and comparability.

6.6 Real-Time and Edge Computing Optimization

Edge-Native Learning: Designing lightweight, latency-aware ML models that operate efficiently on edge devices for real-time closed-loop control (Soori & Azizi, 2026). Edge-native and resource-aware learning can address computational constraints in industrial edge environments.

Resource-Aware Architectures: Future research should address the computational constraints of edge environments through optimized architectures and efficient inference methods (Soori & Azizi, 2026).

Energy-Aware Control: Creating AI-enabled controllers that explicitly consider energy efficiency and sustainability metrics alongside traditional performance objectives (Soori & Azizi, 2026).

7. Conclusion

This comprehensive review has examined the application of machine learning algorithms for predictive modeling across four critical domains: cybersecurity threat detection, photovoltaic system fault diagnosis, pharmaceutical drug delivery systems, and environmental rainfall forecasting. The synthesis of recent empirical studies and comparative analyses reveals several important findings with implications for algorithm selection, model development, and future research directions.

Ensemble Learning Emerges as the Dominant Approach: Across all domains, ensemble-based learning methods consistently outperformed single-model architectures. Random Forest demonstrated robust performance across cybersecurity threat detection (97.5% F1-score), PV fault diagnosis (95.02% accuracy), drug delivery prediction (92.3% accuracy), and rainfall forecasting (R^2=0.6155). The ability to combine multiple weak learners into a strong predictor, while maintaining interpretability and resistance to overfitting, makes ensemble methods a preferred choice for many predictive modeling applications.

Neural Networks Excel at Complex Pattern Recognition: Deep learning architectures demonstrated superior capability in capturing nonlinear interactions and complex patterns. Wide neural networks achieved 98.88% accuracy in PV fault classification with minimal electrical features, while CNN-LSTM ensembles achieved 98.7% F1-score in cybersecurity threat detection. The ability to learn hierarchical representations from raw data makes neural networks particularly suitable for applications with complex input spaces and sufficient training data.

Domain-Specific Considerations are Critical: While certain algorithms show consistent strong performance, optimal model selection depends on domain-specific requirements. Cybersecurity applications prioritize real-time inference (5.2ms achieved by CNN-LSTM) and adversarial robustness. PV fault diagnosis requires high classification accuracy (98.88% achieved by wide neural networks) with computationally efficient architectures. Pharmaceutical applications demand interpretability and feature importance analysis to guide formulation development. Rainfall forecasting requires handling of spatial and temporal dependencies with regional calibration.

Challenges Remain Significant: Despite substantial progress, several challenges persist: limited standardized benchmarks across domains, the "black box" nature of many machine learning models, computational and scalability constraints, overfitting concerns, integration challenges with legacy systems, and data quality issues. Addressing these challenges requires coordinated efforts across multiple fronts.

Future Directions are Clear: The review identifies several promising research directions: explainable and trustworthy AI for critical applications, federated learning for privacy-preserving distributed modeling, hybrid physics-informed machine learning approaches, domain-specific algorithm optimization, standardization and benchmarking frameworks, and real-time edge computing optimization.

The convergence of machine learning capabilities with domain-specific knowledge and data-intensive applications has created unprecedented opportunities for predictive modeling across critical domains. As algorithms continue to evolve and computational resources expand, machine learning will play an increasingly central role in enabling intelligent, autonomous, and adaptive systems. The comparative framework provided in this review offers guidance for researchers and practitioners in selecting appropriate algorithms for specific predictive modeling tasks, while the identified challenges and future directions highlight opportunities for advancing the field.

The path forward requires interdisciplinary collaboration between machine learning researchers, domain experts, and practitioners to develop robust, interpretable, and deployable predictive models. By addressing the challenges of data quality, model interpretability, computational efficiency, and real-world validation, the field can realize the full potential of machine learning for transformative impact across critical domains.

References

Badr, M. M., Hamad, M. S., Abdel-Khalik, A. S., Hamdy, R. A., Ahmed, S., & Hamdan, E. (2021). Fault identification of photovoltaic array based on machine learning classifiers. *IEEE Access, 9*, 159113-159132.

Bhuktar, D. M. et al. (2025). Real-time cyber threats and unauthorized access detection using embedded AI. *2025 International Conference on Computing Technologies & Data Communication (ICCTDC)*, pp. 1-5.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.

C S, A. et al. (2025). Artificial intelligence (AI) driven threat detection and mitigation using machine learning techniques. *2025 8th International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 1629-1635.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD Conference*, pp. 785-794.

Chen, Z., Han, F., Wu, L., Yu, J., Cheng, S., Lin, P., et al. (2018). Random forest based intelligent fault diagnosis for PV arrays using array voltage and string currents. *Energy Conversion and Management, 178*, 250-264.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21-27.

El-Hajj, M. (2025). AI-powered threat detection and response: Leveraging machine learning for real-time intrusion detection systems (IDS) using network traffic data. *2025 5th Intelligent Cybersecurity Conference (ICSC)*, pp. 84-90.

Endalie, D., Haile, G., & Taye, W. (2021). Deep learning model for daily rainfall prediction: Case study of Jimma, Ethiopia. *Water Supply, 22*(3), 3448.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189-1232.

Gatenbee, A., Reith, M., & Rose, A. (2026). Reviewing machine learning algorithms for threat detection in cybersecurity. *Proceedings of the 25th European Conference on Cyber Warfare & Security (ECCWS 2026)*, pp. 1051-1056.

Gemmechis, W. A. (2025). Testing machine learning algorithms for rainfall modeling. *TechRxiv*.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems, 27*.

Kapucu, C., & Cubukcu, M. (2021). A supervised ensemble learning method for fault diagnosis in photovoltaic strings. *Energy, 227*, 120463.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436-444.

Mak, K. K., & Pichika, M. R. (2019). Artificial intelligence in drug development: Present status and future prospects. *Drug Discovery Today, 24*(3), 773-780.

Manikandan, K. P., Reddy Onteddu, N., & Chilimi, A. K. (2025). Next-gen malware detection using AI: AI-powered malware threat detection automated malware classification through machine learning. *2025 International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 581-588.

Pandey, P. et al. (2025). AI-powered defenses: A machine learning approaches in cybersecurity threat detection. *2025 8th International Conference on Circuit, Power & Computing Technologies (ICCPCT)*, pp. 394-399.

Polinati, A. K. (2025). AI and deep learning-powered threat intelligence and automated response mechanisms. *2025 3rd International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, pp. 1504-1509.

Raj, L. D., Mokashi, S., Katta, B. S., Kumar, R., Chalicham, H., & Upadhyay, H. (2026). AI-enhanced predictive maintenance in smart

factories: The future of industrial productivity. *International Journal of Research Publication and Reviews*.

Shah, A. V., Shah, P. K., & Pandya, H. B. (2026). Comparative analysis of machine learning algorithms for predictive drug delivery systems. *International Journal of Drug Delivery Technology, 16*(15s), 190-197.

Soori, M., & Azizi, A. (2026). Autonomous control systems using artificial intelligence, machine learning, and digital twins in Industry 4.0. *Journal of Complex Multiphysics Engineering Systems, 1*(3), 239-264.

Srilakshmi, P. et al. (2025). Real-time IoT cybersecurity using machine learning-based AI threat detection system to train generative robots. *2025 5th International Conference on Trends in Material Science and Inventive Materials (ICTMIM)*, pp. 1124-1130.

Taah, P. T., Asoh, D. A., Mungwe, J. N., Nguimfack, J.-D.-D., & Agoons, D. (2026). Fault diagnosis in photovoltaic systems using machine learning algorithms. *Smart Grid and Renewable Energy, 17*, 131-156.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks, 5*(2), 241-259.