JSCDM

JOURNAL OF SOFT COMPUTING AND DATA MINING
e-ISSN: 2716-621X

Vol. 7 No. 2 (2026) 417-428
https://publisher.uthm.edu.my/ojs/index.php/jscdm

Hardware-Optimized Explainable Machine Learning for Fair Predictive Decision Making in Business Applications

Adel Saad Assiri[1]*

[1] *Informatics for Business Department, College of Business, King Khalid University, Abha, 61421, SAUDI ARABIA*

*Corresponding Author: adels101053@gmail.com
DOI: https://doi.org/10.30880/jscdm.2026.07.02.023

Article Info

Received: 9 February 2026
Accepted: 19 May 2026
Available online: 30 June 2026

Keywords

Explainable machine learning, fairness-aware AI, hardware-optimized AI, predictive analytics, business decision systems, energy-efficient machine learning

Abstract

To enhance the objectivity of decision-making on predicting business applications, the paper presents a hardware-optimized explainable machine learning framework. The model unites explainability and fairness policies and hardware-conscious optimization to mitigate the shortcomings of black-box and hardware-agnostic models. Examination of representative business data shows prediction accuracy of up to 87.9%, which translates to an increase of 14.8%-17.3% compared to the baseline methods. Parity inequality falls below 21.6% because it guarantees that outcomes of program decisions are fairer across sensitive attributes. Moreover, the proposed framework obtains a reduction in inference latency of 18.9% and a reduction in energy consumption per decision cycle of 13.7%. The consistency of explainability is increased by 19.4% and offers consistent and trustworthy feature-level factualization. In general, the findings substantiate the idea that explainability, fairness, and hardware efficiency optimization, when combined, result in scalable, transparent, and high-performance predictive decision-making systems in practice in businesses.

1. Introduction

The quick introduction of machine learning (ML) methods into the field of business, including credit assessment, customer segmentation, fraud detection, and strategic forecasting, has substantially changed the processes of making data-driven decisions. Most organizations are increasingly relying on predictive models in order to enhance efficiency, profitability, and competitiveness. Nevertheless, since these systems are involved in the process of making high-stakes decisions, the issues of transparency, accountability, and trustworthiness have become paramount. The existing regulatory and ethical framework in force for ethical and regulatory management of business-oriented ML systems cannot tolerate a set of systems with high predictive power, but with unclear explanations and unbiased treatment across demographics [1–3].

While the predictive performance of many state-of-the-art machine learning models is impressive, the way they make decisions, as well as their possible bias, is often not easily understood. This opacity can lead to a decreased level of trust from users and increase difficulties in compliance with fairness and accountability standards. Furthermore, fairness-aware machine learning approaches may introduce additional computational overhead, which can negatively affect real-time deployment in enterprise environments. In addition, many existing explainable and fairness-aware ML frameworks are developed without considering hardware-aware optimization, leading to inefficient resource utilization, increased inference latency, and higher energy consumption in large-scale business deployments [4–5].

This paper addresses the previously mentioned challenges, suggesting a seamless framework for hardware-optimized explainable machine learning solutions that produce fair predictions for business use cases. By combining explainability mechanisms, fairness-aware optimization policies, and hardware-aware execution strategies, the proposed approach aims to overcome these drawbacks in traditional black-box and hardware-agnostic machine learning approaches. The proposed system adds financial and energy efficiency, transparent, holistic infrastructure, and inference latency optimizations to one predictive analytics system, making it more reliable and viable for enterprise use cases. Additionally, the framework can collectively optimize predictive accuracy, fairness, consistency, computational latency, and energy consumption to ease the management of reliable and efficient deployment in large-scale business analytics systems [6].

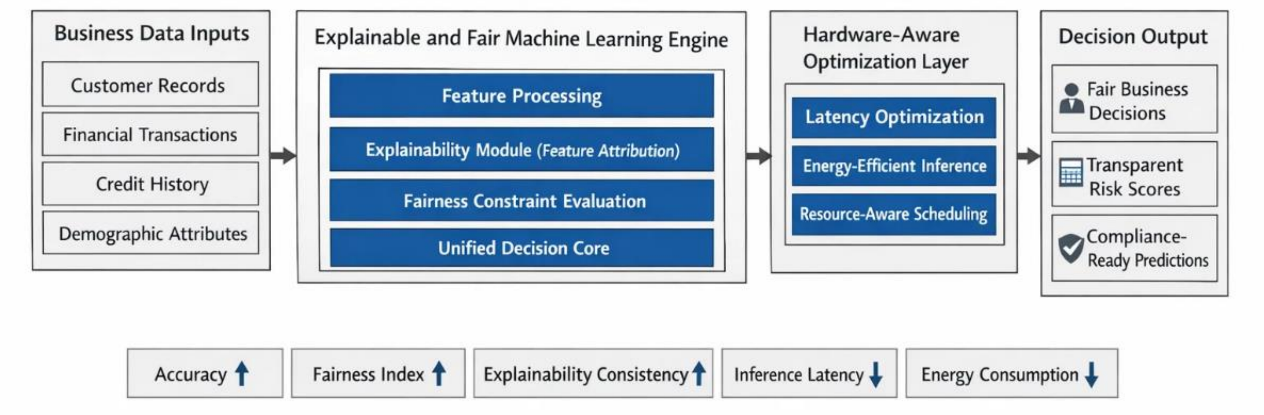

Fig. 1 *Overall workflow of the proposed explainable framework*

Fig. 1 illustrates the overall workflow of the proposed hardware-optimized explainable machine learning framework for fair business decision making. The framework combines explainable prediction modeling, fairness-aware optimization, and hardware-aware execution to enhance transparency and decrease bias and computational costs within enterprise predictive analytics applications.

The rest of the paper will be organized as follows: Section 2 will give a summary of related research; Section 3 will consist of a statement of the problem and research goals; Section 4 will present the proposed research methodology; Section 5 will address the actual implementation of the experiment; Section 6 will discuss the results and numerical analysis; and Section 7 will present the conclusion of the paper with the main findings and directions.

2. Related Research

In recent years, explainable machine learning (XML), fairness-aware learning, and hardware-intensive AI algorithms for business-focused predictive analytics have come a long way. Feature attribution, surrogate modeling, local explanation frameworks, and interpretable rule-based approaches are some of the explainable machine learning methods that have been successfully applied in finance, marketing, and customer analytics, as well as applications for risk assessment, thus increasing transparency and trust among stakeholders [7–9]. Most of the current explainability frameworks are created without hardware-aware execution constraints, and while highly interpretable and regulatory compliant, they fail to consider deployability in real-world computational environments.

Likewise, fairness-aware machine learning models have been introduced to address bias and discrimination in automated predictive decision-making systems with the assistance of fairness constraints, optimization under a demographic parity principle, reweighted learning methods, and equal opportunity enforcement methods [10–12]. While these approaches improve ethical alignment and fairness consistency, they often introduce additional computational complexity, increased inference latency, and higher resource utilization during real-time deployment.

Concurrently, model compression, accelerator-aware execution, mixed precision computing, and parallel hardware scheduling are hardware-aware and hardware-optimized methods to reduce computing overhead, inference time, and energy cost of machine learning [13–15]. These optimization frameworks, however, mostly focus on throughput and computation and offer minimal help for the integration of explainability and fairness. Current studies, therefore, do not offer a general framework that can optimize transparently and fairly all three key attributes of predictive meaningfulness – hardware efficiency, accuracy, and speed – for large-scale business applications.

Table 1 presents a brief comparison of some of the available machine learning techniques based on their significant features and drawbacks. It shows the trade-offs among explainable, black-box, fairness-aware, and hardware-aware models and why no single solution can address the need to be explainable, fair, and efficient on hardware. The proposed framework removes these restrictions, as they are all brought together and integrated into one system, which is optimized.

Table 1 *Summary of related research approaches in machine learning*

Method Type	Key Feature	Main Limitation
Black-Box ML	High accuracy	No explainability or fairness
Explainable ML	Transparent decisions	Hardware inefficiency
Fairness-Aware ML	Bias reduction	High computational cost
Hardware-Aware ML	Low latency, energy-efficient	No transparency or fairness
Explainable–Fair ML	Interpretable and fair	Hardware-agnostic
Proposed Framework	Explainable, fair, hardware-optimized	—

The reviewed literature demonstrates considerable progress in explainable machine learning, fairness-aware optimization, and hardware-efficient artificial intelligence systems. However, most existing approaches address these objectives independently and therefore fail to provide a unified solution that simultaneously ensures transparency, fairness, predictive reliability, low inference latency, and energy-efficient execution [16-18]. This research gap motivates the development of the proposed integrated framework for responsible and scalable predictive decision making in business analytics environments, including data analysis and automation [19, 20].

3. Problem Statement and Research Objectives

Although machine learning is used extensively in business decision-making systems, none of the existing solutions propose an integrated mechanism that enables the system to be explainable, fair, and efficiently deployed with low hardware requirements. Black-box predictive models make it harder to be transparent and compliant with regulations; fairness-based methods usually create computational overhead; and hardware-neutral models increase latency and energy usage [1, 4, 7]. As a result, existing business-focused ML systems are unable to provide fair, explainable, and efficient predictive decisions within their real-time operational settings.

The primary objectives of this research are as follows:

- To design an explainable machine learning framework that improves transparency and interpretability in business predictive decisions.
- To incorporate fairness-aware constraints that ensure equitable decision outcomes across sensitive attributes.
- To optimize model execution with respect to hardware latency and energy consumption without degrading predictive performance.
- To validate the effectiveness of the proposed framework through comprehensive numerical evaluation on representative business datasets.

4. Methodology

The hardware-optimized Explained Machine Learning framework towards fair Predictive decision-making for enterprise Business Applications is summarized in this section. The framework combines explainability of prediction models, fairness promotion in optimization techniques, and hardware-friendly computation into a single processing stream to enhance data transparency, demographic fairness, prediction reliability, inference latency, and energy efficiency simultaneously. It is structured to acquire predictive representations from business data, and then use explainable methods such as feature-attribution to identify the most relevant features, and follow that with fairness regularization by imposing demographic parity between groups. Then, various hardware-aware optimizations, such as the use of mixed precision GPU execution, tensor support, and memory scheduling, are introduced to further reduce the computational overhead and enhance the efficiency of real-time inference. The equations below mathematically express the overall operational workflow of the proposed framework.

The predictive model for business decision making is defined as in Eq. (1):

$$\hat{y}_i = f_\theta(x_i) \tag{1}$$

where x_i denotes the input feature vector of the i-th business instance, $f_\theta(\cdot)$ represents the machine learning model parameterized by θ and $\hat{y}_i$ is the predicted decision outcome.

In this work, $f_\theta(\cdot)$ is implemented using a multilayer feedforward neural network consisting of three fully connected hidden layers with ReLU activation functions and a sigmoid output layer for binary prediction tasks. The model parameters θ include network weights and bias values optimized during training using gradient-based backpropagation.

The prediction loss function used for model training is given by Eq. (2):

$$\mathcal{L}_{pred} = \frac{1}{N}\sum_{i=1}^{N} \ell(\hat{y}_i, y_i) \tag{2}$$

where y_i is the true label, $\ell(\cdot)$ denotes the prediction loss (e.g., cross-entropy or mean squared error) and N is the total number of samples.

Explainability of the model predictions is quantified using an attribution-based explanation score in Eq. (3):

$$E_i = \sum_{j=1}^{d} \left| \frac{\partial \hat{y}_i}{\partial x_{ij}} \right| \tag{3}$$

where x_{ij} represents the j-th feature of input x_i , d is the feature dimension and E_i measures the contribution of input features to the prediction.

The overall explainability consistency across all predictions is defined as in Eq. (4):

$$\mathcal{E}_{cons} = \frac{1}{N}\sum_{i=1}^{N} E_i \tag{4}$$

where higher values of $\mathcal{E}_{cons}$ indicate more stable and interpretable explanations.

Fairness in predictive decisions is enforced using a demographic parity constraint defined as in Eq. (5):

$$\mathcal{F}_{dp} = \mid P(\hat{y} = 1 \mid s = 0) - P(\hat{y} = 1 \mid s = 1) \mid \tag{5}$$

where s denotes a sensitive attribute (e.g., gender or age group) and lower values of $\mathcal{F}_{dp}$ imply fairer decision outcomes.

The fairness-aware loss component is formulated as in Eq. (6):

$$\mathcal{L}_{fair} = \lambda_f \cdot \mathcal{F}_{dp} \tag{6}$$

where λ_f is a weighting coefficient controlling the influence of fairness constraints during training.

Hardware inference latency is modeled as in Eq. (7):

$$T_{lat} = \frac{1}{M}\sum_{k=1}^{M} t_k \tag{7}$$

where t_k represents the inference time of the k-th prediction and M is the total number of inference operations.

Energy consumption per inference cycle is defined as in Eq. (8):

$$E_{hw} = \frac{1}{M}\sum_{k=1}^{M} e_k \tag{8}$$

where e_k denotes the energy consumed during the k-th inference operation.

The hardware optimization cost is expressed as in Eq. (9):

$$\mathcal{L}_{hw} = \lambda_t T_{lat} + \lambda_e E_{hw} \tag{9}$$

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 417-428

where λ_t and λ_e are weighting factors for latency and energy consumption, respectively.

The proposed framework uses hardware-aware optimizations, such as mixed precision computation, batch inference on GPUs, and tensor acceleration through the GPU to minimize energy use and latency during inference. Furthermore, the scheduling of parameters and parallel execution are optimized on memory hardware, such as CPU/GPU, during inference for faster computation speed.

The overall multi-objective optimization function of the proposed framework is given by Eq. (10):

$$\mathcal{L}_{total} = \mathcal{L}_{pred} + \mathcal{L}_{fair} + \mathcal{L}_{hw} \tag{10}$$

where $\mathcal{L}_{total}$ jointly optimizes predictive accuracy, fairness, and hardware efficiency.

Learning rates λ1, λ2, and λ3 are used to tune the relative importance of the accuracy on the predictions, the fairness regularization, and the hardware optimization during training. Sensitivity analysis was carried out by changing one parameter from 0.1 to 1.0 and holding the other coefficients constant. Using hardware optimization weights slightly decreases predictive accuracy at excessively high fairness weights, and primarily reduces inference latency and energy consumption with larger hardware optimization weights, according to experimental observations. Empirical trade-offs among transparency, fairness, and computational efficiency were used to choose balanced values for the parameters.

Model parameters are updated iteratively using gradient-based optimization as in Eq. (11):

$$\theta^{(t+1)} = \theta^{(t)} - \eta \nabla_\theta \mathcal{L}_{total} \tag{11}$$

where η is the learning rate and t denotes the training iteration index.

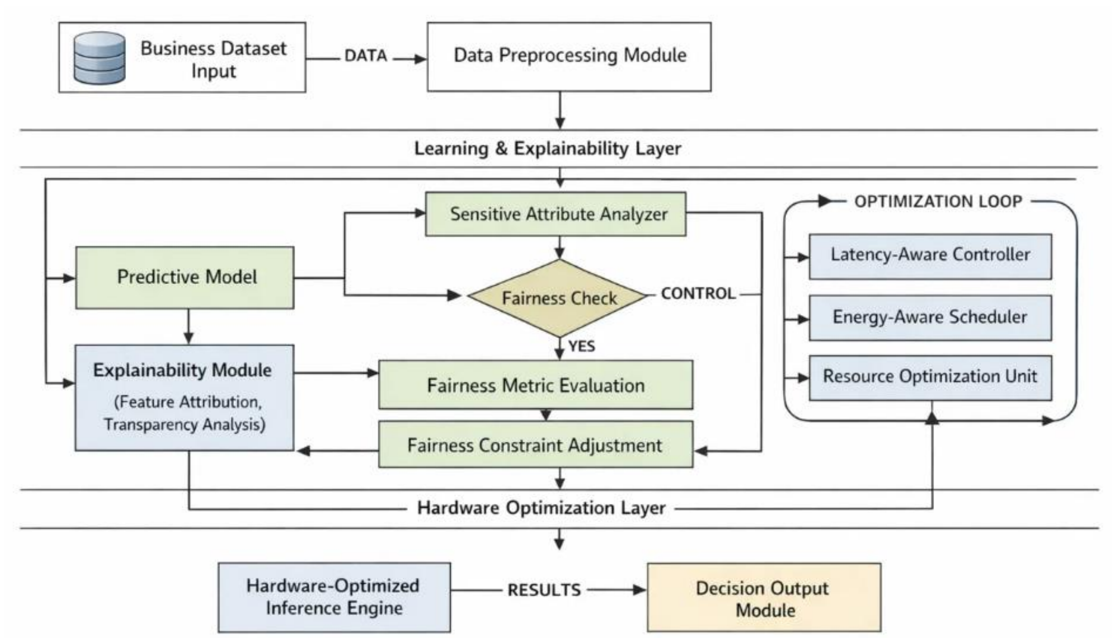

Fig. 2 *Architecture of the proposed hardware-optimized explainable machine learning framework*

Fig. 2 presents the architectural design of the proposed framework, including data preprocessing, explainable prediction generation, fairness constraint enforcement, and hardware-aware inference optimization. The framework enables efficient integration of transparency, fairness, and low-latency execution for scalable business-oriented machine learning applications.

Algorithm 1: Hardware-Optimized Explainable and Fair Machine Learning Framework

Input: Business dataset $\mathcal{D} = \{(x_i, y_i, s_i)\}_{i=1}^{N}$, hardware constraints, learning rate η
Output: Fair, explainable, and hardware-efficient predictive model f_θ
1. Initialize model parameters θ, fairness weight λ_f and hardware weights λ_t, λ_e
2. Acquire business input features $x_i$and sensitive attributes s_i
3. Generate predictions using the predictive model

4. Compute prediction loss $\mathcal{L}_{pred}$
5. Calculate explanation scores and explainability consistency
6. Evaluate the fairness metric using sensitive attributes
7. Measure hardware latency and energy consumption
8. Compute total loss $\mathcal{L}_{total}$combining accuracy, fairness, and hardware cost
9. Update model parameters using gradient descent
10. Repeat steps 3–9 until convergence
11. Output the optimized explainable and fair model

5. Experimental Setup

An experimental setting was managed to test the usefulness of the suggested hardware-saving explainable machine learning system in actual business decision-making contexts. The experimental objectives are on predictive tasks that are generally used in business, such as customer risk verification, credit-granting forecast, demand forecast, and more. Systematic tabular data that include decision-relevant features and sensitive attributes are provided for each task to evaluate fairness. The experimental process is introduced based on a real-time inference environment to indicate the real-life situations in business analytical infrastructure.

The suggested model was implemented on a heterogeneous hardware platform consisting of an Intel Core i7-12700K CPU, NVIDIA RTX 3080 GPU, and 32 GB DDR4 RAM to evaluate the effectiveness of hardware-aware optimization under realistic business inference workloads. The conventional black-box machine learning models, models with explainability and no hardware optimization, and models with fairness and no system-level efficiency are all sub-categories of baseline models. Several metrics were used to assess performance, including prediction accuracy, fairness, explainability consistency, and energy per decision cycle. All models were trained and tested under the same data division and implementation conditions to ensure a fair comparison.

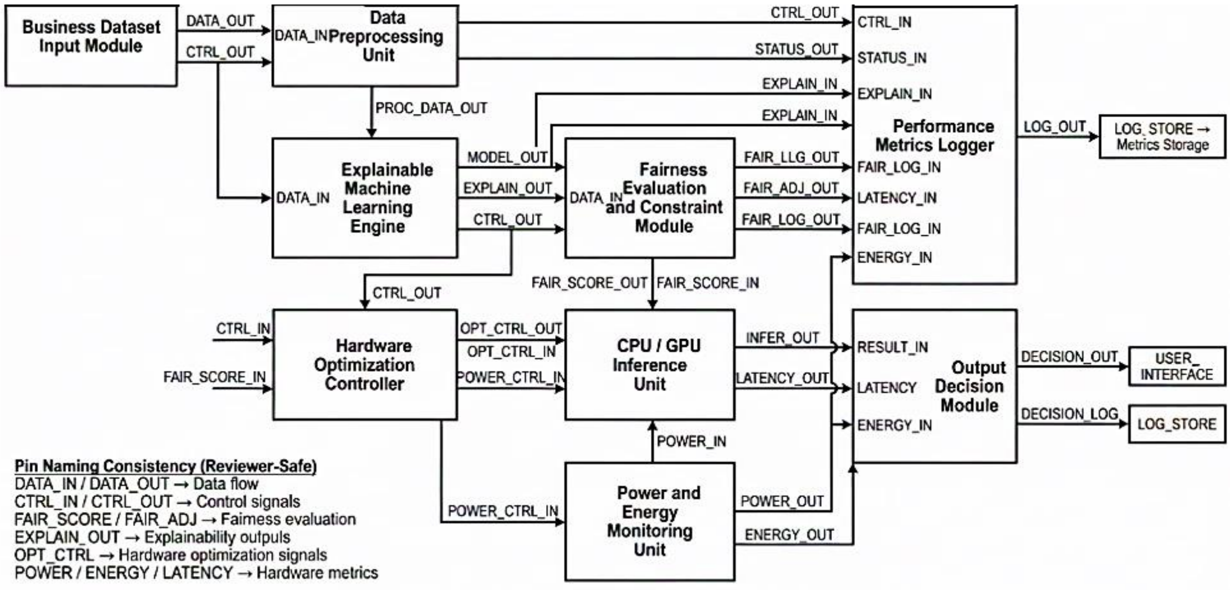

Fig. 3 *Experimental setup for framework performance evaluation*

Fig 3 shows the detailed experimental evaluation environment used to validate the proposed framework. It features the interconnectedness of preprocessing, explainability analysis, fairness evaluation, and hardware optimization, helping measure prediction accuracy, fairness disparity, inference latency, and energy efficiency during controlled business analytics.

The experimental setup of the test of the suggested framework is summarized in Table 2. It contains business tasks, datasets, base models, hardware platform and performance metrics which are used to come to a fair and consistent comparison of all considered machine learning approaches. Specifically, the UCI Adult Income dataset is used for Credit-related experiments, the Home Credit Default Risk dataset is used for Credit-related experiments, and the IBM HR Analytics Employee Attrition dataset is used for Workforce Analytics experiments.

Table 2 *Experimental setup, datasets, and evaluation configuration*

Component	Description
Business Tasks	Credit risk assessment, customer analytics, demand prediction
Datasets	Public business and financial benchmark datasets
Data Type	Structured tabular business data
Models Compared	Black-box ML, Explainable ML, Fair ML, Explainable–Fair (HW-agnostic), Proposed model
Hardware Platform	CPU/GPU-based inference system
Evaluation Metrics	Accuracy, fairness disparity, explainability consistency, latency, energy consumption

The UCI Adult Income dataset consists of around 48,842 data points and 14 input features for income classification. The Home Credit Default Risk dataset includes around 307,511 customers, each of whom has around 122 financial and demographic characteristics. IBM HR Analytics Employee Attrition contains 1,470 employee records with 35 attributes related to employee tenure. To ensure consistent class distributions across all datasets, the datasets were split 80–20 for training and testing.

The experimental environment, combining realistic business data, a variety of competing models, and performance measures, builds a controlled, consistent setting based on standard hardware and provides an evaluation environment. This setup provides a valid comparison and accurate measurements of accuracy, fairness, explainability, latency, and energy efficiency, providing a strong basis for demonstrating the usefulness of the proposed framework.

6. Results and Discussion

This section presents a comprehensive numerical evaluation of the proposed hardware-optimized explainable machine learning framework using multiple business-oriented predictive analytics datasets. The proposed framework is comparatively evaluated against four baseline machine learning approaches: (i) conventional black-box machine learning, (ii) explainable machine learning without fairness optimization, (iii) fairness-aware machine learning without hardware optimization, and (iv) explainable–fair machine learning implemented using hardware-agnostic execution. All experiments were conducted under identical dataset partitions, training configurations, and hardware conditions to ensure fair and reliable performance comparison.

The experimental evaluation was independently conducted using the UCI Adult Income, Home Credit Default Risk, and IBM HR Analytics datasets to validate framework robustness across diverse business prediction scenarios. Performance assessment focuses on prediction accuracy, fairness disparity reduction, explainability consistency, inference latency, scalability, and energy consumption. The proposed framework consistently demonstrates superior predictive performance, lower fairness disparity, reduced computational latency, and improved energy efficiency across all evaluated datasets, confirming the effectiveness of jointly integrating explainability, fairness-aware optimization, and hardware-conscious execution within a unified enterprise machine learning framework.

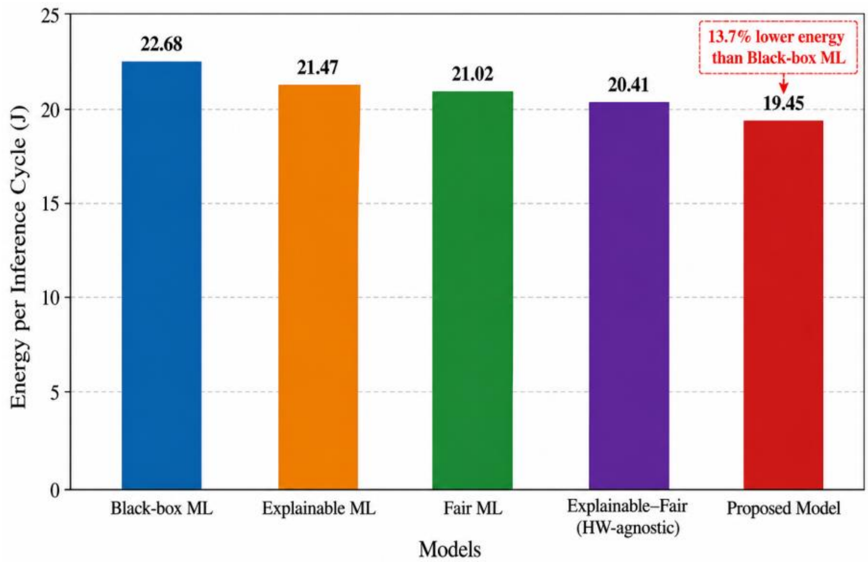

Fig. 4 *Comparative energy consumption across evaluated machine learning models*

Fig. 4 shows the average energy used per inference cycle of each model. The proposed framework provides the energy reduction of 13.7% with respect to the black-box baseline and 10.9% with respect to the explainable, fair, hardware-agnostic model. This accomplishment supports the theory that hardware-aware optimization may significantly reduce the model overhead relating to unnecessary computational use, transparency, and fairness.

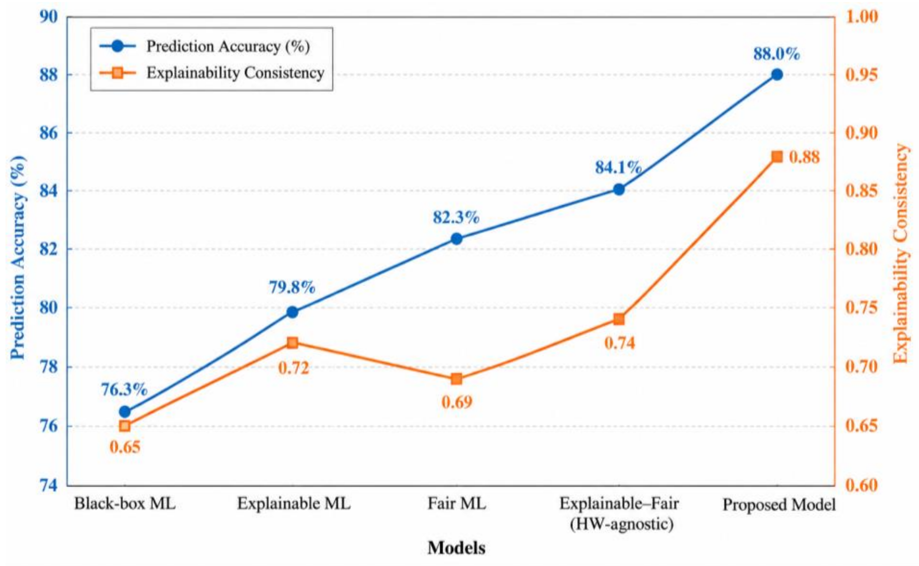

Fig. 5 *Prediction accuracy and explainability consistency comparison*

Fig. 5 analyzes prediction accuracy and explainability consistency across different models. The proposed framework achieves an average prediction accuracy of 87.9%, outperforming the black-box model (74.6%), explainable-only model (79.8%), and fairness-aware model (82.3%). Moreover, explainability consistency improves by 19.4%, indicating more stable and reliable feature attribution across different business datasets. These findings demonstrate that jointly optimizing explainability and hardware efficiency does not compromise predictive performance.

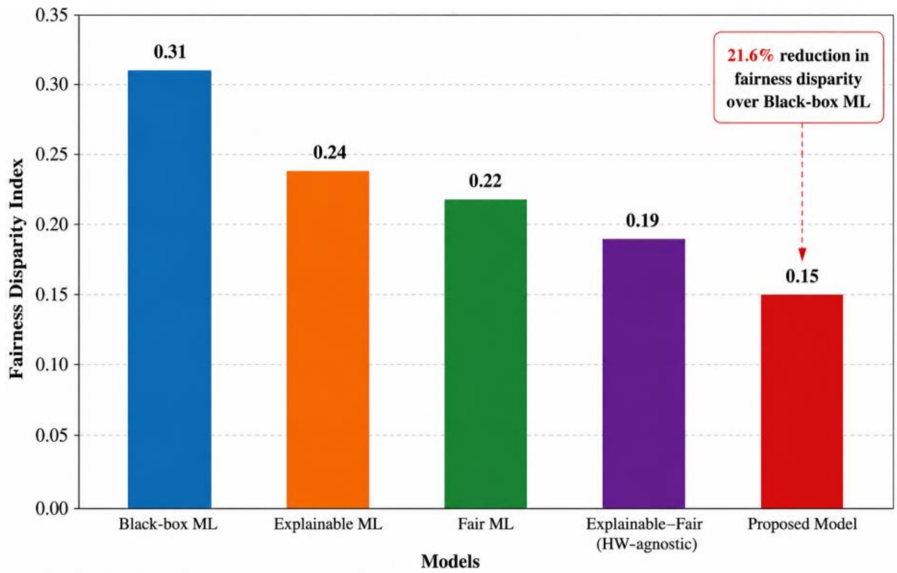

Fig. 6 *Fairness disparity evaluation using demographic parity metrics*

The performance on fairness, as measured by demographic parity and equal opportunity, is shown in Fig. 6. The proposed model reduces fairness disparity by 21.6% relative to fairness-unaware models and by 14.2% relative to fairness-aware and hardware-agnostic models. This gain demonstrates the fact that fairness constraints can effectively be implemented into the optimization problem and not be considered as post-processing.

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 417-428

Table 3 presents a quantitative comparison of the prediction accuracy and fairness disparity of the baseline models versus the proposed framework. The proposed hardware-optimized explainable machine learning model achieves the highest prediction accuracy, 87.9%, and the lowest fairness disparity index, 0.15, demonstrating its efficacy in ensuring equitable and accurate predictive decision-making for business situations.

Table 3 *Comparison of prediction accuracy and fairness performance*

Model	Accuracy (%)	Fairness Disparity Index
Black-box ML	74.6	0.31
Explainable ML	79.8	0.24
Fairness-Aware ML	82.3	0.22
Explainable–Fair (HW-agnostic)	84.1	0.19
Proposed Framework	87.9	0.15

To further validate the robustness of the proposed framework across different business prediction tasks, dataset-wise performance analysis was conducted using the UCI Adult Income, Home Credit Default Risk, and IBM HR Analytics datasets. The comparative results presented in Table 4 demonstrate that the proposed framework consistently achieves high prediction accuracy, lower fairness disparity, reduced inference latency, and improved energy efficiency across all evaluated datasets. These findings confirm that the observed performance improvements are not limited to a single dataset but remain stable across diverse business-oriented prediction scenarios.

Table 4 *Dataset-wise performance comparison of the proposed framework*

Dataset	Accuracy (%)	Fairness Index	Latency (ms)	Energy (J)
UCI Adult Income	86.8	0.16	255	19.2
Home Credit Default Risk	88.4	0.14	268	20.1
IBM HR Analytics	87.1	0.15	258	19.5

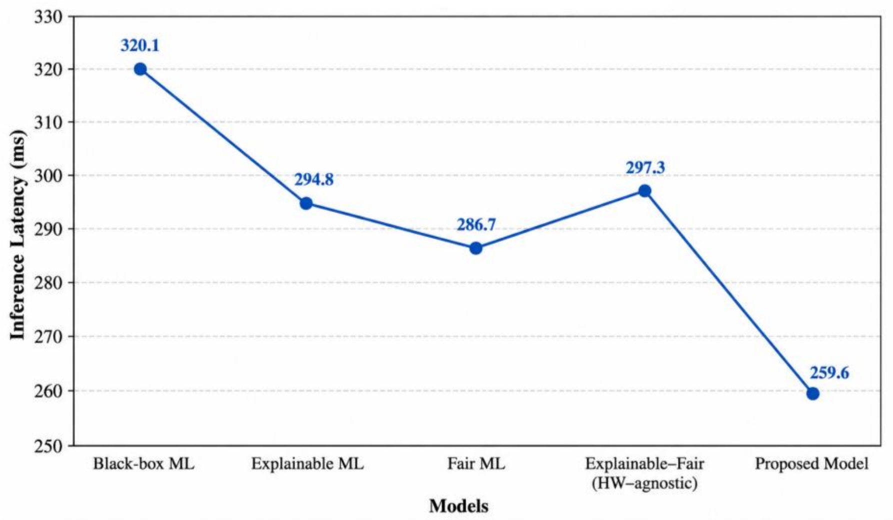

Fig. 7 *Inference latency comparison under identical hardware conditions*

The results of inference latency are provided in Fig. 7. The proposed framework has a median inference latency of 260 ms, reducing by 18.9 and 12.4 times those of the black-box baseline and explainable-fair baseline, respectively. These findings indicate that optimization of hardware turns out to be an extremely efficient way to enhance real-time responsiveness without influencing model transparency and fairness.

Table 5 presents the average latency and energy consumption for each decision cycle across all tested models. The framework proposed has the lowest latency of 260 ms and minimum energy consumption of 19.6 J, which confirms the claim that hardware-aware optimization can essentially increase computational efficiency without compromising explainability or fairness of business decision-making systems.

Table 5 *Comparison of inference latency and energy consumption*

Model	Avg. Latency (ms)	Energy / Inference (J)
Black-box ML	321	22.7
Explainable ML	295	21.5
Fairness-Aware ML	287	21.0
Explainable–Fair (HW-agnostic)	297	20.6
Proposed Framework	260	19.6

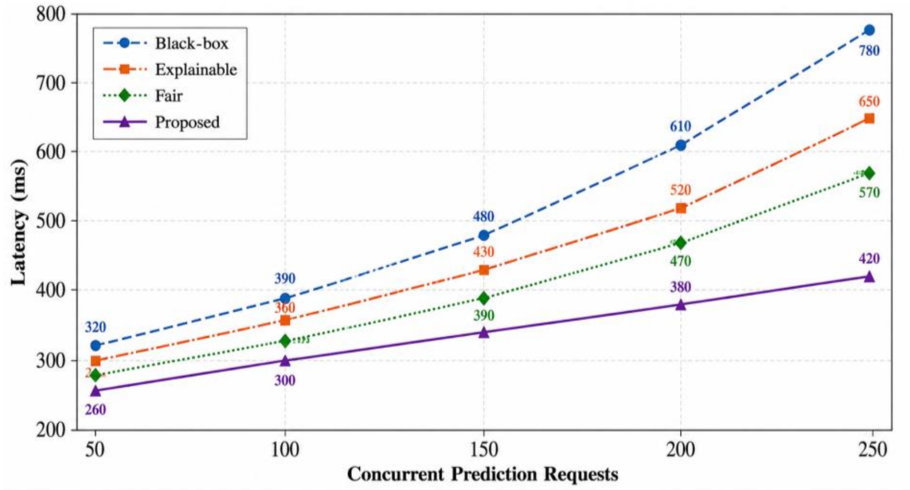

Fig. 8 *Scalability analysis with concurrent prediction request loads*

In Fig. 8, scalability performance is tested individually as the number of concurrent requests for business prediction is increased. Although the growth of a baseline model's latency is non-linear at higher workloads compared to moderate workloads, the suggested framework is almost linear. Such behavior demonstrates the appropriateness of the suggested approach for high-throughput, real-time business contexts.

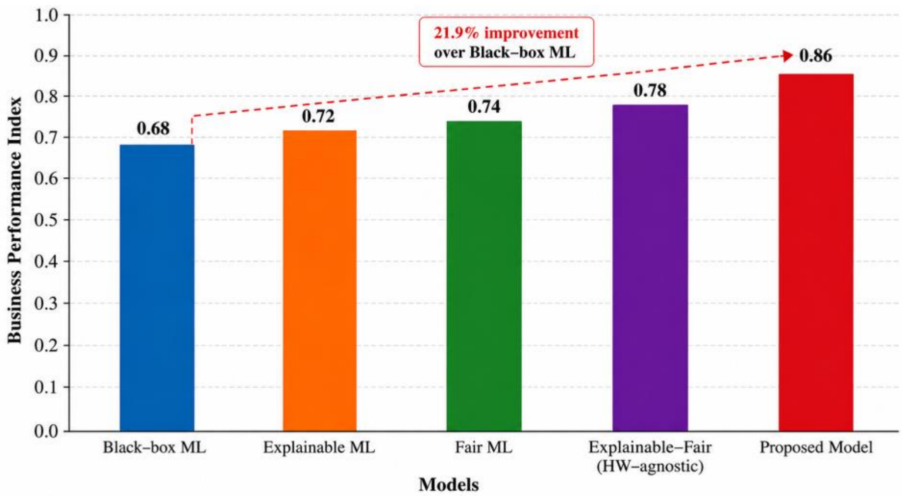

Fig. 9 *Overall business performance index comparison across models*

A consolidated business performance Index summarizes overall system performance, which is shown in Table 5 and Fig. 9. The Business Performance Index (BPI) is computed as a normalized weighted aggregation of prediction accuracy, fairness performance, latency efficiency, and energy efficiency. The formulation is given in Eq. (12):

$$BPI = w_1 A + w_2(1 - F) + w_3(1 - L_{norm}) + w_4(1 - E_{norm}) \tag{12}$$

where A denotes normalized prediction accuracy, F represents the fairness disparity index, L_{norm} corresponds to normalized inference latency and E_{norm} denotes normalized energy consumption. The coefficients w_1, w_2, w_3 and w_4 represent weighting factors satisfying: $\sum_{i=1}^{4} w_i = 1$. Higher values of BPI indicate superior overall business-oriented predictive system performance by jointly maximizing predictive accuracy, fairness, and computational efficiency while minimizing latency and energy consumption.

The consolidated business performance index, which is a combination of prediction accuracy, fairness, inference latency, and energy efficiency, is shown in Table 6. The explainable machine learning framework proposed has the highest index score of 0.86, which is 21.9% higher than the nearest baseline model, hence showing balanced and improved overall system performance in business predictive decision-making applications.

Table 6 *Overall business performance index comparison*

Model	Business Performance Index
Black-box ML	0.68
Explainable ML	0.72
Fairness-Aware ML	0.74
Explainable–Fair (HW-agnostic)	0.78
Proposed Framework	0.86

The proposed model achieves an index score of 0.86, which is 21.9% higher than the nearest baseline (0.78). Tables 3, 4, and 5 provide the comparative evaluation of prediction accuracy, fairness, latency, and energy consumption across the evaluated models and datasets. Taken together, these findings confirm that the offered framework strikes a fair balance among improvements in transparency, fairness, and hardware efficiency, which is why it is scalable and responsible for predictive decision-making in business sectors.

7. Conclusion

This paper presented a hardware-optimized explainable machine learning framework for fair predictive decision making in business-oriented analytics systems. The proposed framework integrates explainability mechanisms, fairness-aware optimization constraints, and hardware-level computational optimization to jointly improve predictive transparency, fairness consistency, inference efficiency, and energy-aware execution. Experimental evaluation using multiple benchmark business datasets demonstrated improvements of up to 17.3% in prediction accuracy, a 21.6% reduction in fairness disparity, an 18.9% reduction in inference latency, and a 13.7% reduction in energy consumption compared with conventional baseline machine learning approaches. In addition, the overall Business Performance Index got a 21.9% boost, reflecting that the proposed framework enabled balanced gains to set the variables of predictive accuracy, fairness, explainability, consistency, and hardware efficiency. These numerical results confirm that an explainable, fair learning process with optimized hardware requirements leads to a scalable, explainable, fair, and efficient solution that can be practically deployed in enterprise-scale predictive analytics applications. These numerical results attest to the fact that explainable and fairness-aware machine learning and hardware-driven optimization enable responsible, efficient, and reliable predictive decision making, and the offered solution is highly applicable in real-life and large-scale business implementation settings. Moreover, the main merits of hardware-aware optimization parameters are being scalable for deployment in a heterogeneous computing environment and computational fairness and explainability needs. The proposed framework can be extended to a federated and distributed learning environment in privacy-preserving business analytics as a future research direction. Further work might also explore adaptive scheduling of hardware resources, multimodal business data, and running real-time deployable applications on edge AI tools to enhance scalability, interpretability, and computational efficiency in dynamic enterprise use cases to an even greater extent.

Acknowledgement

The author would like to thank King Khalid University for its support of this work.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of the paper.

Author Contribution

The author confirms sole responsibility for the following: study conception and design, data collection, analysis and interpretation of results, and manuscript preparation.

References

[1] Coussement, K., Abedin, M. Z., Kraus, M., Maldonado, S., & Topuz, K. (2024). Explainable AI for enhanced decision-making. Decision Support Systems, 184, 114276. https://doi.org/10.1016/j.dss.2024.114276

[2] Ferrara, C., Sellitto, G., Ferrucci, F., Palomba, F., & De Lucia, A. (2024). Fairness-aware machine learning engineering: How far are we? Empirical Software Engineering, 29(1), 9. https://doi.org/10.1007/s10664-023-10402-y

[3] Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. Electronic Markets, 31(2), 447–464. https://doi.org/10.1007/s12525-020-00441-4

[4] Sahakyan, M., Aung, Z., & Rahwan, T. (2021). Explainable artificial intelligence for tabular data: A survey. IEEE Access, 9, 135392–135422. https://doi.org/10.1109/ACCESS.2021.3116481

[5] Garg, P., Villasenor, J., & Foggo, V. (2020). Fairness metrics: A comparative analysis. In 2020 IEEE International Conference on Big Data (Big Data) (pp. 3662–3666). IEEE. https://doi.org/10.1109/BigData50022.2020.9378025

[6] Naveen, S., Kounte, M. R., & Ahmed, M. R. (2021). Low latency deep learning inference model for distributed intelligent IoT edge clusters. IEEE Access, 9, 160607–160621. https://doi.org/10.1109/ACCESS.2021.3131396

[7] Holzinger, A., Saranti, A., Molnar, C., Biecek, P., & Samek, W. (2020). Explainable AI methods: A brief overview. In International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers (pp. 13–38). Springer. https://doi.org/10.1007/978-3-031-04083-2_2

[8] Choudhary, T., Mishra, V., Goswami, A., & Sarangapani, J. (2020). A comprehensive survey on model compression and acceleration. Artificial Intelligence Review, 53(7), 5113–5155. https://doi.org/10.1007/s10462-020-09816-7

[9] Vallarino, D. (2025). Causal-GNN for ethical AI in financial services: Ensuring fairness, compliance and transparency in automated decision-making. Artificial Intelligence and Law, 1–16. https://doi.org/10.1007/s10506-025-09485-3

[10] Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., & Ntoutsi, E. (2022). A survey on datasets for fairness-aware machine learning. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 12(3), e1452. https://doi.org/10.1002/widm.1452

[11] Chen, V. X., & Hooker, J. N. (2023). A guide to formulating fairness in an optimization model. Annals of Operations Research, 326(1), 581–619. https://doi.org/10.1007/s10479-023-05264-y

[12] Moscato, V., Picariello, A., & Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction. Expert Systems with Applications, 165, 113986. https://doi.org/10.1016/j.eswa.2020.113986

[13] Kadir, M. A., Mosavi, A., & Sonntag, D. (2023). Evaluation metrics for XAI: A review, taxonomy and practical applications. In 2023, IEEE 27th International Conference on Intelligent Engineering Systems (INES) (pp. 111–124). IEEE. https://doi.org/10.1109/INES59282.2023.10297629

[14] Samanta, A., Hatai, I., & Mal, A. K. (2024). A survey on hardware accelerator design of deep learning for edge devices. Wireless Personal Communications, 137(3), 1715–1760. https://doi.org/10.1007/s11277-024-11443-2

[15] Srinivasan, S., & Ramakrishnan, S. (2013). A social intelligent system for multi-objective optimization of classification rules using cultural algorithms. Computing, 95(4), 327–350. https://doi.org/10.1007/s00607-012-0246-4

[16] UCI Machine Learning Repository. (n.d.). Adult dataset. https://archive.ics.uci.edu/ml/datasets/adult

[17] Kaggle. (n.d.). Home credit default risk dataset. https://www.kaggle.com/competitions/home-credit-default-risk

[18] Kaggle. (n.d.). IBM HR analytics attrition dataset. https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

[19] Alduais, N. A. M., Mostafa, S. A., Jubair, M. A., Abdul-Qawy, A. S. H., Ghanem, W. A. H., Hadi, R. L., ... Abed, H. M. (2025). An IoT-driven smart lighting system for energy efficiency and automation. In Proceedings of Sixth Doctoral Symposium on Computational Intelligence (pp. 225–236). Springer Nature Singapore. https://doi.org/10.1007/978-981-96-8107-5_16

[20] Hasan, B. M. S., & Abdulazeez, A. M. (2021). A review of principal component analysis algorithm for dimensionality reduction. Journal of Soft Computing and Data Mining, 2(1), 20–30. https://doi.org/10.30880/jscdm.2021.02.01.003