DOI: https://doi.org/10.15276/hait.08.2025.14
UDC 004.81:159.953.5

A multi-criteria scoring metric for evaluating deep learning models in bitcoin price forecasting

Mykola M. Zlobin[1)]
ORCID: https://orcid.org/0009-0000-7653-6109; mykolay.zlobin@gmail.com. Scopus Author ID: 59337918100
Volodymyr M. Bazylevych[1)]
ORCID: https://orcid.org/0000-0001-8935-446X; bazvlamar@stu.cn.ua. Scopus Author ID: 57214432127
[1)] Chernihiv Polytechnic National University, 95, St. Shevchenko. Chernihiv, 14030, Ukraine

ABSTRACT

The increasing computational demands of deep learning have raised concerns about the environmental sustainability of artificial intelligence applications, particularly in high-frequency domains such as financial forecasting. This paper addresses the need for more holistic evaluation criteria by proposing a multi-criteria scoring metric for deep learning models used in Bitcoin price forecasting. The purpose of the study is to develop a performance metric that balances predictive accuracy with computational efficiency and environmental impact. The method involves combining traditional accuracy measures with training time, energy consumption, and carbon emissions into a unified performance score, calculated using a logistic scoring function. The metric was validated by applying it to forty-two configurations of Long Short-Term Memory models trained on historical Bitcoin price data. Each configuration was assessed for its forecasting accuracy, energy use and emissions (measured using a carbon-tracking tool). The results show that simpler Long Short-Term Memory models can offer competitive accuracy while significantly reducing training time and emissions. The highest-performing model achieved a balance of all criteria, while deeper architectures with marginal accuracy gains incurred disproportionate environmental costs. The study concludes that the proposed scoring metric offers a practical and scalable solution for selecting deep learning models under sustainability constraints, supporting more responsible Artificial Intelligence deployment in real-world settings.

Keywords: Deep learning; time series forecasting; Long Short-Term Memory; Performance Metric, Sustainability

For citation: **Zlobin M. M., Bazylevych V. M. "A multi-criteria scoring metric for evaluating deep learning models in bitcoin price forecasting". *Herald of Advanced Information Technology*. 2025; Vol.8 No.2: 221–232. DOI: https://doi.org/10.15276/hait.08.2025.14**

INTRODUCTION

In recent years, the demand for accurate forecasting in volatile financial markets has increased substantially, specifically with the emergence of cryptocurrencies like Bitcoin. Bitcoin, the most capitalized digital asset in global finance and is used by over 300 million users worldwide. Due to its extreme price volatility and 24/7 trading cycle, Bitcoin presents unique challenges for investors, regulators, and algorithmic traders. Machine learning, and deep learning in particular, has shown strong promise in capturing complex patterns in such non-linear markets. However, most existing evaluations of deep learning models focus narrowly on prediction accuracy. This limited view is increasingly problematic in an era where energy efficiency and sustainability are critical concerns for Artificial Intelligence (AI) deployment. The training of deep neural networks often consumes significant computational resources, leading to considerable energy use and CO_2 emissions. These environmental costs are especially relevant for high-frequency financial forecasting, where models are retrained frequently. As such, there is growing demand for evaluation methods that account not only for prediction performance but also for environmental and operational efficiency. Such demands reflect a broader shift toward sustainable AI practices. Developing tools that balance accuracy, resource consumption, and environmental impact is thus a timely and relevant challenge. This analysis responds to this need by focusing on the case of Bitcoin price forecasting using deep learning.

In the time series forecasting, particularly within volatile financial markets such as cryptocurrency, deep learning models have been used successfully. However, the trade-off between accuracy and resource efficiency remains underexplored in practical settings. Traditional evaluation metrics such as mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE) do not capture training time, energy consumption, or environmental impact, all of which are important for sustainable AI development. Consequently, there is growing interest in developing multi-criteria performance metrics that can balance technical accuracy with resource and environmental efficiency. Originally introduced in 2009, Bitcoin has grown to be one of the major

digital assets for international finance. Over 560 million people globally own cryptocurrency as of 2024; of these, Bitcoin makes 54 % of the $3.7 trillion cryptocurrency market [1], [2], [3]. Despite its growth, Bitcoin remains highly volatile [4], [5]. Informed financial decision-making requires accounting for this volatility. In dynamic markets, accurate forecasts support timely risk mitigation and informed investment decisions.

In recent years, the cryptocurrency market has emerged as a highly dynamic with characterized by volatility and a rich, yet challenging dataset for testing forecasting models. Among cryptocurrencies, Bitcoin remains the most widely studied and traded asset, characterized by sharp price swings, limited historical data, non-linear behavior, and high sensitivity to investor sentiment and external events. For example, in a week in May 2022, Solana value dropped by 41 %, Cardano by 35 %, Bitcoin's by 20%, Ethereum by 26 %. On the other hand, during the same period, the financial assets such as the Nasdaq tech stock index and the financial times stock exchange (FTSE) 100 had smaller declines of 7.6 % and 3.6 %, respectively [6]. In addition, investor sentiment, strongly influenced by news and social media, significantly affects cryptocurrency values. Studies have indicated that higher search volumes and online debates can drive more trading activity and price volatility. However, incorporating these qualitative elements into predictive models remains challenging [7], [8]. Thirdly is the limited historical data. With Bitcoin first launched in 2009, the rather short history of cryptocurrencies offers little historical data for the paper. This lack makes it challenging to spot long-term trends necessary for reliable projection. Furthermore, the fast-changing market makes historical data less likely to always reflect future patterns. Unique market structure is positioned in fourth place. Operating 24/7, unique market structures for cryptocurrencies are distinguished by high-frequency trading and notable liquidity variances. These factors contribute to non-linear price movements and complex dependencies, setting challenges for traditional predictive models. Research highlights that positive market returns at high-frequency levels can increase price volatility, a phenomenon less common in traditional financial markets [9]. All things considered, the high volatility of cryptocurrencies, the major impact of market sentiment, lack of historical data, and special market structures define the difficulties in predicting their prices. Addressing these issues requires new modeling approaches capable of capturing the dynamic nature of cryptocurrency markets.

Notably, Long Short-Term Memory (LSTM) networks have been widely applied in predicting Bitcoin prices due to their ability to model long-range temporal relationships. However, LSTM networks are computationally intensive, often requiring significant training time and energy consumption. These characteristics raise concerns not only about model efficiency but also about their carbon footprint, especially when deployed at scale. In contrast, simpler models such as Recurrent Neural Networks (RNNs) may offer comparable predictive performance while demanding fewer computational resources. This makes the trade-off between model complexity and sustainability an important question in financial domains.

Most existing studies compare models based on accuracy alone. Only a few consider performance frameworks that also include training efficiency and environmental impact. This paper addresses this important gap by proposing a novel performance score that integrates predictive accuracy (via MAE, MSE, RMSE), training time, energy consumption, and carbon dioxide (CO_2) emissions into a unified metric. By assigning empirically tuned weights to each component, this score allows researchers to compare model architectures not just on how well they predict but also on how efficiently and sustainably they operate. The score is specifically designed to reflect real priorities where trade-offs must often be made between performance and energy/resource constraints. To validate this approach, the proposed metric was applied to multiple LSTM configurations. The models were trained on a historical Bitcoin price series aggregated from minute-level data spanning 2012 to 2025. After appropriate normalization and preprocessing, a range of LSTM architectures were trained and evaluated under different configurations, including varying numbers of layers, units, dropout rates, batch sizes, and training epochs. Metrics such as training time, energy consumed, and CO_2 emissions were measured using the CodeCarbon Python library, which estimates computational emissions from central processing unit (CPU) and memory usage. The results of this evaluation show that although deeper LSTM architectures can marginally improve predictive accuracy, they sustain significantly higher training costs and environmental impacts. In contrast, simpler architectures, e.g., single-layer LSTMs with 50 units and minimal dropout, achieved the highest composite scores, reflecting an optimal trade-off between accuracy and sustainability. These findings support the broader hypothesis that performance optimization in deep learning should extend beyond accuracy, specifically

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

in applications where sustainability is a growing concern.

1. LITERATURE ANALYSIS

Recent findings in deep learning have reshaped the landscape of cryptocurrency price forecasting. Studies done by [9], [10] provide a comprehensive comparative evaluation of convolutional neural network (CNNs), LSTM variants, and Transformer models for cryptocurrency price forecasting under different volatility regimes. Their findings show that multivariate convolutional LSTM architectures outperform univariate models in terms of RMSE and MAE, especially during turbulent periods such as the COVID-19 pandemic. Similarly, the research [10] proposed a hybrid gated recurrent unit–long short-term memory (GRU-LSTM) model with parent-coin dependency analysis, demonstrating high prediction accuracy for Litecoin and Zcash. This approach highlights the importance of capturing interdependencies between cryptocurrencies, which most standalone models overlook. The paper [11] addressed the challenge of limited historical data by using feature engineering combined with a shallow bidirectional long short-term memory (Bi-LSTM) network. Their results reveal that even low-complexity models can outperform deep architectures when feature selection is optimized. Moreover, the authors [12] proposed improving LSTM models with change point detection techniques like the pruned exact linear time (PELT) algorithm, significantly improving their adaptability to sudden market shifts. Despite these findings, many existing models focus on minimizing prediction error without accounting for training cost, energy efficiency, or environmental impact. They often lack unified metrics that consider both technical performance and sustainability, an increasingly important issue in AI deployment. This analysis will fill this gap by introducing a novel multi-criteria scoring metric that evaluates deep learning models holistically across accuracy, resource usage, and carbon footprint.

Bitcoin price forecasting increasingly relies on advanced machine learning and deep learning models. These models are designed to handle the inherent volatility and complexity of digital asset markets. One of the most frequently used deep learning models is the LSTM network. Long Short-Term Memory networks are effective in capturing temporal dependencies in sequential data for time series forecasting. The paper that used LSTM networks to forecast Bitcoin values, for example, shows that LSTMs efficiently learn underlying patterns in past price data, hence increasing prediction accuracy [13]. Furthermore, also applied for Bitcoin prediction are convolutional neural networks (CNNs). Convolutional neural networks benefit market trend analysis by allowing spatial information from data to be extracted. Convolutional neural networks models have proved to be able to anticipate the financial data of many businesses, therefore showing their potential in ranking credit and price forecasting [14]. Other examples, used to improve prediction performance are hybrid models incorporating many deep learning architectures. One instance is the combination of Extreme Gradient Boosting (XGBoost) and Transformer models. This method uses XGBoost's expertise in managing structured data and the Transformer's capacity to record long-range dependencies. Such hybrid models have been shown to have lower MAE and RMSE in Bitcoin price forecasts than baseline models [15].

As deep learning models grow in complexity, so too do their computational and environmental costs. Recent research has highlighted the significant carbon footprint and energy consumption associated with training large neural networks, especially when deployed at scale or used for high-frequency forecasting tasks such as cryptocurrency prediction. To address this, some works have begun incorporating computational efficiency metrics into their evaluations. For instance, studies have used training time, number of parameters, or floating-point operations (FLOPs) as proxies for model complexity and cost. However, these metrics are often insufficient to capture the true energy or environmental burden of training deep learning models. More efforts, like the use of the CodeCarbon Python library, allow for real-time estimation of energy consumption and CO_2 emissions during model training, offering a more precise lens for evaluating model sustainability [16], [17]. Despite these findings, the evaluation frameworks that synthesize prediction accuracy with energy efficiency and environmental impact still remain under researched enough in the literature. A research gap persists in the lack of integrated performance scoring systems that can simultaneously account for prediction error, training time, energy usage, and carbon emissions. Most existing studies either optimize for accuracy or for speed, without offering a systematic way to balance these competing factors. This limits their applicability in resource-constrained or

environmentally sensitive scenarios. This paper addresses that gap by introducing a novel multi-criteria scoring metric designed to evaluate deep learning models holistically. The proposed metric aggregates classical performance indicators (MAE, MSE, RMSE) with real-time measurements of computational efficiency (training time) and environmental sustainability (energy consumption, CO_2 emissions). The metric employs a logistic scoring function to normalize and weight each component, enabling flexible trade-off configurations aligned with specific research or deployment priorities. To validate the metric, we apply it to evaluate various LSTM configurations trained on Bitcoin price data. Bitcoin serves as a compelling case study due to its extreme price volatility, short historical window, and growing institutional relevance. By comparing model variants across both technical and environmental dimensions, this paper provides a comprehensive framework for selecting forecasting models that are not only accurate but also sustainable.

2. THE PURPOSE AND OBJECTIVES OF THE RESEARCH

The purpose of this paper is the development and research of a novel multi-criteria performance metric for evaluating deep learning models used in cryptocurrency price forecasting, with a focus on balancing predictive accuracy, computational efficiency, and environmental sustainability.

To achieve this goal, the following tasks must be solved.

1. Develop a composite performance score that integrates traditional accuracy metrics (MAE, MSE, RMSE) with training time, energy consumption, and CO_2 emissions.

2. Design a logistic scoring function capable of normalizing heterogeneous indicators into a single value ranging from 0 to 1.

3. Implement and validate the proposed metric by applying it to 42 different configurations of LSTM models trained on Bitcoin price data.

4. Quantify and analyze the trade-offs between model accuracy and environmental impact across various LSTM configurations.

5. Determine the optimal weighting scheme for the scoring metric, reflecting user-defined priorities such as speed, accuracy, and sustainability.

In terms of priority, this paper emphasizes forecasting accuracy and environmental impact as primary goals, assigning them greater importance than training time. While minimizing training time is desirable, it is treated as a secondary objective due to its one-time cost, in contrast to the lasting implications of energy use and emissions. Therefore, greater weight is assigned to MAE, MSE, and CO_2-related components in the final performance score.

3. PERFORMANCE SCORE CALCULATION

A novel performance score is defined as follows:

$$P_{score} = \frac{1}{1+e^{-x}} \tag{1}$$

$$x = \frac{\omega_1}{MAE + MSE + RMSE + \varepsilon} - \omega_2 \frac{T}{T_{\max}} - \omega_3 \frac{E}{E_{\max}} - \omega_4 \frac{C}{C_{\max}} + shift \tag{2}$$

The performance score from (1) is computed using a logistic regression function to bounds scores in range between 0 and 1 and emphasize relative differences of scores. Variable x from (1) is calculated using the proposed equation (2).

The first term $\frac{\omega_1}{MAE + MSE + RMSE + \varepsilon}$ is the accuracy term and rewards low error rates (MAE; MSE; RMSE). Here, a smaller error makes the denominator smaller, leading to a higher performance score.

The second term $\omega_2 \frac{T}{T_{\max}}$ is the time penalty and penalizes longer training times. The next one is the energy penalty $\omega_3 \frac{E}{E_{\max}}$ which penalizes higher energy usage, while $\omega_4 \frac{C}{C_{\max}}$ should suppress higher values of CO_2 emissions. Finally, the shift term is a constant added to *x* with the purpose of ensuring that scores are not too low. The variables from (2) can be introduced as follows: the weighting factors $\omega_1, \omega_2, \omega_3, \omega_4$ scale the relative importance of accuracy metrics (*MAE*, *MSE*, *RMSE*), training time (*T*), energy consumption (*E*), and CO_2 emissions (*C*) in the performance score, where *MAE*, *MSE*, and *RMSE* quantify prediction errors, while *T*, *E*, and *C* represent computational costs normalized by their maximum observed values (*Tmax*, *Emax*, *Cmax*) to ensure comparable scales. The term ε (a small constant) prevents division by zero in the accuracy term.

The proposed metric aims to balance model accuracy against resource efficiency and environmental impact, with weights like $\omega_1 = 10^2$

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

prioritizing accuracy and $\omega_4 = 10^2$ heavily penalizing emissions to align with sustainability goals. By adjusting the weights from (2), it is possible to control the trade-off between accuracy and efficiency.

In practice, the weighting factors in the performance score should be adjusted iteratively through empirical testing, where different model architectures are evaluated to assess their trade-offs between accuracy and computational efficiency. The optimal weights depend on user-defined priorities – whether higher accuracy or lower energy/emissions is more critical – and should be calibrated to align with specific research objectives, such as deploying models in energy-constrained environments or maximizing predictive performance. Since no single weighting scheme fits all use cases, this metric provides a flexible framework that can be tailored through experimentation, ensuring the final score reflects the desired balance between model performance and operational constraints while adhering to domain-specific goals.

In the next section, the utility value of the proposed metric will be practically demonstrated.

4. METHODOLOGY

The dataset used in this research is acquired using the automation tool from [18], [19] which fetches Bitcoin trading data from Bitstamp API. It includes Bitcoin trading information, including the timestamps of records, opening, highest, lowest, and closing prices of the cryptocurrency at 1-min intervals (Fig. 1). In addition, the dataset includes the traded volumes of coins and the corresponding trading volumes in USD. For the purpose of this research, the focus was made on the closing prices of BTC, for a randomly selected period of time which covers an interval between 2012 and 2025. For the selected period, each day was summarized with a single closing price as the mean value of all daily closing prices.

Next, in the preprocessing phase, the closing data is split into training and test data sets. The training data includes all the data point except for the last two months (60 days), which are utilized as the test data. To ensure that these sets have a consistent range, MinMaxScaler was utilized to scale the features, normalizing values in the range from 0 to 1. Next, to train the model correctly, sequences are created from the data using a sliding window approach to generate input features. Here, each sequence includes the previous 60-time steps (keeping in mind that the prediction task will be to forecast a 60-time-step sequence of BTC prices), while the corresponding target is the value of the next time step. Finally, the input data is transformed into a 3D array to be suitable for models to be trained.

The LSTM model which will be utilized for the simulation purposes represents an advanced form of RNN that overcomes the vanishing gradient problem. It introduces memory cells and gating mechanisms to control the flow of information. LSTM unit consists of multiple gates controlling the flow of information (Fig. 2).

Forget gate that determines which information should be removed from the cell state:

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right). \quad (3)$$

Input gate controls which information should be added to the cell state:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i). \quad (4)$$

The input modulation gate generates new candidate values to be added to the cell state:

$$g_t = \phi\left(W_g x_t + U_g h_{t-1} + b_g\right). \quad (5)$$

Cell State Update updates the cell state c_t using the forget and input gates:

$$c_t = f_t \cdot c_{t-1} + i_t \cdot g_t. \quad (6)$$

The output gate decides what the next hidden state h_t should be:

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o). \quad (7)$$

	Timestamp	Open	High	Low	Close	Volume
6711275	1.731197e+09	76706.0	76760.0	76706.0	76760.0	0.279492
6711276	1.731197e+09	76765.0	76810.0	76763.0	76810.0	0.011713
6711277	1.731197e+09	76786.0	76796.0	76779.0	76786.0	0.264600
6711278	1.731197e+09	76712.0	76722.0	76707.0	76722.0	0.250061
6711279	1.731197e+09	76707.0	76752.0	76707.0	76749.0	0.340405

Fig. 1. **A sample of initial Bitcoin trading data**
Source: **compiled by the authors**

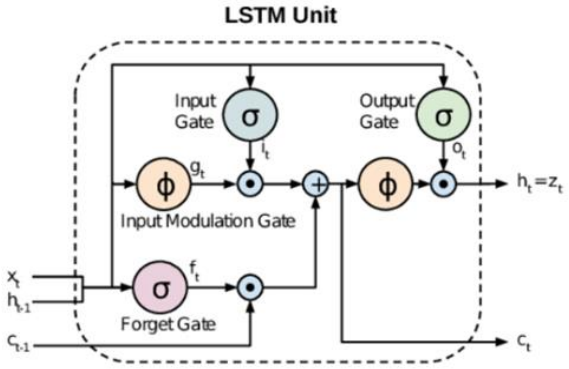

Fig. 2. **Long Short-Term Memory Unit**
Source: **compiled by the [20]**

Hidden State Update computes the final hidden state:

$$h_t = o_t \cdot \phi(c_t), \quad (8)$$

where: c_{t-1} is previous hidden state; f_t, i_t, g_t, o_t are forget, input, modulation, and output gate activations respectively; W_f, W_i, W_g, W_o and U_f, U_i, U_g, U_o are weight matrices; b_f, b_i, b_g, b_o are biases; σ is sigmoid function; ϕ is usually the tanh activation function.

The exploited LSTM model is based on Python Keras Sequential API. The first LSTM layer contains 10 units and uses the ReLU activation function within each unit. The layer processes the input data with an unspecified time step, allowing the model to handle sequences with different lengths. Its main function is to capture temporal dependencies in input data, which makes the model suitable for forecasting tasks. A dense layer follows the LSTM layer, implemented with a single output unit, producing the predicted value for the following time step. To provide a similar structure for both models and enable fair comparisons, the LSTM model is also based on Adam optimizer and MSE loss function.

4. RESULTS AND ANALYSIS

The model is trained and tested in Google Colab environment using two 64-core processors AMD EPYC 7B12 at 2.2 GHz and 14 Gb RAM memory. Both models are trained for two different values of training epochs: 50 and 100, with two different possibilities for the batch size for each training process: 32, 64. Here, a batch size defines the number of training samples which are processed together in one forward and backward pass through a model. Next, the tested model's structures are realized with three different numbers of units (10, 20, 50), while the number of layers is one or two. Finally, the Dropout parameter as a regularization hyperparameter that randomly deactivates a fraction of neurons (during each training iteration) to prevent overfitting, is restricted to one of these two values: 0.1, 0.2. Through the considered values that will define the structure of the LSTM model, a total of 42 different structures were initialized that will be trained and tested in terms of performance, training time, energy consumption and CO2 emissions (Table 1). To note, as the entire simulation data table is too comprehensive and large, only fragments of the entire material are shown in Table 1 to give insight into the information and results. The presentation of results in this section follows the objectives outlined in Section 2. Each subsection addresses a specific research task in the development and validation of the proposed multi-criteria metric.

4.1. Metric application across Long Short-Term Memory configurations

This subsection presents results from applying the proposed scoring metric (Task 1 and 3), using MAE, MSE, RMSE, training time, energy consumption, and CO_2 emissions across 42 LSTM configurations.

The analysis presented below in Table 1 is based on measuring the training times and using three common evaluation metrics to estimate achieved performances: MAE, MSE, and RMSE. In addition, to evaluate the sustainability factors of using the proposed structures of the LSTM model, energy consumption and CO_2 emissions are measured using the CodeCarbon Python library. The library is designed to estimate the carbon footprint of AI and ML algorithms and computational workloads of CPU, GPU, and RAM. By analyzing the efficiency and sustainability of various structures simultaneously, the table offers the calculated values for the proposed performance score in (1) as the overall value of usability level of each LSTM structure.

4.2. Trade-off analysis between accuracy and sustainability

Aligned with Task 4, this subsection interprets results based on the metric's ability to identify optimal trade-offs. Configurations that balance forecasting accuracy and low environmental impact receive higher scores, consistent with the assigned task priority.

By analyzing the results from Table 1 is verified that the performance score effectively synthesizes the critical trade-offs between model accuracy, computational efficiency, and

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

Table 1. **Testing results**

Units	Layers	Dropout	Batch Size	Epochs	MAE	MSE	RMSE	Training Time (s)	Energy Consumed (kWh)	Emissions (kg CO2)	**Performance Score**
20	1	0.2	64	50	3214.236	14281095	3779	98.3020	0.00129	0.000368	**0.86**
50	1	0.1	64	50	1395.4	326740	1807.5	113.78	0.00149	0.00042	**0.86**
20	1	0.1	64	50	1764.4	504773	2246.7	97.380	0.00127	0.00036	**0.86**
10	1	0.1	64	50	4276.1	229921	4795.0	68.478	0.00089	0.00025	**0.85**
50	1	0.2	64	50	2095.6	694978	2636.2	119.24	0.00156	0.00044	**0.85**
20	2	0.2	64	50	2770.5	11908987	3450.9	159.84	0.00209	0.00059	**0.84**
50	2	0.1	64	50	1854.7	5450726	2334.6	236.24	0.0031	0.00088	**0.84**
. . .											
10	1	0.2	64	100	5620.9	35260185	5938.02	138.02	0.00181	0.0005	**0.8**
50	2	0.2	32	50	2183.2	7603652	2757.47	433.18	0.00568	0.001	**0.79**
. . .											
10	2	0.1	32	100	4465.5	24286103	4928	541.49	0.00710	0.00202	**0.67**
10	2	0.2	32	100	11881	1.45E+08	12060	577.44	0.00757	0.00216	**0.67**
20	2	0.1	32	100	5092.9	29079718	5392.5	662.12	0.00868	0.00248	**0.66**
50	2	0.2	32	100	1700.4	4795462	2189.8	893.85	0.01172	0.00334	**0.65**
50	2	0.1	32	100	932.78	2083187	1443.3	945.10	0.01240	0.00354	**0.63**

Source: **compiled by the authors**

environmental impact, proving its value as a comprehensive metric for model evaluation and selection. By analyzing configurations from the table with the highest scores (0.86), such as the 50-unit, single-layer model with dropout=0.1, batch size=64, and 50 epochs, it can be observed that this metric successfully identifies models that achieve an optimal equilibrium–exhibiting low prediction errors (MAE≈1395, RMSE≈1807) while maintaining modest resource consumption (energy ≈ 0.0015kWh, emissions ≈ 0.0004kg CO_2). The score's sensitivity to diminishing returns is evident in its penalization of over-parameterized models (e.g., 2-layer or 100-epoch variants), where marginal accuracy gains fail to justify the disproportionate increase in computational costs and emissions, resulting in lower scores (0.63-0.79). This robust alignment with practical priorities, rewarding balanced performance rather than extreme specialization in any single dimension, validates the score's utility for guiding deployment decisions, ensuring selected models meet both technical and sustainability goals. Thus, the performance score serves as a reliable, actionable tool for comparing architectures and optimizing neural network designs in resource-constrained environments.

4.3. Weight calibration and prioritization justification

Task 5 is addressed here by detailing how the metric's weights were empirically calibrated to prioritize accuracy and emissions over training time, aligning with the sustainability-oriented objective outlined in Section 2.

Regarding the values of weight coefficients from equation (2) which are specified for training purposes, all these free parameters are set empirically. After a series of parameter adjustments during the training process, the optimal values of the weight coefficients are calculated as: $w1$=10, $w2$=1, $w3$=10, $w4$=100. These weights were carefully calibrated through iterative testing to balance model accuracy, computational efficiency, energy consumption, and environmental impact. These values ensure the performance score reflects real-world priorities, where sustainability and accuracy are paramount, while training time is treated as a secondary concern. More specifically, training time ($w2$=1) is assigned a significantly lower weight than

accuracy ($w1$=10) and emissions ($w4$=100) because its impact is less critical in most deployment scenarios. While faster training is desirable, it does not carry the same long-term consequences as high energy consumption or CO_2 emissions. For instance, a model that trains twice as slowly but uses half the energy is often preferable, as energy costs accumulate over time, whereas training time is a one-time expense. Testing confirmed that increasing $w2$ beyond 1.0 disproportionately penalized models for minor time differences without meaningful improvements in sustainability or accuracy. Thus, keeping $w2$ an order of magnitude lower than $w1$ and $w4$ ensures training time influences the score without overriding more critical factors.

It is also empirically shown that the emissions weight ($w4$=100) is the largest to enforce strict penalties for high carbon footprints, aligning with global sustainability goals. In testing, models with even moderately high emissions consistently received poor scores unless they also delivered exceptional accuracy, ensuring that only truly high-value models justify their environmental cost. For example, a model emitting 0.5 kg CO_2 would incur a penalty of 50 (assuming normalized emissions), which is severe enough to override moderate accuracy gains. This design reflected a deliberate prioritization: a model cannot be considered "high-performing" if it exacerbates climate impact, regardless of its predictive power. Empirical adjustments showed that lower values for $w4$ (e.g., 50) failed to sufficiently discourage unsustainable practices, while higher values (e.g., 200.0) made the score overly sensitive to minor emission variations.

To graphically summarize the achieved results, two different approaches are exploited. In the first case, the Pareto frontier is presented in Fig. 3 to illustrate the inherent trade-off between model accuracy (quantified as 1/MAE) and environmental sustainability (quantified as 1/Emissions) across the evaluated configurations.

A few observations can be extracted from the previous figure. First of all, the concave shape of the frontier confirms the competing relationship between accuracy and sustainability. Models achieving high accuracy (for example, high 1/MAE>0.00081/MAE>0.0008) exhibit diminished sustainability (lower 1/Emissions<0.00041/Emissions <0.0004), while highly sustainable models (high 1/Emissions>0.00061/Emissions>0.0006) show reduced accuracy (1/MAE<0.00041/MAE<0.0004). Next, the Pareto-optimal region (upper-right quadrant) contains models balancing both metrics effectively. For instance, configurations with 1/MAE≈0.0006 and 1/Emissions≈0.0006 achieve near-maximal performance scores (>0.8), demonstrating that neither metric need be sacrificed entirely for marginal gains in the other. In addition, the color gradient reveals that the highest performance scores (>0.85>0.85) cluster in regions of moderate accuracy and sustainability, suggesting the scoring metric prioritizes balanced performance over extreme specialization in either dimension. Beyond 1/MAE>0.0008, further accuracy improvements require disproportionately large increases in emissions (steep decline in 1/Emissions), indicating a threshold where computational costs outweigh predictive gains. Finally, the practical implication would be that the models near the "knee" of the frontier (1/MAE=0.0005, 1/Emissions=0.00051) represent optimal choices for real-world deployment, where both accuracy and sustainability are critical. Configurations deviating from this region either underperform in accuracy or incur excessive environmental costs.

As the second graphical presentation of the results, the Fig. 4 is created. The graph reveals a trade-off between model performance and sustainability in training. The main insights demonstrate that models achieving higher accuracy (lower MAE values between 1000–5000) generally incur greater environmental costs, with CO_2 emissions ranging from 0.0005 to 0.0019 kg. Notably, configurations with emissions below 0.001 kg maintain competitive performance scores (0.02–0.07), suggesting that eco-efficient models need not sacrifice substantial predictive power.

Training times (100-300 seconds) show a nonlinear relationship with accuracy, where optimal models balance speed and precision without excessive resource use. For scientific applications, these findings highlights the necessity of incorporating environmental metrics alongside traditional performance indicators to guide sustainable model development – directly promoting the vital need for utilizing similar overall performance metrics as the pone proposed in this paper.

CONCLUSIONS AND PROSPECTS OF FURTHER RESEARCH

Nowadays, the computational demands of AI systems have raised concerns about their environmental footprint, necessitating the development of evaluation frameworks that account for sustainability alongside traditional performance metrics. This work addressed this challenge by introducing a novel performance score that

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

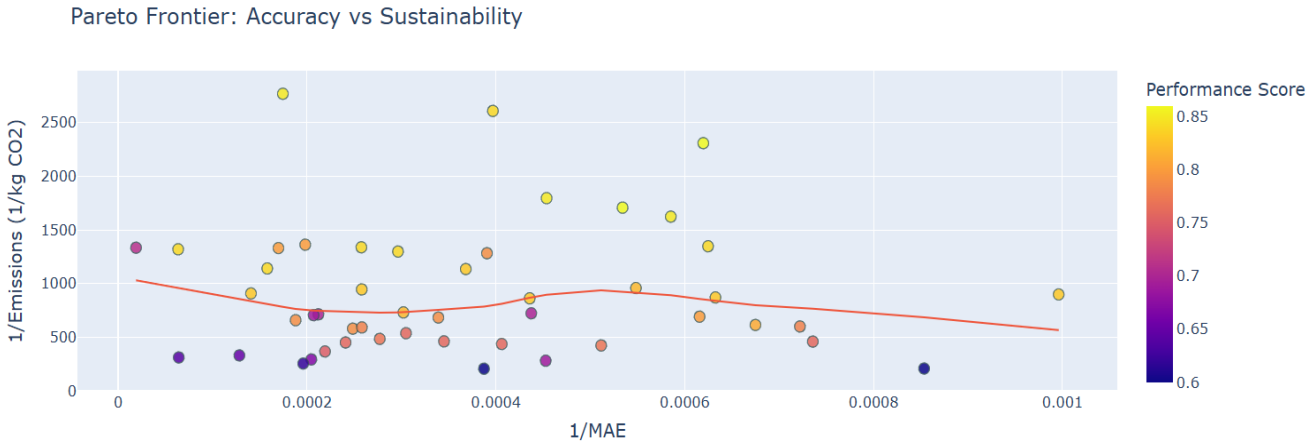

Fig. 3. **Pareto Frontier: Accuracy vs Sustainability**

Source: **compiled by the authors**

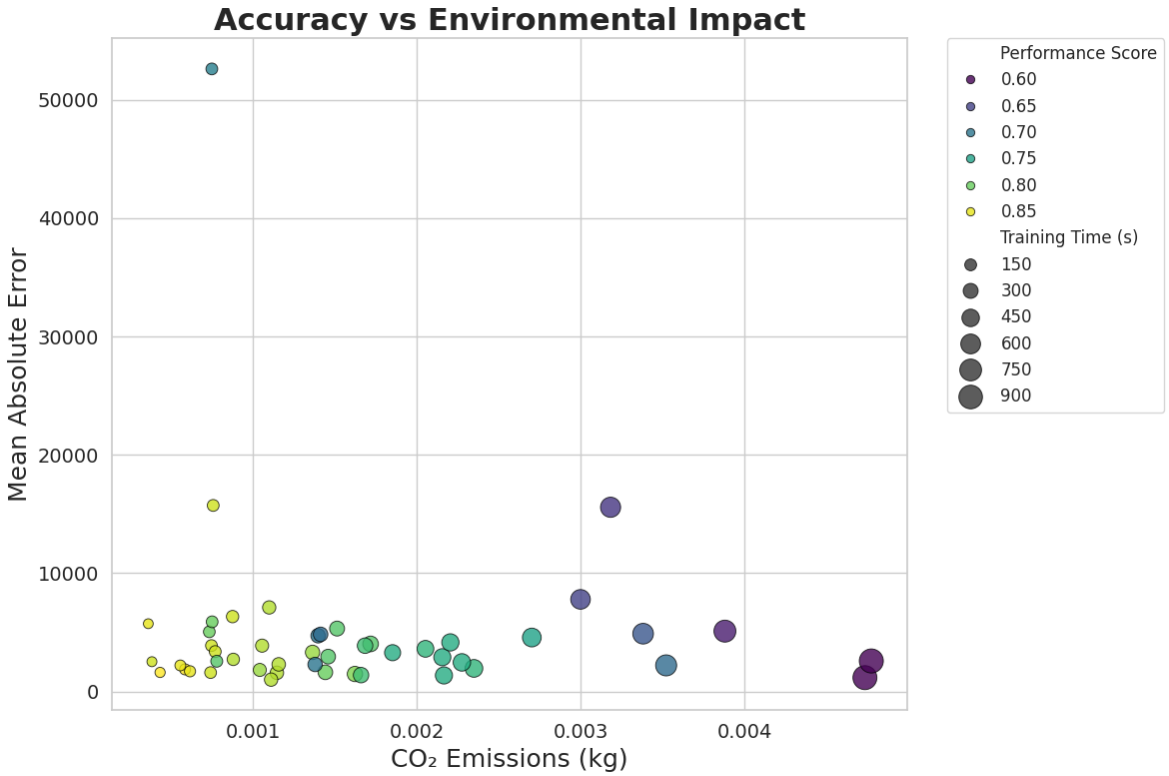

Fig. 4. **Accuracy vs Environmental Impact**

Source: **compiled by the authors**

systematically balances predictive accuracy, computational efficiency, and environmental impact. The evaluation process that is conducted in this research included multiple architectures and training configurations to demonstrate that this novel performance score effectively identifies models that achieve optimal trade-offs between all competing objectives.

The proposed performance score incorporates three components: (1) accuracy metrics (MAE, MSE, RMSE) to assess predictive capability, (2) computational costs (training time), and (3) environmental impact (energy consumption and CO_2 emissions). Through empirical validation, it is established that models with intermediate complexity (LSTM configurations with 50 units, single-layer architectures, and dropout rates equal to 0.1) achieved the highest performance scores (0.86). These configurations maintained high predictive accuracy (MAE≈1395) while minimizing environmental impact (0.0004 kg CO_2 emissions and 0.0015 kWh energy consumption). In continuation,

the analysis revealed clear diminishing returns for more complex models, where additional layers or extended training epochs provided only marginal accuracy improvements at disproportionately high computational and environmental costs.

Through numerous conducted simulations the performance score successfully captured trade-offs between three analyzed performance components, suggesting a single universal metric that aligns with practical deployment considerations. From a broader perspective, the proposed metric could answer to the field of sustainable AI by providing both a methodological framework and empirical evidence for making smart environmentally decisions of building different AI and ML models.

Looking forward, several promising directions emerge from this work. Future research could investigate adaptive weighting schemes that automatically adjust to different operational constraints or environmental policies. The development of more sophisticated normalization approaches could further improve the score's sensitivity to critical thresholds in energy consumption or emissions.

The tasks formulated in Section 2 have been fully addressed through the proposed methodology. A comprehensive performance metric was developed and applied across 42 LSTM configurations. The best-performing model configuration achieved a performance score of 0.86, with an MAE of 1395.4, RMSE of 1807.5, energy consumption of only 0.00149 kWh, and CO_2 emissions of 0.00042 kg. These values show a substantial improvement over more complex models that, despite achieving slightly better accuracy (e.g., MAE=932.78), incurred significantly higher resource costs (energy=0.01240 kWh, emissions=0.00354 kg), resulting in lower composite scores (down to 0.63). Thus, the goal of this research, the development and validation of a multi-criteria scoring metric that balances predictive accuracy, computational efficiency, and environmental impact, has been successfully achieved.

REFERENCES

1. Panda, S. K., Sathya, A. R. & Das, S. “Bitcoin: Beginning of the cryptocurrency era. In: Recent Advances in Blockchain Technology: Real-World Applications”. *Springer International Publishing, Cham*. 2023; 237: 25–58, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85153036481&doi=10.1007%2f978-3-031-228353_2&partnerID=40&md5=1b46c3860a7f08423b15c102b86fc927. DOI: https://doi.org/10.1007/978-3-031-22835-3_2.

2. Arslanian, H. “Bitcoin”. In *The Book of crypto: the complete guide to understanding Bitcoin, cryptocurrencies and digital assets. Cham: Springer International Publishing*. 2022. p. 45–89, www.scopus.com/inward/record.uri?eid=2-s2.0-85173900096&doi=10.1007%2f978-3-030-97951-5&partnerID=40&md5=eb3ed729fdb361e545ce921a7fdba4be. DOI: 10.1007/978-3-030-97951-5_2.

3. Fernstrom, A., Frank, M. M., Lewis, S. A., Matos, P., Macfarlane, J. G., Frank, M. M., Lewis, S. A., Matos, P. & Macfarlane, J. G. “JUST Capital”. *Darden Business Publishing Cases*. 2024. p. 1–31. DOI: https://doi.org/10.1108/case.darden.2024.001844.

4. Janjua, L. R., Gigauri, I., Wójcik-Czerniawska, A. & Pohulak-Żołędowska, E. “Risk management in the area of Bitcoin market development: Example from the USA”. *Risks*. 2024; 12 (4): 67. DOI: https://doi.org/10.3390/risks12040067.

5. Doumenis, Y., Izadi, J., Dhamdhere, P., Katsikas, E. & Koufopoulos, D. “A critical analysis of volatility surprise in Bitcoin cryptocurrency and other financial assets”. *Risks*. 2021; 9 (11): 207. DOI: https://doi.org/10.3390/risks9110207.

6. Kuzior, A., Krawczyk, D., Koibichuk, V. & Mohylna, K. “The Price and market prospects for the ethereum cryptocurrency development”. *Financial Markets, Institutions and Risks*. 2024; 8 (3): 187–205. DOI: https://doi.org/10.61093/fmir.8(3).187-205.2024.

7. Kulbhaskar, A. K. & Subramaniam, S. “Breaking news headlines: Impact on trading activity in the cryptocurrency market”. *Economic Modeling*. 2023; 126: 106397. DOI: https://doi.org/10.1016/j.econmod.2023.106397.

8. Mou, J., Liu, W., Guan, C., Westland, J. C. & Kim, J. “Predicting the cryptocurrency market using social media metrics and search trends during COVID-19”. *Electronic Commerce Research* 2024; 24: 1307–1333. DOI: https://doi.org/10.1007/s10660-023-09801-6.

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

9. Wu, J., Zhang, X., Huang, F., Zhou, H. & Chandra, R. “Review of deep learning models for crypto price prediction: implementation and evaluation”. *arXiv preprint* arXiv:2405.11431. 2024. DOI: https://doi.org/10.48550/arXiv.2405.11431.

10. Tanwar, S., Patel, N. P., Patel, S. N., Patel, J. R., Sharma, G. & Davidson, I. E. “Deep learning-based cryptocurrency price prediction scheme with inter-dependent relations”. *IEEE Access*. 2021; 9: 138633–138646. DOI: https://doi.org/10.1109/ACCESS.2021.3118111.

11. Modi, P. D., Arshi, K., Kunz, P. J. & Zoubir, A. M. “A data-driven deep learning approach for Bitcoin price forecasting”. In *2023 24th International Conference on Digital Signal Processing (DSP)*. 2023. p. 1–4. IEEE. DOI: https://doi.org/10.1109/DSP56456.2023.10247689.

12. Akila, V., Nitin, M. V. S., Prasanth, I., Reddy, S. & Kumar, A. “A cryptocurrency price prediction model using deep learning”. In *E3S Web of Conferences*. 2023; 391: 01112. DOI: https://doi.org/10.1051/e3sconf/202339101112.

13. Akila, V., Nitin, M. V. S., Prasanth, I., Reddy, S. & Kumar, A. “A Cryptocurrency Price Prediction Model using Deep Learning”. In *E3S Web of Conferences.* 2023; 391: 01112. DOI: https://doi.org/10.1051/e3sconf/202339101112.

14. Zhang, J., Cai, K. & Wen, J. “A survey of deep learning applications in cryptocurrency”. *Iscience*. 2024; 27 (1):108509. DOI: https://doi.org/10.1016/j.isci.2023.108509.

15. Bourday, R., Aatouchi, I., Kerroum, M. A. & Zaaouat, A. “Cryptocurrency Forecasting Using Deep Learning Models: A Comparative Analysis”. *HighTech and Innovation Journal*. 2024; 5 (4): 1055–1067. DOI: https://doi.org/10.1016/10.28991/HIJ-2024-05-04-013.

16. Gurgul, V., Lessmann, S. & Härdle, W. K. “Forecasting Cryptocurrency Prices Using Deep Learning: Integrating Financial, Blockchain, and Text Data”. *arXiv preprint arXiv:2311.14759.* 2023. DOI: https://doi.org/10.48550/arXiv.2311.14759.

17. Yousaf, M., Jabbar, M. T. A. “A Comprehensive Survey of Cryptocurrency Forecasting: Methods, Trends, and Challenges”. *SSRN Preprint*. 2024. DOI: https://dx.doi.org/10.2139/ssrn.5234539.

18. Zielinski, M. “Kaggle Bitcoin”. *GitHub*. – Available from: https://github.com/mczielinski/kaggle-bitcoin/ – [Accessed: Jan, 2024].

19. Tripathi A. “What is the Main Difference Between RNN and LSTM | NLP | RNN vs LSTM”. 2021. Available from: https://ashutoshtripathi.com/2021/07/02/what-is-the-main-difference-between-rnn-and-lstm-nlp-rnn-vs-lstm. [Accessed: Jan, 2024].

20. Sherstinsky, A. “Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network”. *Physica D: Nonlinear Phenomena.* 2020; 404, 132306. DOI: https://doi.org/10.1016/j.physd.2019.132306.

Conflicts of Interest: The authors declare that they have no conflict of interest regarding this study, including financial, personal, authorship or other, which could influence the research and its results presented in this article

Received 17.02.2025
Received after revision 23.04.2025
Accepted 14.05.2025

DOI: https://doi.org/10.15276/hait.08.2025.14
УДК 004.81:159.953.5

Багатокритеріальна метрика для оцінки моделей глибокого навчання в прогнозуванні ціни біткойна

Злобін Микола Миколайович[1)]
ORCID: https://orcid.org/0009-0000-7653-6109; mykolay.zlobin@gmail.com. Scopus Author ID: 59337918100
Базилевич Володимир Маркович[1)]
ORCID: https://orcid.org/0000-0001-8935-446X; bazvlamar@stu.cn.ua. Scopus Author ID: 57214432127
[1)] Національний університет «Чернігівська політехніка», вул. Шевченка, 95. Чернігів, 14030, Україна

АНОТАЦІЯ

Зростаючі обчислювальні вимоги глибокого навчання викликали занепокоєння щодо екологічної стійкості застосувань штучного інтелекту, особливо у високочастотних областях, таких як фінансове прогнозування. Стаття розглядає потребу в більш цілісних критеріях оцінки, пропонуючи багатокритеріальну метрику оцінки для моделей глибокого навчання, що використовуються в прогнозуванні ціни біткойна. Метою дослідження є розробка метрики продуктивності, яка збалансує точність прогнозування з обчислювальною ефективністю та впливом на навколишнє середовище. Метод передбачає поєднання традиційних показників точності з часом навчання, споживанням енергії та викидами вуглецю в єдину оцінку продуктивності, розраховану за допомогою логістичної функції оцінки. Метрику було перевірено шляхом її застосування до сорока двох конфігурацій моделей з довгостроковою пам'яттю (LSTM), навчених на історичних даних про ціну біткойна. Кожну конфігурацію було оцінено за точністю прогнозування, споживанням енергії та викидами (виміряними за допомогою інструменту відстеження вуглецю). Результати показують, що простіші моделі довгої короткострокової пам’яті (LSTM) можуть запропонувати конкурентоспроможну точність, водночас значно скорочуючи час навчання та викиди. Найпродуктивніша модель досягла балансу всіх критеріїв, тоді як глибші архітектури з незначним підвищенням точності понесли непропорційні екологічні витрати. У дослідженні зроблено висновок, що запропонована метрика оцінювання пропонує практичне та масштабоване рішення для вибору моделей глибокого навчання в умовах обмежень сталого розвитку, підтримуючи більш відповідальне розгортання штучного інтелекту (ШІ) в реальних умовах.

Ключові слова: глибоке навчання; прогнозування часових рядів; довга короткострокова пам'ять (LSTM); метрика продуктивності (показник ефективності); сталий розвиток (стійкість)

ABOUT THE AUTHORS

Mykola M. Zlobin - PhD student, Chernihiv Polytechnic National University, 95, St. Shevchenko, Chernihiv, 14030, Ukraine
ORCID: https://orcid.org/0009-0000-7653-6109; mykolay.zlobin@gmail.com. Scopus Author ID: 59337918100
***Research field*:** machine learning; bankruptcy prediction, classification tasks; deep learning, time series forecasting; stock price, volatility modeling, mathematical models; overfitting, underfitting, model optimization, hyperparameter tuning; intelligent decision support systems in finance

Злобін Микола Миколайович - аспірант. Національний університет «Чернігівська політехніка», вул. Шевченка, 95. Чернігів, 14030, Україна

Volodymyr M. Bazylevych - PhD, Associate Professor, кафедри Інформаційних та комп’ютерних наук, Head of ESI EIT, Chernihiv Polytechnic National University, 95, St. Shevchenko, Chernihiv, 14030, Ukraine
ORCID: https://orcid.org/0000-0001-8935-446X; bazvlamar@stu.cn.ua. Scopus Author ID: 57214432127
***Research field*:** machine learning; bankruptcy prediction, classification tasks; deep learning, time series forecasting; stock price, volatility modeling, mathematical models; overfitting, underfitting, model optimization, hyperparameter tuning; intelligent decision support systems in finance

Базилевич Володимир Маркович - кандидат економічних наук, доцент кафедри Інформаційних та комп’ютерних наук, директор ННІ ЕІТ. Національний університет «Чернігівська політехніка», вул. Шевченка, 95. Чернігів, 14030, Україна