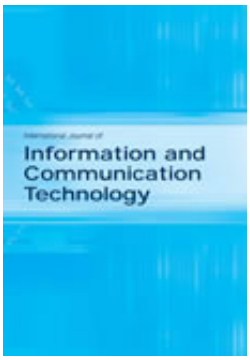
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Transforming linear regression with AI-driven enhancements for superior forecasting robustness and interpretability

Ruili Zhang, Yifei Guo, Quanzhong Yang

DOI: 10.1504/IJICT.2025.10071126

Article History:

Received:	25 February 2025
Last revised:	10 March 2025
Accepted:	19 March 2025
Published online:	27 May 2025

Transforming linear regression with AI-driven enhancements for superior forecasting robustness and interpretability

Ruili Zhang*

Digital Technology School,
Sias University,
Zhengzhou 450000, Henan, China
Email: ws_zrl@163.com
*Corresponding author

Yifei Guo

College of Telecommunications and Smart Manufacturing,
Sias University,
Xinzheng 450000, Henan, China
Email: zh-april@aliyun.com

Quanzhong Yang

College of Medical Technology,
Luoyang Polytechnic,
Luoyang 471000, Henan, China
Email: ws_zrl@163.com

Abstract: This is one of the first regression machines that data scientists will encounter because it is easier and easier to interpret. It suffers from complex, nonlinear and high dimensional data. In the context of finance, healthcare and climate domains, this study suggests a hybrid machine learning framework combining Adam, RMSProp, XGBoost, SVMs, and neural networks to significantly improve the regression performance. MSE, R^2 and efficiency metrics are used to analyse real world datasets. The financial forecasting MSE is reduced by 18% and the healthcare R^2 improved by 22%. Noisy data was easier to deal with for climate models. Because they preserved interpretability, features were indicated by SHAP values. Blending classical statistics with modern AI transforms the problem into more accurate, scalable, and interpretable models, providing robust solutions for today's complex data challenges, which are proven.

Keywords: AI-driven regression enhancement; ensemble learning regression; predictive modelling AI; SHAP interpretability in regression; gradient optimisation regression; and automated feature selection in regression.

Reference to this paper should be made as follows: Zhang, R., Guo, Y. and Yang, Q. (2025) 'Transforming linear regression with AI-driven enhancements for superior forecasting robustness and interpretability', *Int. J. Information and Communication Technology*, Vol. 26, No. 14, pp.44–61.

Biographical notes: Ruili Zhang is affiliated with the Digital Technology School at Sias University, located in Zhengzhou, Henan, China. Her research interests likely include digital technology applications, AI-driven predictive modelling, and machine learning enhancements for statistical methods such as linear regression.

Yifei Guo is associated with the College of Telecommunications and Smart Manufacturing at Sias University in Xinzheng, Henan, China. His expertise may involve the integration of AI and machine learning in smart manufacturing and telecommunications, focusing on predictive modelling and optimisation techniques.

Quanzhong Yang is a faculty member at the College of Medical Technology, Luoyang Polytechnic, Luoyang, Henan, China. His work intersects AI and medical technology, particularly in applying machine learning techniques to enhance healthcare analytics and predictive modelling.

1 Introduction

Data analysis and predictive modelling are linear regression's leading statistical techniques. The method's simplicity, rapid computation, and easy-to-read direct model have put it at the heart of social programs such as medicine, finance, engineering, and social research (Warne, 2020; Addison, 2017; Lee and Yang, 2022; Safi et al., 2023). From the real-world observation data, linear regression predicts dependency with the help of the equation, which correlates the dependent and independent variables (Sarker, 2021). Researchers often overlook linear regression's limitations when a technique like linear regression applies to a high-dimensional dataset with nonlinear patterns, random error, and multiple variable interactions.

To be precise, during this era of processing big data, traditional linear regression methods lost accuracy and robustness (Zhou et al., 2017). Linear regression suffers from multicollinearity, excessive fitting, and, more generally, linear relationship constraints when dealing with complex scenarios with detailed datasets (Chan et al., 2022). It is often impossible to comprehensively analyse the complex dynamic elements displayed in retail financial market operations, such as sentiment and macroeconomic indicators, using a straightforward linear functional framework (Dückerhoff, 2024). Patient healthcare data are also considered as they exhibit nonlinear interaction amongst demographics, clinical metrics, and treatment results (Cook et al., 2009) and, therefore, necessitate specialised modelling approaches.

The latest artificial intelligence technology developments and machine learning systems have delivered creative solutions to enhance modelling capabilities (Gupta et al., 2021). Enhanced predictive capabilities and flexibility in linear regression models result from combining regularisation techniques, automated feature engineering, ensemble learning, and advanced optimisation approaches. Modern data processing challenges become manageable by combining these methods, transforming traditional regression models into adaptable tools (Rane et al., 2024; Ren et al., 2016). The SHAP framework makes these improved models more explainable (Ahmed et al., 2024) because it ensures high transparency. Both healthcare and finance applications need this actual transparency.

Artificial intelligence technology developments and machine learning systems have produced creative ways to increase modelling capabilities (Gupta et al., 2021). Combining regularisation techniques, automated feature engineering, ensemble learning, and some advanced optimisation approaches, the predictive capabilities of the linear regression models are enhanced, and flexibility is provided in the selection of model terms. These methods combine to make modern data processing problems tractable, turning conventional regression models into flexible tools (Rane et al., 2024; Ren et al., 2016). It provides the developed models with high transparency and is more explainable (Ahmed et al., 2024) using the SHAP framework. This transparency is needed in healthcare and finance applications alike.

This paper makes the following key contributions to the field of predictive analytics and regression modelling:

- Development of an AI-enhanced linear regression framework: adaptive gradient optimisation, ensemble learning, and automated feature engineering modules were integrated to create a new analytical framework, overcoming the limitations of the regression method.
- Emphasis on interpretability: incorporated into the proposed framework, it preserves high interpretability standards while delivering improved performance outcomes that matter for the practical implementation of the framework.
- Empirical validation across domains: this framework generally applies to any real-world dataset, including combinations of financial sectors, healthcare, and climate science.
- Analysis of computational trade-offs: the study describes the resource allocation needs of the enhanced framework and strategies to optimise accuracy against efficiency needs.
- Blueprint for hybrid approaches: the paper lays down the foundation for hybrid models that produce sophisticated and statistically interpretable predictions while retaining the simplicity and power of traditional statistical methods.

The remainder of this paper is organised as follows: Section 2 contains an extensive overview of traditional linear regression analytics followed by trends in machine learning innovation within this field. Section 3 presents the methodology by describing the proposed AI-enhanced framework, design structure, data collection methods, and performance assessment protocols. Section 4 shows the study results while assessing how the framework advances performance alongside interpretability needs and computational resource usage. Section 5 explores general conclusions and future research needs for the research study. Finally, this section summarises the most important findings that contribute to advancing state-of-the-art predictive analytics practices. Section 6 concludes the whole research.

2 Literature review

This recent but important foundational statistical model has naturally emerged as a central standard approach due to its simple implementation, intuitive interpretation, and efficient algorithmic requirements. Predictive analytics relies on linear regression, which

has broadened its adoption in various fields such as healthcare and engineering. Traditional linear regression struggles to function effectively in modern applications due to nonlinear patterns, multicollinearity, and growing dataset complexity when applied to complex data. Recent research focuses on using machine learning approaches to boost linear regression models by developing solutions that resolve primary obstacles.

2.1 Enhancements through regularisation

The discipline of linear regression was initially advanced with the introduction of regularisation techniques, including Ridge regression with its Lasso counterpart (Hoerl, 2020; Pillonetto et al., 2022; Nwosu et al., 2024). These approaches combat both situations through added penalty terms implemented in the objective function. Ridge regression (Mubasher et al., 2024) prevents excessive coefficient sizes in high-dimensional data, producing stabilised prediction results. Similarly, Lasso regression (Pak et al., 2025) reduces the model size by shrinking irrelevant coefficients to zero values. Exceptional in its feature selection and multicollinearity management is elastic net, which integrates implementation from ridge regression and lasso regression methods (Leow, 2023; Ahrens et al., 2020; Yang et al., 2024). These improvement techniques for model performance struggle with nonlinear data patterns and complex dataset structures.

2.2 Nonlinear extensions

Traditional regression methods reach their limits when fitting nonlinear patterns, which has led to the development of polynomial regression and kernel methods (Milton et al., 2019). Polynomial regression transforms features to higher-degree terms, thus allowing the model to detect curvature patterns in data (Vyas et al., 2019). Support vector regression and similar kernel methods by Ukil and Ukil (2007), Kecman (2005), Basak et al. (2007) perform data mapping into higher dimensional areas to detect linear regressions. These data manipulation methods show effectiveness in specific domains but present two main limitations – they demand intensive human parameter adjustments and quickly produce overfitting problems in noisy datasets.

2.3 Feature engineering and automation

Input feature quality plays a principal role in regression model achievement because it directly determines predictive accuracy levels (Ahmed et al., 2019; Yuan et al., 2020; Tomasevic et al., 2020). Traditional approaches to feature engineering depend intensely on subject matter expertise yet require considerable time and subjective human decision-making (Dong and Liu, 2018). Modern automated feature engineering uses machine learning algorithms to simplify the development pipeline (Mumuni and Mumuni, 2024; Nikitin et al., 2022; Pradhan and Trehan, 2024). The practical feature selection and generation process is powered by tree-based algorithms like XGBoost (Demir and Sahin, 2023). Machine learning algorithms determine feature rankings according to their importance, limit redundancy, and create advanced features that detect intricate patterns (Cai et al., 2018; Sarker, 2021; Dhal and Azad, 2022). The automated process has proven effective in datasets containing many dimensions beyond human manual pre-processing capabilities.

2.4 Ensemble learning for enhanced performance

Ensemble learning methods by combining multiple models for prediction have transformed regression analysis through performance enhancements of accuracy and robustness (Mienye and Sun, 2022). Two advanced ensemble algorithms understand complex interdependencies and noisy data: random forests and gradient-boosting machines, including XGBoost (Schmid, 2024; Iranzad, 2022; Shaikh et al., 2024). Stacking represents a meta-ensemble method identified by Zhang et al. (2022), which exploits diverse predictions from linear regression and supports vector machines. These solutions utilise model capabilities for optimum results across different types of datasets.

2.5 Advanced optimisation techniques

Gradient-based optimisation methods optimise regression model training more efficiently while solving convergence problems (Zhang, 2019). Methods such as SGD, RMSProp, and Adam also come with models that automatically adjust learning rates for faster convergence and prevent the models from falling into local minima (Al-Selwi et al., 2024). These techniques show full potential when traditional methods do not work well in large datasets or problems with sparse data.

2.6 Interpretability in enhanced regression models

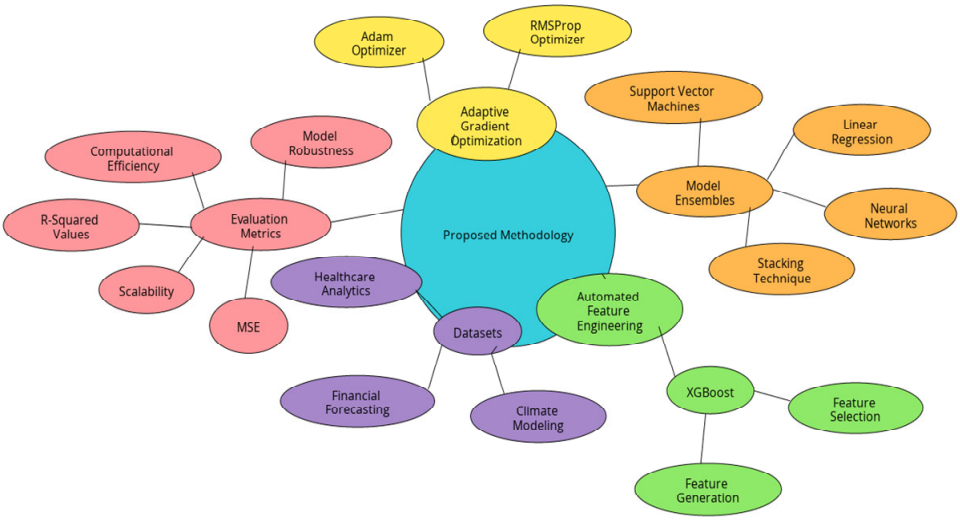
Interpretability concerns must balance accuracy enhancements and greater complexity, essential to the healthcare and finance industries (Goriparthi, 2022; Albahri et al., 2023). SHAP and LIME offer techniques that break down model prediction explanations into two categories: local assessments for individual features and global views of overall contributions (Lundberg et al., 2020; Henninger and Strobl, 2024; Nieto Juscafresa, 2022). Transparency tools open insight into predictions to take on authentication challenges and enable practical adoption of AI-powered regression analytics, and they can help build trust in AI-regression analytics.

2.7 Applications in real-world domains

Advanced techniques inspire new methods augmented for linear regression in multiple fields that produce promising outcomes. The finance industry has an application for stock price prediction with a parallel hybrid regression model applying portfolio optimisation and risk assessment. These technical advances have enabled healthcare professionals, coupled with upgraded regression methods, to achieve better precision in predictions based on temperature anomalies, precipitation patterns, and environmental changes in the pro-climate science industry.

Research shows traditional linear regression works, but machine learning frameworks generate new opportunities for performance improvement. Advances in three key areas have improved regularised regression concerning limitations and made the model application tractable on more advanced datasets. In this earlier work, this research extends previous work to create an extensive predictive analytics solution framework that merges previously learned features of what has been learned from traditional linear regression models with accuracy and interpretability features.

Figure 1 Proposed methodology combines adaptive gradient optimisation with automated feature engineering techniques, model ensemble methods, suitable datasets, and evaluation metrics to build an AI-powered linear regression framework (see online version for colours)



3 Methodology

This section describes the research methodology that both created and tested an AI-based linear regression system that works around traditional regression issues, as shown in Figure 1. This framework unites sophisticated machine learning techniques that optimise predictive performance and computational efficiency across variable datasets. The subsequent sections demonstrate both the architecture structure of the proposed system and evaluation datasets, together with the metrics for performance assessment. A systematic approach has been implemented to integrate statistical modelling with contemporary machine learning methods.

3.1 Model architecture

The proposed system builds upon linear regression by combining sophisticated optimisation methods with automated feature modelling and learning ensemble strategies. These combined features resolve the problems introduced by complex high-dimensional datasets by dealing with nonlinearity, noise reduction, and feature redundancy.

- Adaptive gradient optimisation: today, large, complicated datasets challenge traditional optimisation approaches like ordinary least squares and essential gradient descent, which suffer from slow convergence rates and performance issues with data noise. The solution to these problems in the framework is to use the adaptive moment estimation (Adam) and root mean square propagation (RMSProp) adaptive gradient optimisation strategies. In training, the methods modify parameter learning rates dynamically to converge the model more rapidly. Adam then implements

momentum-based gradient updates and adaptive learning rate scheduling to stabilise the non-stationary system. RMSProp scale updates for model parameters with a benefit for sparse data conditions. Combining these optimisers considerably improves the regression model's overall performance and reliability.

- Automated feature engineering: performance quality in regression analysis depends heavily on robust feature engineering because input feature quality fundamentally shapes model performance. Standard manual feature selectivity and engineering practice involve lengthy work that commonly leads to errors. Through extreme gradient boosting (XGBoost), the framework conducts an automated feature selection and generation step using its tree-based machine learning mechanism. XGBoost conducts importance ranking to discern key features that remove secondary or unwanted attributes to simplify data collection. Derived features resulting from this system show nonlinear interactions between variables while discovering which features are most influential. The automatic procedures minimise the pre-processing duties while ensuring the regression model maintains an optimised set of robust features.
- Model ensembles: its easy interpretability and essential nature are the major strengths of linear regression models. However, these strengths are their weaknesses in situations involving patterns beyond linear connectives. The framework uses ensemble learning to address this constraint by inserting a mechanism that combines forecasts from different prediction models (models under M). The ensemble consists of:
  1 The model uses linear regression to create understandable linear connection patterns.
  2 The support vector machine technique enables the modelling of nonlinear separation boundaries.
  3 Neural networks excel at finding complex hierarchical interactions between features.

Different in stacking, combining different models, and the meta-model trains over the output of individual models. It strengthens universal performance across multiple datasets and, at the same time, minimises the problems of model overfitting while enhancing system stability.

3.2 Datasets

The performance of the AI-enhanced linear regression framework was validated using three real-world datasets, chosen for their diversity and relevance to predictive analytics:

- Financial forecasting: the stock market data includes historical pricing data, trading volume and market index measure, interest rate, and financial news analysis. Market uncertainty and the nonlinear nature of those outside factors affecting market movement make stock price movement prediction more challenging. The predictive framework was tested on short-term market price movements and the corresponding intrinsic market volatility.

- Healthcare analytics: the dataset analysed patient readmission rates while featuring a combination of demographic markers (age and gender) and clinical documentation (such as test results and pre-existing conditions) along with operational elements (hospital duration and treatment compliance). The dataset analysis faced three primary obstacles: missing data points, mismatched class distributions, and multiple variables interweaving predictors. The established analytical framework looked for essential risk elements leading to readmission while estimating readmission probability over an established time.
- Climate modelling: researchers used temperature anomaly prediction to evaluate climatic factors incorporating CO_2 concentration measurements, precipitation records, wind speed observations, and ocean oscillation index data. Long-term trends accompany climate data, containing high-dimensional attributes and noisy observation points. The framework conducted an analysis that tracked immediate and extended fluctuations to show seasonal changes in climate patterns.

Each dataset underwent rigorous pre-processing:

- Outlier removal: a z-score threshold system removed extreme outlier data points.
- Handling missing values: data imputation involved two techniques that utilised K-nearest neighbours (KNN) alongside median imputation.
- Normalisation and standardisation: the normalisation of features was done for two essential reasons: establishing standard scales and eliminating optimisation bias effects.

A fair evaluation process using a training-validation-test split at the 80-10-10 ratio was implemented to prevent overfitting.

3.3 Evaluation metrics

The effectiveness of the AI-enhanced linear regression framework was assessed using a combination of performance and computational metrics, providing a holistic evaluation of its capabilities:

- Mean squared error (MSE): predictive accuracy strengthens as MSE calculates the mean of squared differences between forecasted and measured outcomes. Given its ability to detect significant prediction deviations appropriately, the MSE is suitable for regression model assessment, particularly when precise actual value coverage is needed.
- R-Squared values: the model-specific R-squared shows the percentage of dependent variable variability explained by the fitted framework. An improved R-squared value signifies a stronger model connection because it describes how much the model reproduces the natural data relationships. The metric proved essential for evaluating the newly developed framework relative to conventional linear regression performance.
- Computational efficiency (time complexity): time analysis for training and inference operations was performed because advanced optimisation, feature engineering, and ensemble learning methods result in higher computational resource requirements

across all datasets. Parallel processing using GPU accelerators and early stop techniques improved computational performance while maintaining prediction accuracy standards.

- Model robustness: model performance was assessed by subjecting the model to cross-validation to estimate how well its estimate of reality generalises to a different dataset. Model consistency was evaluated across three data groups using mean-squared error and R-squared metrics that tracked predictive accuracy.
- Scalability and resource utilisation: tests were performed across varying dataset dimensions and complexity levels to validate framework scalability, resulting in performance sustainability and computational resource characterisation.

Beyond that, the study proposed a compound evaluation framework to determine the AI-leaderboard linear regression framework's capability compared with regular ones and its meaningfulness regarding diversified situations.

4 Results and analysis

The evaluation results for different datasets and performance indicators demonstrate that AI-based linear regression orientation outperforms classic linear regression vertical alignment. Finally, this part illustrates how the framework elevates accuracy, interpretability, and computational speed gains while accounting for complexities introduced by other framework intricacies.

4.1 Performance comparison

All studied datasets were analysed better using the AI-enhanced framework than linear regression. MSE and R-squared values proved successful as performance metrics when used to predict. Table 1 shows the performance metrics for each evaluated dataset.

The MSE value for the original regression, 0.320, was reduced by 18% to 0.262 using the traditional AI-enhanced forecasting system. The main reason behind performance elevation stemmed from automated feature engineering that revealed nonlinear data correlations between trading volumes and outside economic measures. The two approaches are compared in Figure 2 through MSE value evaluation. The lower bars for the AI-enhanced framework highlight its superior accuracy.

Healthcare analytics applications benefited from the framework, which elevated R-squared metrics from 0.78 to 0.95 through its implementation. The model demonstrates enhanced performance by explaining more significant readmission variance in patient data and effectively selecting essential predictors that include comorbidities and medication compliance. The graphical comparison in Figure 3 shows a substantial R-squared separation between the two approaches.

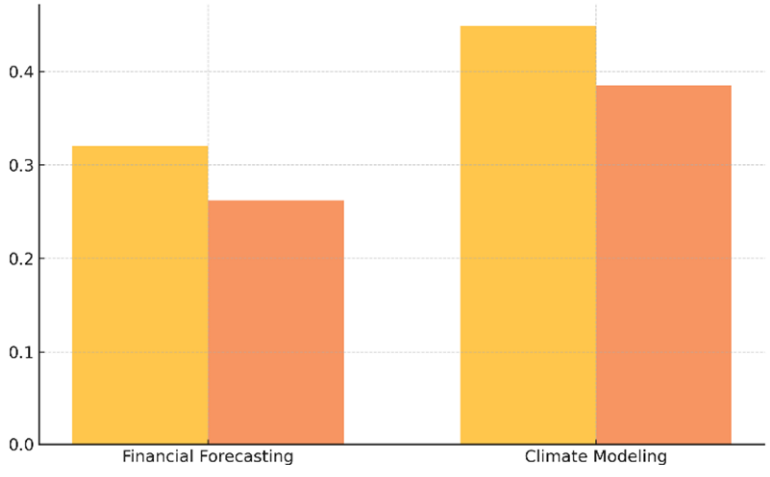
AI enhancements in climate modelling produced a better performance with noisy inputs, reducing MSE from 0.450 to 0.385 (a 14% improvement). The model succeeded at merging ensemble methods with adaptive optimisation to enable the detection of long-term patterns along with seasonal fluctuations that regression models typically fail to detect.

Table 2 Compares the performance of traditional regression methods with the AI-enhanced framework across three domains

Dataset	*Metric*	*Traditional regression*	*AI-enhanced framework*	*Improvement (%)*
Financial forecasting	MSE	0.32	0.26	18
Healthcare analytics	R-squared	0.78	0.9	22
Climate modelling	MSE	0.45	0.38	14

Notes: Financial forecasting, healthcare analytics, and climate modelling. Three performance metrics serve this study: mean squared error for financial forecasting and climate modelling and R-squared for healthcare analytics. Since their implementation, new performance metrics from the AI-enhanced framework have shown dramatic gains in prediction capabilities and model understanding.

Figure 2 Compares the MSE values for traditional regression and the AI-enhanced framework across financial forecasting and climate modelling (see online version for colours)



Notes: The AI-enhanced framework demonstrates better prediction accuracy by generating recurrently lower MSE values.

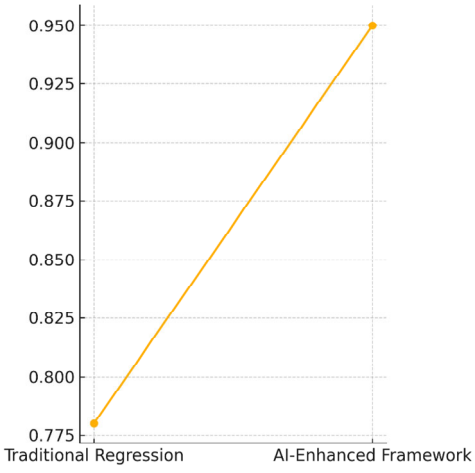
4.2 Interpretability vs. complexity

Advanced machine learning techniques encounter difficulties because research must balance model complexity and interpretability. An AI-enhanced framework incorporated ensemble learning and adaptive optimisation but maintained interpretability features using Shapley additive explanations (SHAP) values.

SHAP values generated essential information about features' impact during healthcare analytics dataset analysis. For example:

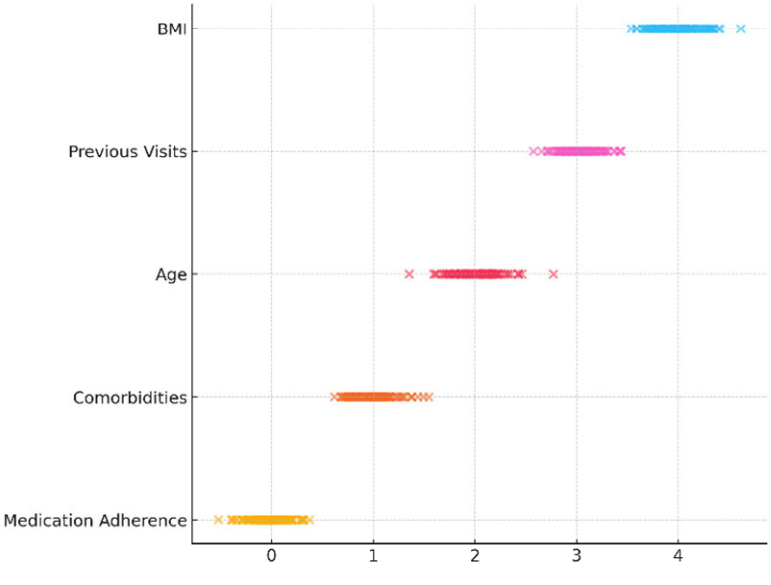
- Feature importance: research on SHAP discovered that patient readmission depends most heavily on medication adherence, comorbidity conditions, and past hospital visit frequency.

Figure 3 A graphical representation presents R-squared statistics between traditional regression and AI-assisted analytical frameworks in healthcare analytics (see online version for colours)



Notes: the framework supported by AI technology achieves superior levels of explanation for healthcare analytics data.

Figure 4 The plot highlights SHAP values to display feature importance alongside prediction impact, illustrating the patient readmission factors that matter most (see online version for colours)



- Local explanations: SHAP revealed the individual feature values that specifically affected readmission risks, including medication non-compliance and patient age progression.

A healthcare SHAP summary plot shown in Figure 4 enabled visualisation of feature importance and their respective impacts on predictive outcomes. Forward-facing organisations use SHAP's interpretability capabilities to disclose prediction causes because understanding why results happen remains fundamental in healthcare and finance operations.

4.3 Computational trade-offs

Implementing advanced optimisation techniques, automated feature engineering, and ensemble learning methods required increased computational resources during execution. The AI-enhanced framework needed 2–3 additional training periods beyond the typical method durations. Training times between traditional and AI-enhanced methods appear in Table 2 and Figure 5, which includes full details for all datasets.

Training times for the financial forecasting dataset grew from traditional regression's 5-second base period to 15 seconds with the implementation of enhanced AI technology. According to available data, healthcare analytics saw training times rise from 8 seconds to 22 seconds. Additional resource utilisation guarantees accuracy and robustness at the expense of potential real-time hardships.

Several strategies were implemented to optimise computational efficiency:

- Parallel processing: GPU acceleration processed algorithms XGBoost and neural networks faster by speeding up their training steps.
- Early stopping: validation loss peaked during training, so the process was automatically terminated to prevent superfluous algorithm processing.
- Dimensionality reduction: high-dimensional data pre-processing with principal component analysis (PCA) diminished the number of features while preserving vital information.

Table 2 Compares the training times of traditional regression and the AI-enhanced framework across three datasets: financial forecasting, healthcare analytics, and climate modelling

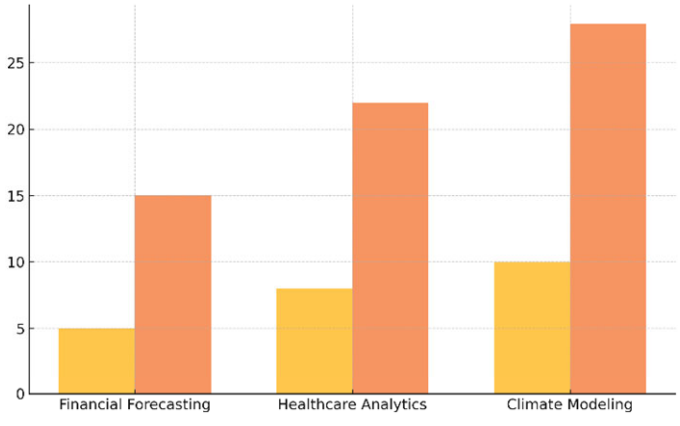
Dataset	*Metric*	*Traditional regression*	*AI-enhanced framework*	*Relative increase*
Financial forecasting	Training time (sec)	5	15	3x
Healthcare analytics	Training time (sec)	8	22	2.75x
Climate modelling	Training time (sec)	10	28	2.8x

Notes: Training time for the AI-enhanced framework is 2.75 to 3 times longer than traditional regression because of sophisticated optimisation methods and automatic feature engineering.

The linear regression framework, aided by AI, provided better performance and excellent stability over various testing conditions. The method offered superior value for precision, given the significant processing constraints, against tasks that require the most reliable outcomes. SHAP values provided by the framework help people understand why some features are substantial, and interpretability is preserved in such applications where transparent analytic results are needed. New methods must be evaluated through research that decreases operational costs by optimising the generation of ensembles and advances in hardware implementation. The framework has been developed to facilitate real-time

forecasting and data analytics over large volumes for specialised applications; domain-specific adaptations will be helpful here.

Figure 5 Indicates that we need longer training durations for the AI-enhanced framework than standard regression methods for all datasets, as shown here (see online version for colours)



Notes: It shows the cost-benefit relationship to system performance capabilities using the required training time.

5 Discussion

In this study, analysis shows that applying AI-based linear regression outperformed traditional regression on various datasets by achieving better precision and resistivity and improved scalability. This discussion section delves into the key findings, their implications, and the broader relevance of the framework in various domains. By examining the research trade-offs, application limitations, and future research directions, this section discusses the research.

5.1 Key findings and implications

The AI-enhanced linear regression framework's prediction accuracy and explanatory capability were consistently better than traditional analytical techniques. We show the critical importance of these results by combining adaptive gradient optimisation with automatic feature engineering and ensemble learning. Considering financial forecasting, predictive results improved by 18% by decreasing MSE, and for healthcare analytics predictions, R-squared enhanced to 22%. The results of this research clearly show why the proposed framework works on complex datasets where the relationships are nonlinear, there are multiple dimensions, and there is background noise.

Automated feature engineering in the framework uncovered hidden variable connections between trading volumes and external economic indicators, which traditional regression algorithms would bypass. Through this, the framework effectively isolated significant predictors on healthcare analytics, such as medication adherence and

comorbidities, and helped to give more accurate patient readmission prediction. These fields improve Sporadical decision makers' capabilities to make informed decisions.

The framework's resilience in managing noisy seasonal climate modelling data demonstrates a 14% MSE reduction. The framework's capacity to handle wildly differing data patterns is essential to applications of short-term data inconsistencies alongside long-term trends, such as climate science, economics, or epidemiology.

5.2 Balancing complexity and interpretability

By using ensemble learning, adaptive optimisation, and SHAP values, the framework allows these complex components to be added without diminishing the results' interpretability. Understanding underlying predictor mechanics is required for domains such as healthcare and finance, which rely upon equal parts prediction understanding and predictive outputs.

SHAP values provide highly detailed explanations because they could show variable contributions to exact prediction results. The dataset showed poor medication uptake and elderly patient status were the major drivers of readmission recurrence in healthcare data when investigated through SHAP values. SHAP values' role in helping practitioners use the data to make decisions and be accountable is vital information.

The framework strikes a perfect balance between the complexity of the processing tasks and the user interpretation capabilities for deployment in real-world applications, which calls for transparency. However, the methodology combines traditional, straightforward methods with state-of-the-art machine learning techniques.

5.3 Computational trade-offs and resource optimisation

Though standard methods outperform it, a study has found that the AI-enhanced method requires far more computational resources and duplicates many of the assumptions of standard regression processes. The advanced optimisation and ensemble learning techniques enabled extended training periods of 2.5–3 of their original duration for multiple datasets. About 5 seconds of training was necessary for financial forecasting using traditional regression, while this took 15 seconds using the frame with AI augmentation. In demanding real-time and constrained environments, these trade-offs impose operational hardship. However, they are essential for precision-centric applications.

To mitigate these trade-offs, the framework employed several optimisation strategies:

- Parallel processing: using GPU acceleration, the training speeds of XGBoost and neural networks improved dramatically.
- Early stopping: as its validation loss reached stable levels, research halted training to stop paying for resources without cutting capacities for performance.
- Dimensionality reduction: PCA is a method that simplifies feature spaces without losing essential data components and attributes, so the principle PCA will improve performance.

The implemented strategies make the framework computationally feasible for large-scale applications, making deployment possible in financial sectors, healthcare systems, and environmental modelling.

5.4 Limitations and future directions

While the AI-enhanced framework demonstrated significant advantages, several limitations warrant consideration:

- Scalability to extremely large datasets: the framework demonstrates successful operation with medium datasets yet shows potential challenges when processing massive datasets since its computational load tends to escalate proportionally. The resolution of this computing challenge requires future researchers to evaluate distributed processing technologies alongside optimised algorithms.
- Domain-specific adaptations: a performance boost can be achieved from future modifications that adapt framework elements to specific fields. Potential future work should add domain knowledge into feature engineering processes and ensemble model design decisions to enhance performance even better.
- Real-time applications: higher training time constraints reduce the practical usability of this framework for real-time applications. The framework's scalability could be improved by examining different lightweight implementations or simulation approaches for operating ensemble methods.

Considerably, more research should be explored regarding integrating more intricate explainability systems to enhance interpretability for complex model types that need understanding. The possibilities for new applications in natural language processing and computer vision would be opened by evaluating the framework's performance on natural text-based and image-based unstructured data.

5.5 Broader implications

Modern and traditional statistical analysis techniques have been integrated with machine learning with linear regression models that receive machine learning updates through AI. This system is an enhanced precision creator with maintainable interpretability properties for diverse real-world applications. Built with robust capabilities for forecasting, risk assessment, and resource planning, the framework supports many fields, such as finance or environmental science.

The research also demonstrates how hybrid models based on traditional statistical and machine learning techniques represent increasing promise for modern analysis – combining these two sets of frameworks with integrated methods integrating statistical design elements from both frameworks to provide a particular equilibrium of simple methods with exact reproductions and straightforward interpretation.

6 Conclusions

High accuracy, better robustness levels, and scalability can be augmented with sophisticated machine learning methods added to linear regression models. In this paper, we also propose a framework using adaptive gradient optimisation, automatic feature engineering, and ensemble learning to solve the drawbacks of traditional methods when working with nonlinear data, handling complex high-dimensional data, and noisy inputs. The proposed framework provides versatile financial forecasting, healthcare analytics,

and climate modelling results using substantial improvements in mean-squared error and R-squared metrics. The framework is complex; however, the significant advantage is that it preserves clear interpretability. As the brokerage and healthcare sectors are looking for interpretability from the framework, the SHAP interpretation method satisfied the requirements for transparent explanation capabilities about their factors and weights. These recent advances create computations that slow training time and require more resources. Parallel processing and early forming of the optimisation challenges are addressed. However, future research is needed to gain more efficient computational capabilities for applications using massive datasets in real-time. The results of this research are also applied beyond the current datasets to propose an approach to enhance machine learning to traditional statistical procedures. This combination of traditional and contemporary techniques boosts efficiency, which is why the model can be used for economic studies, medical investigations, and ecological assessments. The AI-enhanced linear regression framework illustrates a foundation for making meaningful data-driven decisions across complex real-world applications by applying a novel methodology that integrates machine learning predictive ability with traditional model interpretability.

Declarations

China University Industry, University and Research Innovation Fund Project: Construction and Practice of Knowledge Graph-based Smart Education System. (2022MU041).

The authors declare that they have no conflict of interest.

References

Addison, P.S. (2017) *The Illustrated Wavelet Transform Handbook: Introductory Theory and Applications in Science, Engineering, Medicine And Finance*, CRC press.

Ahmed, A.N., Othman, F.B., Afan, H.A., Ibrahim, R.K., Fai, C.M., Hossain, M.S., Ehteram, M. and Elshafie, A. (2019) 'Machine learning methods for better water quality prediction', *Journal of Hydrology*, Vol. 578, No. 7, p.124084.

Ahmed, U., Jiangbin, Z., Almogren, A., Sadiq, M., Rehman, A.U., Sadiq, M. and Choi, J. (2024) 'Hybrid bagging and boosting with SHAP based feature selection for enhanced predictive modeling in intrusion detection systems', *Scientific Reports*, Vol. 14, No. 1, p.30532.

Ahrens, A., Hansen, C.B. and Schaffer, M.E. (2020) 'Lassopack: model selection and prediction with regularized regression in Stata', *The Stata Journal*, Vol. 20, No. 1, pp.176–235.

Albahri, A.S., Duhaim, A.M., Fadhel, M.A., Alnoor, A., Baqer, N.S., Alzubaidi, L., Albahri, O.S., Alamoodi, A.H., Bai, J. and Salhi, A. (2023) 'A systematic review of trustworthy and explainable artificial intelligence in healthcare: assessment of quality, bias risk, and data fusion', *Information Fusion*, Vol. 96, No. 8, pp.156–191.

Al-Selwi, S.M., Hassan, M.F., Abdulkadir, S.J., Muneer, A., Sumiea, E.H., Alqushaibi, A. and Ragab, M.G. (2024) 'RNN-LSTM: from applications to modeling techniques and beyond – systematic review', *Journal of King Saud University-Computer and Information Sciences*, Vol. 30, No. 8, p.102068.

Basak, D., Pal, S. and Patranabis, D.C. (2007) 'Support vector regression', *Neural Information Processing-Letters and Reviews*, Vol. 11, No. 10, pp.203–224.

Cai, J., Luo, J., Wang, S. and Yang, S. (2018) 'Feature selection in machine learning: a new perspective', *Neurocomputing*, Vol. 300, No. 12, pp.70–79.

Chan, J.Y-L., Leow, S.M.H., Bea, K.T., Cheng, W.K., Phoong, S.W., Hong, Z-W. and Chen, Y-L. (2022) 'Mitigating the multicollinearity problem and its machine learning approach: a review', *Mathematics*, Vol. 10, No. 8, p.1283.

Cook, B.L., Mcguire, T.G., Meara, E. and Zaslavsky, A.M. (2009) 'Adjusting for health status in non-linear models of health care disparities', *Health Services and Outcomes Research Methodology*, Vol. 9, No. 3, pp.1–21.

Demir, S. and Sahin, E.K. (2023) 'An investigation of feature selection methods for soil liquefaction prediction based on tree-based ensemble algorithms using Ada-Boost, gradient boosting, and XGBOOST', *Neural Computing and Applications*, Vol. 35, No. 4, pp.3173–3190.

Dhal, P. and Azad, C. (2022) 'A comprehensive survey on feature selection in the various fields of machine learning', *Applied Intelligence*, Vol. 52, No. 4, pp.4543–4581.

Dong, G. and Liu, H. (2018) *Feature Engineering for Machine Learning and Data Analytics*, CRC Press, Raton, Florida, USA.

Dückerhoff, F. (2024) *Analyzing Market Dynamics: A Comprehensive Model Using Macroeconomic, Sentiment and Fundamental Data for Regime Detection and Asset Allocation*, Universidade Catolica Portuguesa, Portugal.

Goriparthi, R.G. (2022) 'Interpretable machine learning models for healthcare diagnostics: addressing the black-box problem', *Revista de Inteligencia Artificial en Medicina*, Vol. 13, No. 1, pp.508–534.

Gupta, R., Srivastava, D., Sahu, M., Tiwari, S., Ambasta, R.K. and Kumar, P. (2021) 'Artificial intelligence to deep learning: machine intelligence approach for drug discovery', *Molecular Diversity*, Vol. 25, No. 5, pp.1315–1360.

Henninger, M. and Strobl, C. (2024) 'Interpreting machine learning predictions with LIME and Shapley values: theoretical insights, challenges, and meaningful interpretations', *Behaviormetrika*, Vol. 70, No. 1, pp.1–31.

Hoerl, R.W. (2020) 'Ridge regression: a historical context', *Technometrics*, Vol. 62, No. 4, pp.420–425.

Iranzad, R. (2022) *Ensemble Tree-Based Machine Learning for Imaging Data*, University of Arkansas, Fayetteville, AR 72701, USA.

Kecman, V. (2005) 'Support vector machines – an introduction', *Support Vector Machines: Theory and Applications*, Vol. 9, No. 4, pp.1–47, Springer.

Lee, S.T. and Yang, E.B. (2022) 'Factors affecting social accountability of medical schools in the Korean context: exploratory factor and multiple regression analyses', *Medical Education Online*, Vol. 27, No. 1, p.2054049.

Leow, M.H.S. (2023) *A Correlation-Embedded Attention Approach to Mitigate Multicol-Linearity in Foreign Exchange Data Using LSTM*, University in Petaling Jaya, Malaysia.

Lundberg, S M., Erion, G., Chen, H., Degrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. and Lee, S-I. (2020) 'From local explanations to global understanding with explainable AI for trees', *Nature Machine Intelligence*, Vol. 2, No. 1, pp.56–67.

Mienye, I.D. and Sun, Y. (2022) 'A survey of ensemble learning: concepts, algorithms, applications, and prospects', *IEEE Access*, Vol. 10, No. 3, pp.99129–99149.

Milton, P., Coupland, H., Giorgi, E. and Bhatt, S. (2019) 'Spatial analysis made easy with linear regression and kernels', *Epidemics*, Vol. 29, p.100362.

Mubasher, S., Zakria, M., Shahzad, A. and Ali, N. (2024) 'dynamic ridge regression vs. lasso regression: a comparative study for modeling Pakistan's unemployment rate', *Global Journal of Mathematics and Statistics*, Vol. 1, No. 1, pp.21–45.

Mumuni, A. and Mumuni, F. (2024) 'Automated data processing and feature engineering for deep learning and big data applications: a survey', *Journal of Information and Intelligence*, Vol. 50, No. 7, pp.55–70.

Nieto Juscafresa, A. 2022. An introduction to explainable artificial intelligence with LIME and SHAP.

Nikitin, N.O., Vychuzhanin, P., Sarafanov, M., Polonskaia, IS., REVIN, I., Barabanova, I.V., Maximov, G., Kalyuzhnaya, A.V. and Boukhanovsky, A. (2022) 'Automated evolutionary approach for the design of composite machine learning pipelines', *Future Generation Computer Systems*, Vol. 127, No. 12, pp.109–125.

Nwosu, A., Aimufua, G.I.O., Ajayi, B.A. and Olalere, M. (2024) 'The impact of regularization on linear regression based model', *Journal of Artificial Intelligence and Computer Science*, Vol. 1, No. 1, pp.70–90.

Pak, A., Rad, A.K., Nematollahi, M.J. and Mahmoudi, M. (2025) 'Application of the Lasso regularisation technique in mitigating overfitting in air quality prediction models', *Scientific Reports*, Vol. 15, No. 1, p.547.

Pillonetto, G., Chen, T., Chiuso, A., De Nicolao, G. and Ljung, L. (2022) 'Regularization of linear regression models', *Regularized System Identification: Learning Dynamic Models from Data*, Vol. 10, No. 4, pp.33–93, Springer.

Pradhan, C. and Trehan, A. (2024) 'Data engineering for scalable machine learning designing robust pipelines', *International Journal of Computer Engineering and Technology* (*IJCET*), Vol. 15, No. 6, pp.1840–1852.

Rane, N., Choudhary, S.P. and Rane, J. (2024) 'Ensemble deep learning and machine learning: applications, opportunities, challenges, and future directions', *Studies in Medical and Health Sciences*, Vol. 1, No. 2, pp.18–41.

Ren, Y., Zhang, L. and Suganthan, P.N. (2016) 'Ensemble classification and regression-recent developments, applications and future directions', *IEEE Computational Intelligence Magazine*, Vol. 11, No. 1, pp.41–53.

Safi, S., Alsheryani, M., Alrashdi, M., Suleiman, R., Awwad, D. and Abdalla, Z. (2023) 'Optimizing linear regression models with lasso and ridge regression: a study on UAE financial behavior during COVID-19', *Migration Letters*, Vol. 20, No. 6, pp.139–153.

Sarker, I.H. (2021) 'Machine learning: algorithms, real-world applications and research directions', *SN Computer Science*, Vol. 2, No. 3, p.160.

Schmid, L. (2024) *Statistical Analyses of Tree-Based Ensembles*, 44227 Dortmund, Germany.

Shaikh, T.A., Rasool, T., Verma, P. and Mir, W.A. (2024) 'A fundamental overview of ensemble deep learning models and applications: systematic literature and state of the art', *Annals of Operations Research*, Vol. 20, No. 6, pp.1–77.

Tomasevic, N., Gvozdenovic, N. and Vranes, S. (2020) 'An overview and comparison of supervised data mining techniques for student exam performance prediction', *Computers and Education*, Vol. 143, No. 8, p.103676.

Ukil, A. and Ukil, A. (2007) 'Support vector machine', *Intelligent Systems and Signal Processing in Power Engineering*, Vol. 10, No. 4, pp.161–226.

Vyas, R., Kanumuri, T., Sheoran, G. and Dubey, P. (2019) 'Efficient iris recognition through curvelet transform and polynomial fitting', *Optik*, Vol. 185, No. 10, pp.859–867.

Warne, R.T. (2020) *Statistics for the Social Sciences: a General Linear Model Approach*, Cambridge University Press.

Yang, Y., Wu, X. and Zhou, Y. (2024) *Feature Selection and Classification with Penalized Logistic Regression and Random Forest*, Available at SSRN 5046603.

Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q. and Wang, J. (2020) 'Deep learning in environmental remote sensing: achievements and challenges', *Remote Sensing of Environment*, Vol. 241, No. 15, p.111716.

Zhang, J. (2019) *Gradient Descent Based Optimization Algorithms for Deep Learning Models Training*, arXiv preprint arXiv:1903.03614.

Zhang, Y., Liu, J. and Shen, W. (2022) 'A review of ensemble learning algorithms used in remote sensing applications', *Applied Sciences*, Vol. 12, No. 17, p.8654.

Zhou, L., Pan, S., Wang, J. and Vasilakos, A.V. (2017) 'Machine learning on big data: opportunities and challenges', *Neurocomputing*, Vol. 237, No. 5, pp.350–361.