machines

MDPI

Article

TBDDQN: Imbalanced Fault Diagnosis for Blast Furnace Ironmaking Process via Transformer–BiLSTM Double Deep Q-Networks

Jinlong Zheng [1], Ping Wu [1,2,*], Ruirui Zuo [1], Xin Su [1], Yinzhu Liu [1] and Nabin Kandel [1]

[1] School of Information Science and Engineering, Zhejiang Sci-Tech University, Hangzhou 310018, China; 2023220705047@mails.zstu.edu.cn (J.Z.); 2024220705060@mails.zstu.edu.cn (R.Z.); 2024210703007@mails.zstu.edu.cn (X.S.); 2024220705018@mails.zstu.edu.cn (Y.L.); 20242107e3001@mails.zstu.edu.cn (N.K.)

[2] Changshan Research Institute, Zhejiang Sci-Tech University, Quzhou 324200, China

* Correspondence: pingwu@zstu.edu.cn

Abstract

The blast furnace ironmaking process (BFIP) is a highly complex and dynamic industrial system where strong spatiotemporal coupling and severe data imbalance pose substantial challenges for fault diagnosis. To address these issues, this study proposes a Transformer–BiLSTM Double Deep Q-Network (TBDDQN) framework for intelligent fault diagnosis. The framework employs a dual-branch architecture that integrates a Transformer-based spatial encoder with a BiLSTM-attention temporal extractor to capture global dependencies and dynamic patterns from multivariate time-series data. To mitigate class imbalance and asymmetric fault costs, a cost-sensitive reinforcement learning scheme based on Double DQN is incorporated, featuring prioritized experience replay and adaptive misclassification penalties. Experiments on real blast furnace datasets show that TBDDQN achieves a macro-averaged precision of 0.970 and a macro-averaged F1-score of 0.929, outperforming conventional CNN, LSTM, and DQN-based baselines. These results demonstrate that TBDDQN offers a robust and interpretable solution for imbalanced industrial fault diagnosis in the BFIP.

Keywords: blast furnace ironmaking process; imbalanced fault diagnosis; Transformer; BiLSTM-attention; Double DQN; cost-sensitive reinforcement learning

1. Introduction

Global steel output exceeds two billion tons per year. Among ironmaking technologies, the BFIP remains the dominant route for large-scale pig iron production due to its high productivity, operational stability, and cost efficiency. A schematic representation of the BFIP is shown in Figure 1. In this process, alternating layers of iron-bearing materials, coke, and flux are continuously charged from the furnace top, while preheated blast air, often enriched with oxygen and auxiliary fuels, is injected through tuyeres located near the furnace hearth. As the burden descends, the iron oxides are progressively reduced by ascending reducing gases generated from coke gasification, forming molten iron and slag. The molten iron collects in the hearth and is periodically tapped, whereas top gas is recycled after dust removal and heat recovery [1].

The BFIP is essentially a highly coupled, nonlinear, and multiphase chemical-physical reactor characterized by strong spatiotemporal variability, high temperatures, and intense

Check for updates

Academic Editor: Ahmed Abu-Siada

Received: 5 February 2026
Revised: 27 February 2026
Accepted: 27 February 2026
Published: 2 March 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

energy exchange. These complexities pose significant challenges for process optimization, energy conservation, and environmental protection. Given the substantial energy consumption and emissions associated with ironmaking, improving process efficiency and ensuring operational safety are key priorities for the metallurgical and chemical process industries [2,3].

In practical blast furnace ironmaking operations, fault scenarios such as furnace collapse, furnace temperature rise or down, and channeling events frequently occur under complex and strongly coupled process conditions. These faults not only differ in their physical mechanisms and observable patterns but also exhibit highly imbalanced occurrence frequencies. More importantly, misclassification of certain fault types, especially missed detection of severe faults, may lead to substantial safety risks, equipment damage, and economic losses. Therefore, fault diagnosis in BFIP is not merely a pattern recognition task but a risk-sensitive decision-making problem that must account for asymmetric misclassification costs. This practical characteristic motivates the formulation of BFIP fault diagnosis as a cost-sensitive sequential decision problem in this study.

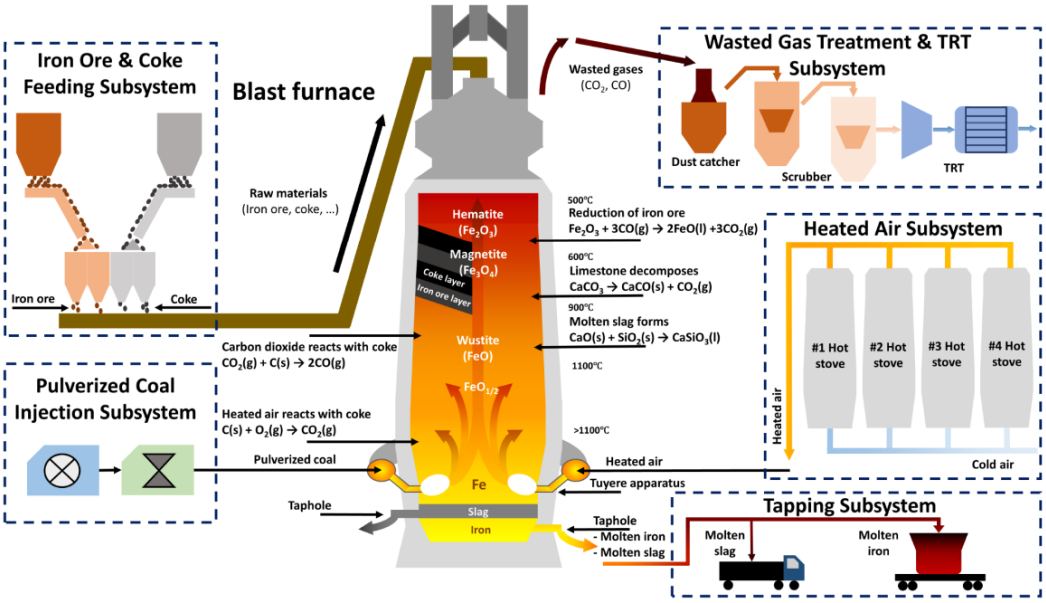


Figure 1. Schematic diagram of the blast furnace ironmaking process.

Fault diagnosis technology is essential for maintaining stable furnace operation and avoiding catastrophic failures. Fault diagnosis methods are broadly categorized as either model-based or data-driven. Model-based techniques depend on accurate mathematical models based on first-principles mechanisms. However, constructing an accurate mathematical model for a process as complex as the BFIP, where gas–solid reactions, heterogeneous heat and mass transfer, and dynamic burden descent coexist, is difficult [4,5]. A vast amount of process data is now continuously collected from blast furnace instrumentation systems. This enables the rise of data-driven fault diagnosis, in which knowledge of the system's health state is inferred directly from historical and real-time measurements [6,7]. Consequently, data-driven diagnosis has become a promising research frontier for improving operational reliability and efficiency in the BFIP [3].

Traditional machine learning algorithms were widely applied in this context due to their relatively straightforward implementation and interpretability. For instance, Zhou et al. [8] applied principal component analysis (PCA)-based methods for process monitoring of the iron-making process in a blast furnace. Zhang et al. [9] proposed a two-stage PCA for fault diagnosis in the blast furnace ironmaking process. Tian et al. [10] developed a support vector machine (SVM) ensemble for fault diagnosis in blast furnaces.

Liu et al. [11] introduced an optional SVM approach for fault diagnosis of blast furnaces handling imbalanced data. Pan et al. [12] presented robust principal component pursuit for fault detection in blast furnace processes.

Despite these developments, traditional machine learning methods were limited by their dependence on manually engineered features and their shallow representational capacity. Deep learning (DL) has therefore emerged as a powerful alternative, offering automated extraction of deep hierarchical features from high-dimensional process data [13]. For example, Jang et al. [14] proposed an adversarial autoencoder-based approach to learn features for fault detection in industrial processes. Wang et al. [15] applied a stacked supervised auto-encoder for fault-relevant feature extraction and classification. Yu et al. [16] introduced a supervised convolutional autoencoder for fault-relevant feature learning in fault diagnosis. Zhu et al. [17] explored variational autoencoders for fault detection and diagnosis in industrial processes. Xu et al. [18] employed a hybrid framework combining long short-term memory (LSTM) networks and convolutional neural network (CNN) to extract deep temporal and spatial features from time series and image data, respectively. Building on LSTM architectures, bidirectional long short-term memory (BiLSTM) models have also been widely used to capture bidirectional temporal dependencies in industrial time series data analysis [19]. In addition to recurrent models, the Transformer architecture [20], powered by multi-head self-attention, has recently been adopted for industrial fault diagnosis due to its strong capability in capturing long-range global dependencies.

However, two practical challenges persist in blast furnace fault diagnosis. The first is data imbalance, in which abnormal conditions occur infrequently and result in scarce labeled faulty data. The second is asymmetric misclassification cost, where false negatives (missed fault detection) can lead to serious operational or safety risks, unnecessary process interruptions, and productivity losses. These issues hinder the deployment of conventional classification models for reliable industrial applications. Reinforcement learning (RL) offers a promising solution to overcome these limitations by optimizing long-term rewards through continuous interaction with the environment, eliminating the need for extensive labeled datasets [21]. Its capability to account for penalties, uncertainties, and deferred consequences positions RL as an ideal method for decision-making in sophisticated industrial environments [22]. By framing fault diagnosis within a Markov Decision Process (MDP) framework, RL's effectiveness is amplified in that the agent progressively refines its diagnostic policy and integrates both diagnostic results and operational expenses [23]. This methodology is highly compatible with fault diagnosis in the BFIP, where dynamic decisions are required to balance precision, operational safety, and cost-effectiveness [24]. Nevertheless, traditional RL still struggles with the rarity of fault samples during learning. Prioritized Experience Replay (PER) [25] addresses this issue by sampling experiences in proportion to their temporal-difference error, enabling the agent to focus more heavily on informative but infrequent fault-related transitions. Furthermore, to better handle the asymmetric misclassification costs inherent in industrial diagnosis, an adaptive penalty mechanism [26] can be incorporated into the reward design, dynamically adjusting penalty magnitudes based on the evolving diagnostic performance to achieve a more balanced tradeoff between safety and false alarms.

Although cost-sensitive learning can alternatively be achieved within supervised learning frameworks through techniques such as class-weighted cross-entropy or focal loss, these approaches typically rely on fixed and predefined penalty structures. In contrast, the reinforcement learning paradigm adopted in this study provides a more flexible and policy-oriented solution. By formulating the fault diagnosis task as a cost-sensitive Markov decision process, the diagnostic policy is optimized through a reward-driven learning mechanism that explicitly integrates adaptive misclassification penalties and

prioritized experience replay. This formulation allows the learned policy to account for long-term decision consequences and asymmetric fault risks, rather than optimizing only instantaneous classification loss. Consequently, the model can place greater emphasis on high-consequence fault categories while maintaining overall diagnostic effectiveness under normal operating conditions. This risk-aware decision formulation is particularly suitable for the blast furnace ironmaking process, where fault occurrences are highly imbalanced and the costs associated with different misclassification types vary substantially.

Considering the characteristics of blast furnace process data, effective fault diagnosis requires models that can simultaneously capture global correlations among multivariate variables and fine-grained temporal evolution patterns. Transformer architectures, equipped with multi-head self-attention, are well suited for extracting long-range contextual dependencies, while BiLSTM networks provide strong capabilities in modeling forward–backward temporal dynamics in sequential industrial data. However, even with powerful feature extractors, the challenges of data imbalance and asymmetric misclassification costs remain difficult to address using purely supervised learning. Double Deep Q-Network (DDQN) offers a principled solution by enabling cost-aware decision-making through interaction with the environment and by incorporating mechanisms such as prioritized experience replay to emphasize rare fault samples. Recent studies have made notable progress in this domain. Lou et al. [27] proposed a deep correlation network representation regression framework for modeling and monitoring blast furnace ironmaking processes. Wang et al. [28] introduced multiple local manifold learning for fault detection in blast furnace systems. Domain-adaptation and knowledge-driven methods have also been developed to handle limited fault samples and distribution shifts in blast furnace fault diagnosis [29,30]. Meanwhile, deep reinforcement learning approaches with reward adjustment for minority classes and cost-sensitive classifiers have been explored to address class imbalance in industrial fault diagnosis [31]. Despite these advancements, existing hybrid architectures (e.g., CNN-LSTM, Transformer-based models) often lack effective joint spatiotemporal modeling for strongly coupled process variables. Furthermore, most reinforcement learning-based methods still fail to explicitly address asymmetric misclassification costs under extreme class imbalance. Building upon the above discussions, this study proposes a Transformer–BiLSTM DDQN framework for imbalanced fault diagnosis in the BFIP. The proposed method integrates a Transformer-based feature extraction branch to capture global contextual dependencies with a bidirectional LSTM-attention branch for dynamic temporal pattern learning. The outputs of both branches are fused to construct discriminative spatiotemporal representations. On this basis, a reinforcement learning strategy based on a DDQN is employed, incorporating prioritized experience replay and an adaptive cost-sensitive reward mechanism to handle data imbalance and optimize diagnostic decision-making dynamically. The main contributions of this work are summarized as follows:

- A dual-branch spatiotemporal feature extraction framework is proposed, in which a Transformer-based spatial encoder captures global inter-variable dependencies, while a BiLSTM-attention module models temporal dynamics in BFIP data.
- A cost-sensitive reinforcement learning formulation for fault diagnosis is developed by modeling the task as a Markov decision process and embedding adaptive misclassification penalties into a Double DQN architecture with prioritized experience replay, enabling effective handling of class imbalance and asymmetric fault costs.
- Extensive experiments conducted on real-world blast furnace datasets demonstrate that the proposed TBDDQN achieves competitive performance compared with conventional deep learning and reinforcement learning baselines in terms of weighted-averaged and macro-averaged evaluation metrics.

The remainder of this paper is organized as follows. Section 2 reviews related work on Classification Markov Decision Process, DDQN learning strategies, and Transformer encoders. Section 3 presents a detailed description of the proposed framework and methodology. Section 4 presents the experimental validation on real-world BFIP data and performance comparisons. Finally, Section 5 concludes the paper and outlines potential future research directions.

2. Related Work

2.1. Classification Markov Decision Process

The Classification Markov Decision Process (CMDP) extends the conventional MDP paradigm to address classification-oriented learning tasks [32]. Unlike standard MDPs designed for sequential decision-making under uncertainty, CMDP reformulates the supervised classification problem into a sequential decision process, allowing reinforcement learning agents to iteratively improve classification accuracy through interaction with labeled data.

The CMDP can be expressed as a tuple $\{S, A, P, R, \gamma\}$, built upon a training dataset $\mathcal{D}_{\text{train}} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K\}$, where each subset $\mathcal{D}_k$ represents samples belonging to the kth class and contains labeled pairs (x_i, l_i). The key components of the CMDP are defined as follows.

- Action Space ($\mathcal{A}$): For a classification task involving K categories, the action set is given by $A = \{0, 1, 2, \ldots, K-1\}$, where each action corresponds to the agent's predicted class label. Selecting an appropriate action thus reflects the agent's classification decision at each step.
- State Space ($\mathcal{S}$): Each state in the CMDP corresponds to one labeled sample drawn from $\mathcal{D}_{\text{train}}$. The initial state s_0 is associated with the first sample in the training dataset, and the agent transitions through subsequent samples as it progresses through the classification episode.
- Transition Probability (P): The environment transitions deterministically between samples in a fixed sequential order. Specifically,

$$P(s_{t+1}|s_t, a_t) = \begin{cases} 1, & \text{if } s_{t+1} \text{ is the next sample in } \mathcal{D}_{\text{train}}, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

  This deterministic setting ensures that each classification attempt is treated as a sequential step rather than a stochastic sampling process.
- Reward Function ($\mathcal{R}$): To evaluate the correctness of classification actions, the reward function is defined as

$$R(s_t, a_t, l_t) = \begin{cases} 1, & \text{if } a_t = l_t, \\ -1, & \text{if } a_t \neq l_t. \end{cases} \tag{2}$$

  This binary reward design provides immediate feedback on prediction accuracy. However, such a formulation overlooks the issue of class imbalance, where misclassifying minority samples incurs the same penalty as misclassifying majority ones, an important limitation motivating later extensions to cost-sensitive CMDP.
- Discount Factor (γ): The parameter $\gamma \in (0, 1]$ determines the weight of future rewards, maintaining consistency with standard reinforcement learning frameworks [33].

A complete CMDP trajectory can thus be expressed as $s_1, a_1, r_1, s_2, a_2, r_2, \ldots, s_T, a_T, r_T$, where T denotes the terminal step. Each episode terminates either when all samples in the training dataset are processed or when the agent makes an incorrect prediction, marking the end of a classification sequence.

2.2. Double Deep Q-Network

The DDQN algorithm is developed to address the systematic overestimation of action values in Deep Q-Network (DQN). In the standard DQN framework, a single neural network is used for both selecting and evaluating actions, which often leads to overly optimistic value estimates due to the coupling of these two processes.

In DQN, the target value is defined as:

$$y_t^{\mathrm{DQN}} = r_{t+1} + \gamma \max_a Q(s_{t+1}, a; \bar{\theta}) \tag{3}$$

where $\bar{\theta}$ represents the parameters of the target network. The max operator in this formulation tends to favor actions with overestimated Q-values, resulting in biased learning and unstable convergence.

To mitigate this issue, the DDQN framework introduces a crucial modification by decoupling the action selection and evaluation processes. DDQN retains the core advantages of DQN, specifically leveraging deep neural networks in Q-learning, while substantially reducing the overestimation bias. Specifically, two separate networks with identical architectures are maintained, which are the main network with parameters θ and the target network with parameters $\bar{\theta}$. The main network selects the optimal action for the next state, while the target network evaluates the value of that chosen action. Then, it leads to a revised target definition for DDQN:

$$y_t^{\mathrm{DDQN}} = r_{t+1} + \gamma Q\left(s_{t+1}, \arg\max_a Q(s_{t+1}, a; \theta); \bar{\theta}\right) \tag{4}$$

Here, the main network determines the action $\arg\max_a Q(s_{t+1}, a; \theta)$, and the target network evaluates its Q-value. This separation effectively eliminates the overestimation induced by the max operator, yielding a more accurate and stable learning signal.

Similar to DQN, the DDQN framework also incorporates two essential mechanisms which are Experience Replay (ER) and a target network to improve sample efficiency and stabilize training. Experience tuples (s_t, a_t, r_t, s_{t+1}) gathered from interactions with the environment are stored in a replay buffer, from which mini-batches are sampled for training.

The network parameters are optimized by minimizing the mean squared error (MSE) loss function:

$$L(\theta) = \mathbb{E}_{(s_t, a_t, r_t, s_{t+1}) \sim ER}\left[w_t \left(y_t^{\mathrm{DDQN}} - Q(s_t, a_t; \theta)\right)^2\right] \tag{5}$$

and updated via gradient descent:

$$\theta \leftarrow \theta - \alpha \nabla_\theta L(\theta) \tag{6}$$

where α denotes the learning rate. The target network parameters $\bar{\theta}$ are updated from the main network either periodically or through a continuous averaging process to ensure smoother convergence and training stability.

In summary, DDQN effectively inherits the representational power of deep Q-learning while mitigating the overestimation bias of DQN through the decoupling of action selection and evaluation. This improvement enhances both the reliability and robustness of the reinforcement learning process.

2.3. Transformer Encoder

The Transformer model, originally proposed by Vaswani et al. [20], introduces a novel network architecture based solely on attention mechanisms, representing a paradigm shift in sequence modeling by eschewing traditional recurrent and convolutional architectures in favor of a purely attention-based mechanism. This architectural innovation

addresses fundamental limitations of recurrent neural networks (RNNs), particularly their sequential computation nature that precludes parallelization and struggles with long-range dependencies.

The core component of the Transformer Encoder is the multi-head self-attention mechanism, which allows the model to jointly attend to information from different representation subspaces at different positions. The self-attention function maps a query and a set of key-value pairs to an output, which is computed as a weighted sum of the values, where the weights are determined by the compatibility between queries and keys. The scaled dot-product attention formulation is expressed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7}$$

where Q, K, and V represent queries, keys, and values respectively, and d_k is the dimensionality of the keys. The scaling factor $\frac{1}{\sqrt{d_k}}$ prevents the softmax function from entering regions of extremely small gradients.

The encoder is composed of multiple identical layers stacked together, with each layer containing two sub-layers which are a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. Residual connections [34] are employed around each sub-layer, followed by layer normalization [35]. The feed-forward network applies two linear transformations with a Rectified Linear Unit (ReLU) activation in between:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{8}$$

A critical innovation in the Transformer architecture is the positional encoding, which injects information about the relative or absolute position of tokens in the sequence since the model contains no recurrence or convolution. This design allows the model to extrapolate to sequence lengths longer than those encountered during training and enables learning of relative positional relationships.

The overall architecture of the Transformer Encoder is illustrated in Figure 2, which shows how each encoder layer consists of a multi-head self-attention block followed by a feed-forward network, enabling global dependency modeling across multivariate features.

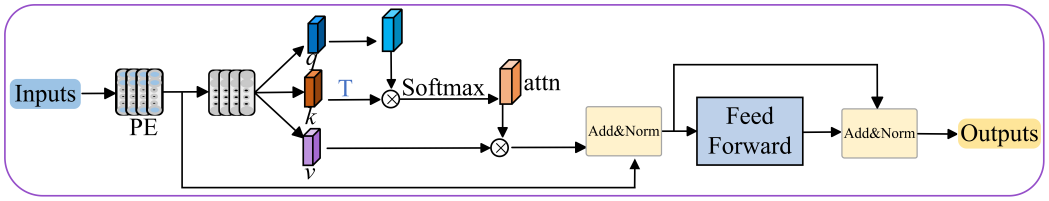


Figure 2. Architecture of the Transformer-Based Spatial Encoder module. Each encoder layer consists of a multi-head self-attention block followed by a feed-forward network, enabling global dependency modeling across multivariate features.

Compared to recurrent and convolutional layers, the Transformer Encoder offers significant advantages, including enabling full parallelization during training and demonstrating strong performance in capturing long-range dependencies. These properties make the Transformer Encoder a foundational building block in numerous subsequent architectures across various domains beyond natural language processing.

3. Proposed Method

This section introduces a novel intelligent fault diagnosis framework for blast furnaces based on deep reinforcement learning. The proposed method formulates the fault classification task as a Cost-Sensitive Markov Decision Process (CSMDP), which explicitly

incorporates the asymmetric costs associated with different types of misclassifications. This allows the model to address real-world class imbalance by accounting for the different consequences of false positives and false negatives.

To effectively model complex spatial and temporal dependencies inherent in industrial time-series data, a dual-branch architecture based on Double Deep Q-Networks is proposed. The TBDDQN model integrates a Transformer module for spatial feature extraction and a BiLSTM-attention network for temporal modeling. Additionally, to improve sample efficiency and training stability, a PER mechanism is incorporated. The overall architecture of the TBDDQN model is shown in Figure 3.

After introducing the overall architecture of the proposed TBDDQN framework, we briefly discuss its implementation considerations from an engineering perspective.

From an implementation perspective, the proposed TBDDQN framework adopts a modular design. The Transformer-based spatial encoder and BiLSTM-attention temporal extractor follow task-oriented deep learning implementations and can be readily developed using existing libraries. Compared with traditional supervised fault classifiers, the additional implementation effort primarily arises from the reinforcement learning-based decision layer, including the formulation of the diagnostic process as a Markov decision process and the design of cost-sensitive rewards. Nevertheless, this layer operates independently of the sensing and preprocessing modules, enabling the proposed framework to be integrated into existing industrial monitoring systems with moderate additional effort while providing enhanced robustness under class imbalance and asymmetric fault costs.

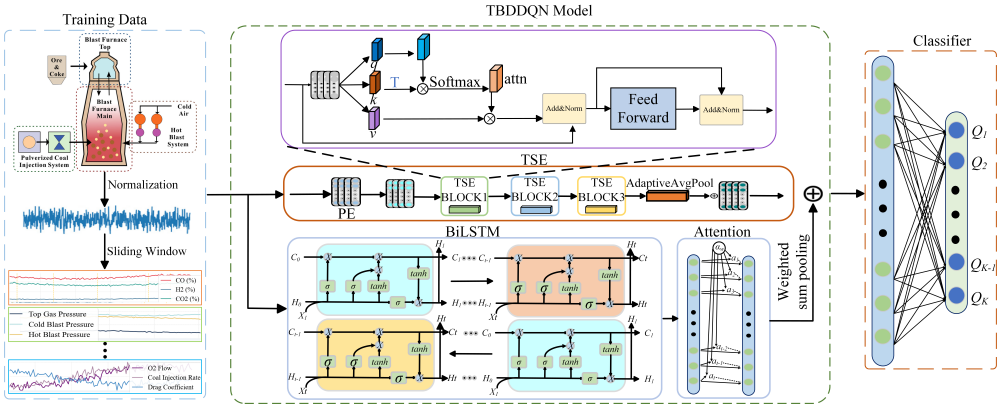


Figure 3. Architecture of the proposed TBDDQN framework, consisting of Transformer-based spatial encoding, BiLSTM-attention temporal modeling, and a fusion layer for integrated decision-making.

3.1. Spatiotemporal Feature Extraction

The proposed TBDDQN framework adopts a dual-branch structure for spatiotemporal feature extraction, designed to capture both spatial correlations among process variables and temporal dependencies within sequential data. As depicted in Figure 3, the dual-branch design consists of two complementary components, each responsible for extracting different aspects of the data.

The first branch employs a Transformer-based spatial encoder (TSE) to model the global dependencies between multivariate features. The input time-series segment is first linearly projected into a latent embedding space, then combined with learnable positional encodings. The data are processed through multiple stacked Transformer encoder layers, each consisting of a multi-head self-attention mechanism and a feed-forward subnetwork with residual connections. This structure enables the model to adaptively capture cross-feature relationships across time, providing a comprehensive spatial understanding of the process state. Finally, a global pooling layer aggregates these contextual representations into a compact spatial embedding.

The second branch focuses on capturing temporal dependencies within the data using a BiLSTM network, enhanced with an attention pooling mechanism. The BiLSTM captures both forward and backward temporal dependencies, enabling the model to represent the complete sequence of events. The attention mechanism assigns adaptive weights to each time step based on its relevance to the overall state, thus producing a more informative temporal representation.

The outputs of both branches are concatenated along the feature dimension, resulting in a fused spatiotemporal representation. This joint feature vector is then passed through a fully connected layer, which estimates Q-values for all possible actions, enabling the TBDDQN model to make integrated spatiotemporal decisions.

Beyond representation learning, the proposed dual-branch architecture also provides intrinsic interpretability. Specifically, the Transformer-based spatial encoder employs multi-head self-attention to explicitly model the relative importance and interactions among different process variables, enabling the identification of critical variables that contribute most to fault discrimination. Meanwhile, the attention mechanism embedded in the BiLSTM branch assigns adaptive weights to different time steps, highlighting key temporal segments that are most informative for fault diagnosis. As a result, the fused spatiotemporal representation not only improves diagnostic performance but also offers interpretable insights into both variable-level and temporal-level contributions underlying the decision-making process.

3.2. Cost-Sensitive Markov Decision Process

To address the challenge of asymmetric misclassification costs and class imbalance in blast furnace fault diagnosis, this study reformulates the condition classification problem as a CSMDP. Within this framework, the learning agent interacts with the environment to make diagnostic decisions while explicitly incorporating the varying severities of misclassification through a cost-weighted reward mechanism.

At each time step t, the system state s_t is represented as a sliding window W_{len} containing consecutive process observations:

$$s_t = \{x_{t-W_{len}+1}, x_{t-W_{len}+2}, \ldots, x_t\} \tag{9}$$

where x_t denotes the multivariate feature vector at time t, and W_{len} represents the sliding window length. The action $a_t \in \mathcal{A}$ corresponds to the predicted operating condition or fault category.

To model the asymmetric impact of different diagnostic outcomes, a cost-sensitive reward function $R(s_t, a_t, l_t)$ is defined as:

$$R(s_t, a_t, l_t) = \begin{cases} +1.0, & a_t = l_t, \\ -p_{l_t,a_t}, & a_t \neq l_t, \end{cases} \tag{10}$$

where p_{l_t,a_t} represents the predefined penalty coefficient derived from a misclassification cost matrix. This matrix encodes the relative severity of various error types, ensuring that high-risk misclassifications yield greater penalties and promoting cost aware decision-making behavior.

To further enhance robustness under imbalanced data conditions and mitigate the persistent misclassification of rare fault types, the penalty term is dynamically adjusted according to historical error frequencies:

$$p'_{l_t,a_t} = p_{l_t,a_t}\left(1 + \frac{\text{error_counts}[l_t, a_t]}{\text{total_counts}[l_t] + \epsilon}\right) \tag{11}$$

Here, error_counts$[l_t, a_t]$ denotes the number of times class l_t has been misclassified as a_t, total_counts$[l_t]$ represents the total occurrences of class l_t, and ϵ is a small positive constant introduced to prevent division by zero. This adaptive reward mechanism encourages the agent to focus more on high-cost or frequently misclassified categories, thereby achieving balanced learning and improved diagnostic precision, particularly for minority and critical fault classes.

Unlike conventional supervised classification models, the diagnostic decisions of the proposed TBDDQN are explicitly guided by a cost-sensitive reward design. The learned Q-values therefore encode not only classification confidence but also the potential operational risk associated with different misclassification types. This explicit coupling between reward and decision-making allows the diagnostic policy to be interpreted as a rational trade-off between accuracy and fault-related risk, which cannot be directly achieved by static loss reweighting strategies.

Unlike the standard CMDP framework that assumes deterministic temporal transitions, the proposed CSMDP adopts a stochastic transition mechanism to enhance exploration diversity. Notably, in our formulation, the transition to the next state is independent of the agent's action because the diagnostic decision does not alter the underlying blast furnace process. Therefore, after each interaction, the subsequent state s_{t+1} is randomly sampled from the training dataset $\mathcal{D}_{\text{train}}$ instead of following a fixed sequence. This randomized sampling strategy enhances exploration by exposing the agent to diverse states and reduces overfitting to local temporal correlations that are common in industrial process data, while preserving the Markov property for decision learning and preventing bias toward temporally clustered fault patterns.

The agent ultimately learns an optimal policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$ that maximizes the expected cumulative discounted reward under the cost-sensitive constraints, thereby achieving an intelligent and risk-aware fault diagnosis strategy that balances diagnostic accuracy and operational safety.

3.3. Prioritized Experience Replay

To enhance sample efficiency and improve the stability of reinforcement learning in fault diagnosis, the proposed TBDDQN framework integrates the PER mechanism [25]. Unlike conventional experience replay, which samples past interactions uniformly from the replay buffer, PER assigns higher sampling probabilities to experiences with greater learning potential, thereby allowing the agent to focus more on transitions that are critical for value estimation improvement.

In standard experience replay, all stored transitions are treated equally, leading to inefficient use of valuable experiences and potentially slowing convergence. To address this limitation, PER evaluates the importance of each transition based on its temporal-difference (TD) error, defined as:

$$p_i = \left(|\delta_i| + \epsilon\right)^{\alpha} \tag{12}$$

where δ_i denotes the TD error of the ith transition, ϵ is a small positive constant ensuring non-zero probability, and α controls the degree of prioritization. A higher TD error indicates a larger discrepancy between the predicted and target Q-values, implying a higher learning potential. The normalized sampling probability for each transition is computed as:

$$P(i) = \frac{p_i}{\sum_j p_j} \tag{13}$$

Since prioritized sampling introduces bias into the training distribution, importance-sampling (IS) weights are used to compensate for this effect and maintain unbiased gradient updates:

$$w_t = \frac{(NP(t))^{-\beta}}{\max_k (NP(k))^{-\beta}} \tag{14}$$

Here, N is the buffer size, $P(t)$ is the sampling probability of transition t, and β adjusts the strength of bias correction. These weights ensure that less frequently sampled transitions receive proportionally higher importance during the optimization process, promoting balanced and effective learning.

To further stabilize the training process, the proposed framework incorporates several reinforcement learning techniques in conjunction with PER. Specifically, a soft target-network update [36] is employed to prevent abrupt changes in Q-value estimation:

$$\bar{\theta} \leftarrow \tau\,\theta + (1-\tau)\,\bar{\theta} \tag{15}$$

where $\tau \ll 1$ denotes the soft update coefficient controlling the rate of parameter blending. Meanwhile, an exponentially decaying ϵ-greedy exploration strategy [21] dynamically balances exploration and exploitation as training progresses, while gradient clipping [37] prevents gradient explosion and improves numerical stability.

Through the integration of PER, soft target updates, and adaptive exploration, the TBDDQN agent achieves a robust learning process with enhanced convergence efficiency and fault diagnosis accuracy. The overall training procedure is summarized in Algorithm 1.

Algorithm 1 Cost-Sensitive Reinforcement Learning Environment for BFIP Diagnosis

Require: Training subset $\mathcal{D}_{\text{train}}$, base cost matrix $\mathbf{C}$, adaptive flag adapt_flag
Ensure: Environment state s_t, reward r_t

Initialize:
  Randomly select index $i \leftarrow \text{rand}(|\mathcal{D}_{\text{train}}|)$
  $\mathbf{C}_{\text{base}} \leftarrow \mathbf{C}$, $\mathbf{C}_{\text{adapt}} \leftarrow \mathbf{C}_{\text{base}}$
  error_counts $\leftarrow \mathbf{0}$, total_counts $\leftarrow \mathbf{0}$
  Small constant $\epsilon \leftarrow 10^{-8}$
procedure RESET
  $i \leftarrow \text{rand}(|\mathcal{D}_{\text{train}}|)$
  $s_0 \leftarrow \mathcal{D}_{\text{train}}[i][0]$ ▷ Return initial state sample
  return s_0
end procedure
procedure STEP(a_t)
  Obtain true label $l_t \leftarrow \mathcal{D}_{\text{train}}[i][1]$
  if $a_t = l_t$ **then**
    $r_t \leftarrow 1.0$ ▷ Correct classification reward
    total_counts$[l_t] \leftarrow$ total_counts$[l_t] + 1$
  else
    total_counts$[l_t] \leftarrow$ total_counts$[l_t] + 1$
    error_counts$[l_t, a_t] \leftarrow$ error_counts$[l_t, a_t] + 1$
    if adapt_flag = True **then**
      $\text{err_ratio} \leftarrow \dfrac{\text{error_counts}[l_t, a_t]}{\text{total_counts}[l_t] + \epsilon}$
      $\mathbf{C}_{\text{adapt}}[l_t, a_t] \leftarrow \mathbf{C}_{\text{base}}[l_t, a_t] \times (1 + \text{err_ratio})$
    end if
    $r_t \leftarrow -\mathbf{C}_{\text{adapt}}[l_t, a_t]$ ▷ Penalty based on cost sensitivity
  end if
  $i \leftarrow \text{rand}(|\mathcal{D}_{\text{train}}|)$
  $s_{t+1} \leftarrow \mathcal{D}_{\text{train}}[i][0]$
  return (r_t, s_{t+1})
end procedure

3.4. Proposed Imbalanced Fault Diagnosis via TBDDQN

The proposed TBDDQN-based imbalanced fault diagnosis framework is designed to achieve accurate and stable classification in the BFIP, where the inherent class imbalance

and strong spatiotemporal dependencies pose major challenges. As shown in Figure 4, the framework integrates Transformer-based spatial encoding, BiLSTM-attention temporal modeling, PER, and a CSMDP into a unified reinforcement learning paradigm.

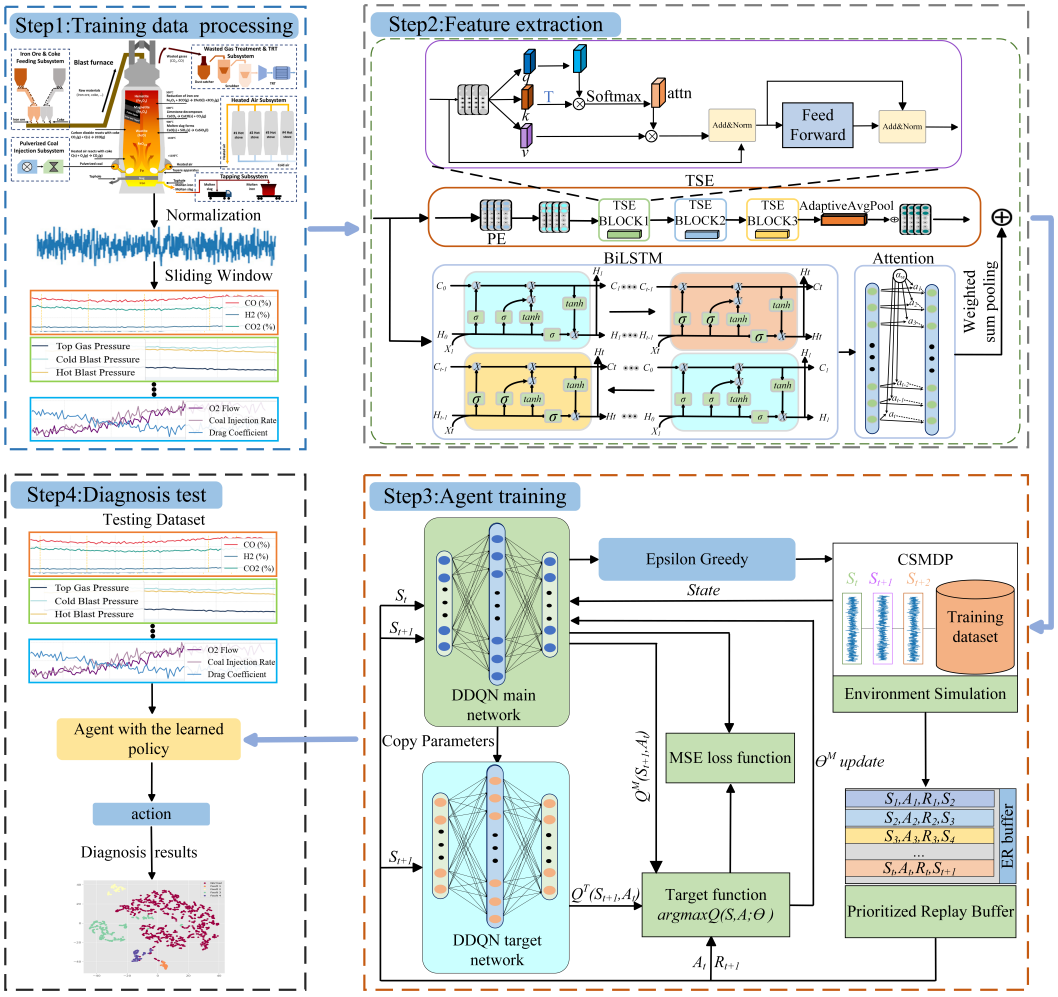


Figure 4. Flowchart of the proposed TBDDQN-based imbalanced fault diagnosis framework.

Step 1 Data preprocessing. To preserve temporal continuity, each sample is constructed by applying a sliding window to consecutive process measurements. All variables are standardized to eliminate scale discrepancies.

Step 2 Spatiotemporal feature representation. The preprocessed sequence is simultaneously fed into two parallel encoding branches. The Transformer-based spatial encoder captures global inter-variable correlations and dependencies, while the BiLSTM-attention branch learns bidirectional temporal dynamics within each window. The feature embeddings from both branches are concatenated to form a unified spatiotemporal representation, which is subsequently passed through fully connected layers to estimate Q-values corresponding to candidate fault categories.

Step 3 Adaptive learning under cost-sensitive MDP. To address class imbalance, the learning process is formulated as a CSMDP. In this setting, rewards are adaptively scaled according to class-specific misclassification costs, thereby mitigating bias toward majority classes. Meanwhile, PER is incorporated to improve sampling efficiency by emphasizing transitions with higher TD errors. Together, these mechanisms enable the agent to focus on rare but informative fault patterns. A DDQN strategy is adopted to decouple action selection and evaluation, thereby reducing overestimation bias. The target network parameters are updated softly as $\bar{\theta} \leftarrow \tau\theta + (1-\tau)\bar{\theta}$ to ensure stable convergence. Moreover,

ϵ-greedy exploration with exponential decay maintains a balance between exploration and exploitation during training.

Step 4 Fault classification. Once training is complete, the agent operates in a greedy mode, selecting the action with the maximum estimated Q-value for each input sequence. The resulting policy enables accurate and balanced fault identification across both minority and majority fault classes, demonstrating the effectiveness of the TBDDQN framework in handling imbalanced industrial diagnosis tasks.

4. Case Study

To evaluate the capability and effectiveness of the proposed method, datasets were collected from the No. 2 blast furnace system at Guangxi Liuzhou Iron and Steel Company Ltd., located in Southern China. The datasets reflect the typical characteristics of BFIP, including strongly coupled multivariate variables, complex operating conditions, and highly imbalanced fault occurrences. Based on the expert knowledge, 10 critical process variables were selected for fault diagnosis as listed in Table 1. The proposed TBDDQN framework is developed with these blast furnace-specific data characteristics in mind, aiming to address the challenges of imbalanced fault diagnosis and asymmetric misclassification risks in BFIP. Although the dataset cannot be publicly released due to industrial confidentiality constraints, the methodological design is applicable to other industrial processes with similar characteristics.

Table 1. Monitored Variables for BFIP Fault Diagnosis.

No.	Variable (Unit)
V1	CO Volume (%)
V2	H_2 Volume (%)
V3	CO_2 Volume (%)
V4	Enriching oxygen flow (m^3/h)
V5	BF top gas pressure (MPa)
V6	Cold blast pressure (MPa)
V7	Total pressure drop (MPa)
V8	Hot blast pressure (MPa)
V9	Drag coefficients
V10	Actual coal injection rate (ton/h)

According to the operational log, a total of 4343 samples were collected across 5 distinct operational scenarios, comprising 3018 normal cases, 125 furnace collapse incidents, 216 instances of furnace temperature rise, 739 channeling events, and 245 cases of furnace temperature down. These faults correspond to distinct operational risks in BFIP. Furnace collapse incidents may lead to burden instability and emergency shutdowns; furnace temperature rise affect hot metal quality and refractory lifespan; channeling events result in uneven gas distribution and reduced reaction efficiency while furnace temperature down is often related to abnormal fuel injection or blast conditions, directly impacting productivity and energy efficiency. The severe consequences and highly imbalanced occurrence frequencies of these faults further motivate the adoption of cost-sensitive decision modeling. The dataset was split into training and testing sets in an 8:2 ratio.

4.1. Parameter Settings

For comparative analysis, several baseline models are employed, including CNN, LSTM, CNN-LSTM, CNN-DQN, and CNN-LSTM-DQN. The baseline models are configured as follows:

- CNN model: Four convolutional layers with channel sizes of 1, 4, 16, and 32, followed by two fully connected layers with 100 and 5 neurons, respectively, for spatial feature compression and classification.
- LSTM model: A single LSTM with 64 hidden units, where the final hidden state is used to capture temporal dynamics.
- CNN-LSTM model: A hybrid framework integrating two convolutional layers (with 4 and 8 channels) as the CNN feature extractor and an LSTM with 64 hidden units.
- CNN-DQN model: A CNN backbone comprising two 1D convolutional layers (64 and 128 channels) and a fully connected layer with 256 neurons for Q-value prediction.
- CNN-LSTM-DQN model: Extends the CNN-DQN by adding a parallel LSTM branch with 128 hidden units to capture temporal dependencies.

All baseline models are trained for 100 epochs using the Adam optimizer with a learning rate of 0.001. For the DQN-based models, the discount factor is set to $\gamma = 0.99$, and an ϵ-greedy strategy is employed. Each epoch contains 512 steps, and the target networks are updated every 1000 steps to maintain training stability.

In the proposed TBDDQN model, the sliding window size is set to 10. In the Transformer encoder branch, the input is projected to 256 dimensions using a linear layer, combined with learnable positional encodings, and processed by three Transformer encoder layers with eight attention heads each, generating 256-dimensional spatial features. In the LSTM branch, a single-layer, bidirectional LSTM with a hidden size of 128 is used, enhanced with an attention mechanism that adaptively weights all time steps, producing a 256-dimensional temporal feature vector. These representations are concatenated to form a 512-dimensional feature vector (256 from Transformer + 256 from BiLSTM), which is subsequently passed through a fully connected layer with five outputs corresponding to the action space.

The discount factor is set to $\gamma = 0.99$, and an ϵ-greedy exploration strategy is adopted, where $\epsilon_{\text{start}} = 1.0$ and $\epsilon_{\text{end}} = 0.01$ with a decay rate of 0.05. The Adam optimizer is used with a learning rate of 0.001. Training is performed over 100 epochs with 512 steps per epoch, and the target network update rate is set to $\tau = 0.001$. For prioritized experience replay, the buffer capacity is set to 10,000, with prioritization parameters $\alpha = 0.6$ and $\beta = 0.4$ (annealed to 1.0 during training). The cost-sensitive reward is constructed using the following base cost matrix:

$$\mathbf{C} = \begin{bmatrix} 0.0 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.0 & 0.2 & 0.2 & 0.1 \\ 0.1 & 0.2 & 0.0 & 0.3 & 0.2 \\ 0.1 & 0.2 & 0.3 & 0.0 & 0.2 \\ 0.1 & 0.1 & 0.2 & 0.2 & 0.0 \end{bmatrix}$$

To further refine the reward design, an adaptive penalty adjustment mechanism is introduced, defined as Equation (11) which adaptively penalizes frequently misclassified actions based on historical error statistics.

4.2. Results Analysis

This subsection evaluates the diagnostic performance using precision, recall, F1-score, t-SNE visualizations, and confusion matrices to demonstrate TBDDQN's effectiveness in handling imbalance. In multi-class classification tasks, the performance metrics such as precision, recall, and F1-score are calculated for each class and then aggregated to evaluate the overall performance. To assess the fault diagnosis capability of different models, we employ two widely adopted aggregation strategies, macro-averaging and weighted-averaging, as discussed by Sokolova and Lapalme [38]. The macro-averaged treats all classes equally, providing a balanced view across classes, while the weighted-averaged

considers the support (sample size) of each class, reflecting the model's performance under class imbalance.

Macro-averaging treats all classes equally by computing the unweighted mean across all classes, providing a fair reflection of performance on minority fault categories. In contrast, weighted-averaging accounts for class imbalance by computing the average weighted by the number of instances per class, which better reflects the model's overall performance dominated by frequent classes. The corresponding formulations are expressed as:

$$\text{Macro-averaged}(M) = \frac{1}{K}\sum_{k=1}^{K} M_k, \tag{16}$$

$$\text{Weighted-averaged}(M) = \frac{\sum_{k=1}^{K} n_k M_k}{\sum_{k=1}^{K} n_k}, \tag{17}$$

where M_k represents the metric (precision, recall, or F1-score) of class k, n_k is the number of samples in class k, and K denotes the total number of classes.

Tables 2–4 present the diagnostic performance in terms of precision, F1-score, and recall, respectively. As shown in Table 2, the CNN model achieves a weighted-averaged precision of 0.873 and a macro-averaged precision of 0.835. LSTM improves both metrics to 0.902 and 0.835, respectively. The CNN-LSTM model, by combining spatial and temporal feature extraction, further enhances performance, achieving a weighted-averaged precision of 0.931 and a macro-averaged precision of 0.865.

Reinforcement learning-based models exhibit more significant improvements. CNN-DQN attains a weighted-averaged precision of 0.929 and a macro-averaged precision of 0.899, indicating its ability to handle class imbalance effectively. The CNN-LSTM-DQN model achieves slightly higher results, with a weighted-averaged precision of 0.925 and a macro-averaged precision of 0.903. In contrast, the proposed TBDDQN achieves the best performance, with a weighted-averaged precision of 0.934 and a macro-averaged precision of 0.970, clearly outperforming all other baselines.

Table 2. Precision diagnosis performance of the comparative models.

Fault Type	CNN	LSTM	CNN-LSTM	CNN-DQN	CNN-LSTM-DQN	TBDDQN
Normal	0.862	0.906	**0.935**	0.927	0.918	0.910
Fault 1 (Collapse)	0.714	0.686	0.625	0.656	0.774	**1.000**
Fault 2 (Furnace temperature rise)	**1.000**	0.978	**1.000**	**1.000**	**1.000**	**1.000**
Fault 3 (Channeling)	**1.000**	**1.000**	**1.000**	0.952	0.983	**1.000**
Fault 4 (Furnace temperature down)	0.600	0.606	0.766	**0.957**	0.840	0.939
Macro-averaged precision	0.835	0.835	0.865	0.899	0.903	**0.970**
Weighted-averaged precision	0.873	0.902	0.931	0.929	0.925	**0.934**

Note: Bold values indicate the best performance in each row.

Table 3. F1-score diagnosis performance of the comparative models.

Fault Type	CNN	LSTM	CNN-LSTM	CNN-DQN	CNN-LSTM-DQN	TBDDQN
Normal	0.904	0.923	0.944	**0.959**	0.948	0.951
Fault 1 (Collapse)	0.652	0.800	0.769	0.737	0.857	**1.000**
Fault 2 (Furnace temperature rise)	**0.989**	**0.989**	0.914	0.646	0.761	0.976
Fault 3 (Channeling)	0.577	0.792	0.866	**0.949**	0.868	0.768
Fault 4 (Furnace temperature down)	0.744	0.696	0.867	0.611	0.848	**0.948**
Macro-averaged F1-score	0.773	0.840	0.872	0.780	0.856	**0.929**
Weighted-averaged F1-score	0.836	0.888	0.920	0.916	0.917	**0.922**

Note: Bold values indicate the best performance in each row.

Table 4. Recall diagnosis performance of the comparative models.

Fault Type	CNN	LSTM	CNN-LSTM	CNN-DQN	CNN-LSTM-DQN	TBDDQN
Normal	0.950	0.940	0.954	0.993	0.980	**0.995**
Fault 1 (Collapse)	0.600	0.960	**1.000**	0.840	0.960	**1.000**
Fault 2 (Furnace temperature rise)	0.977	**1.000**	0.841	0.477	0.614	0.952
Fault 3 (Channeling)	0.405	0.655	0.764	**0.946**	0.777	0.623
Fault 4 (Furnace temperature down)	0.980	0.816	**1.000**	0.449	0.857	0.958
Macro-averaged recall	0.783	0.874	**0.912**	0.741	0.838	0.906
Weighted-averaged recall	0.851	0.889	0.920	0.924	0.920	**0.928**

Note: Bold values indicate the best performance in each row.

Similarly, Table 4 shows that TBDDQN achieves the highest weighted-averaged recall of 0.928 and macro-averaged recall of 0.906, respectively, reflecting a comparative advantage in sensitivity and robustness to class imbalance. The CNN-LSTM model also demonstrates high recall, with a weighted-averaged recall of 0.920 and a macro-averaged recall of 0.912, verifying the benefit of fusing temporal-spatial dependencies. As illustrated in Table 3, TBDDQN also exhibits the best overall F1-score, achieving a weighted-averaged F1-score of 0.922 and a macro-averaged F1-score of 0.929. These results indicate that TBDDQN effectively balances precision and recall while maintaining stability across both major and minor fault classes.

Furthermore, in terms of computational complexity, the proposed TBDDQN contains approximately 745,478 trainable parameters (0.75M). The forward pass complexity is 0.010 GFLOPs. Training on the BFIP dataset (approximately 3474 training samples) for 100 epochs takes about 20 min on an NVIDIA RTX 3080 GPU. The average inference latency is 1.75 ms on GPU and 3.48 ms on CPU per sample, which is suitable for real-time monitoring in industrial control systems. Compared with baseline models, TBDDQN provides a favorable trade-off between diagnostic accuracy and computational efficiency.

To further verify the representational capability, t-distributed stochastic neighbor embedding (t-SNE) is used to visualize the features learned by different models. As shown in Figure 5, features extracted by TBDDQN display clearer cluster separations than those of CNN, LSTM, CNN-LSTM, CNN-DQN, and CNN-LSTM-DQN. The improved discriminability indicates that the dual-branch structure and adaptive cost-sensitive reward mechanism enable the model to learn more distinct class boundaries in imbalanced industrial scenarios.

Figure 6 presents the confusion matrices of the six models across five operational states. The diagonal elements represent the precision for each class, indicating the proportion of correctly classified samples among all samples predicted as that class, while the off-diagonal elements illustrate the misclassification patterns between different classes. Several critical observations can be made as follows. First, TBDDQN demonstrates effective rare-fault detection, achieving 100% precision for both Fault 2 (Furnace temperature rise) and Fault 3 (Channeling). This validates the effectiveness of the adaptive cost-sensitive reward design in addressing severe class imbalance. Second, CNN-LSTM shows an improvement in capturing temporal-spatial correlations but still exhibits a precision drop for Fault 1 (Collapse), suggesting insufficient robustness under data scarcity. Lastly, CNN-LSTM-DQN achieves good balance but exhibits slightly lower performance on certain rare faults due to limited samples. Overall, the proposed TBDDQN achieves the highest precision consistency across all fault categories, confirming its strong performance in intelligent blast furnace fault diagnosis.

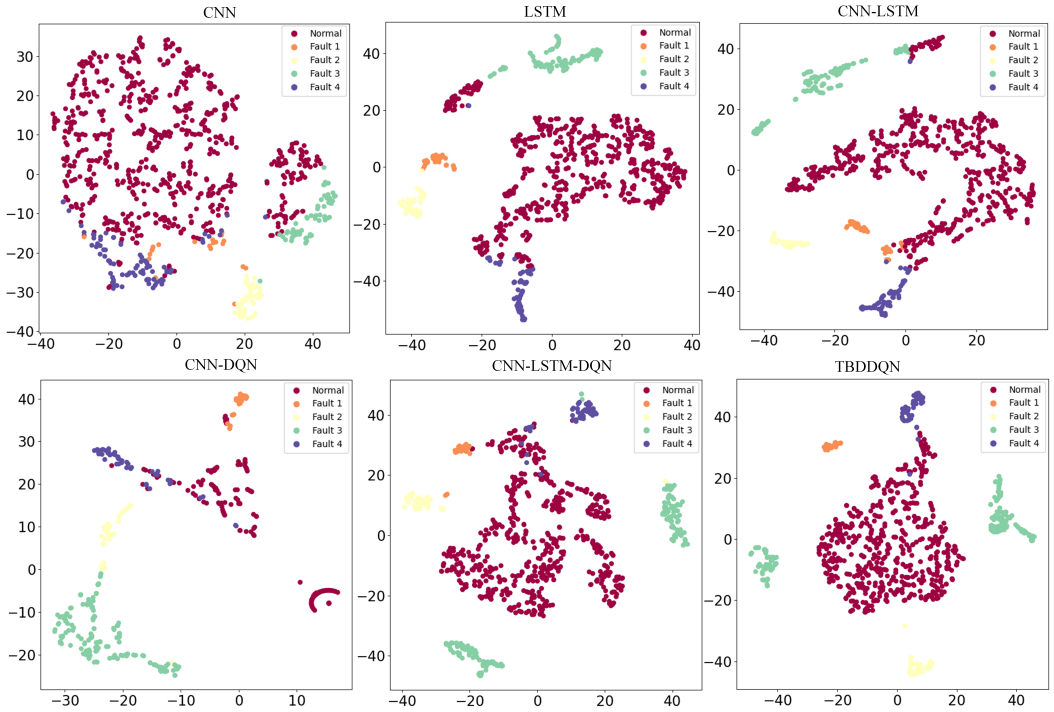


Figure 5. t-SNE visualization of the comparative models.

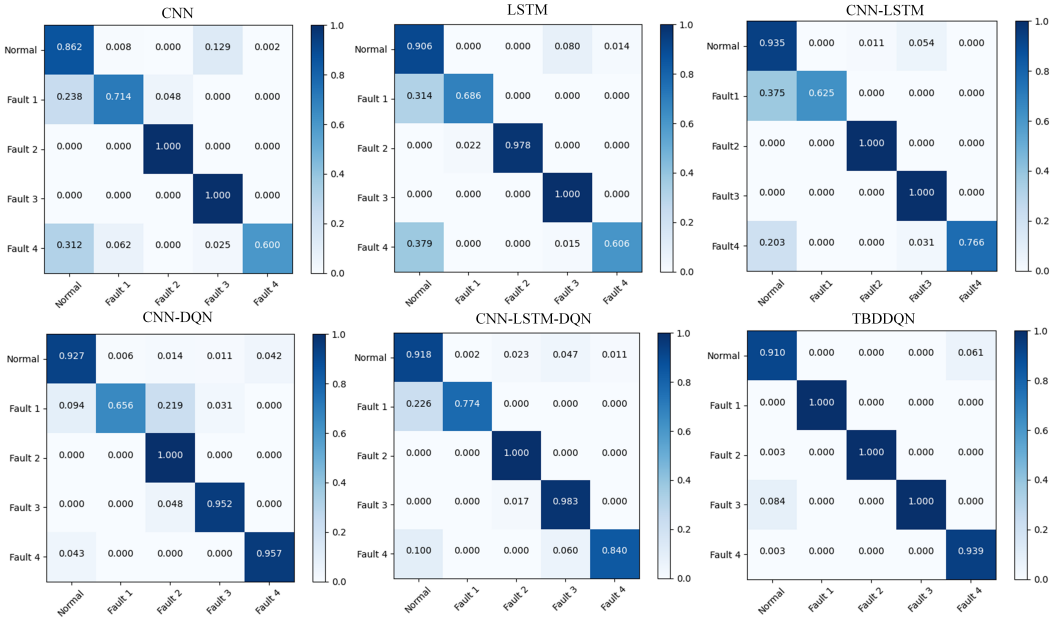


Figure 6. Confusion matrices of the comparative diagnostic models.

4.3. Ablation Study

To thoroughly investigate the contribution of each key component in the proposed TBDDQN framework, we conducted a comprehensive ablation study. Five variants were evaluated: (1) the full TBDDQN model, (2) w/o BiLSTM-attention (Transformer branch only), (3) w/o Transformer (BiLSTM-attention branch only), (4) w/o PER (replaced with uniform replay buffer), and (5) w/o Adaptive Cost (fixed cost matrix). Performance was assessed using per-class recall, F1-score, and precision, with macro-averaged and weighted-averaged metrics reported to account for class imbalance in blast furnace fault diagnosis. The results are summarized in Tables 5–7.

Table 5. Precision diagnosis performance of the ablated models.

Fault Type	w/o BiLSTM-Attention	w/o Transformer	w/o PER	w/o Adaptive Cost	TBDDQN
Normal	**0.940**	0.904	0.891	0.908	0.910
Fault 1 (Collapse)	0.500	0.920	0.960	0.889	**1.000**
Fault 2 (Furnace temperature rise)	0.872	**1.000**	**1.000**	0.977	**1.000**
Fault 3 (Channeling)	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
Fault 4 (Furnace temperature down)	0.710	0.887	**0.939**	0.922	**0.939**
Macro-averaged precision	0.804	0.942	0.958	0.939	**0.970**
Weighted-averaged precision	0.922	0.925	0.919	0.927	**0.934**

Note: Bold values indicate the best performance in each row.

Table 6. F1-score diagnosis performance of the ablated models.

Fault Type	w/o BiLSTM-Attention	w/o Transformer	w/o PER	w/o Adaptive Cost	TBDDQN
Normal	0.948	0.944	0.940	0.947	**0.951**
Fault 1 (Collapse)	0.571	0.939	0.980	0.941	**1.000**
Fault 2 (Furnace temperature rise)	0.921	0.963	0.909	0.988	**0.976**
Fault 3 (Channeling)	**0.975**	0.752	0.719	0.741	0.768
Fault 4 (Furnace temperature down)	0.557	0.931	0.948	**0.949**	0.948
Macro-averaged F1-score	0.795	0.906	0.899	0.913	**0.929**
Weighted-averaged F1-score	0.919	0.911	0.902	0.914	**0.922**

Note: Bold values indicate the best performance in each row.

Table 7. Recall diagnosis performance of the ablated models.

Fault Type	w/o BiLSTM-Attention	w/o Transformer	w/o PER	w/o Adaptive Cost	TBDDQN
Normal	0.957	0.987	0.993	0.988	**0.995**
Fault 1 (Collapse)	0.667	0.958	**1.000**	**1.000**	**1.000**
Fault 2 (Furnace temperature rise)	0.976	0.929	0.833	**1.000**	0.952
Fault 3 (Channeling)	**0.952**	0.603	0.562	0.589	0.623
Fault 4 (Furnace temperature down)	0.458	0.979	0.958	**0.979**	0.958
Macro-averaged recall	0.802	0.891	0.869	**0.911**	0.906
Weighted-averaged recall	0.921	0.918	0.911	0.921	**0.928**

Note: Bold values indicate the best performance in each row.

As shown in Table 5, the complete TBDDQN model achieves the highest macro-averaged precision of 0.970, outperforming all ablated variants. Removing the BiLSTM-attention branch results in the most pronounced degradation, with macro-averaged precision dropping sharply to 0.804. In particular, the precision for Fault 1 (Collapse) decreases sharply from 1.000 to 0.500, indicating that the BiLSTM-attention module is crucial for capturing discriminative temporal patterns associated with sudden and high-risk collapse events. In contrast, removing the Transformer branch leads to a moderate decline in macro-averaged precision to 0.942, highlighting the importance of global inter-variable dependency modeling provided by the Transformer encoder. Additionally, both the w/o PER and w/o Adaptive Cost variants exhibit noticeable reductions in macro-averaged precision and weighted-averaged precision, confirming the effectiveness of prioritized experience replay in emphasizing rare fault samples and adaptive cost-sensitive rewards in penalizing high-risk misclassifications.

Table 6 reports the F1-score comparison among different variants. The full TBDDQN achieves the highest macro-averaged F1-score of 0.929 and weighted-averaged F1-score of 0.922, demonstrating a well-balanced diagnostic capability across all fault categories. Removing the BiLSTM-attention branch again causes the most severe performance drop, with macro-averaged F1-score decreasing by 0.134 to 0.795. Notably, the F1-score for Fault 1 (Collapse) drops from 1.000 to 0.571, and for Fault 4 (Furnace temperature down) from 0.948 to 0.557, underscoring the critical role of attention-enhanced temporal modeling

in distinguishing subtle and imbalanced fault signatures. Excluding the Transformer branch also leads to a clear decline in macro-averaged F1-score to 0.906, confirming its effectiveness in capturing long-range dependencies within the sliding time window. Furthermore, disabling prioritized experience replay reduces the macro-averaged F1-score to 0.899, while replacing adaptive cost-sensitive rewards with fixed penalties yields a macro-averaged F1-score of 0.913, demonstrating that both mechanisms contribute to improving learning efficiency and mitigating bias toward the majority class.

The recall comparison in Table 7 further corroborates these observations. The complete TBDDQN achieves the highest weighted-averaged recall of 0.928, proving to be the most effective approach in achieving overall fault coverage. Removing the BiLSTM-attention branch leads to the most significant degradation, with macro-averaged recall dropping to 0.802 and the recall for Fault 4 (Furnace temperature down) decreasing sharply from 0.958 to 0.458. This confirms that the BiLSTM-attention mechanism is indispensable for capturing critical temporal dependencies and long-term fault evolution patterns in multivariate time-series data. Conversely, removing the Transformer branch severely affects the detection of Fault 3 (Channeling), where recall drops to 0.603, demonstrating the Transformer's advantage in modeling global contextual relationships across strongly coupled process variables. The effects of prioritized experience replay and adaptive cost-sensitive rewards are also evident. Eliminating PER reduces the weighted-averaged recall to 0.911, indicating that prioritizing high-error and rare fault transitions is beneficial for improving fault coverage under class imbalance. Although removing the adaptive cost mechanism yields a slightly higher macro-averaged recall, the full TBDDQN achieves the highest weighted-averaged recall, which is more representative of real-world diagnostic risk. This demonstrates that the adaptive cost-sensitive reward effectively emphasizes safety-critical and high-frequency fault misclassifications during training, thereby improving the overall robustness of the diagnosis system under imbalanced industrial conditions.

Overall, the results in Tables 5–7 consistently demonstrate that each component of the proposed TBDDQN framework contributes positively to fault diagnosis performance. The Transformer branch enhances global dependency modeling, the BiLSTM-attention module captures discriminative temporal dynamics, prioritized experience replay improves learning efficiency for rare faults, and the adaptive cost-sensitive reward mechanism effectively addresses asymmetric misclassification risks. Their synergistic integration enables TBDDQN to achieve well-balanced diagnostic performance that surpasses that of any individual component alone.

5. Conclusions

This paper proposes a novel dual-branch DDQN, termed TBDDQN, to address the challenge of imbalanced fault diagnosis in the BFIP. The proposed framework integrates a Transformer-based spatial encoder and a BiLSTM-attention-based temporal extractor to jointly capture spatiotemporal dependencies from multivariate time-series data. To mitigate the adverse effects of class imbalance and asymmetric misclassification costs, a cost-sensitive reinforcement learning strategy is introduced, incorporating prioritized experience replay and adaptive penalty mechanisms to dynamically optimize the diagnostic policy. Conceptually, the TBDDQN framework advances the field by shifting from traditional supervised classification to a dynamic, reward-optimized decision-making paradigm that better aligns with real-world industrial requirements, such as handling rare faults and minimizing high-cost errors. This approach not only improves overall diagnostic performance but also provides interpretable insights into process behaviors, facilitating better operational understanding and risk management in metallurgical processes.

https://doi.org/10.3390/machines14030276

Despite these advances, several promising research directions warrant further investigation. First, while the BiLSTM-attention mechanism captures temporal dependencies, it may struggle with very long-term patterns. Future work could explore multi-scale attention mechanisms to better capture long-range dependencies across different time horizons. Second, the interpretable feature maps generated by the model have the potential to be used for root cause analysis. Integrating graph-based or causal inference techniques could leverage these representations to trace fault propagation pathways and develop diagnostic systems with causal reasoning capabilities. Finally, extending the current framework to an online learning setting, where the diagnostic agent continuously adapts to newly arriving process data, would be a significant step toward practical deployment in dynamic industrial environments.

Author Contributions: J.Z.: Writing—original draft, Software, Data curation, Validation. P.W.: Writing—review and editing, Methodology, Supervision, Conceptualization, Funding acquisition. R.Z.: Software, Data curation, Validation. X.S.: Software, Data curation, Validation. Y.L.: Writing—review and editing. N.K.: Writing—review and editing, Validation. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Fundamental Research Funds of Zhejiang Sci-Tech University (25222139-Y).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no potential conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BFIP	Blast Furnace Ironmaking Process
PCA	Principal Component Analysis
SVM	Support Vector Machine
DL	Deep Learning
LSTM	Long Short-Term Memory
BiLSTM	Bidirectional Long Short-Term Memory
CNN	Convolutional Neural Network
DQN	Deep Q-Network
DDQN	Double Deep Q-Network
RL	Reinforcement Learning
MDP	Markov Decision Process
PER	Prioritized Experience Replay
CMDP	Classification Markov Decision Process
RNN	Recurrent Neural Network
ReLU	Rectified Linear Unit
CSMDP	Cost-Sensitive Markov Decision Process
TD	Temporal-Difference
ER	Experience Replay
IS	Importance-Sampling
MSE	Mean Squared Error
t-SNE	t-distributed Stochastic Neighbor Embedding

References

1. Lou, S.; Yang, C.; Zhang, X.; Zhang, H.; Wu, P. Data/mechanism hybrid-driven modeling of blast furnace smelting system and global sequential optimization. *J. Process Control* **2024**, *139*, 103235. [CrossRef]
2. Takahashi, K.; Nouchi, T.; Sato, M.; Ariyama, T. Perspective on progressive development of oxygen blast furnace for energy saving. *ISIJ Int.* **2015**, *55*, 1866–1875. [CrossRef]

3. Lu, B.; Wang, S.; Tang, K.; Chen, D. An all-factors analysis approach on energy consumption for the blast furnace iron making process in iron and steel industry. *Processes* **2019**, *7*, 607. [CrossRef]
4. Venkatasubramanian, V.; Rengaswamy, R.; Yin, K.; Kavuri, S.N. A review of process fault detection and diagnosis: Part I: Quantitative model-based methods. *Comput. Chem. Eng.* **2003**, *27*, 293–311. [CrossRef]
5. Gao, Z.; Cecati, C.; Ding, S.X. A survey of fault diagnosis and fault-tolerant techniques—Part I: Fault diagnosis with model-based and signal-based approaches. *IEEE Trans. Ind. Electron.* **2015**, *62*, 3757–3767. [CrossRef]
6. Abhale, P.B.; Viswanathan, N.N.; Saxén, H. Numerical modelling of blast furnace–Evolution and recent trends. *Miner. Process. Extr. Metall.* **2020**, *129*, 166–183. [CrossRef]
7. Hou, Y.; Mi, J.; Lu, J.; Wu, P.; Yang, S.; Huang, B.; Huang, W.; Wu, D. A Deep Transferable Convolution Neural Network for Unsupervised Cross-Domain Diagnostics of Rolling Bearings in Industrial Internet of Things. *IEEE Internet Things J.* **2025**, *12*, 17316–17332. [CrossRef]
8. Zhou, B.; Ye, H.; Zhang, H.; Li, M. Process monitoring of iron-making process in a blast furnace with PCA-based methods. *Control Eng. Pract.* **2016**, *47*, 1–14. [CrossRef]
9. Zhang, T.; Ye, H.; Wang, W.; Zhang, H. Fault diagnosis for blast furnace ironmaking process based on two-stage principal component analysis. *ISIJ Int.* **2014**, *54*, 2334–2341. [CrossRef]
10. Tian, H.; Wang, A. A novel fault diagnosis system for blast furnace based on support vector machine ensemble. *ISIJ Int.* **2010**, *50*, 738–742. [CrossRef]
11. Liu, L.; Wang, A.; Sha, M.; Sun, X.; Li, Y. Optional SVM for fault diagnosis of blast furnace with imbalanced data. *ISIJ Int.* **2011**, *51*, 1474–1479. [CrossRef]
12. Pan, Y.; Yang, C.; An, R.; Sun, Y. Robust principal component pursuit for fault detection in a blast furnace process. *Ind. Eng. Chem. Res.* **2018**, *57*, 283–291. [CrossRef]
13. Qian, J.; Song, Z.; Yao, Y.; Zhu, Z.; Zhang, X. A review on autoencoder based representation learning for fault detection and diagnosis in industrial processes. *Chemom. Intell. Lab. Syst.* **2022**, *231*, 104711. [CrossRef]
14. Jang, K.; Hong, S.; Kim, M.; Na, J.; Moon, I. Adversarial autoencoder based feature learning for fault detection in industrial processes. *IEEE Trans. Ind. Inform.* **2021**, *18*, 827–834. [CrossRef]
15. Wang, Y.; Yang, H.; Yuan, X.; Shardt, Y.A.; Yang, C.; Gui, W. Deep learning for fault-relevant feature extraction and fault classification with stacked supervised auto-encoder. *J. Process Control* **2020**, *92*, 79–89. [CrossRef]
16. Yu, F.; Liu, J.; Liu, D.; Wang, H. Supervised convolutional autoencoder-based fault-relevant feature learning for fault diagnosis in industrial processes. *J. Taiwan Inst. Chem. Eng.* **2022**, *132*, 104200. [CrossRef]
17. Zhu, J.; Jiang, M.; Liu, Z. Fault detection and diagnosis in industrial processes with variational autoencoder: A comprehensive study. *Sensors* **2021**, *22*, 227. [CrossRef]
18. Xu, S.; Pan, D.; Chen, Z.; Fang, Y.; Qiu, X.; Jiang, Z. Abnormal Condition Recognition of Blast Furnace Ironmaking Process Based on Time Series—Image Jointly Driven Deep Neural Network. *IEEE Trans. Ind. Inform.* **2025**, *21*, 6198–6209. [CrossRef]
19. Graves, A.; Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw.* **2005**, *18*, 602–610. [CrossRef]
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
21. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 1998; Volume 1.
22. Nian, R.; Liu, J.; Huang, B. A review on reinforcement learning: Introduction and applications in industrial process control. *Comput. Chem. Eng.* **2020**, *139*, 106886. [CrossRef]
23. Liu, X.; Zhao, Z.; Yang, F.; Liang, F.; Bo, L. Environment adaptive deep reinforcement learning for intelligent fault diagnosis. *Eng. Appl. Artif. Intell.* **2025**, *151*, 110783. [CrossRef]
24. Tian, J.; Yin, M.; Jiang, J. Fault self-healing: A biological immune heuristic reinforcement learning method with root cause reasoning in industrial manufacturing process. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108553. [CrossRef]
25. Schaul, T.; Quan, J.; Antonoglou, I.; Silver, D. Prioritized experience replay. *arXiv* **2015**, arXiv:1511.05952.
26. Garcıa, J.; Fernández, F. A comprehensive survey on safe reinforcement learning. *J. Mach. Learn. Res.* **2015**, *16*, 1437–1480.
27. Lou, S.; Yang, C.; Liu, Z.; Zhang, H.; Liu, C.; Wu, P. From Shallow to Deep: A Novel Correlation Network Representation Regression Framework for Modeling and Monitoring MIQ-Driven Blast Furnace Ironmaking Processes. *IEEE/CAA J. Autom. Sin.* **2026**, *13*, 281–299.
28. Wang, K.; Wu, P.; Lou, S.; Pan, H.; Gao, J. Fault detection-based multiple local manifold learning and its application to blast furnace ironmaking process. *Electronics* **2023**, *12*, 4773. [CrossRef]
29. Gao, D.; Yang, C.; Tang, X.Y.; Zhu, X.; Huang, X. Fault diagnosis of blast furnace based on incomplete multi-source domain adaptation with feature fusion. *Adv. Eng. Inform.* **2024**, *62*, 102946. [CrossRef]

30. Huang, X.; Yang, C.; Zhang, H.; Lou, S.; Gao, D.; Kong, L. Data and knowledge collaborative-driven fault identification and self-healing control action inference framework for blast furnace. *Expert Syst. Appl.* **2024**, *245*, 123040. [CrossRef]
31. Zhong, X.; Zhang, L.; Ban, H. Deep reinforcement learning for class imbalance fault diagnosis of equipment in nuclear power plants. *Ann. Nucl. Energy* **2023**, *184*, 109685. [CrossRef]
32. Wiering, M.A.; van Hasselt, H.; Pietersma, A.D.; Schomaker, L. Reinforcement learning algorithms for solving classification problems. In Proceedings of the 2011 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL), Paris, France, 11–15 April 2011; pp. 91–96. [CrossRef]
33. Wang, H.; Zhou, Z.; Yan, R.; Zhang, L. Multiscale deep attention reinforcement learning for imbalanced fault diagnosis of gearbox under multi-working conditions. In Proceedings of the 2023 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), Kuala Lumpur, Malaysia, 22–25 May 2023; pp. 1–6.
34. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
35. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450. [CrossRef]
36. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous control with deep reinforcement learning. *arXiv* **2015**, arXiv:1509.02971.
37. Pascanu, R.; Mikolov, T.; Bengio, Y. On the difficulty of training recurrent neural networks. In *Proceedings of the International Conference on Machine Learning, Atlanta, GA, USA, 16–21 June 2013*; ICML: San Diego, CA, USA, 2013; pp. 1310–1318.
38. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.