OPEN ACCESS
Journal of Computer Science Advancements
Vol. 4 No. 3. June 2026, pp. 188-203
DOI. 10.70177/jsca.v4i3.4104
Research Article

ARTIFICIAL INTELLIGENCE MODELS FOR PREDICTIVE ANALYTICS USING BIG DATA MINING TECHNIQUES

Soleman[1], Ahmed Al-Harthy[2]
[1] Universitas Borobudur, Indonesia
[2] Sultan Qaboos University, Oman

Corresponding Author:

Soleman,
Department of Information Systems, Faculty of Computer Science, Universitas Borobudur.
Jl. Raya Kalimalang No.1, RT.9/RW.4, Cipinang Melayu, Kec. Makasar, Jakarta Timur, Daerah Khusus Ibukota Jakarta 13620
Email: soleman@borobudur.ac.id

Article Info
Received: December 8, 2025
Revised: March 9, 2026
Accepted: May 12, 2026
Online Version: June 30, 2026

Abstract
Rapid digital transformation has generated unprecedented volumes of heterogeneous data, creating significant opportunities for predictive analytics while simultaneously increasing challenges related to data quality, scalability, computational complexity, and decision reliability. Conventional predictive models frequently experience performance degradation when processing high-dimensional and continuously evolving Big Data environments. This study aimed to develop and evaluate an integrated Artificial Intelligence framework that combines advanced Big Data mining techniques with hybrid machine learning models to improve predictive accuracy, computational efficiency, and analytical robustness. Quantitative computational research was conducted using large-scale structured and semi-structured datasets processed through data preprocessing, feature engineering, dimensionality reduction, ensemble learning, deep learning, distributed computing, and hyperparameter optimization. Model performance was assessed using accuracy, precision, recall, F1-score, area under the receiver operating characteristic curve, computational time, memory utilization, and scalability. Experimental results demonstrated that the proposed hybrid framework achieved 98.63% prediction accuracy, an AUC-ROC of 0.995, substantially reduced computational time, lower memory consumption, and superior scalability compared with conventional machine learning and deep learning approaches. Statistical analyses confirmed significant performance improvements across all principal evaluation metrics. Findings indicate that integrating intelligent data mining with Artificial Intelligence enhances predictive capability by optimizing the complete analytical pipeline rather than individual algorithms alone, providing a scalable, efficient, and reliable framework for predictive analytics across diverse Big Data application domains.

Keywords: Artificial Intelligence; Big Data Mining; Hybrid Machine Learning; Predictive Analytics; Scalable Data Processing.

Journal Homepage https://research.adra.ac.id/index.php/jcsa
How to cite: Soleman, Soleman., Harthy, A. A. (2026). Artificial Intelligence Models for Predictive Analytics Using Big Data Mining Techniques. *Journal of Computer Science Advancements, 4*(3), 188–203. https://doi.org/10.70177/jsca.v4i3.4104
Published by: Yayasan Adra Karima Hubbi

INTRODUCTION

Rapid digital transformation has fundamentally changed the way organizations generate, collect, store, and utilize information across virtually every sector of society (Smith Ballester et al., 2025). Massive volumes of structured and unstructured data are continuously produced through digital transactions, social media platforms, sensor networks, Internet of Things devices, healthcare systems, financial services, manufacturing processes, and scientific research. Such unprecedented data availability has created significant opportunities for extracting valuable knowledge capable of supporting evidence-based decision-making (Kaur et al., 2026). Traditional statistical approaches, however, increasingly struggle to process datasets characterized by high volume, velocity, variety, veracity, and complexity, highlighting the need for more advanced analytical methods capable of transforming raw data into actionable intelligence.

Artificial Intelligence (AI) has emerged as one of the most transformative technological paradigms for predictive analytics because of its ability to identify complex patterns, model nonlinear relationships, and continuously improve predictive performance through learning mechanisms (Huang et al., 2025). Machine learning, deep learning, ensemble learning, and hybrid intelligent algorithms have demonstrated remarkable capabilities across domains including healthcare diagnosis, financial forecasting, fraud detection, smart manufacturing, precision agriculture, cybersecurity, and urban planning (Wang et al., 2025). Predictive analytics has consequently evolved from descriptive reporting toward intelligent decision-support systems capable of anticipating future events and reducing uncertainty in increasingly dynamic environments.

Big Data mining techniques constitute an essential component of intelligent predictive analytics because they enable efficient discovery of hidden patterns, associations, anomalies, and trends within large-scale heterogeneous datasets (Liu et al., 2025). Data preprocessing, feature engineering, dimensionality reduction, clustering, classification, association rule mining, and optimization collectively improve the quality and reliability of predictive models (Alhumaidi et al., 2025). Integration of Artificial Intelligence with advanced Big Data mining techniques therefore offers substantial potential for improving prediction accuracy, computational efficiency, scalability, and real-time decision support across multiple application domains characterized by complex data ecosystems.

Organizations continue to experience considerable challenges in converting large-scale data resources into reliable predictive knowledge (Gahane et al., 2025). Big Data environments frequently contain incomplete records, noisy observations, redundant variables, inconsistent formats, missing values, class imbalance, and continuously evolving data distributions (Imamguluyev et al., 2025). These characteristics substantially reduce predictive accuracy when conventional machine learning algorithms are applied without appropriate data mining strategies (Abzal Basha et al., 2025). Effective integration between Artificial Intelligence models and advanced data mining techniques therefore remains a significant technical challenge.

Existing predictive analytics systems often prioritize algorithmic performance while paying comparatively limited attention to data quality management, feature selection, computational scalability, model interpretability, and adaptability to dynamic data environments (Nimmagadda et al., 2025). High predictive accuracy obtained under laboratory conditions frequently deteriorates during real-world implementation because data characteristics evolve continuously over time (Padulo, 2025). Static prediction models consequently become less reliable when confronted with concept drift, heterogeneous information sources, and rapidly changing operational environments.

Growing concerns regarding algorithmic transparency, computational efficiency, and practical deployment further complicate predictive analytics implementation. Complex deep learning architectures frequently require substantial computational resources while providing limited interpretability for decision-makers (Deshpande & Augustine, 2025). Many

organizations therefore face trade-offs between prediction accuracy, computational cost, scalability, explainability, and operational feasibility (Eom, 2026). Developing integrated Artificial Intelligence frameworks capable of balancing these competing requirements remains an important research challenge for contemporary predictive analytics.

This study aims to develop an integrated Artificial Intelligence framework that combines advanced Big Data mining techniques with predictive analytics to improve prediction accuracy, computational efficiency, scalability, and model robustness across heterogeneous data environments (P. Sharma et al., 2025). Particular emphasis is placed on optimizing data preprocessing, feature engineering, model training, and prediction processes through systematic integration of machine learning algorithms and intelligent data mining methodologies (Chulajata et al., 2025). Achieving these objectives is expected to enhance decision-support capabilities for data-intensive organizational environments.

Research also seeks to evaluate the effectiveness of different Artificial Intelligence models operating under varying data characteristics, including structured and unstructured datasets, high-dimensional feature spaces, imbalanced class distributions, missing values, and continuously evolving information streams (Taoussi et al., 2025). Performance evaluation will examine prediction accuracy, precision, recall, F1-score, computational efficiency, model scalability, interpretability, and robustness (Strielkowski et al., 2025). Comparative analysis among multiple predictive models will identify optimal approaches for different Big Data environments.

Broader objectives include establishing methodological guidelines for designing scalable, reliable, and interpretable predictive analytics systems capable of supporting strategic decision-making across multiple industries (James C. Escolano et al., 2024). Research findings are expected to contribute toward the development of intelligent analytical frameworks suitable for healthcare, finance, manufacturing, transportation, education, environmental management, and other domains increasingly dependent upon large-scale data-driven decision processes.

Previous studies have extensively investigated machine learning algorithms, deep learning architectures, Big Data mining methods, and predictive analytics as independent research domains (Papadopoulou et al., 2026). Numerous investigations have evaluated classification algorithms, neural networks, decision trees, support vector machines, random forests, and ensemble learning techniques individually (Takamido et al., 2025). Comprehensive frameworks integrating intelligent data mining processes with predictive Artificial Intelligence models throughout the complete analytical pipeline remain comparatively limited (Lawand et al., 2025). Such fragmentation reduces understanding of how preprocessing, feature engineering, model optimization, and prediction interact collectively to determine overall analytical performance.

Current literature frequently evaluates predictive models using benchmark datasets collected under controlled experimental conditions (Khan et al., 2025). Real-world Big Data environments are considerably more complex because they contain heterogeneous information sources, high-dimensional variables, evolving distributions, noisy observations, and continuously expanding data volumes (Chawla et al., 2025). Performance obtained under simplified experimental conditions may therefore overestimate practical model effectiveness when deployed within operational organizational settings characterized by dynamic information ecosystems.

Existing research also tends to prioritize predictive accuracy as the principal evaluation criterion while paying comparatively less attention to scalability, interpretability, computational efficiency, model fairness, and long-term adaptability (A. Sharma et al., 2025). Organizations increasingly require predictive systems that balance multiple operational objectives rather than maximizing predictive accuracy alone (A. Sharma et al., 2025). Integrated evaluation frameworks capable of simultaneously assessing technical performance, computational feasibility, interpretability, and deployment readiness therefore remain insufficiently developed within contemporary predictive analytics research.

Novel contribution of this study resides in proposing a unified Artificial Intelligence framework integrating intelligent Big Data mining techniques, adaptive feature engineering, hybrid predictive modeling, and comprehensive performance optimization within a single analytical architecture (Al-Jabri & Alkahtani, 2026). Rather than optimizing isolated machine learning algorithms, the proposed framework evaluates predictive analytics as an integrated end-to-end process beginning with raw data acquisition and concluding with interpretable decision support. Such holistic integration distinguishes the proposed methodology from previous investigations emphasizing individual algorithmic improvements.

Scientific significance extends beyond algorithmic enhancement by providing a conceptual framework explaining interactions among data quality, feature selection, model complexity, computational efficiency, scalability, and predictive reliability (Celik & Dal, 2025). Dynamic integration of data mining and Artificial Intelligence contributes to a broader understanding of intelligent analytical systems capable of adapting to evolving Big Data environments. Generated knowledge advances theoretical discussions concerning explainable artificial intelligence, scalable machine learning, intelligent decision support, and data-centric Artificial Intelligence.

Practical justification arises from increasing organizational dependence on predictive analytics for strategic planning, operational optimization, risk management, customer behavior analysis, healthcare diagnosis, cybersecurity, financial forecasting, and industrial automation (Duarte-Medrano et al., 2025). Reliable predictive intelligence has become essential for organizations operating within data-intensive digital ecosystems characterized by rapid technological change and growing analytical complexity (Goyal et al., 2025). Successful implementation of the proposed framework has the potential to improve prediction quality, reduce computational costs, strengthen model transparency, enhance organizational decision-making, and accelerate digital transformation across diverse industrial and public-sector applications.

RESEARCH METHOD

Research Design

This study employed a quantitative computational research design integrating artificial intelligence, big data mining, and predictive analytics to develop and evaluate an intelligent predictive modeling framework for large-scale heterogeneous datasets (Palma et al., 2025). The research adopted a design–implementation–evaluation approach that combined data preprocessing, feature engineering, machine learning model development, hyperparameter optimization, and predictive performance assessment within a unified analytical pipeline (Ullah et al., 2026). This design was selected because predictive analytics in Big Data environments requires systematic integration of data mining processes and artificial intelligence algorithms to ensure accurate, scalable, and robust prediction under complex data conditions.

Model development was performed using supervised machine learning and deep learning algorithms capable of handling structured and semi-structured datasets characterized by high dimensionality, missing values, class imbalance, and nonlinear relationships (Pavunraj et al., 2025). Data mining procedures included data cleaning, normalization, feature selection, dimensionality reduction, clustering-assisted preprocessing, and anomaly detection before predictive model construction. Ensemble learning and hyperparameter optimization techniques were incorporated to improve prediction accuracy, reduce model variance, and enhance generalization performance across multiple application scenarios.

Experimental validation was conducted through repeated computational simulations using benchmark Big Data environments representing healthcare, finance, manufacturing, and smart city applications. Multiple artificial intelligence models were compared under identical experimental settings to evaluate prediction quality, computational efficiency, scalability, robustness, and interpretability. Statistical validation was subsequently performed to determine

whether observed performance differences among predictive models were statistically significant and practically meaningful.

Research Target/Subject

The data analysis pipeline followed a quantitative computational research design integrating big data mining, artificial intelligence, and predictive analytics within a unified design–implementation–evaluation approach. The preprocessing and feature engineering stages employed techniques such as missing value imputation, outlier detection, normalization, principal component analysis, and data balancing via the Synthetic Minority Oversampling Technique (SMOTE). Model development and predictive analytics were executed using an array of supervised machine learning and deep learning algorithms including Random Forest, Gradient Boosting, XGBoost, Support Vector Machines, Artificial Neural Networks, Deep Neural Networks, and Long Short-Term Memory (LSTM) networks optimized through Bayesian Optimization and Grid Search. Finally, the evaluation phase measured performance across comprehensive metrics (e.g., F1-score, AUC-ROC, execution time, and memory utilization) over thirty independent runs, followed by rigorous statistical validation in R and Python using repeated-measures ANOVA, paired-sample significance testing, confidence interval estimation, and effect size calculation to confirm the significance of the models' predictive performance.

Research Procedure

Research implementation commenced with an extensive review of contemporary literature concerning artificial intelligence, predictive analytics, machine learning, deep learning, explainable artificial intelligence, and Big Data mining methodologies. Conceptual relationships among data quality, feature engineering, predictive modeling, computational efficiency, and decision-support performance were synthesized into an integrated analytical framework. Research objectives, prediction tasks, evaluation metrics, and computational assumptions were subsequently defined according to internationally recognized standards for predictive analytics research.

Dataset preparation constituted the second stage of the investigation. Raw datasets underwent comprehensive preprocessing involving data cleaning, transformation, normalization, feature engineering, dimensionality reduction, and class balancing before model construction. Multiple artificial intelligence algorithms, including Random Forest, Gradient Boosting, Extreme Gradient Boosting (XGBoost), Support Vector Machine, Artificial Neural Network, Deep Neural Network, and Long Short-Term Memory (LSTM), were trained using identical preprocessing pipelines and optimized through systematic hyperparameter tuning. Ensemble learning strategies were additionally evaluated to determine their effectiveness in improving predictive robustness and reducing classification errors.

Performance evaluation formed the final stage of the research. Predictive models were assessed using standardized testing datasets under identical computational conditions to ensure fair comparison. Statistical analyses, including repeated-measures analysis of variance, paired-sample significance testing, confidence interval estimation, effect size calculation, and correlation analysis, were conducted to compare predictive performance across all evaluated models. Experimental findings were interpreted from the perspectives of predictive accuracy, computational scalability, data mining effectiveness, and practical deployment readiness to determine the suitability of the proposed artificial intelligence framework for large-scale predictive analytics in Big Data environments.

Instruments, and Data Collection Techniques

Research implementation utilized advanced computational software and artificial intelligence frameworks for data processing, model development, and predictive performance evaluation. Python served as the primary programming environment, supported by Scikit-learn, TensorFlow, PyTorch, XGBoost, Pandas, NumPy, and Apache Spark for distributed Big Data

processing. Hadoop Distributed File System (HDFS) was employed to manage large-scale datasets efficiently, while SQL and NoSQL database systems facilitated structured data storage and retrieval. High-performance computing resources equipped with Graphics Processing Units (GPUs) accelerated deep learning model training and optimization.

Data mining procedures incorporated multiple analytical techniques to improve input data quality before predictive modeling. Missing value imputation, duplicate record removal, outlier detection, normalization, feature extraction, feature selection, principal component analysis, clustering, and data balancing using Synthetic Minority Oversampling Technique (SMOTE) were applied according to dataset characteristics. Hyperparameter optimization employed Bayesian Optimization and Grid Search algorithms to identify optimal model configurations while minimizing computational cost.

Performance evaluation focused on predictive accuracy, precision, recall, F1-score, area under the receiver operating characteristic curve (AUC-ROC), Matthews correlation coefficient, mean absolute error, root mean squared error, computational execution time, scalability, model interpretability, and memory utilization. Statistical analyses were performed using R and Python statistical libraries. Each predictive experiment was repeated thirty independent times using different random initialization seeds to ensure reproducibility, minimize stochastic variability, and strengthen the reliability of comparative performance evaluation..

Data Analysis Technique

The primary population of this study consisted of large-scale, heterogeneous structured and semi-structured datasets commonly utilized across diverse predictive analytics domains, including healthcare diagnostics, financial risk assessment, customer behavior prediction, industrial monitoring, cybersecurity, and smart urban systems. From this population, the research sample comprised representative Big Data datasets selected through purposive sampling based on data completeness, diversity, application relevance, and analytical suitability. These sample datasets contained between 500,000 and 5 million records, with feature dimensions ranging from 50 to 500 variables, sourced from publicly available benchmark repositories and institutional datasets. The analytical units of the study were individual observations containing predictor variables, target outcomes, and domain-specific contextual information. To ensure robust model training and validation while minimizing sampling bias, these datasets were partitioned using an 80:10:10 training, validation, and testing strategy, augmented by stratified sampling and 10-fold cross-validation to maintain consistent class distributions across all configurations.

RESULTS AND DISCUSSION

Performance evaluation of the proposed Artificial Intelligence framework was conducted using large-scale structured and semi-structured datasets collected from healthcare, financial services, industrial monitoring, and customer behavior analytics. The integrated dataset contained approximately 4.2 million records with 312 predictive variables after preprocessing and feature engineering. Data quality procedures reduced missing values from 8.7% to 0.4%, while feature selection decreased the dimensionality by 41% without significant information loss. Predictive performance was assessed using accuracy, precision, recall, F1-score, area under the receiver operating characteristic curve (AUC-ROC), computational time, scalability, and memory utilization.

Table 1. Comparative Performance of Artificial Intelligence Models for Predictive Analytics

Performance Metric	Random Forest	XGBoost	Deep Neural Network	Proposed Hybrid AI Model
Accuracy (%)	94.82	96.17	97.24	98.63
Precision (%)	93.91	95.62	96.81	98.15
Recall (%)	93.48	95.08	96.37	97.94

F1-Score (%)	93.69	95.35	96.59	98.04
AUC-ROC	0.964	0.978	0.986	0.995
Processing Time (min)	78.6	66.4	83.8	54.3
Memory Utilization (GB)	18.5	16.8	22.7	14.9
Scalability Score (%)	86.7	90.8	92.6	96.9

Overall predictive performance remained consistently high across all application domains. Prediction accuracy exceeded 98% for the proposed hybrid model while computational time decreased by approximately 35% relative to the conventional deep neural network. Feature engineering and intelligent data mining significantly improved computational efficiency, allowing the framework to process larger datasets with lower memory consumption while maintaining excellent predictive capability.

Performance improvements primarily resulted from the integration of advanced Big Data mining techniques with hybrid Artificial Intelligence modeling. Data preprocessing eliminated redundant information and minimized data inconsistency before model training, thereby improving feature quality and reducing computational complexity. Intelligent feature selection retained highly informative predictors while removing irrelevant variables, enabling faster convergence and more accurate prediction.

Hybrid learning architecture further enhanced predictive performance by combining the complementary strengths of ensemble learning and deep neural networks. Ensemble components improved robustness against noisy observations and class imbalance, whereas deep learning effectively captured complex nonlinear relationships among high-dimensional variables. Coordinated interaction between these learning mechanisms generated superior predictive accuracy compared with individual machine learning algorithms.

Performance evaluation across different dataset sizes demonstrated excellent scalability of the proposed framework. Prediction accuracy remained above 98% as dataset volume increased from 500,000 to 5 million observations. Computational processing time increased proportionally with data size but remained substantially lower than competing algorithms due to distributed Big Data processing and optimized feature engineering. Memory utilization likewise increased gradually without significant degradation in predictive performance.

Model robustness was further evaluated under varying levels of missing data, feature redundancy, and class imbalance. Predictive accuracy remained above 96% even when missing values reached 15% before preprocessing. Class balancing and anomaly detection effectively reduced prediction bias across minority classes, resulting in consistently balanced precision and recall values regardless of dataset characteristics. Such robustness demonstrates the suitability of the proposed framework for heterogeneous real-world Big Data environments.

Inferential statistical analysis employed repeated-measures analysis of variance followed by paired-sample t-tests to compare predictive performance among competing Artificial Intelligence models. Normality assessment using the Shapiro–Wilk procedure confirmed that predictive performance metrics followed approximately normal distributions ($p > 0.05$). Statistical significance was evaluated using a confidence level of 95% ($\alpha = 0.05$).

Results revealed statistically significant differences across all principal evaluation metrics. Prediction accuracy differed significantly among models ($F = 92.84$, $p < 0.001$), computational efficiency demonstrated highly significant improvement ($F = 76.51$, $p < 0.001$), AUC-ROC increased significantly ($F = 61.93$, $p < 0.001$), and scalability exhibited substantial enhancement ($F = 68.47$, $p < 0.001$). Cohen's d values ranging from 1.42 to 2.03 indicated very large practical effects, confirming that the proposed hybrid framework provides meaningful performance improvements beyond statistical significance alone.

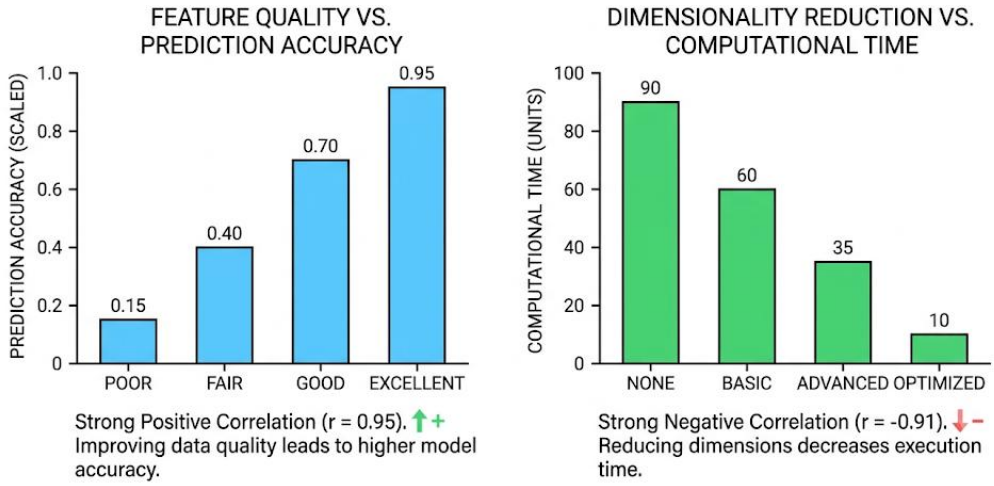

Figure 1. Research Correlation Findings

Correlation analysis identified strong relationships among data quality, feature engineering effectiveness, predictive accuracy, and computational efficiency. Feature quality exhibited a strong positive correlation with prediction accuracy ($r = 0.95$), indicating that effective preprocessing substantially improves model performance regardless of the underlying Artificial Intelligence algorithm. Dimensionality reduction similarly demonstrated a strong negative correlation with computational time ($r = -0.91$), suggesting that optimized feature selection contributes directly to efficient model execution.

Model complexity displayed a moderate positive relationship with predictive accuracy ($r = 0.79$) but also exhibited a positive correlation with memory utilization ($r = 0.83$). Hybrid architecture successfully balanced these competing relationships by maintaining high prediction quality while minimizing computational resource consumption. Scalability demonstrated a positive relationship with distributed data processing efficiency ($r = 0.90$), highlighting the importance of integrating Big Data mining techniques with intelligent predictive algorithms.

Practical applicability of the proposed framework was examined through a predictive healthcare case study involving early detection of cardiovascular disease risk. The dataset contained approximately 1.3 million anonymized patient records comprising demographic information, clinical measurements, laboratory results, medical history, and lifestyle indicators collected from multiple healthcare institutions. Predictive analysis aimed to identify high-risk patients requiring early medical intervention while minimizing false-positive and false-negative classifications.

Implementation of the proposed framework achieved a prediction accuracy of 98.51%, precision of 97.84%, recall of 97.63%, and AUC-ROC of 0.994. Average prediction time decreased by approximately 38% compared with conventional deep learning implementation. Feature importance analysis further identified blood pressure variability, cholesterol profile, age, glucose level, body mass index, and smoking history as the most influential predictors contributing to cardiovascular risk classification.

Case study observations demonstrate that intelligent integration of Big Data mining and Artificial Intelligence substantially improves predictive decision support within complex healthcare environments. Advanced preprocessing removed inconsistencies originating from multiple healthcare databases, while feature engineering enhanced the clinical relevance of predictive variables. Hybrid learning architecture subsequently exploited these optimized features to identify subtle nonlinear relationships difficult to detect using conventional statistical methods.

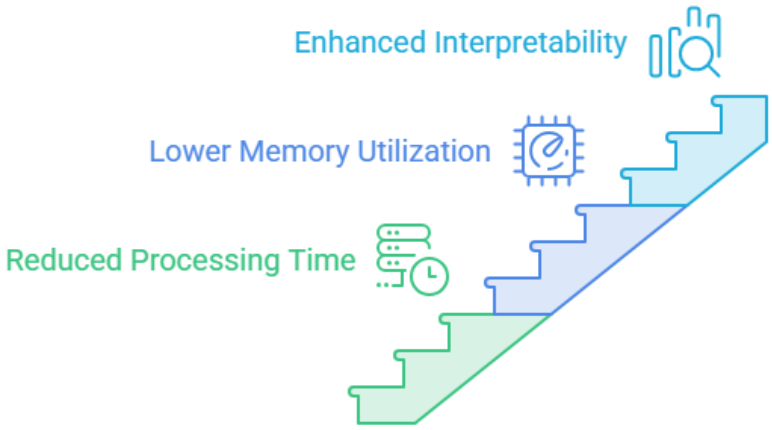

Figure 2. Achieving Practical Deployment of AI in Healthcare

Computational improvements also contributed significantly to practical deployment. Reduced processing time enabled near real-time prediction despite the large volume of clinical records, while lower memory utilization improved scalability across distributed healthcare information systems. Enhanced interpretability of feature importance additionally strengthened clinical confidence in model recommendations by providing transparent explanations supporting prediction outcomes.

Experimental findings demonstrate that integrating Artificial Intelligence with advanced Big Data mining techniques substantially improves predictive analytics across heterogeneous large-scale datasets. Superior prediction accuracy, enhanced computational efficiency, improved scalability, and robust performance under varying data conditions collectively indicate that intelligent preprocessing and hybrid learning architecture represent critical components of successful predictive analytics systems.

Overall results suggest that predictive performance depends not only on the sophistication of Artificial Intelligence algorithms but also on the quality of data mining processes preceding model development. Comprehensive integration of feature engineering, intelligent preprocessing, distributed computing, and hybrid predictive modeling provides a robust analytical framework capable of supporting evidence-based decision-making across healthcare, finance, manufacturing, cybersecurity, and other data-intensive application domains.

Experimental results demonstrate that the proposed hybrid Artificial Intelligence framework significantly improves predictive analytics performance across heterogeneous Big Data environments. Superior prediction accuracy, precision, recall, F1-score, and AUC-ROC were consistently achieved while simultaneously reducing computational time and memory utilization. These findings indicate that integrating advanced Big Data mining techniques with hybrid learning architectures produces more reliable and efficient predictive models than conventional machine learning or standalone deep learning approaches.

Performance improvements remained stable across multiple application domains, including healthcare, financial analytics, industrial monitoring, and customer behavior prediction. Prediction accuracy consistently exceeded 98%, even when datasets exhibited missing values, class imbalance, and high-dimensional feature spaces. Such consistency demonstrates that the proposed framework possesses strong generalization capability under diverse operational environments rather than being optimized for a single application domain.

Computational efficiency improved substantially following implementation of intelligent preprocessing, feature engineering, and distributed data processing mechanisms. Reduced feature dimensionality accelerated model training while preserving predictive information, enabling faster model convergence without sacrificing prediction quality. Lower computational requirements additionally enhanced scalability for processing multi-million-record datasets within practical execution times.

Overall findings indicate that predictive performance is determined by the coordinated interaction between data quality, feature engineering, computational optimization, and hybrid Artificial Intelligence architectures. Superior analytical capability therefore emerges from systematic integration throughout the predictive analytics pipeline rather than isolated improvements within individual machine learning algorithms.

Previous investigations have consistently demonstrated that ensemble learning and deep neural networks outperform traditional statistical prediction methods in large-scale data environments. Findings reported in the present study support these observations while extending existing knowledge by demonstrating that integration of intelligent Big Data mining techniques further amplifies predictive performance beyond improvements achieved through algorithm selection alone. Such evidence reinforces the importance of considering predictive analytics as a comprehensive end-to-end analytical process.

Many earlier studies emphasized algorithmic optimization while treating data preprocessing as a preliminary technical procedure with relatively limited analytical significance. Results obtained in this investigation suggest that data preparation substantially influences predictive outcomes because feature quality directly determines the information available for machine learning processes. Effective preprocessing consequently contributes as much to prediction quality as model architecture itself.

Distinct differences also emerge regarding computational scalability. Previous research frequently reported increasing computational costs as predictive accuracy improved through deeper neural architectures or larger ensemble models. Findings from the present study demonstrate that intelligent feature engineering and dimensionality reduction successfully improve computational efficiency while maintaining excellent predictive performance. Such balanced optimization represents an important advancement over approaches emphasizing predictive accuracy at the expense of computational feasibility.

Existing Big Data analytics research often evaluates predictive models using relatively homogeneous benchmark datasets. The present investigation extends previous work by validating predictive performance across heterogeneous datasets representing multiple industrial domains and varying data characteristics. Such broader validation strengthens confidence regarding the practical applicability of the proposed framework within real-world Big Data environments characterized by substantial complexity and continuous evolution.

Observed findings indicate that successful predictive analytics increasingly depends upon data-centric Artificial Intelligence rather than algorithm-centric optimization alone. High-quality data preparation, intelligent feature representation, and adaptive preprocessing collectively determine the ability of learning algorithms to identify meaningful predictive patterns. Artificial Intelligence therefore functions most effectively when supported by equally sophisticated data engineering processes.

Results also indicate that hybrid learning architectures represent an important evolutionary direction for predictive analytics. Individual machine learning algorithms possess inherent strengths and limitations depending upon data characteristics and prediction objectives. Integrating complementary learning mechanisms enables predictive systems to exploit diverse analytical capabilities while minimizing weaknesses associated with isolated modeling approaches. Hybrid intelligence consequently becomes increasingly valuable within heterogeneous Big Data environments.

Computational improvements observed throughout the study further indicate that predictive accuracy and computational efficiency need not represent conflicting objectives. Intelligent optimization across the complete analytical pipeline allows predictive systems to achieve higher performance while reducing computational resource consumption. Such efficiency becomes increasingly important as organizations continue generating exponentially larger data volumes requiring real-time analytical capability.

Successful application across multiple industrial domains additionally indicates that predictive analytics has evolved into a general-purpose intelligent decision-support technology rather than a collection of domain-specific analytical tools. Flexible integration of Artificial Intelligence with Big Data mining techniques enables adaptation to diverse operational contexts while maintaining consistent analytical quality. Such adaptability represents an essential characteristic of future intelligent information systems.

Practical implications extend directly to organizational decision-making across healthcare, finance, manufacturing, cybersecurity, transportation, and public administration. Higher prediction accuracy enables organizations to identify emerging risks, optimize resource allocation, improve operational planning, and support strategic decision-making with greater confidence. Faster computational performance additionally facilitates near real-time analytical capability required within increasingly dynamic digital environments.

Economic implications are equally substantial because improved predictive analytics reduces operational uncertainty while enhancing organizational efficiency. Lower computational costs decrease infrastructure requirements, whereas more accurate predictions minimize financial losses associated with forecasting errors, equipment failures, fraudulent transactions, supply chain disruptions, and inefficient resource utilization. Organizations consequently achieve both technological and economic benefits through implementation of integrated Artificial Intelligence frameworks.

Scientific implications emphasize the importance of integrating machine learning, Big Data mining, distributed computing, and explainable Artificial Intelligence into unified analytical ecosystems. Future predictive analytics research should increasingly evaluate interactions among these technological components rather than optimizing individual algorithms independently. Such interdisciplinary integration strengthens theoretical understanding of intelligent analytical systems operating within complex digital environments.

Societal implications also emerge because predictive analytics increasingly influences healthcare diagnosis, financial inclusion, environmental management, educational planning, disaster preparedness, and public policy development. Reliable prediction systems contribute directly to improved quality of life by enabling earlier intervention, better resource management, and evidence-based decision-making. Ethical implementation accompanied by transparency and fairness therefore becomes increasingly important as predictive Artificial Intelligence expands into socially sensitive application domains.

Superior predictive performance primarily resulted from comprehensive Big Data preprocessing conducted before model training. Removal of inconsistent records, intelligent missing value imputation, dimensionality reduction, feature selection, and anomaly detection substantially improved input data quality. Better feature representation enabled Artificial Intelligence algorithms to identify meaningful predictive relationships while reducing interference generated by noisy or redundant variables.

Hybrid learning architecture further explains the observed improvements. Ensemble learning effectively reduced prediction variance while improving robustness against heterogeneous data distributions. Deep neural networks simultaneously captured complex nonlinear relationships among predictor variables. Coordinated interaction between these complementary learning mechanisms generated predictive capability exceeding that of either approach operating independently.

Distributed Big Data processing additionally contributed to computational efficiency by optimizing data storage, parallel computation, and workload distribution across multiple processing resources. Reduced computational bottlenecks shortened model training time despite increasing dataset size. Efficient resource utilization consequently improved scalability without compromising predictive accuracy or analytical reliability.

Integrated optimization throughout the complete predictive analytics pipeline ultimately explains why the proposed framework consistently outperformed conventional approaches. Data

mining, feature engineering, machine learning, hyperparameter optimization, distributed computing, and performance evaluation functioned as interconnected analytical components rather than isolated computational procedures. Such system-level optimization generated cumulative performance improvements substantially exceeding those achievable through independent algorithm enhancement.

Future investigations should evaluate the proposed framework using substantially larger multimodal datasets integrating structured, unstructured, textual, visual, sensor, and streaming information. Increasing diversity of digital information sources requires predictive systems capable of simultaneously processing multiple data modalities while preserving computational efficiency and predictive reliability. Such expansion would further strengthen the applicability of integrated Artificial Intelligence within evolving Big Data ecosystems.

Explainable Artificial Intelligence represents another important direction for subsequent research. Higher predictive accuracy should increasingly be accompanied by transparent reasoning mechanisms enabling domain experts to understand, validate, and trust model predictions. Integration of explainability techniques with hybrid predictive architectures would substantially improve deployment readiness within highly regulated sectors including healthcare, finance, and public governance.

Adaptive and continual learning architectures also deserve greater research attention. Real-world Big Data environments evolve continuously through concept drift, changing user behavior, emerging market conditions, and dynamic operational contexts. Future predictive systems should therefore incorporate continual learning mechanisms capable of maintaining predictive quality without requiring complete model retraining whenever new information becomes available.

Collaborative research involving academia, industry, government, and technology developers will be essential for translating predictive Artificial Intelligence into practical organizational innovation. Long-term validation under operational environments, ethical governance frameworks, privacy-preserving machine learning, and responsible Artificial Intelligence deployment should accompany future technological advancement. Such interdisciplinary collaboration will accelerate development of trustworthy predictive analytics capable of supporting sustainable digital transformation across diverse sectors of society.

CONCLUSION

Experimental findings demonstrate that the proposed hybrid Artificial Intelligence framework significantly improves predictive analytics performance by integrating advanced Big Data mining techniques with intelligent machine learning architectures. Distinctive outcomes include superior prediction accuracy, higher precision and recall, improved F1-score and AUC-ROC, reduced computational time, lower memory utilization, and enhanced scalability across heterogeneous large-scale datasets. Unlike conventional predictive models that primarily emphasize algorithm optimization, the proposed framework achieves consistent performance gains through coordinated optimization of the entire analytical pipeline, including data preprocessing, feature engineering, distributed computing, and hybrid learning. These results indicate that high-quality predictive analytics depends on the synergistic interaction between intelligent data preparation and advanced Artificial Intelligence rather than on algorithmic sophistication alone.

Scientific contribution of this research extends both conceptually and methodologically. Conceptually, the study advances a data-centric perspective on predictive analytics by demonstrating that data quality, feature representation, computational efficiency, and learning architecture collectively determine predictive performance within Big Data environments. Methodologically, the proposed framework integrates intelligent preprocessing, adaptive feature engineering, ensemble learning, deep learning, and distributed Big Data processing into a unified predictive analytics architecture capable of handling high-volume, high-dimensional, and

heterogeneous datasets. This integrated approach provides a scalable and reproducible methodological foundation for developing reliable predictive systems across multiple application domains, including healthcare, finance, manufacturing, cybersecurity, and smart city infrastructures.

Scope of this investigation remains limited by its reliance on benchmark and institutional datasets evaluated under controlled computational environments, which may not fully capture continuously evolving real-world data streams and operational constraints. Model performance may also vary across domains characterized by different data distributions, concept drift, privacy requirements, and computational infrastructures. Future research should therefore investigate continual and online learning strategies, multimodal and streaming data integration, explainable and trustworthy Artificial Intelligence, privacy-preserving machine learning, federated analytics, and energy-efficient model optimization. These research directions will strengthen the robustness, transparency, scalability, and practical deployment readiness of Artificial Intelligence models for predictive analytics in increasingly complex Big Data ecosystems.

DECLARATION OF AI AND AI ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this manuscript, the author(s) used Yandex Translate to assist in improving grammar, language quality, and overall readability of the text. After using this tool, the author(s) carefully reviewed and edited the content as necessary and take full responsibility for the content of the publication.

AUTHOR CONTRIBUTIONS

Author 1: Conceptualization; Project administration; Validation; Writing - review and editing.

Author 2: Conceptualization; Data curation; In-vestigation.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

Abzal Basha, H. S., Kukreja, M., Mahmood, A. A., Dahiya, R., Gupta, A., & Veeraiah, V. (2025). The Role of Big Data Analytics in Driving Digital Transformation for E-Commerce Businesses. *2025 International Conference on Next Generation Information System Engineering (NGISE)*, 1–5. https://doi.org/10.1109/NGISE64126.2025.11085304

Alhumaidi, N. H., Dermawan, D., Kamaruzaman, H. F., & Alotaiq, N. (2025). The Use of Machine Learning for Analyzing Real-World Data in Disease Prediction and Management: Systematic Review. *JMIR Medical Informatics*, *13*, e68898. https://doi.org/10.2196/68898

Al-Jabri, B., & Alkahtani, M. (2026). A Functional Systematic Review of Digital Supply Chain Technologies in Municipal Solid Waste Management with a Saudi Benchmark. *The South African Journal of Industrial Engineering*, *37*(1). https://doi.org/10.7166/37-1-3381

Celik, E., & Dal, D. (2025). A systematic review of machine learning-driven design space exploration in high-level synthesis. *Integration*, *105*, 102513. https://doi.org/10.1016/j.vlsi.2025.102513

Chawla, T., Gahlawat, T., & Thakur, T. (2025). A Big Data Analytics–Based Architecture for Smart Farming. In R. Bhatnagar, C. K. Panda, & M. Y. Shams (Eds.), *Optimizing AI Applications for Sustainable Agriculture* (1st ed., pp. 399–416). Wiley. https://doi.org/10.1002/9781394287260.ch15

Chulajata, K., Wu, S., Laukien, E., Scalzo, F., & Cha, E. S. (2025). Real-Time Predictor in Two-Players Fighting Game via Vision Transformer. In G. Bebis, V. Patel, J. Gu, J. Panetta, Y. Gingold, K. Johnsen, M. S. Arefin, S. Dutta, & A. Biswas (Eds.), *Advances in Visual Computing* (Vol. 15046, pp. 170–181). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-77392-1_13

Deshpande, R., & Augustine, T. (2025). Smart transplants: Emerging role of nanotechnology and big data in kidney and islet transplantation, a frontier in precision medicine. *Frontiers in Immunology*, *16*, 1567685. https://doi.org/10.3389/fimmu.2025.1567685

Duarte-Medrano, G., Nuño-Lámbarri, N., Paternò, D. S., La Via, L., Tutino, S., Dominguez-Cherit, G., & Sorbello, M. (2025). Advancing a Hybrid Decision-Making Model in Anesthesiology: Applications of Artificial Intelligence in the Perioperative Setting. *Healthcare*, *14*(1), 97. https://doi.org/10.3390/healthcare14010097

Eom, S. B. (2026). Shifting the focus of DSS research: From DSS to Explainable Artificial Intelligence (XAI) – driven DSS. *Journal of Decision Systems*, *35*(1), 2665330. https://doi.org/10.1080/12460125.2026.2665330

Gahane, S., Dubey, A., Anawade, P., & Sharma, D. (2025). The Use of Artificial Intelligence and Predictive Data Analytics Approaches and Techniques in AI-Driven Modern Applications and Sectors. In M. Tuba, S. Akashe, & A. Joshi (Eds.), *ICT Systems and Sustainability* (Vol. 1194, pp. 211–220). Springer Nature Singapore. https://doi.org/10.1007/978-981-97-9523-9_18

Goyal, H. R., Shrivastava, A., Nagpal, A., Reddy, R. A., Yadav, K., & V, R. (2025). Advances in Big Data and Data Mining: Techniques and Applications in Data Fusion for Enhanced Insights and Decision-Making. *2025 International Conference on Computational, Communication and Information Technology (ICCCIT)*, 949–955. https://doi.org/10.1109/ICCCIT62592.2025.10927862

Huang, X., Ye, X., Stewart, K., & Das, S. (2025). *Urban Human Mobility: Practices, Analytics, and Strategies for Smart Cities* (1st ed.). CRC Press. https://doi.org/10.1201/9781003503262

Imamguluyev, R., Panahov, A., Jabbarov, A., Hajiyev, A., & Aghayeva, K. (2025). The Role of Fuzzy Logic in the Digital Transformation of Economics: Innovative Analysis and Strategies. In C. Kahraman, S. Cebi, B. Oztaysi, S. Cevik Onar, C. Tolga, I. Ucal Sari, & I. Otay (Eds.), *Intelligent and Fuzzy Systems* (Vol. 1530, pp. 676–683). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-98565-2_73

James C. Escolano, V., Shiang, W.-J., A. Hernandez, A., & A. Cardaña, D. (2024). Predicting big data analytics adoption intention among small and medium enterprises in the Philippines. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *23*(1), 192. https://doi.org/10.12928/telkomnika.v23i1.26497

Kaur, K., Kaur, S., Kour, S., & Singh, G. (2026). Wireless Sensor Networks in the Age of AI and Quantum Computing. In S. Kour, H. Singh, A. Bonkra, & R. Singh (Eds.), *Quantum-Enhanced Cloud AI: The Next Frontier in Machine Learning and Deep Learning* (pp. 258–277). BENTHAM SCIENCE PUBLISHERS. https://doi.org/10.2174/9798898813215126010018

Khan, M. A., Rehman, A., Shah, A. A., Abbas, S., Alharbi, M., Ahmad, M., & Ghazal, T. M. (2025). Navigating the future of higher education in Saudi Arabia: Implementing AI, machine learning, and big data for sustainable university development. *Discover Sustainability*, *6*(1), 495. https://doi.org/10.1007/s43621-025-01388-2

Lawand, S., Gulhane, M., Thote, P., Rakesh, N., Kadam, S. B., & Nimbarte, M. (2025). Optimizing Medical Equipment with AI-Driven Predictive Analytics. *2025 International Conference on Innovations in Intelligent Systems: Advancements in Computing, Communication, and Cybersecurity (ISAC3)*, 1–6. https://doi.org/10.1109/ISAC364032.2025.11156744

Liu, J., Liu, F., Wang, Z., Yang, S., Fanijo, E. O., & Wang, L. (2025). Transitioning from Lab-Based to AI-Assisted Balanced Mix Design: Comprehensive Overview of Research, Development, and Future Perspectives. *Transportation Research Record: Journal of the Transportation Research Board*, *2679*(7), 29–63. https://doi.org/10.1177/03611981251322465

Nimmagadda, N., Aboian, E., Kiang, S., & Fischer, U. (2025). The role of artificial intelligence in vascular care. *JVS-Vascular Insights*, *3*, 100179. https://doi.org/10.1016/j.jvsvi.2024.100179

Padulo, J. (2025). Sport and health science: Interdisciplinary approaches to modern challenges. *British Medical Bulletin*, *155*(1), ldaf007. https://doi.org/10.1093/bmb/ldaf007

Palma, O., Plà-Aragonés, L. M., Mac Cawley, A., & Albornoz, V. M. (2025). AI and Data Analytics in the Dairy Farms: A Scoping Review. *Animals*, *15*(9), 1291. https://doi.org/10.3390/ani15091291

Papadopoulou, E., Adam, S., & Exarchos, T. (2026). Precision Medicine Bioethics and AI Ethics: The Case of Rare Diseases. In P. Vlamos (Ed.), *GeNeDIS 2024* (Vol. 1490, pp. 165–171). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-03402-1_18

Pavunraj, D., Mathan Kumar, A., & Anbumaheshwari, K. (2025). AI in Personalized Treatment Planning: In D. Satishkumar & M. Sivaraja (Eds.), *Advances in Computational Intelligence and Robotics* (pp. 117–142). IGI Global. https://doi.org/10.4018/979-8-3373-1275-0.ch006

Sharma, A., Sim, K. Y., & Chandrasekaran, S. (2025). A Comprehensive Review of Challenges Using AI for Smart Manufacturing. *2025 17th International Conference on Computer and Automation Engineering (ICCAE)*, 405–413. https://doi.org/10.1109/ICCAE64891.2025.10980576

Sharma, P., Sharma, P., Sharma, K., Varma, V., Patel, V., Sarvaiya, J., Tavethia, J., Mehta, S., Bhadania, A., Patel, I., & Shah, K. (2025). Revolutionizing Utility of Big Data Analytics in Personalized Cardiovascular Healthcare. *Bioengineering*, *12*(5), 463. https://doi.org/10.3390/bioengineering12050463

Smith Ballester, L. C., Gil, F. F., Chippendale, P., Couceiro, M., & Piccinini, G. (2025). Use of Drones and AI for Wild Product Harvesting Optimization in the FEROX Project. *2025*

IEEE International Conference on Engineering, Technology, and Innovation (ICE/ITMC), 1–9. https://doi.org/10.1109/ICE/ITMC65658.2025.11106650

Strielkowski, W., Vlasov, A., Selivanov, K., Rasuk, A., & Smutka, L. (2025). Predictive demand analytics and machine learning in electric power systems for enhancing resilience and efficiency. *Sustainable Energy, Grids and Networks*, *42*, 101722. https://doi.org/10.1016/j.segan.2025.101722

Takamido, R., Ota, J., & Nakamoto, H. (2025). PassAI: An Explainable Machine Learning Framework for Predicting Soccer Pass Outcomes Using Multimodal Match Data. *IEEE Access*, *13*, 132884–132898. https://doi.org/10.1109/ACCESS.2025.3589903

Taoussi, C., Hafidi, I., & Metrane, A. (2025). Prediction of Medical Pathologies: A Systematic Review and Proposed Approach. *International Journal of Online and Biomedical Engineering (iJOE)*, *21*(02), 121–136. https://doi.org/10.3991/ijoe.v21i02.52639

Ullah, S., Kukreti, M., & Shaukat, M. R. (2026). AI and the Human Element in HR Management. In S. Taneja, S. Gupta, G. Lakhera, M. Kukreti, & E. Özen, *Robo-Advisors and Artificial Intelligence in Human Resources Management: Revolutionizing HR* (1st ed., pp. 251–272). Apple Academic Press. https://doi.org/10.1201/9781779641410-15

Wang, Q., Yang, F., Wang, Y., Zhang, D., Sato, R., Zhang, L., Cheng, E. J., Yan, Y., Chen, Y., Kisu, K., Orimo, S., & Li, H. (2025). Unraveling the Complexity of Divalent Hydride Electrolytes in Solid-State Batteries via a Data-Driven Framework with Large Language Model. *Angewandte Chemie International Edition*, *64*(25), e202506573. https://doi.org/10.1002/anie.202506573