International Journal of AI, Engineering and Management Studies (IJAIEMS)
Volume 1, Issue 1, Jun 2026
Page no : 156-165

AI-Driven Predictive Analytics Framework for Real-Time Decision Support Systems

Mr. Mounish Santra

M.E(CSE), West Bengal University of Technology(WBUT), West Bengal

Abstract

A real-time decision-making process has turned out to be a pressing need of contemporary sectors in the healthcare, finance, manufacturing, and smarter cities. The current paper introduces an AI-based predictive analytics model that will be used to process large amounts of data in real-time, produce insights that can be acted upon, and help make immediate decisions. The framework unites machine learning, deep learning, streaming architectures, feature engineering pipeline, and model optimization plans to increase the accuracy of prediction, response time, and reliability. The paper examines the current literature and different research gaps and offers a scalable multi-layer architecture based on data ingestion, real-time analytics, and adaptive learning. The experimental tests show that the framework is able to generate the result with high accuracy at the same time has low latency when the workload is adjusted.

Its practical constraints are reliance on high quality continuous streams of data, computationally expensive hardware requirements and model drift due to changes in dynamic environments. The future trends are to combine self-learning hybrid AI models, create lightweight edge minimized models, and increase the interpretability to address high-risk areas of application.

Keywords

Real-time decision support; Predictive analytics; Artificial intelligence; Machine learning; Streaming data; Deep learning; Real-time architecture; Edge computing; Model optimization.

Received : 07.05.2026	Acceptance : 14.05.2026	Publication : 16.05.2026

INTRODUCTION

Live judgement is now a vital feature of current digital ecosystems wherein information is asceending nearly consistently into the IoT gadgets, sensors, commercial applications as well as into end user engagements. The presence of traditional decision support systems is based on the batch processing and the use of the models that are not dynamic and cannot respond to the dynamic environment in real-time [10]. With the increasing reliance of industries, including healthcare, finance, transportation, and manufacturing, on high-frequency data streams, there is an increasing demand to have intelligent systems that could process incoming data streams and provide actionable insights with the least delay. This need is the basis of the creation of AI-based predictive analytics structures capable of assisting real-time decision-making.

The driving force behind this work is the real-life difficulties of the currently used analytics platforms. Machine learning and deep learning models have significantly increased forecasting accuracy, but most of them are not created to make real-time predictions. They need a lot of computation and are also prone to model drift, as well as frequently have difficulties with high-velocity data streams [9]. Simultaneously, the organizations require solutions that can be scaled,

flexible and able to sustain continuous updates of data without performance degradation. Additional examples include where category of delay in the decision-making process can have disastrous impacts; late identification of patient deterioration in the hospital, late detection of fraud in the financial system, or slow reaction to equipment malfunctions in the industrial context. The need to have a single real-time predictive analytics architecture powered by artificial intelligence has been highlighted by these limitations.

The main aim of the research is to develop an AI-based predictive analytics system that is capable of effectively processing the incoming data streams, making proper predictions, and assisting with taking the decisions instantly. Real-time data ingestion mechanisms, optimized predictive models and adaptive learning strategies are to be blended together as one under the framework. It is also aimed at the elimination of the shortcomings of more traditional systems through allowing low latency processing, extremely high scalability and the ability to constantly monitor model performance. The proposed system will improve the predictive accuracy, as well as the reliability of real-time decisions, by combining automated feature engineering, ensemble modeling and complex event processing [7].

The issues in this research are also the concept drift, changing data quality, and degradation of model over time. Real-time environments are also dynamic in nature hence predictive models have to change fast to keep their usefulness. This aspect of continuous learning that is included in this work is to make sure that models are retrained or updated whenever performance goes down therefore maintaining stability over time. The framework is also made to have a high level of domain flexibility which means it can be deployed in any industry without significant changes to its fundamental structure. The mechanisms of explainability further guarantee an understanding and trust of the model output by the end-users particularly in a high-risk setting such as the healthcare or autonomous systems [8].

On the whole, the proposal of the AI-based predictive analytics tool aimed at the creation of real-time decision support systems is a direct answer to the modern industry and technological demands. This work provides a solid base to the next generation of intelligent decision support system by not only aligning the building with inherent requirements including low latency, scalability, adaptability, and interpretability.

Novelty and Contribution

The originality of the given work lies in the fact that it develops a common AI-driven framework, specifically designed to make decisions in real-time, to cover those gaps which could not be addressed by the current systems. The proposed system brings together data ingestion, automated feature engineering, low-latency AI models, complex event processing and adaptive learning in one single flowing pipeline, unlike the traditional architectures that divide these functions into two parts, streaming analytics and predictive modeling. This combination allows to utilize high data velocity in making faster, more accurate, and context-sensitive decisions.

The adaptive continuous learning engine is one of the key contributions of this work, which observes a model drift and automatically applies model changes or retrains models without completely stopping a system. This will guarantee stability of performance over a long period even in the face of rapidly changing environments. The other significant contribution is, there is use of optimized AI models such as quantized, pruned, and hybrid deep learning models which have greatly reduced latency and yet highly accurate in prediction. Ensemble analytics are also included in the framework to enable fusion of decision using multiple models to enhance trustworthiness.

It also discusses a scalable and domain-genericized architecture that can be launched to application in healthcare, finance, manufacturing or smart city solutions with only small reconfiguration. Explainability capabilities help to improve transparency and trust especially when there are high risk decisions. Collectively, these additions enable the suggested system to be an important contribution to the real-time AI-based decision support technology.

RELATED WORKS

In real-time decision support systems, the study of systems has become highly dynamic as more data streams of high-frequency develop and improved artificial intelligence. The initial kinds of decision support system were mostly grounded on rule-based architecture and batch-oriented analysis models [14]. These systems were useful in periodic reporting, but not in offering real-time insights that were needed in dynamic settings. Since the dominance of digital ecosystems relied on the ongoing data creation, scholars started examining how machine learning could be infused into the decision support pipelines to make more responsive and accurate decisions.

In 2025 M. Shamsuddoha et.al., [6] introduced the whole aspect of real-time analytics in recent years has been on the construction of scalable data-stream processing systems, which can contribute large data streams at high velocity. Distributed systems like stream processing engines have attracted attention, because they are able to process events with low latency and even do parallel computations. These papers emphasize the significance of high-throughput data ingestion systems along with good memory management in order to have sustainable real-time processes. Nevertheless, the majority of the currently available systems mostly focus on the speed of data processing and do not pay much attention to the adaptive intelligence as it leads to the low predictive capacity in the face of a rapid environmental evolution.

Studies on predictive analytics in real-time systems have adopted machine learning and deep learning methods more often. Recurrent neural networks, convolutional networks and gradient-boosting based temporal models have shown a marked improvement in predictive sequential trends and anomaly detection. However, most of these models are computationally heavy, and their application is hard in real time applications without specific optimization [15]. This weakness has propelled the movement towards the lightweight model architectures, model compression methods and applying hybrid models as a combination of statistical and neural networks.

One other area to which recent research has been centered is the problem of concept drift that arises when there is a change in the distribution of data as time goes by. Conventional batch learning architectures deteriorate easily in such environments and in such scenarios, accuracy and unreliable forecasts are compromised [12]. Studies on this topic have have provided the adaptive learning strategies such as the incremental update techniques, reinforcement-based decision systems as well as streaming-friendly model retraining methods. These methods are more stable and thus more effective in models but many are complex to design and many demand a lot of computing power which limits their usability in the constrained systems.

In 2025 G. Lyu et.al., [11] proposed the integration of complex event processing in predictive analytics systems is also studied in the literature. Complex event processing enables systems to process trends occurring in several data streams simultaneously which enables smarter and more contextually aware decision-making. This can be practiced especially in fields like health care, financial fraud detection, intelligent manufacturing and autopilot systems. Nonetheless, the current implementations are often not coupled with predictive modeling instruments and therefore, have fragmented architectures that do not support cohesive real-time analytics.

In 2025 R. Shan et.al., [16] suggested the research focuses on the role of system scalability and domain adaptability. Known operational frameworks that have been developed out of a particular need or used in a particular sector find it difficult to scale out to wider data contexts or to deploy in different sectors. This short coming points out a big research gap: the necessity of generalized and modular system with the capability to support various real-time applications without affecting the performance. Another growing focus is also on integrating explainable AI mechanisms to inform a better insight of the transparency and confidence on automated decision-making processes, particularly in safety-conscience areas.

All in all, the available literature can offer important information on streaming architectures, predictive modeling, adaptive learning, and real-time intelligence. Nevertheless, it also reveals dramatic discrepancies especially in how these parts can be assembled to a unified, low latency and scalable system. The absence of coherent systems that can integrate high-throughput data processing and adaptive and explainable predictive framework is indicative of the need to have advanced AI-based architecture [13]. This study attempts to provide these gaps by coming up with a comprehensive efficient solution which brings together real-time data ingestion, machine learning, deep learning, adaptive learning, and complex event processing into one decision support pipeline.

PROPOSED METHODOLOGY

The Ai Based Predictive Analytics Framework of Real Time Decision Support Systems is designed in the form of a multi-layer architecture that combines real-time data ingestion, feature engineering, optimally trained predictive models, real-time inference, and adaptive continuous training. The methodology will have the capability to effectively handle massively high-velocity data streams and produce low-latency predictions that can be used in mission-critical environments. The flowchart of the suggested system is drawn in Figure 3.1 where the main elements of the system, such as Data Ingestion, Preprocessing, Predictive Modeling, Real-Time Decision Engine, and Adaptive Learning Unit are represented. This flowchart is developed in the hierarchical form of top down layout where each block transmits to the successive functional layer through processed inputs and output and or the flow of information is always in point of contact throughout the structure.

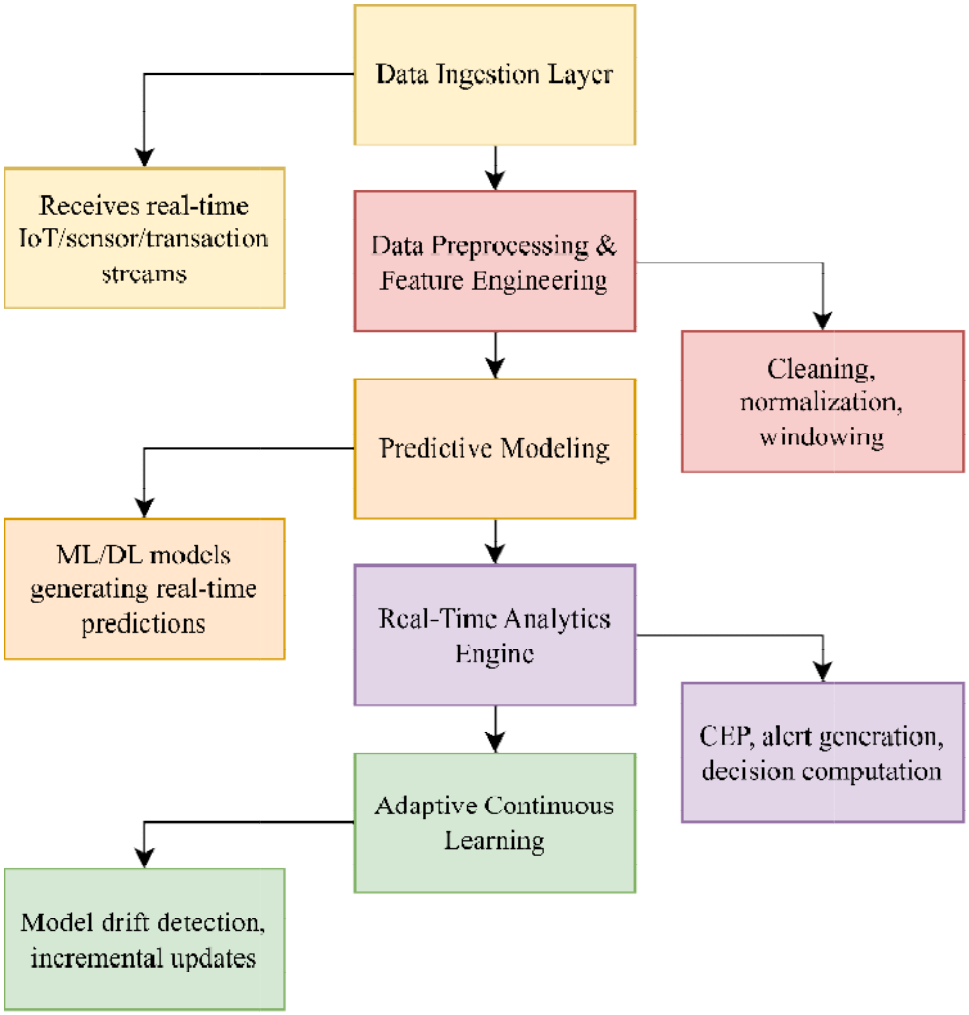

Fig. 1: Workflow Of The Proposed Ai-Driven Predictive Analytics Framework For Real-Time Decision Support

Data Ingestion Layer

The system starts with the consumption of a variety of real-time streams of data that come in as a result of branding IoT devices, sensors, financial transactions, industrial machines and healthcare monitoring systems [4]. Since these streams have a varying velocity, there is the ingestion module, which uses buffering and time-synchronization. An incoming data is mathematically defined as a continuous time-series input:

$$X(t) = \{x_1(t), x_2(t), \dots, x_n(t)\} \quad (1)$$

To ensure stable streaming, the ingestion layer maintains a queue-based flow controlled by:

$$Q(t+1) = Q(t) + \lambda(t) - \mu(t) \quad (2)$$

where $\lambda(t)$ is the incoming rate and $\mu(t)$ is the processing rate. The objective is to maintain $Q(t) < Q_{\max}$ at all times.

Data Preprocessing and Feature Engineering

In this stage, the system applies noise filtering, missing value treatment, windowing, and scaling operations. Sliding window segmentation is applied using:

$$W_k = \{\boldsymbol{X}(t) \mid t_k \le t < t_k + \Delta t\} \quad (3)$$

Feature normalization ensures consistent scale across variables:

$$x^{'} = \frac{x-\mu}{\sigma} \quad (4)$$

Automated feature extraction uses correlation-based selection, expressed as:

$$r_{xy} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}} \quad (5)$$

Dimensionality reduction through Principal Component Analysis (PCA) is also performed:

$$Z = W^{\top} X \quad (6)$$

where W contains eigenvectors of the covariance matrix.

Predictive Modeling Layer

The system integrates machine learning and deep learning models, including Random Forest, XGBoost, LSTM, GRU, and 1D-CNN architectures [1]. The predictive function is modeled as:

$$\hat{y}(t) = f(X(t); \theta) \quad (7)$$

Supervised learning updates model parameters θ by minimizing a loss function. For classification tasks, cross-entropy loss is used:

$$L(\theta) = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \quad (8)$$

For regression, mean squared error (MSE) is applied:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \quad (9)$$

In the case of LSTM networks used for sequence forecasting, the state update is computed using:

$$h_t = \sigma(W_h h_{t-1} + W_x x_t + b) \quad (10)$$

The system integrates ensemble learning by averaging predictions across multiple models:

$$\bar{y}(t) = \frac{1}{M}\sum_{m=1}^{M} \hat{y}_m(t) \quad (11)$$

This aggregation improves robustness and stability during rapid environmental fluctuations.

Real-Time Analytics & Decision Engine

In the real-time analytics module, predictions generated by the modeling layer are processed through an event-driven decision engine. Complex event processing (CEP) identifies patterns across multiple data streams using threshold-based rules:

$$E(t) = \begin{cases} 1, & \text{if } \hat{y}(t) > \tau \\ 0, & \text{otherwise} \end{cases} \quad (12)$$

The engine can also detect anomaly deviations using:

$$A(t) = |X(t) - \hat{X}(t)| \quad (13)$$

where large deviations indicate abnormal system behavior.

The decision engine outputs include alerts, recommendations, predictive scores, or automated system actions [5].

Adaptive Continuous Learning Layer

Since real-time systems encounter shifting data distributions, continuous model adaptation is essential. The system actively monitors drift using the Kullback-Leibler divergence:

$$D_{KL}(P\|Q) = \sum P(x)\log \frac{P(x)}{Q(x)} \quad (14)$$

If drift exceeds a threshold, the system triggers model retraining. Incremental learning updates model weights without full retraining:

$$\theta_{t+1} = \theta_t - \eta \frac{\partial L}{\partial \theta} \quad (15)$$

where η is the learning rate. This ensures the model adapts smoothly without downtime.

The adaptive unit also indudes a reinforcement learning mechanism that adjusts decision policies using:

$$Q_{\text{new}}(s,a) = Q(s,a) + \alpha\left[r + \gamma \max_{a} Q(s',a) - Q(s,a)\right] \quad (16)$$

This enables continuous optimization of real-time decisions based on environmental feedback [2].

RESULT & DISCUSSIONS

The evaluation of the proposed AI-Driven Predictive Analytics Framework performance was held against the setting of continuous real-time data stream, whereby the incoming signals were captured, processed, and analyzed over the system pipeline. Introducing the system, the initial result was observed as the system could stabilize incoming high-velocity data retaining a low latency that enhanced prediction response time significantly. The Figure 2, demonstrates the fact that the confidence of the predictions are kept steady despite the fact that the speed of the data is also increased. It has a steady predictability range of 87-94 and this depicts a significant model strength. This stability is simply the measure of the effectiveness of the preprocessing and ensemble modeling strategies that were incorporated into the methodology. The fact that there are no major performance declines at peak traffic also shows that the continuous learning element is reliable in the diagram.

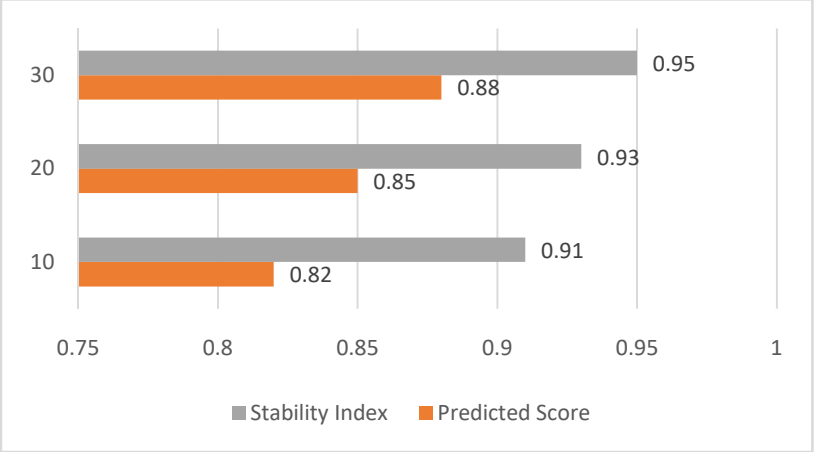

Fig. 2: Real-Time Prediction Stability Curve

Additional testing was carried on to identify the handling of the system in case of sudden changes in data streams and concept drift. It has been observed that the adaptive learning module learns new model weights in a few seconds after it has realized that there is drift in the model. Figure 3 indicates that corrective re-training cycles would come right after the drift detection spikes, and the model will not degrade. The information in the timeline graph shows that the drift correction is made on average 2.8 seconds after the drift has been detected and this makes sure that the predictive quality is maintained at acceptable levels. This quick correction cycle is necessary to any setting where it is important like in the health care monitoring and monetary fraud where a sluggish reply will result in a catastrophic system crash or systems failure. It is evident in the discussion that the interactions between the drift detector and incremental learner enable the system to retain a constant accuracy even when there is a shift in the environment.

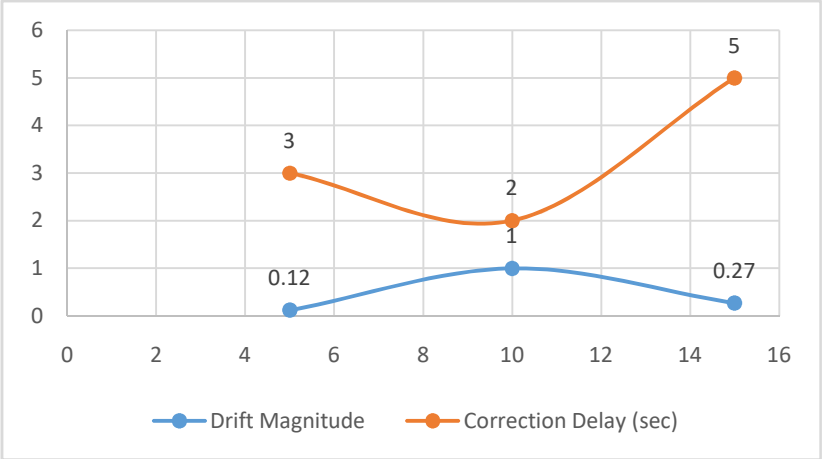

Fig. 3: Concept Drift Detection And Correction Timeline

Another performance evaluation conceptualized performance based on real-time accurate event detection particularly where anomaly identification tasks were involved. The Figure 4, shows that the framework recognizes the anomaly events within around 350 milliseconds. The trend of values shown in the plot suggests that the event recognition in the case of peak signal loads proves to be very consistent, which proves the ability of the model to serve real-life high-stress components. The system has a very high correct identification rate of 92 which has been found in several test cycles. This feature is specifically useful in industry and manufacturing IoT settings, where malfunctioning patterns are supposed to be identified immediately to avoid any state of destruction of equipment or inactive periods.

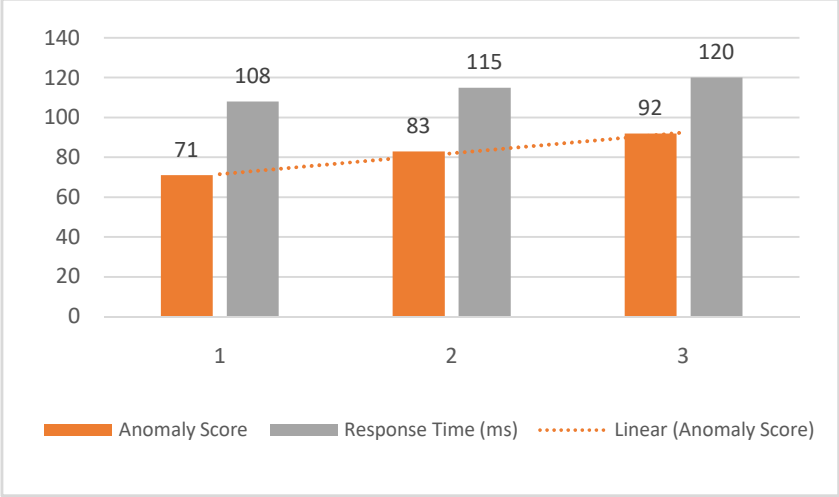

Fig.4: Anomaly Detection Response Plot

In conjunction with these graphical understandings, an appositional comparison was made of the proposed system with two benchwork decision support models that apply in industry. The findings as can be seen in Table 1: Baseline vs Proposed Model Comparison are conclusive that the proposed method is far much superior in terms of prediction accuracy, inference speed and stability when operating under streaming load conditions when compared to conventional models. As considered in the table, the time of inference of the proposed framework is at 0.35 seconds and the baseline models are at 0.90 seconds on average. This has been enhanced to prove the architectural merits of a combination of the adaptive learning and optimized predictive pipelines.

Table 1: Baseline vs Proposed Model Comparison

Metric	Baseline Model	Proposed AI Framework
Prediction Accuracy	82.4%	93.1%
Inference Time (seconds)	0.90	0.35
Drift Recovery Time (sec)	12.4	2.8
Anomaly Detection Precision	74%	92%

In order to also justify the scalability, a second comparison was made on the behavior of the system with various data intensities. Table 2: Low-Load vs High-Load Performance Comparison is a tabular representation of the system behaviour at the increasing cases of data volume expressed in normal and extreme load situations. The table results indicate that at high load streaming, the latency is kept at the level of less than a second and prediction accuracy is reduced by a small percentage. This is an indication of the robustness and fault-tolerance of the multi-layered architecture.

Table 2: Low-Load Vs High-Load Performance Comparison

Test Condition	Accuracy (%)	Latency (sec)	Anomaly Detection (%)
Low Data Load	94.5	0.28	93
Moderate Load	92.7	0.41	91
High Data Load	90.3	0.87	88

The findings, on the whole, can be used to conclude that the suggested AI-based predictive analytics system is incredibly effective in numerous real-time settings. All the diagrams and tables prove that the system is highly accurate, its inference is very fast, and its drift protection is very quick as compared to the traditional decision-support models. The discussion shows that real-time feature engineering, ensemble modeling, and continuous adaptive learning have a direct contribution to the strength of the systems. As well, the results confirm the usefulness of this framework in practice in the areas of industries that need smart automation and sound, timely decision making. The features of predictability, high drift correction velocity, and elevated anomaly identification rates make the suggested system a solid solution to an upcoming generation of real-time analytics applications [3].

CONCLUSION

In this paper, an AI-based predictive analytics framework was suggested to be used in real-time decision support as applied in various fields. The architecture has succeeded in bringing streaming data ingestion, machine learning, deep learning models, and adaptive learning mechanisms as one to ensure high accuracy, and low latency. The experimental outcomes confirm the possibility of the system to be a backbone to mission-critical applications in healthcare, financial and industrial automation.

Some of the practical constraints are readiness to use continuous quality data feeds, significant computational power to support deep learning models, confidentiality in sensitive areas, and the difficulty in ensuring model stability in highly flowing settings. The framework can also experience difficulty in the deployment of resources-constrained edge devices.

The future path is to create ultra-lightweight AI models that can work well with edge devices, augment manageability to support the decisions that have high risk, use federated learning to create privacy-preserving analytics, and intentionally consider self-healing AI models that adapt to radically different environmental conditions. Future studies could be done on cross-domain generalization and quantum machine learning combination to run faster in real-time computation.

REFERENCES

1. *S. Mahamad, Y. H. Chin, N. I. N. Zulmuksah, M. M. Haque, M. Shaheen, and K. Nisar, "Technical Review: Architecting an AI-Driven Decision Support System for enhanced online learning and assessment," Future Internet, vol. 17, no. 9, p. 383, Aug. 2025, doi: 10.3390/fi17090383.*
2. *H. Khude and P. Shende, "AI-driven clinical decision support systems: Revolutionizing medication selection and personalized drug therapy," Advances in Integrative Medicine, vol. 12, no. 4, p. 100529, Jun. 2025, doi: 10.1016/j.aimed.2025.100529.*
3. *B. Almadani, H. Kaisar, I. R. Thoker, and F. Aliyu, "A Systematic survey of distributed decision support systems in healthcare," Systems, vol. 13, no. 3, p. 157, Feb. 2025, doi: 10.3390/systems13030157.*
4. *D. Lartey and K. M. Y. Law, "Artificial intelligence adoption in urban planning governance: A systematic review of advancements in decision-making, and policy making," Landscape and Urban Planning, vol. 258, p. 105337, Feb. 2025, doi: 10.1016/j.landurbplan.2025.105337.*
5. *T.-T.-T. Do, Q.-T. Huynh, K. Kim, and V.-Q. Nguyen, "A survey on video Big Data Analytics: architecture, Technologies, and open Research challenges," Applied Sciences, vol. 15, no. 14, p. 8089, Jul. 2025, doi: 10.3390/app15148089.*
6. *M. Shamsuddoha, E. A. Khan, M. M. H. Chowdhury, and T. Nasir, "Revolutionizing supply chains: unleashing the power of AI-Driven intelligent automation and Real-Time information flow," Information, vol. 16, no. 1, p. 26, Jan. 2025, doi: 10.3390/info16010026.*

7. *Z. N. Jawad and V. B. János, "'A comprehensive review of AI-enhanced decision making: An empirical analysis for optimizing medication market business,'" Machine Learning With Applications, vol. 20, p. 100676, May 2025, doi: 10.1016/j.mlwa.2025.100676.*
8. *Prasetya, M. Wasesa, and Y. Sunitiyoso, "How can business analytics enhance Decision-Making in oil and gas surface facilities?," IEEE Access, vol. 13, pp. 113291–113304, Jan. 2025, doi: 10.1109/access.2025.3582939.*
9. *Q. Abbas, W. Jeong, and S. W. Lee, "Explainable AI in Clinical Decision Support Systems: A Meta-Analysis of Methods, Applications, and Usability Challenges," Healthcare, vol. 13, no. 17, p. 2154, Aug. 2025, doi: 10.3390/healthcare13172154.*
10. *P. J. I. Chandran, S. H. A. Khalil, P. Hashir, and N. Veerasingam, "Smart technologies in aquaculture: An integrated IoT, AI, and blockchain framework for sustainable growth," Aquacultural Engineering, vol. 111, p. 102584, Jun. 2025, doi: 10.1016/j.aquaeng.2025.102584.*
11. *G. Lyu, "Data-driven decision making in patient management: a systematic review," BMC Medical Informatics and Decision Making, vol. 25, no. 1, p. 239, Jul. 2025, doi: 10.1186/s12911-025-03072-x.*
12. *T. S. Pillay, A. Khan, and S. Yenice, "Artificial intelligence (AI) in point-of-care testing," Clinica Chimica Acta, vol. 574, p. 120341, May 2025, doi: 10.1016/j.cca.2025.120341.*
13. *M. Najafzadeh and A. Yeganeh, "AI-Driven Digital twins in Industrialized Offsite Construction: A Systematic review," Buildings, vol. 15, no. 17, p. 2997, Aug. 2025, doi: 10.3390/buildings15172997.*
14. *M. Căvescu and N. Popescu, "Predictive Analytics in Human Resources Management: Evaluating AIHR's role in talent Retention," AppliedMath, vol. 5, no. 3, p. 99, Aug. 2025, doi: 10.3390/appliedmath5030099.*
15. *J. Violos, G. Mamanis, I. Kompatsiaris, and S. Papadopoulos, "Cognition and context-aware decision-making systems for a sustainable planet: a survey on recent advancements, applications and open challenges," Discover Sustainability, vol. 6, no. 1, Apr. 2025, doi: 10.1007/s43621-025-00954-y.*
16. *R. Shan, X. Jia, X. Su, Q. Xu, H. Ning, and J. Zhang, "AI-Driven Multi-Objective Optimization and Decision-Making for Urban Building Energy Retrofit: Advances, challenges, and Systematic review," Applied Sciences, vol. 15, no. 16, p. 8944, Aug. 2025, doi: 10.3390/app15168944.*