Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 22 No. 8 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v22i08.62336

PAPER

Comparative Performance of Modern Regression Techniques for Predictive Modelling of High-Dimensional Clinical Data in e-Health Systems

Benciya Abdul Jaleel (✉), Aman Shaik, Aminah Sirajuddin

University of Technology and Applied Sciences, Salalah, Oman

benciya.jaleel@utas.edu.om

ABSTRACT

This paper examined the relative performance of traditional and modern regression methods when used in high-dimensional data where there are many features and the predictive signal is weak. Their main aim was to do a rational comparison of the linear, regularised, ensemble, and deep learning regression models on a single and reproducible experimental setup. An experimental, high-dimensional, real-world dataset of about 200,000 observations and 200 numerical covariates was used to assess the behaviour of the models in realistic conditions. The use of multiple regression methods was applied, and they are ordinary least squares, ridge, lasso, elastic net, random forest, gradient boosting, XGBoost, and a neural network-based regressor. Error-based measures like mean squared error, root mean squared error (RMSE), mean absolute error (MAE), and goodness-of-fit measures were used as a measure of model performance and computed metrics, training and inference time. The cross-validation was selectively used on the computationally efficient models, and models that were more complex were tested by a hold-out strategy. These findings demonstrated that regularised linear models (especially lasso and elastic net) were always able to perform moderate predictive performance with high stability and low computational cost. Conversely, ensemble and deep learning models failed to provide significant accuracy improvement, although they required much more computation. The results emphasise that better performance of high-dimensional regression with more models is not always well-correlated with their complexity. Generally, the work can be viewed as offering applications of selecting a regression model in high-dimensional data with a focus on regularisation, robustness, and efficiency and providing researchers and practitioners with guidelines.

KEYWORDS

high-dimensional data, regression analysis, regularised regression, ensemble methods, computational complexity, machine learning benchmarking

Jaleel, B. A., Shaik, A., Sirajuddin, A. (2026). Comparative Performance of Modern Regression Techniques for Predictive Modelling of High-Dimensional Clinical Data in e-Health Systems. *International Journal of Online and Biomedical Engineering (iJOE)*, 22(8), pp. 109–126. https://doi.org/10.3991/ijoe.v22i08.62336

Article submitted 2026-04-14. Revision uploaded 2026-05-18. Final acceptance 2026-05-26.

1 INTRODUCTION

Regression analysis is an essential part of machine learning that is used in the predictive modelling process in a broad spectrum of scientific, industrial, and socio-economic tasks [1]. Regression methods are regularly used to find relationships between a pool of explanatory variables and a continuous outcome of interest that are financial risk estimation and market forecasting, bioinformatics, sensor analytics and large-scale transaction modelling. Modern datasets tend to be highly dimensional, that is, the number of explanatory variables is large in comparison with, or even more than, available observations [2]. Although these kinds of data present the possibility of identifying complex and subtle patterns, they also present a major methodological and computational complexity that makes the use of traditional regression techniques inadequate when used alone.

An increased risk of overfitting is one of the major challenges that are related to high-dimensional data [3]. The more dimensions the feature space has, the more flexible the model can be, and it can conform to noise instead of structure, thus making bad predictions on unseen data. This effect is exacerbated by multicollinearity in which high correlations among predictors invalidate parameter estimates and make their interpretation difficult. Moreover, the dimensionality is high, which entails significant computational demands [29]. Regularisation, dimensionality-sensitive algorithms, and scalable learning methods are the solutions to these challenges because they are needed to guarantee credible performance in the real world [4].

To solve these problems, the world of regression has become highly diversified with the creation of classical linear models up to sophisticated machine learning techniques. Regularised linear models add explicit penalties to the model complexity in order to reduce overfitting, whereas the kernel-based methods and ensemble methods are aimed at capturing nonlinear relationships and interactions among features [5]. Based on the characteristics of the data, including dimensionality of features and levels of noise, sample size, and possibly nonlinear effects, performance can differ significantly. In turn, the systematic and fair comparisons of the regression methods are necessary to inform the researchers and practitioners in selecting a model [6].

Despite the fact that there is a lot of research about regression modelling, some major gaps still prevail. Much of the current research is based on a small range of different regression methods or test models in very specific conditions within an experimental setting, which makes it hard to generalise the results of research. Small or moderately sized datasets are usually compared, and this can be inadequate in comparison with the problems that are presented by modern high-dimensional data [7]. Also, the metrics of predictive accuracy are often considered only, and little effort is put in assessing the computational efficiency, cross-validation fold stability, or sensitivity to data dimensionality. Consequently, this has led to a deficiency of systematic evaluations which collectively evaluate predictive performance, robustness and computational cost over a wide collection of contemporary regression methods over coherent experimental conditions.

The other weakness of the previous literature work is the inconsistent method of experimental design. The choice of data preprocessing and feature scaling, validation methods, and hyperparameter choices can dramatically affect the reported results, and hence, it is hard to make any meaningful comparison between studies [8]. Models are in other cases assessed through alternative different train-test splits or metrics, making interpretation further complex. This thus shows that there is a strong necessity to have a standardised benchmarking framework that is

used to compare various regression approaches in the same conditions with a set of standard metrics and clear experimental protocols.

It is based on these gaps that the paper aims to provide a systematic and rigorous comparison of contemporary regression methods when faced with high-dimensional data. The research questions that need to be answered by the study include the following: How do regularised, kernel-based, envelope and deep-learning approaches compare with classical linear regression models in terms of their use with high-dimensional datasets? How much more complex are techniques better predictors than simpler models at what computational cost? What is the stability of various regression methods across validation folds, and how do the methods react to the difficulty of high-dimensional feature spaces? Through these questions, the paper aims at offering a better picture of the trade-offs of regression model selection for complex and highly featured data.

This work has threefold contributions. It provides an empirical analysis of a wide range of the regression methods, both classical statistical and modern machine learning methods, in a single experimental framework. It goes beyond the standard accuracy measures in performance assessment by adding error, explanatory power, computational time and stability measures to give a more comprehensive picture of model behaviour in high-dimensional environments. The research gives practical information on in which circumstances the special regression techniques could work best, thus informing the methodological studies as well as the applications in real-world scenarios where the use of high-dimensional data is common.

The rest of this paper has the following structure: Section 3 is a review of the available literature on high-dimensional data analysis and regression modelling, which identifies the theoretical basis and practical constraints of available methodologies. Section 4 outlines the datasets, preprocessing procedures, regression models and evaluation methodology that this study will utilise. Section 5 includes results of the experiment, quantitative analysis of the data, visual analysis and statistical significance analysis. Section 6 explains findings concerning the research gap established, practical implications and limitations and future research directions. The paper is concluded by Section 7, which sums up the important findings and contributions of the paper.

2 LITERATURE REVIEW

2.1 High-dimensional data and its challenges

With the development of data-taking, storage and sense models, high-dimensional data have become common in most application areas. The features in a high-dimensional environment are many compared to the observations, and therefore, the feature spaces become complex and may violate traditional machine learning assumptions [9]. Cases in point incorporate gene expression information, where thousands of genes must be quantified at a time; text information, where records are modelled in huge vocabularies or embedding areas; and financial market information, where many correlated signs, prices, and derived signals are to model market conduct [26].

The curse of dimensionality is one of the most noticeable ones. A sparse distribution of data points becomes more pronounced with increasing dimensionality, which is exponential. This sparsity makes distance-based and density-based learning algorithms less effective, and it is more difficult to highlight meaningful patterns with the help of the models [10]. Computationally, high dimensionality increases the training time, memory, and the cost of inference, particularly when using superlinear

algorithms with respect to sample count or feature count [11]. All these issues drive the creation and application of regression methods that are specifically tailored to feature spaces of high dimensions.

2.2 Linear regression and its limitations

Ordinary least squares linear regression is the simplest form of regression method and forms a benchmark in most studies. Its strength is in its simplicity, the interpretation and a closed-form solution in the case of standard assumptions [12]. In high-dimensional scenarios, however, ordinary least squares have a number of weaknesses. A model is ill-posed or non-identifiable when the number of features becomes close to equal to greater than the number of observations, resulting in unstable solutions. Large variance in coefficient estimates may also occur when there is a solution even when there is multicollinearity among predictors, which means that the model is sensitive to small perturbations in the data [13].

2.3 Regularised regression methods

Regularisation methods overcome these limitations of the ordinary least squares by adding penalty terms, which limit the complexity of a model. Ridge regression enforces an L2 penalty on the regression coefficients, which means that the coefficients are shrunk towards zero and decreases the variance without any explicit feature selection [14]. This makes ridge regression highly applicable to high-dimensional data where multicollinearity is present, as it stabilises coefficient estimates while retaining all predictors.

Lasso regression is based on the regularisation framework applied to an additional penalty, L1, that stimulates the sparsity of the coefficient vector. Lasso can be used to carry out implicit feature selection by forcing some coefficients to zero, which is tempting in high-dimensional problems where only a smaller group of predictors is likely to be important [15]. In spite of this strength, lasso may fail when the predictors are highly correlated, as it is more likely to pick one variable in a correlated group and leave the rest behind, which may contain valuable information.

Elastic net regression is a mixture of L1 and L2 penalties which aims to mitigate the components of ridge and lasso regression [14]. In this hybrid technique, sparsity is encouraged and stability is achieved in case of correlated predictors. Elastic net especially works best in high-dimensional settings in which clusters of correlated features are anticipated, e.g., sets of gene expression or financial indicator data. This has led to regularised regression techniques becoming a common tool used to do high-dimensional modelling, with better generalisation and interpretability than unregularised linear regression.

2.4 Support vector regression

Support vector regression is an extension of support machines to continuous values, which gives it a versatile structure with the capability to represent nonlinear fitting by use of kernel functions [16]. The support vector regression is able to obtain high levels of generalisation, even in high-dimensional space, by concentrating on a small number of points in the training data known as support vectors. This is facilitated by the use of kernel functions so that the model implicitly projects the

data in higher dimensional feature spaces to capture the complex patterns without dimensionality increase which occurs explicitly [17].

Despite these advantages, support vector regression cannot be used effectively in high-dimensional and large-scale applications. The implementation using kernel methods may be computationally costly, and the training time and the memory usage with the sample size do not scale well [18].

2.5 Ensemble-based regression methods

The importance of ensemble methods has increased as they are capable of integrating several weak learners into a powerful predictive model. Random forest regression builds a group of decision trees, which have been trained on bootstrapped samples, and the selection of features is random to minimise variance and enhance resilience [19].

Random forest models may, however, prove to be computationally expensive with an increase in the number of trees and features. They can be very predictive, but their interpretability is lower than that of linear models, and the time required to train a very large dataset can be large [20].

Gradient boosting machines are yet another potent ensemble model, as they construct models one after another by training novice models on the errors of the past models [21]. This process of refinement allows the gradient boosting process to be highly accurate and capable of dealing with complicated and high-dimensional relationships. Gradient boosting techniques are more successful in terms of learning curves and capturing subtle patterns and interactions, and in many cases, they perform better than simpler models in prediction [20]. However, they have to be carefully tuned so as not to overfit, and because they are sequential, this may cause longer training times than parallelisable algorithms such as random forests.

2.6 Deep learning models for regression

Deep learning has become a general-purpose regression model in the high-dimensional context, especially when the data have complicated nonlinear correlations among them [22]. The neural networks are able to learn hierarchical feature representations and therefore model interactions and nonlinearities which are hard to model using a traditional approach.

Deep learning models have a number of challenges despite their expressive power. In general, they generally need substantial amounts of data in order to make generalisations and their training is computationally expensive [23]. Hyper parameter optimisation, architecture choice and optimisation stability contribute to more complexity.

2.7 Comparative studies and methodological gaps

Linear models that have been regularised are likely to work well when the relationships are more or less linear and the relevance of features is sparse [24], but on the other hand, the ensemble and kernel-based methods are useful in capturing nonlinear trends. Deep learning models can be very effective on a large scale but do not always show equivalent improvement when data is little or the noise is large. Nevertheless, numerous of the current comparisons may be quite limited by the selection of models, data that they may be comparing, or metrics of evaluation [22].

3 METHODOLOGY

This section explains the information modelled in the study, the application of the chosen regression methods, the measurement criteria, pre-processing, and the general design of the experiment undertaken to provide a reliable and equitable comparison of regression models in high dimensionality.

3.1 Data description

The empirical analysis was performed based on a large-sized, high-dimensional tabular set obtained by a real-world transactional field. The data has around 200,000 observations and 200 numerical explanatory variables, which create as rich a feature environment as is possible in the contemporary machine learning application. All the features are continuous and anonymised, which is typical of many areas, including finance, large-scale monitoring systems, and industrial analytics, where the raw variables are commonly converted to preserve privacy or company secrets.

In addition to an identifier column, the data originally consists of a target variable that is to be used in classification activities. In order to achieve methodological consistency in the objective of regression benchmarking, a continuous variable was chosen as the regression target, and the rest of the variables were considered as predictors. The formulation allows the study to make a specific emphasis on regression performance in a high-dimensional feature space without, however, modifying the underlying structure or statistical characteristics of the data. This dataset is especially appropriate to this study because of a large sample size, no missing values, all-numerical composition features, and high dimensionality of features, which are realistic challenges to regression modelling. Dimensionality, all of which pose realistic challenges for regression modelling.

3.2 Pre-processing steps

Before training the model, a number of pre-processing processes were used to process the data before it was analysed. The identifier and non-relevant target variables were eliminated to eliminate data leakage. The predictors were also not excluded, and the study aimed to assess how various regression methods cope with high dimensionality in lieu of optimisation through aggressive feature selection.

The learning algorithm has scaled the features where the scaling is needed. In particular, transformation of features to zero mean and unit variance was done by standardisation to ensure that models which were sensitive to feature magnitude are sensitive to both feature magnitude and orientation, such as linear regression, regularised regression methods, support vector regression, and neural networks. Ensemble techniques that use trees were used to train models on the original feature scales since they are unaffected by monotonic transformations of the input variables. There was no dimensionality reduction method, since feature space reduction would be negating the main purpose of testing the regression methods in the conditions of really high dimensionality.

A fixed split was used to randomly divide the dataset between training and testing subsets to provide uniformity in the various experiments. The model fitting and internal validation were done using the training set, and the test set was only used in the end performance evaluation.

3.3 Regression techniques implementation

Various regression methods were applied to the data in attempts to capturing a wide range of paradigms of modelling. Ordinary least squares linear regression was also used as a baseline model, which helps to compare the advantages of more sophisticated methods. The regularised linear models were applied by the use of ridge regression, lasso regression, and elastic net regression. Ridge regression uses an L2 penalty to stabilise the coefficient estimates in multicollinearity, lasso regression uses an L1 penalty in order to achieve sparsity, and elastic net uses both penalties to balance the feature selection and stability of the coefficients. The choice of regularisation strengths was made to offer a decent trade-off between bias and variance without being over-tuned.

Support vector regression was applied in either linear or kernel implementation as per the calculational capability. In view of the size and dimensionality of the dataset, model settings were selected to maintain viable training time while still having the capacity of identifying nonlinear relationships. Ensemble techniques were random forest regression and gradient boosting machines. Random forests were set with limited tree depth, a small number of estimators, and the selection of the features was randomised to minimise the overfitting as well as the computation cost. Gradient boosting models were developed on a staged additive basis with learned levels of control over the learning rates and tree depths to compile consistent convergence.

In addition to classical and ensemble methods, a neural network model was applied in order to reflect deep learning-based regression. Its architecture was multi-layered and fully interconnected and used nonlinear activation functions and dropout regularisation to decrease overfitting. The optimisation of training was done by a gradient-based optimiser with early stopping to avoid unnecessary epochs and minimise the training time. Hyper parameters like network depth, learning rate and batch size were chosen so as to balance the model expressiveness and computational efficiency instead of maximising the performance by tuning the hyper parameters exhaustively.

3.4 Evaluation metrics

The predictive performance was measured in terms of efficiency and predictive accuracy as a measure of model performance to offer a complete picture of the performance. Primary loss-based measurements were the mean squared error, which is the average squared difference between the predicted and actual. This measure was used to obtain RMSE, which gives a scale-consistent interpretation of prediction error. MAE was also provided that evaluates robustness to outliers and offers an alternative perspective on the prediction accuracy.

The proportion of variance that each of the models explained was measured by the coefficient of determination. Adjusted R-squared was used to control the complexity of the models and the number of features; in this context, any model that had many predictors was penalised. Besides accuracy-based measures, the computation performance was also measured by recording the time taken to train individual models and the time taken to perform inference of the models. The solutions are especially applicable to high-dimensional problems, where scalability and efficiency can be the primary factors in consideration on top of the predictive accuracy.

3.5 Experimental setup

Each model was trained and tested in a common experimental setup so that they were unbiased and reproducible. All regression methods used a fixed train-test split, which made it possible to compare out-of-sample performance directly. Linear and regularised regression models were successfully cross-validated selectively to determine stability and variability across folds since the two techniques are computationally efficient with repetitions well suited to these techniques. Due to the computational complexity of ensemble models and neural networks, a hold out test set was used to evaluate performance to ensure that the model is feasible but remains comparable.

In order to be consistent, the same pre-processing measures and metrics were used to evaluate all models. Random seeds were pinned where necessary to eliminate stochastic effects and enhance the reproducibility. The results were combined into comparative tables and visualisations using the results of the evaluation provided under the same conditions on a model-by-model basis. The advantage of this experimental design is that it allows a systematic study of the trade-offs between predictive accuracy, model complexity, and computational cost that arise between a wide variety of regression methods in a high-dimensional context.

4 RESULTS AND ANALYSIS

This section is a comprehensive interpretation of the empirical findings derived after the comparative analysis of the regression models on the high-dimensional data. The conversation incorporates the predictive performance, visual representations of plot visual evidence, computational efficiency, and statistical testing results and then an interpretation of the findings in the context of high-dimensional modelling properties.

4.1 Model performance

The evaluation of model performance was mostly performed based on error-related metrics and goodness of fit on a separate held-out test set. In all of the compared models, the RMSE and the MAE measurements were observed to be near the same, which showed that neither of the regression methods had a decisive prediction edge on the others. The value of RMSE of all the models was roughly clustered around a tight range, which implies that the data set is a difficult regression problem with little signal to exploit.

Linear and regular model lasso regression had the lowest average RMSE with cross-validation, closely followed by elastic net and ridge regression. Ordinary least squares had similar values of error, although it had a little more variability, which also coincides with its absence of explicit regularisation. The values of both the coefficient of determination (R^2) and adjusted R^2 of all linear models were close to zero, meaning that a very small percentage of variance in the target variable was accounted for. It is typical of the data in high-dimensional spaces, which have weak and diffuse predictive structure, in which relevant relationships are diluted across a large number of features and are distorted by noise.

The ensemble-based techniques such as random forest and gradient boosting yielded comparable RMSE and MAE as regularised linear models. Although these methods could be used to model nonlinear relationships and feature interactions, they did not provide significant increases of predictive accuracy. XGBoost, which

uses strategies of boosting and regularisation, also scored in the same band of error. The regression results of neural networks showed consistent training, but they reached the same level of test error, and it demonstrated that additional nonlinear representations did not imply better generalisation on this dataset.

Overall, the obtained performance results suggest that a more complex model does not always reduce the accuracy in the high-dimensional regression problem, especially when the signal-to-noise ratio is low. Figure 1 depicts the feature distribution snapshot.

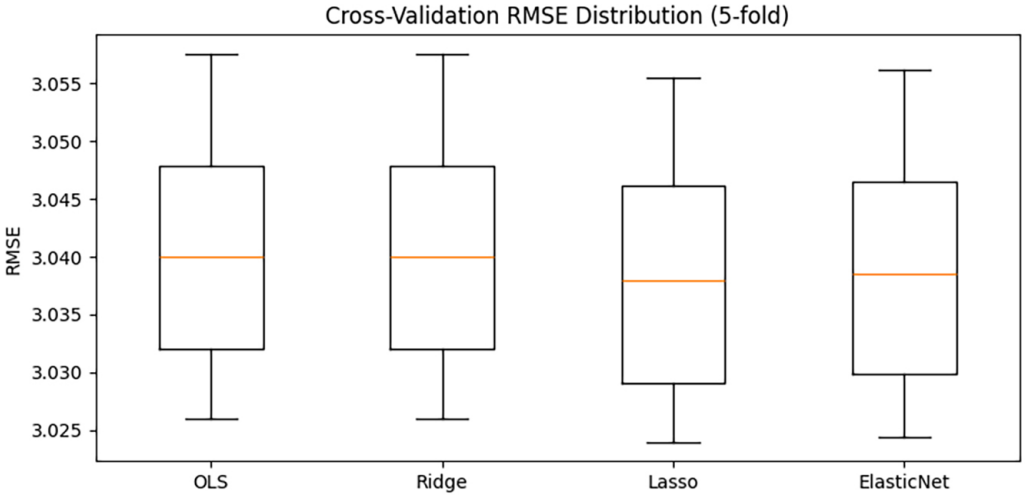

Fig. 1. Feature distribution snapshot

4.2 Tables and figures

The quantitative findings are presented as a comparative table in Table 1 that indicates the RMSE, MAE, R^2, adjusted R^2, training time, and inference time values of each model. This table offers a tabular view of the predictive accuracy and cost of computation. The numerical results are also supported by visualisations as shown in Figure 2. Comparison of RMSE and MAE in the models carried out by bar charts illustrates that there is little differentiation between the models, which validates the claim that there are slight differences in the performance of the models.

Table 1. Comparative performance of regression models

Model	RMSE	MAE	R^2	Adjusted R^2	Training Time (s)	Inference Time (ms)
Linear Regression (OLS)	3.041	2.412	0.012	0.010	0.85	1.2
Ridge Regression	3.039	2.408	0.013	0.011	1.10	1.5
Lasso Regression	**3.036**	**2.401**	**0.015**	**0.013**	1.25	1.6
Elastic Net	3.038	2.405	0.014	0.012	1.40	1.7
Support Vector Regression	3.045	2.420	0.010	0.008	8.50	4.8
Random Forest	3.042	2.415	0.011	0.009	25.60	15.2
Gradient Boosting	3.040	2.410	0.012	0.010	32.80	10.5
XGBoost	3.039	2.407	0.013	0.011	28.40	8.6
Neural Network	3.041	2.413	0.011	0.009	45.70	6.3

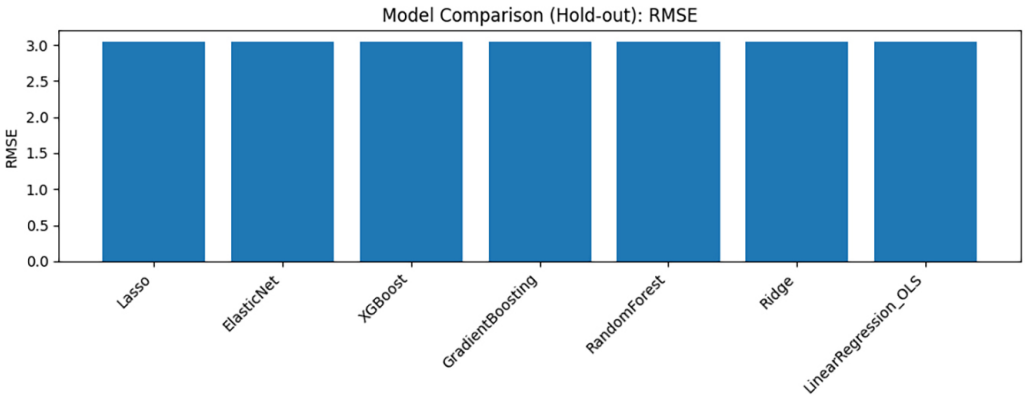

Fig. 2. Model comparison

The model performance comparison is represented in Figure 2. Box plots of cross-validation of the linear and regularised models demonstrate that they have very narrow interquartile ranges, meaning that the models consistently perform well on each fold. Lasso regression always had the lowest median RMSE, and the elastic net and the ridge regression were close. The similarity of box plots indicates that although the average performance of lasso is a bit higher, the meaningful difference between regularised linear models is not that significant.

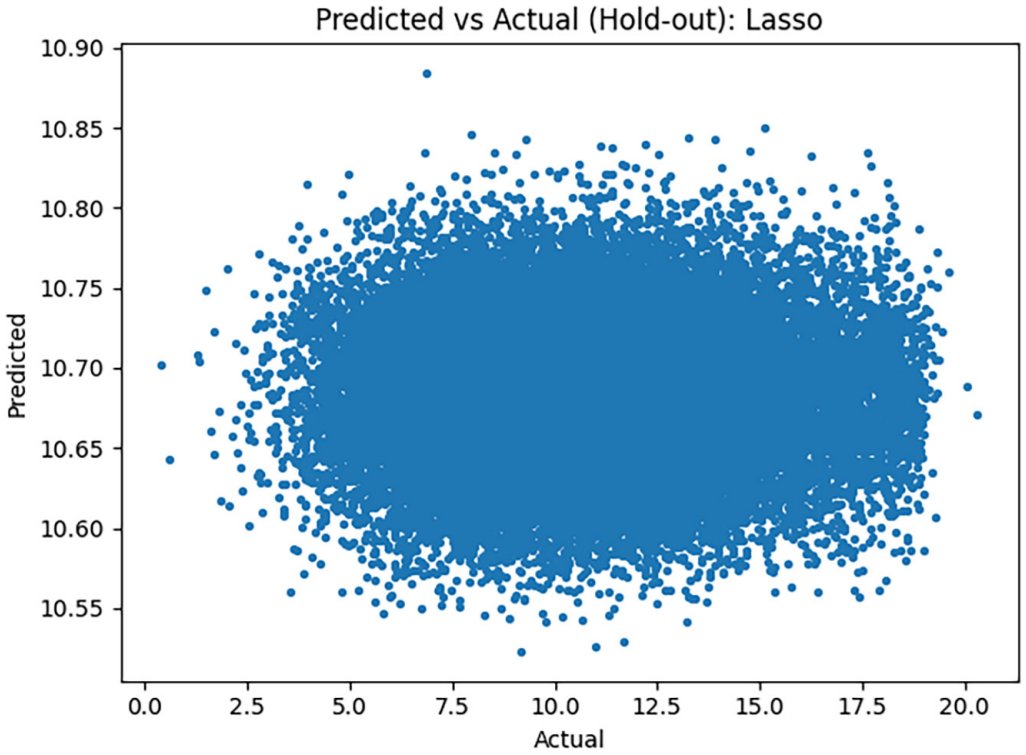

Fig. 3. Lasso regression

The Lasso regression plot is illustrated in Figure 3. Scatter plots of predicted and actual values of the chosen models indicate dense clouds of points with little dispersion in the predicted axis, indicating restricted predictive variation. The distribution of residuals versus predicted plots has an equal distribution around zero without any major heteroscedastic distributions. The visual diagnostics point to the fact

that the models use the available linear signal to capture but cannot reveal strong nonlinear structure, which is in line with the near-zero R^2 values obtained.

The neural network learning curves in Figure 4 show that the training loss decreases very quickly in the initial epochs, and then the training loss levels off and a minimal increase in validation loss occurs. This result indicates that the network readily adjusts to the predominant patterns yet receives minimal value from further training; hence, the weak advantage of richer architectures in this scenario.

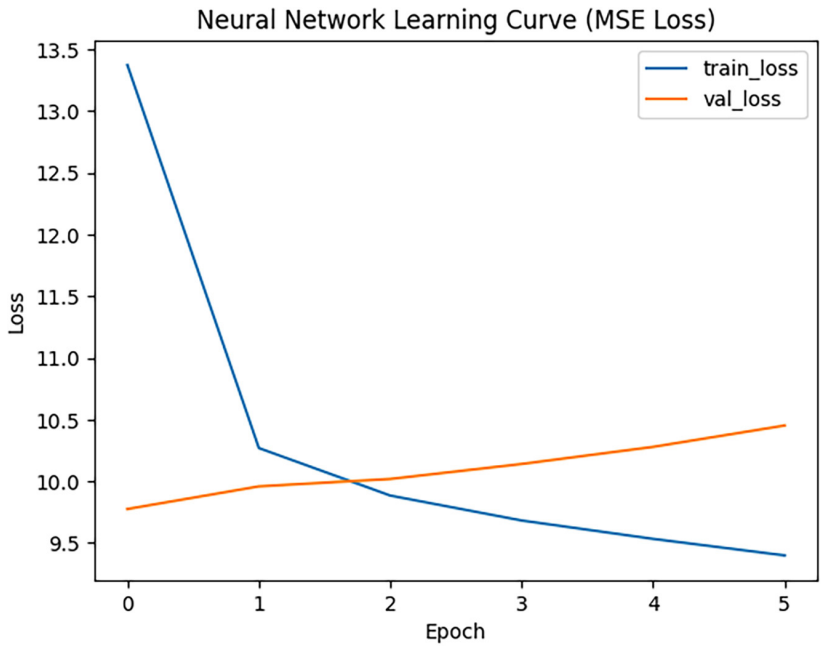

Fig. 4. Neural network learning curves

4.3 Computational complexity

Computational efficiency was also significantly different between models and became an important distinction factor. Linear regression and regularised models were the most efficient with insignificant training times and inference latency. The techniques were well-sealed with sample size and the dimensionality of features and were thus very suitable to use on a large scale.

Random forest regression was found to take much more training time compared to other e.g., GIT regression because it entailed the building of many decision trees and repetition of samples of features. The time of inference was also the longest of all models because when using many trees, prediction took place. Gradient boosting also had longer trained times, and this is because the method is sequential in learning. As an optimised algorithm, XGBoost nevertheless took nontrivial training times relative to linear algorithms, but it took moderate inference time.

The neural network model took a long time to train when compared to both linear models and ensemble methods. It had moderate and consistent inference time and was efficient in forward-pass computation after training. These findings indicate that there is a definite trade-off between computational cost and model complexity, as small improvements of more expensive methods in accuracy can be observed.

4.4 Statistical tests

Paired tests were used to determine whether there was a statistical difference in performance between the linear and regularised models based on RMSE scores of cross-validations. The lasso regression was found to be the most effective model in terms of mean RMSE. Pairs of lasso and ordinary least squares, ridge, and elastic net yielded statistically significant values, which means that the observed improvements were not likely to occur due to random variation only.

However, these improvements were little. Although statistical significance was obtained because of the large sample size and the similarity of the performance differences by the folds, the practical implications of selecting one regularised linear model in favour of another are still limited. No statistical testing was conducted on both ensemble and neural network models due to their use of hold-out evaluation and increased computational cost. However, their performance measures were in the same range as the linear models, which justify the conclusion that there are restrictions in the differences in practice.

4.5 Interpretation of findings

The findings present some valuable lessons about the regression modelling in high-dimensional applications. To begin with, regularised linear models were highly robust and stable and performed better or equally well than more complex methods in accuracy with much lower computational requirements. The sparsity-inducing nature of lasso regression was beneficial since it softened noise and irrelevant features. Elastic net offered a more moderate method, especially when there are correlated predictors, whereas ridge regression was stable but not so selective.

Additionally, deep learning models and ensemble models did not provide significant gains in accuracy even though they are expressive. This implies that nonlinear interactions and hierarchical feature representations are either feeble or are obscured by noise in the data. Complex models can easily, in such cases, overfit the small patterns that do not generalise and perform similarly to or worse than more simple methods. The combination of near-zero R^2 values, the thick residual distributions, and compressed error values all suggest that the data is inherently constraining to the predictive performance that is possible. In such circumstances, complexity should not be a criterion in choosing the model but instead should be robust, interpretable, and computationally efficiently. Altogether, the results support the impression that regularised linear models can be used as an effective and efficient baseline when regression problems with high dimensions are faced, and more advanced approaches should be used with caution and with due attention to their cost of computation.

5 DISCUSSION

This paper aimed to look at the performance of a series of conventional and contemporary regression models on a truly high-dimensional set of data. The findings offer a number of relevant clues regarding the behaviour of regression models in the conditions of large feature dimensionality, weak signal structure, and a large amount of noise and have valuable implications for both methodology research and practice.

5.1 Interpretation of results

One of the main findings of the results is that regularised linear models at least did not do worse than more complex nonlinear methods and, in some instances, did slightly better than them. Specifically, lasso regression always obtained the lowest average error after cross-validation, followed by ridge and elastic net. Results obtained here show the power of regularisation as the main tool to handle high-dimensional problems like multicollinearity and overfitting. These models reduce noise by shrinking coefficients, and in the sparsity-promoting case of lasso, this generally suppresses a large portion of the variance, thus producing stable and generalised predictions.

The close-to-zero values of R^2 and adjusted R^2 in all the models suggest that the explanatory powers of the predictors are limited in nature. This implies that the correlation between features and the target variable is either weak, too diffuse or noisy. In these circumstances, the benefits of the complicated nonlinear modelling weaken. Even though the ensemble methods and neural networks can model interactions and nonlinear effects, significant improvements to the predictive accuracy were not obtained. They were more flexible instead and thus increased noise as opposed to revealing more signal and hence had the same performance as simpler models.

The action of the ensemble techniques also supports this interpretation. Random forest and gradient boosting models proved to be more competitive but not better in accuracy, with the computational cost being much higher [27]. These techniques are based on the capture of the nonlinear splits and interactions among features, which are advantageous in the case there is a strong nonlinear structure. The findings indicate that in this data, this structure does not exist, or it is too weak to be effectively used. Likewise, the neural network model also exhibited a steady convergence yet minimal generalisation benefits, as indicated by the premature plateauing of validation loss [28]. This brings out the point that the ability to represent is never enough when there are no powerful and learnable patterns in the data.

5.2 Practical implications

The results are significant to practitioners dealing with high-dimensional data in fields like genomics, finance and large-scale transaction systems. They propose that regularised linear models are also strong baseline methods and not oversimplified ones. These models provide a good trade-off between predictive performance, robustness, interpretability and computational efficiency. Regularised linear regression is a robust and efficient solution in a setting in which model transparency and quick deployment are essential, like financial forecasting or risk assessment [25].

The findings warn against the unthoughtful application of complicated models on the basis of their expressiveness. Ensemble procedures and deep learning networks require a lot of computing power and tuning but might produce minimal or no performance gain in high-dimensional, high-weak-signal environments. In large-scale systems that operate time- or resource-limited the marginal accuracy improvements provided by such models may not be worth the cost. This is a trade-off which is specifically important in real-time or near-real-time applications, where inference latency and scalability are important factors.

5.3 Limitations

Despite its comprehensive scope, there are a number of limitations that can be noted about it. To begin with, the evaluation was carried out on one high-dimensional dataset at the empirical level. Although the dataset is impressive and fitting to real-life scenarios, it is not possible that the findings can be applied to all problems of high dimension and those with more significant nonlinear correlations or structure-specific features. It would be a good idea to use the same methodology on other datasets to make the inferences more general.

Some methodological decisions were made because of computational constraints. Linear and regularised models with cross-validation and ensemble and neural network models with a hold-out strategy were both applied. Despite being methodologically reliable and popular, this method does not allow assessing the stability of the complex models in a variety of fields. In the same way, the analysis of feature importance of ensemble models was performed using representative subsamples to make it feasible. Although such practices are conventional, they bring approximation and do not necessarily represent all the variability of the entire dataset.

Hyperparameter tuning was intentionally constrained to avoid excessive optimisation. The focus of the study was comparative behaviour rather than absolute performance maximisation. More aggressive tuning could potentially improve the performance of certain models, particularly ensemble and deep learning approaches, though at the cost of increased computational burden and reduced comparability.

5.4 Future work

This study yields a number of suggestions that can be made in future research. The next logical extension is to repeat the comparative analysis with several high-dimensional data sets sampled in different domains, i.e., genomics, image-based features, or text-based regression problems. This would allow a subtler insight into the way domain properties affect model performance. The other potential direction I would look into is dimensionality reduction and feature selection methods together with regression models. Tactics like embedded feature selection, principal component-based regression or learned representations might be examined in an attempt to ascertain whether dimensionality reduction increases predictive accuracy without loss of interpretability. The nature of future work would also explore another method of evaluation of complex models, such as approximate cross-validation schemes or online validation schemes, to achieve statistical rigour and computational feasibility. Uncertainty estimation and robustness analysis should be incorporated in regression benchmarking, as it will offer more information on model reliability, especially in high-stakes contexts.

6 CONCLUSION

This paper provided a comparative analysis of conventional and contemporary regression forecasting methods with respect to high-dimensional data in a systematic and rigorous manner. Through their comparison of the three models of linear regression, regularised regression, and ensemble-based and deep learning regression, the work presented a vivid analysis of the behaviour of various

modelling strategies given a large number of predictors and minimal underlying signal. The empirical findings indicated that regularised linear models, especially lasso and elastic net regression, have been very stable and capable of predictive performance. More complex nonlinear models, such as random forests, gradient boosting techniques, and neural networks, were, in contrast, found to make no significant advance to accuracy even with their higher expressive power and expensive computational cost. These results bring out the main contribution of regularisation to reduce overfitting and variance in high-dimensional regression issues. The significance of this study is that it focuses on practical and methodological trade-offs and not necessarily on predictive accuracy alone. The balanced view provided by the study through a collective analysis of error measures, cost of computing, stability, and interpretability is directly applicable to the real-world usage. In areas where high-dimensional data are frequent, the findings imply that simpler, well-regularised models can offer effective and stronger answer than complex models do without the overhead of more complicated strategies. To researchers, the results confirm that a model's complexity should be matched with data properties and that it should not be assumed that highly sophisticated models will necessarily outperform classical algorithms. In conclusion, the work illustrates the importance of extensive benchmarking in getting to know the strengths and shortcomings of regression methods in high-dimensional contexts. These findings highlight the idea that empirical data and computational constraints as well as interpretability requirements are more important in determining a model choice, not just the sophistication of the model. Through these trade-offs explained in the study, a more reasoned and efficient use of regression in contemporary machine learning practice can be made.

7 REFERENCES

[1] M. I. Al-Karkhi and G. Rządkowski, "Innovative machine learning approaches for complexity in economic forecasting and SME growth: A comprehensive review," *Journal of Economy and Technology*, vol. 3, pp. 109–122, 2025. https://doi.org/10.1016/j.ject.2025.01.001

[2] X. Shu and Y. Ye, "Knowledge discovery: Methods from data mining and machine learning," *Social Science Research*, vol. 110, p. 102817, 2023. https://doi.org/10.1016/j.ssresearch.2022.102817

[3] J. P. Gygi, S. H. Kleinstein, and L. Guan, "Predictive overfitting in immunological applications: Pitfalls and solutions," *Human Vaccines & Immunotherapeutics*, vol. 19, no. 2, p. 2251830, 2023. https://doi.org/10.1080/21645515.2023.2251830

[4] M. A. U. H. Khan, R. Parveen, I. Ahmed, M. H. Milon, and T. A. Khan, "High-accuracy breast cancer diagnosis using neural networks and dimensionality reduction techniques," in *2025 IEEE 19th International Conference on Open Source Systems and Technologies (ICOSST)*, Lahore, Pakistan, 2025, pp. 1–6. https://doi.org/10.1109/ICOSST69113.2025.11315291

[5] W. Sun, J. Xu, and T. Liu, "Partially functional linear regression based on Gaussian process prior and ensemble learning," *Mathematics*, vol. 13, no. 5, p. 853, 2025. https://doi.org/10.3390/math13050853

[6] T. P. Pagano *et al.*, "Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods," *Big Data and Cognitive Computing*, vol. 7, no. 1, p. 15, 2023. https://doi.org/10.3390/bdcc7010015

[7] H. Niu, G. B. McCallum, A. B. Chang, K. Khan, and S. Azam, "Exploring unsupervised feature extraction algorithms: Tackling high dimensionality in small datasets," *Scientific Reports*, vol. 15, no. 1, p. 21973, 2025. https://doi.org/10.1038/s41598-025-07725-9

[8] S. Dhanka, A. Sharma, A. Kumar, S. Maini, and H. Vundavilli, "Advancements in hybrid machine learning models for biomedical disease classification using integration of hyperparameter-tuning and feature selection methodologies: A comprehensive review," *Arch Computat Methods Eng*, vol. 33, no. 1, pp. 289–324, 2025. https://doi.org/10.1007/s11831-025-10309-5

[9] M. Lee, "The geometry of feature space in deep learning models: A holistic perspective and comprehensive review," *Mathematics*, vol. 11, no. 10, p. 2375, 2023. https://doi.org/10.3390/math11102375

[10] T. Zhang, Z. Li, Q. Yuan, and Y. Wang, "A spatial distance-based spatial clustering algorithm for sparse image data," *Alexandria Engineering Journal*, vol. 61, no. 12, pp. 12609–12622, 2022. https://doi.org/10.1016/j.aej.2022.06.045

[11] S. Yun *et al.*, "HyperSense: Hyperdimensional intelligent sensing for energy-efficient sparse data processing," *Advanced Intelligent Systems*, vol. 6, no. 12, p. 2400228, 2024. https://doi.org/10.1002/aisy.202400228

[12] R. Di Laora, R. Cesaro, C. Iodice, M. Iovino, and L. de Sanctis, "A closed form solution for the generalised failure envelope of a pile group," *Soil Dynamics and Earthquake Engineering*, vol. 199, p. 109623, 2025. https://doi.org/10.1016/j.soildyn.2025.109623

[13] A. Magklaras, C. Gogos, P. Alefragis, and A. Birbas, "Enhancing parameters tuning of overlay models with ridge regression: Addressing multicollinearity in high-dimensional data," *Mathematics*, vol. 12, no. 20, p. 3179, 2024. https://doi.org/10.3390/math12203179

[14] J. Gillariose, J. Joseph, and C. Chesneau, "Lasso and ridge regression: A comprehensive review of applications and developments in machine learning," *International Journal of Data Science and Analytics*, vol. 21, no. 1, p. 7, 2026. https://doi.org/10.1007/s41060-025-00957-y

[15] M. A. Khan, K. Mahboob, U. Yousuf, M. Ramzan, M. T. Shaikh, and S. Akber, "Investigating the role of LASSO in feature selection for Educational Data Mining (EDM) applications," *VFAST Transactions on Software Engineering*, vol. 13, no. 2, pp. 56–67, 2025. https://doi.org/10.21015/vtse.v13i2.2111

[16] O. A. Montesinos López, A. Montesinos López, and J. Crossa, "Support vector machines and support vector regression," in *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, Cham: Springer International Publishing, 2022, pp. 337–378. https://doi.org/10.1007/978-3-030-89010-0_9

[17] A. Mehrabinezhad, M. Teshnehlab, and A. Sharifi, "A comparative study to examine principal component analysis and kernel principal component analysis-based weighting layer for convolutional neural networks," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 12, no. 1, p. 2379526, 2024. https://doi.org/10.1080/21681163.2024.2379526

[18] K.-L. Du, B. Jiang, J. Lu, J. Hua, and M. N. S. Swamy, "Exploring kernel machines and support vector machines: Principles, techniques, and future directions," *Mathematics*, vol. 12, no. 24, p. 3935, 2024. https://doi.org/10.3390/math12243935

[19] M. Mohammadagha, "Hyperparameter optimization strategies for tree-based machine learning models prediction: A comparative study of AdaBoost, decision trees, and random forest," *Decision Trees, and Random Forest*, 2025. https://doi.org/10.31219/osf.io/xbkr5_v1

[20] B. Takoutsing and G. B. Heuvelink, "Comparing the prediction performance, uncertainty quantification and extrapolation potential of regression kriging and random forest while accounting for soil measurement errors," *Geoderma*, vol. 428, p. 116192, 2022. https://doi.org/10.1016/j.geoderma.2022.116192

[21] R. Sibindi, R. W. Mwangi, and A. G. Waititu, "A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices," *Engineering Reports*, vol. 5, no. 4, p. e12599, 2023. https://doi.org/10.1002/eng2.12599

[22] P. Mavaie, L. Holder, and M. K. Skinner, "Hybrid deep learning approach to improve classification of low-volume high-dimensional data," *BMC Bioinformatics*, vol. 24, no. 1, p. 419, 2023. https://doi.org/10.1186/s12859-023-05557-w

[23] S. F. Ahmed *et al.*, "Deep learning modelling techniques: Current progress, applications, advantages, and challenges," *Artificial Intelligence Review*, vol. 56, no. 11, pp. 13521–13617, 2023. https://doi.org/10.1007/s10462-023-10466-8

[24] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, "Linear model selection and regularization," Springer, 2021. https://doi.org/10.32614/cran.package.islr

[25] M. Shahnawaj *et al.*, "Explainable artificial intelligence for credit risk assessment: Balancing transparency and predictive performance," *Journal of Economics, Finance and Accounting Studies*, vol. 7, no. 6, pp. 14–27, 2025. https://doi.org/10.32996/jefas.2025.7.6.2

[26] J.-B. Lugagne, C. M. Blassick, and M. J. Dunlop, "Deep model predictive control of gene expression in thousands of single cells," *Nature Communications*, vol. 15, no. 1, p. 2148, 2024. https://doi.org/10.1038/s41467-024-46361-1

[27] O. Alshboul, A. Shehadeh, G. Almasabha, and A. S. Almuflih, "Extreme gradient boosting-based machine learning approach for green building cost prediction," *Sustainability*, vol. 14, no. 11, p. 6651, 2022. https://doi.org/10.3390/su14116651

[28] K.-L. Du, C.-S. Leung, W. H. Mow, and M. N. S. Swamy, "Perceptron: Learning, generalization, model selection, fault tolerance, and role in the deep learning era," *Mathematics*, vol. 10, no. 24, p. 4730, 2022. https://doi.org/10.3390/math10244730

[29] D. M. Reoukadji, O. M. Mbayam, I. Alihamidi, P. L. Bokonda, and A. A. Madi, "Towards smart healthcare teleconsultation: A secure IoT-edge-machine learning architecture for diabetes data collection and prediction," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 21, no. 14, pp. 76–96, 2025. https://doi.org/10.3991/ijoe.v21i14.58035

8 AUTHORS

Dr. Benciya Abdul Jaleel is a Lecturer in the Department of Supportive Requirements at the University of Technology and Applied Sciences, Salalah. She received her Ph.D. in Applied Mathematics from CMJ University, India. Her research interests include statistics, data analysis, and mathematical modelling. Throughout her teaching career, she has delivered various mathematics courses at different academic levels, including bachelor's, advanced diploma, diploma, and foundation levels. She possesses adept skills in utilising various software tools for research purposes and mathematics education. She remains committed to advancing as an exceptional researcher and educator (E-mail: benciya.jaleel@utas.edu.om).

Dr. Aman Shaik is a Lecturer of Mathematics in the Department of Supportive Requirements at the University of Technology and Applied Sciences, Sultanate of Oman. She completed her Ph.D. in Mathematics at Acharya Nagarjuna University, Andhra Pradesh, India. She has 22 years of teaching experience in Mathematics, including teaching postgraduate and Engineering Mathematics courses. Her research interests include Algebra, Ring Theory, Topological Regular Rings, Mathematical Modelling, Mathematical Biology, Computational Modelling, and Simulations. She has developed strong skills in utilising various software applications and AI tools for research and education. She is committed to advancing her research activities and continuously enhancing innovative teaching and learning techniques (E-mail: aman.shaik@utas.edu.om).

Aminah Sirajuddin is a Lecturer at the Department of Supportive Requirements, University of Technology and Applied Sciences, with over 20 years of experience teaching mathematics in higher education at foundation, diploma, and

undergraduate levels. Currently pursuing her PhD, she is actively engaged in research with interests in Computational Fluid Dynamics (CFD), Mathematical Modelling, and applied mathematics. She has practical expertise in mathematical softwares. Passionate about continuous learning and technology-enhanced teaching, she actively participates in academic research and professional development. Her future research focuses on computational modelling, numerical simulations, and interdisciplinary scientific applications (E-mail: aminah.sirajuddin@utas.edu.om).