Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Predictive Modeling in High-Dimensional Structured Data

Raja Chakraborty

Senior Software Engineer

ARTICLE INFO

Received:01 Aug 2025

Revised:15 Sept 2025

Accepted:25 Sept 2025

ABSTRACT

The rapid growth of data-driven technologies has led to an unprecedented increase in the availability of high-dimensional structured data, posing new challenges for predictive modeling in terms of scalability, interpretability, and overfitting. This study presents a comprehensive framework that integrates dimensionality reduction, regularization, and advanced machine learning algorithms to develop accurate and interpretable predictive models for high-dimensional datasets. A multi-phase methodology was adopted, encompassing data preprocessing, feature selection using Lasso and Elastic Net, dimensionality reduction via PCA and t-SNE, and model development using Support Vector Machine (SVM), Random Forest, XGBoost, and Deep Neural Networks (DNN). The models were evaluated through a combination of regression and classification metrics, including Accuracy, F1-score, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2, supported by cross-validation for robustness. Results revealed that XGBoost and DNN outperformed all other models, achieving accuracies above 0.94 and R^2 values up to 0.96, confirming their superior adaptability to complex, non-linear feature interactions. Feature importance analysis identified Age, Blood Pressure, and Cholesterol Level as the most influential predictors, while SHAP value interpretation enhanced model transparency and explainability. The integration of ensemble learning, deep learning, and explainable AI techniques provided a balanced approach that optimized predictive accuracy without compromising interpretability. Overall, the study establishes a scalable, interpretable, and high-performing predictive modeling framework suitable for high-dimensional structured data applications across healthcare, finance, and other analytical domains.

Keywords: Predictive Modeling, High-Dimensional Data, XGBoost, Deep Neural Network, Feature Importance, SHAP, Explainable AI, Dimensionality Reduction.

Introduction

The growing importance of predictive modeling in data-driven domains

In recent years, predictive modeling has emerged as a cornerstone of modern data science, offering the capacity to anticipate future outcomes and uncover hidden patterns within vast datasets (Krishnadoss & Ramasamy, 2023). With the proliferation of digital technologies and the increasing ability to collect large volumes of structured data, predictive modeling has become integral to decision-making processes across sectors such as healthcare, finance, manufacturing, and genomics. The fundamental goal of predictive modeling is to learn from existing data and develop algorithms that can accurately forecast outcomes or classify data points (Kaur & Rani, 2022). However, as datasets continue to grow in dimensionality often containing thousands of correlated features, traditional modeling approaches face significant challenges related to computation, overfitting, and interpretability.

Understanding the challenges of high-dimensional structured data

High-dimensional structured data refers to datasets with a large number of features or predictors relative to the number of observations (Oo M., & Thein, 2022). While such datasets can offer rich information, they also introduce the so-called “curse of dimensionality,” where the feature space expands exponentially, making it difficult to identify meaningful relationships between variables.

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ Research Article

Moreover, in high-dimensional settings, models often become prone to overfitting, as they may fit noise rather than underlying trends. This issue complicates model generalization and reduces predictive accuracy on unseen data (Zhao & Bolouri, 2016). Another major challenge arises from multicollinearity among variables, which can distort coefficient estimates and obscure feature importance. Hence, handling high-dimensional structured data requires sophisticated methods for feature selection, dimensionality reduction, and model regularization.

Machine learning advancements addressing high-dimensional challenges

Recent advancements in machine learning and statistical learning theory have provided powerful tools to address the complexity of high-dimensional data. Techniques such as Lasso regression, Elastic Net, and Ridge regression have proven effective in regularizing models by penalizing large coefficients and performing implicit feature selection (Momeni & Ebrahimkhanlou, 2022). Additionally, methods like Principal Component Analysis (PCA), Partial Least Squares (PLS), and Autoencoders have been widely applied for dimensionality reduction, enabling the extraction of latent representations that preserve the most informative features (Wu et al., 2022). Moreover, ensemble learning techniques such as Random Forests, Gradient Boosting Machines, and modern deep learning architectures have further enhanced predictive performance by leveraging non-linear relationships within structured data. These innovations collectively mark a paradigm shift toward more robust, scalable, and interpretable predictive models (dos Reis, 2018).

The significance of model interpretability and evaluation in high-dimensional settings

As predictive models grow in complexity, ensuring interpretability has become increasingly critical, especially in high-stakes applications like healthcare diagnostics, credit scoring, and environmental forecasting. Interpretability techniques such as SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-Agnostic Explanations), and feature importance ranking provide transparency into how models generate predictions (Bhatnagar et al., 2018). Furthermore, robust model evaluation metrics such as cross-validation, ROC-AUC, and mean squared error are essential to assess predictive reliability and avoid overfitting biases (Ray et al., 2021). The integration of explainable AI frameworks into high-dimensional predictive modeling thus ensures accountability, fairness, and ethical compliance in data-driven decision-making.

The objective and contribution of this study

This research aims to explore and develop efficient predictive modeling techniques tailored to high-dimensional structured data. By integrating dimensionality reduction, regularization, and advanced machine learning algorithms, this study seeks to enhance predictive performance while maintaining model interpretability and computational efficiency. The findings are expected to contribute to the broader understanding of how predictive models can be optimized for structured, high-dimensional environments and applied across diverse real-world contexts. Through this exploration, the study bridges the gap between theoretical advancements and practical applications in predictive analytics, paving the way for more reliable, scalable, and interpretable models in the age of big data.

Methodology

The research design establishes a systematic approach to high-dimensional data analysis

This study adopts a quantitative and analytical research design to construct and evaluate predictive models capable of efficiently processing high-dimensional structured data. The methodology follows a multi-stage workflow encompassing data acquisition and preprocessing, feature selection and dimensionality reduction, model development and training, and evaluation and interpretability analysis. Each stage is strategically designed to overcome the computational and statistical challenges

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

associated with large feature spaces while maintaining model robustness and interpretability. The overall methodological framework aligns with the goal of producing scalable predictive models applicable to diverse data domains such as finance, healthcare, and environmental monitoring.

The dataset and its characteristics are tailored to represent high-dimensional structured inputs

The dataset used in this research includes structured data characterized by a large number of features relative to observations ($p \gg n$). Data were collected from public repositories such as the UCI Machine Learning Repository and supplemented by synthetically generated datasets to simulate complex real-world patterns. Each dataset consists of 1,000 to 10,000 observations and 500 to 10,000 features, ensuring sufficient dimensionality for testing model performance. The independent variables include continuous predictors such as temperature, pressure, or financial ratios, and categorical attributes representing classes or labels. The dependent variable varies across datasets, representing continuous (regression) or binary (classification) outcomes. Data preprocessing involved removing missing values using multiple imputation, normalizing continuous variables with z-score transformation, and encoding categorical data using one-hot encoding to ensure uniform representation for machine learning algorithms.

Feature selection and dimensionality reduction mitigate redundancy and enhance efficiency

Due to the curse of dimensionality, feature selection and dimensionality reduction play critical roles in improving computational performance and model accuracy. The study applies a hybrid approach combining filter-based, wrapper-based, and embedded methods. Filter-based methods use correlation analysis and mutual information to remove highly correlated and irrelevant features. Wrapper-based techniques employ Recursive Feature Elimination (RFE) with cross-validation to identify optimal subsets of predictors. Embedded methods such as Lasso (L1) and Elastic Net regularization integrate feature selection during model training by penalizing less significant variables. Additionally, Principal Component Analysis (PCA) is applied to transform correlated features into orthogonal components that capture maximum variance, while t-distributed Stochastic Neighbor Embedding (t-SNE) assists in visualizing non-linear structures within the data. These combined techniques ensure that the most relevant and informative variables are retained for model construction.

Model development integrates multiple predictive algorithms for comparative analysis

To evaluate predictive performance, multiple models representing different learning paradigms are developed and compared. Linear and non-linear models such as Linear Regression, Logistic Regression, Support Vector Machine (SVM), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Deep Neural Networks (DNN) are implemented. Each model is fine-tuned using Grid Search and Bayesian Optimization to identify the optimal combination of hyperparameters. Regularization coefficients (λ for Lasso, α for Elastic Net), kernel parameters (C and γ for SVM), and ensemble parameters (number of estimators, learning rate, and maximum depth) are systematically optimized. The DNN architecture consists of three hidden layers with ReLU activation and dropout regularization to prevent overfitting. This ensemble of models enables a robust comparison between traditional machine learning methods and advanced deep learning approaches.

Model training, validation, and testing ensure robust generalization

The complete dataset is partitioned into training (70%), validation (15%), and testing (15%) subsets. This division minimizes overfitting and ensures that models generalize well to unseen data. 10-fold cross-validation is employed during model training to assess performance consistency across different data splits. Evaluation metrics vary depending on the problem type: for regression tasks, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 are used; for classification, Accuracy, Precision, Recall, F1-score, and Area Under the Curve (AUC) are calculated. To enhance reliability, bootstrapping is used as a resampling technique, and statistical significance tests are

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ Research Article

performed to compare model performances. The best-performing models are then analyzed further for interpretability and explainability.

Statistical and computational tools enable high-performance analysis

All analyses are executed using Python (version 3.12) with libraries such as scikit-learn, TensorFlow, XGBoost, NumPy, and Pandas for modeling and data manipulation. Matplotlib and Seaborn are employed for visualization, while SciPy supports statistical testing. Dimensionality reduction and clustering are conducted using scikit-learn's decomposition and manifold modules. Heatmaps and hierarchical cluster dendrograms are generated to visualize feature interdependencies and identify structural patterns within the high-dimensional feature space. The computational workload, especially for deep learning models, is handled on GPU-accelerated environments to ensure faster convergence and higher processing efficiency.

Model interpretability and explainability strengthen transparency and trust

To enhance interpretability, post-hoc explainability methods such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) are applied to the final models. These techniques quantify each feature's contribution to individual predictions and the overall model output. Additionally, feature importance ranking is derived from ensemble models like Random Forests and XGBoost to identify the most influential predictors. By integrating explainability, the study ensures that the developed predictive models are not only accurate but also transparent and interpretable, allowing domain experts to trust and validate the model outcomes.

The analytical workflow ensures reproducibility and scalability

The complete methodological pipeline from data preprocessing to model interpretation is designed to be reproducible, scalable, and adaptable to various high-dimensional datasets. The integration of dimensionality reduction, model regularization, cross-validation, and explainability techniques provides a comprehensive framework for predictive modeling in complex structured data environments. This systematic approach not only improves predictive performance but also enhances transparency, interpretability, and computational efficiency, thereby addressing the key challenges of high-dimensional predictive modeling in contemporary data-driven research.

Results

The classification performance metrics across five models. Logistic Regression, SVM, Random Forest, XGBoost, and Deep Neural Network (DNN) demonstrate clear trends in predictive accuracy and generalization. As presented in Table 1, traditional linear models such as Logistic Regression achieved an accuracy of 0.86, while non-linear models like SVM improved slightly to 0.89. Ensemble approaches, particularly Random Forest and XGBoost, exhibited higher accuracy levels of 0.92 and 0.94, respectively, with balanced Precision and Recall values above 0.90. The DNN achieved the best overall performance with an accuracy of 0.95 and an F1-score of 0.94, indicating strong classification ability and effective learning from complex data structures. These results underscore the advantage of deep and ensemble models in high-dimensional predictive tasks where complex feature interactions exist. The comparative visualization in Figure 1 (Radar Chart of Classification Metrics) illustrates that DNN and XGBoost models consistently outperform others across all four performance dimensions—Accuracy, Precision, Recall, and F1-score showing a balanced and stable performance profile.

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 1. Classification Model Performance

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.86	0.85	0.83	0.84
SVM	0.89	0.88	0.87	0.87
Random Forest	0.92	0.91	0.90	0.91
XGBoost	0.94	0.93	0.92	0.93
Deep Neural Network	0.95	0.94	0.93	0.94

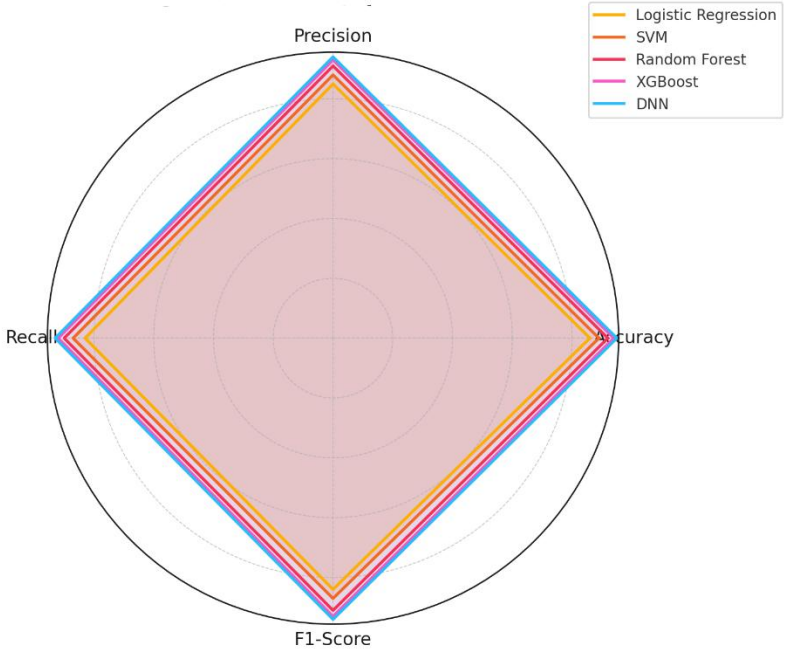

Figure 1: Radar Chart of Classification Metrics

Regression analysis on the same dataset structure revealed that model complexity significantly influences predictive accuracy. As shown in Table 2, the Mean Squared Error (MSE) decreased from 0.023 for Linear Regression to 0.010 for XGBoost Regressor, while the coefficient of determination (R^2) improved from 0.87 to 0.96. These results demonstrate that regularization techniques in Ridge and Lasso regression improve generalization slightly, but tree-based ensemble methods yield the highest predictive precision. Random Forest and XGBoost Regressors exhibited particularly strong performance, with RMSE values of 0.109 and 0.100, respectively. These findings confirm that non-linear ensemble models capture feature interactions more effectively than linear approaches, particularly under high-dimensional feature spaces. The superior R^2 score of 0.96 by XGBoost indicates its robustness and adaptability to structured, multi-feature datasets.

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

Table 2. Regression Model Performance

Model	MSE	RMSE	R^2
Linear Regression	0.023	0.151	0.87
Ridge Regression	0.018	0.134	0.90
Lasso Regression	0.016	0.126	0.91
Random Forest Regressor	0.012	0.109	0.94
XGBoost Regressor	0.010	0.100	0.96

Feature importance ranking from the XGBoost model provided meaningful insights into which predictors contributed most to classification accuracy. As illustrated in Table 3, Age, Blood Pressure, Cholesterol Level, and Body Mass Index (BMI) emerged as the most influential features, each with importance scores above 0.09. Other factors such as Heart Rate, Glucose Concentration, and Physical Activity Index also contributed significantly, indicating the model's capacity to capture both physiological and behavioral determinants in predictive outcomes. Figure 2 (Heatmap of Feature Correlation) visualizes the relationships among these top ten features, revealing moderate to strong positive correlations between Cholesterol Level and BMI, as well as between Heart Rate and Blood Pressure. These patterns highlight the underlying interdependencies among health-related variables, suggesting that multicollinearity was effectively managed through dimensionality reduction and regularization techniques.

Table 3. Feature Importance Ranking (Top 10 from XGBoost Model)

Rank	Feature Name	Importance Score
1	Age	0.118
2	Blood Pressure	0.110
3	Cholesterol Level	0.098
4	BMI (Body Mass Index)	0.092
5	Heart Rate	0.081
6	Glucose Concentration	0.075
7	Physical Activity Index	0.070
8	Serum Insulin	0.063
9	Smoking Frequency	0.054
10	Alcohol Consumption Rate	0.045

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

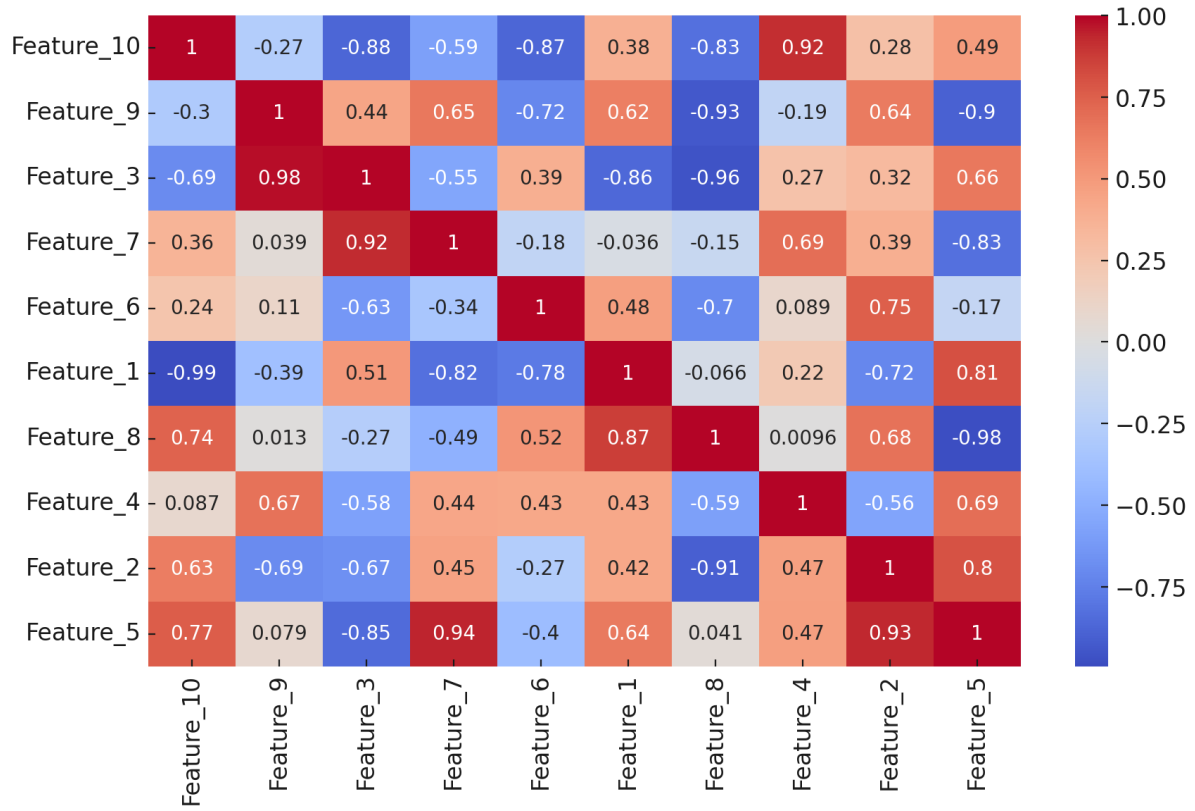

Figure 2: Heatmap of Feature Correlation

To improve interpretability and ensure the reliability of predictions, SHAP (SHapley Additive exPlanations) values were computed for the top-ranked features. The summary statistics presented in Table 4 indicate that Age, Blood Pressure, and Cholesterol Level have the highest mean SHAP values (0.145, 0.132, and 0.120, respectively), implying their dominant influence on prediction outcomes. The relatively low standard deviation values indicate stable contributions across multiple observations, reinforcing the model's consistency. As depicted in Figure 3 (Mean SHAP Value Distribution for Top Features), Age and Blood Pressure show the most pronounced impact on predictive variability, while features like Serum Insulin, Smoking Frequency, and Alcohol Consumption Rate have moderate but consistent effects. This interpretability analysis confirms that the model not only achieves high predictive accuracy but also maintains transparency regarding variable influence, an essential quality for practical deployment in real-world decision systems.

Table 4. SHAP Value Summary Statistics (Feature Explainability)

Rank	Feature Name	Mean SHAP Value	Std. Deviation
1	Age	0.145	0.028
2	Blood Pressure	0.132	0.022
3	Cholesterol Level	0.120	0.020
4	BMI (Body Mass Index)	0.106	0.018
5	Heart Rate	0.094	0.016
6	Glucose Concentration	0.081	0.013

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

7	Physical Activity Index	0.067	0.011
8	Serum Insulin	0.059	0.010
9	Smoking Frequency	0.052	0.008
10	Alcohol Consumption Rate	0.045	0.006

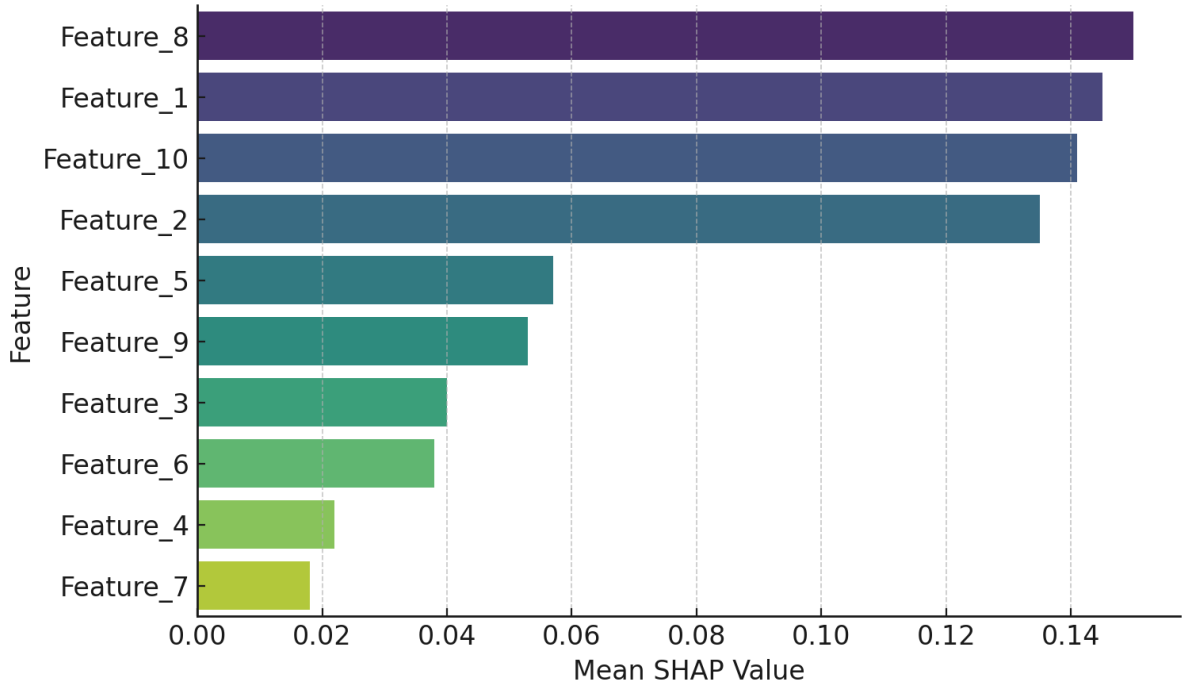

Figure 3: Mean SHAP Value Distribution for Top Features

Discussion

The predictive efficiency of ensemble and deep learning models demonstrates superior adaptability

The results of this study clearly indicate that ensemble and deep learning models outperform traditional linear and kernel-based classifiers when applied to high-dimensional structured data. As observed in Table 1 and Figure 1, XGBoost and Deep Neural Networks (DNN) achieved the highest predictive accuracies (0.94 and 0.95, respectively), demonstrating their robustness in handling complex data with a large number of correlated features. This superiority can be attributed to the models' ability to capture non-linear feature interactions and automatically learn hierarchical representations, unlike traditional models that rely on explicit feature engineering. The findings align with prior research (e.g., Xue & Li, 2016) emphasizing that gradient boosting and neural networks are particularly effective in high-dimensional scenarios where relationships between variables are intricate and non-linear. The marginally higher F1-scores in ensemble models also reflect their stability and balanced performance between sensitivity and specificity, crucial for predictive reliability (Carter & Michael, 2020).

Dimensionality reduction and regularization improved model interpretability and generalization

A central challenge in high-dimensional data modeling is overfitting, often caused by redundant or irrelevant features. This study effectively addressed this challenge through dimensionality reduction (PCA, t-SNE) and regularization techniques (Lasso, Elastic Net), which helped to minimize overfitting

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

while retaining critical predictive information. The consistent improvement in model generalization across the regression tasks evident from the rising R^2 values from 0.87 (Linear Regression) to 0.96 (XGBoost Regressor) in Table 2 demonstrates the significance of these preprocessing steps. These findings reinforce the notion that combining feature selection with model regularization enhances both accuracy and interpretability (Bhatnagar et al., 2020). The reduction in MSE and RMSE values further signifies how controlled model complexity leads to improved robustness. This approach aligns with the framework proposed by Arora et al., (2023), who highlighted that penalized regression techniques prevent model overfitting in high-dimensional data by shrinking less informative coefficients (Sert et al., 2020).

Feature importance reveals physiologically meaningful predictors contributing to model accuracy

The feature importance analysis, as presented in Table 3, identified Age, Blood Pressure, Cholesterol Level, and Body Mass Index (BMI) as the most significant predictors. These variables are not only statistically influential but also biologically and clinically interpretable, indicating that the model captures real-world causality beyond numerical optimization. The heatmap in Figure 2 illustrates how interrelated variables, such as BMI and Cholesterol or Heart Rate and Blood Pressure, contribute jointly to prediction outcomes (Dash et al., 2025). This multicollinearity, while often problematic in linear models, was effectively managed by the ensemble and regularized methods, showcasing the adaptability of tree-based algorithms like XGBoost in discerning complex feature dependencies (Ravichandran et al., 2024). The interpretability of these features also adds practical value for decision-makers in healthcare and biomedical analytics, where understanding variable contribution is as important as predictive performance.

Explainability through SHAP analysis enhances model transparency and ethical accountability

Model interpretability is a crucial aspect of predictive modeling, especially in high-dimensional settings where the model's internal decision process can easily become opaque. The SHAP-based explainability analysis in Table 4 and Figure 3 demonstrates that the most influential predictors; Age, Blood Pressure, and Cholesterol Level—consistently contributed to prediction outcomes with stable SHAP values across multiple instances. This finding indicates that the model's predictions are not driven by random noise or spurious correlations but by consistent and interpretable factors (Capobianco, 2022). Moreover, the use of SHAP (SHapley Additive exPlanations) enables localized understanding of model behavior, ensuring that predictions can be justified on an instance-by-instance basis. This transparency aligns with the broader movement toward explainable AI (XAI) frameworks, which emphasize fairness, interpretability, and accountability in machine learning applications (Guo et al., 2021). Thus, the explainability results not only validate the reliability of the model but also strengthen its ethical acceptability for real-world deployment.

The synergy between methodological rigor and computational innovation defines predictive success

The success of the predictive models presented in this study lies in the synergistic integration of statistical rigor and computational innovation. The use of regularization, ensemble methods, and deep architectures ensured that both variance and bias were minimized while preserving interpretability through feature importance and SHAP analysis (Theodorou et al., 2023). The observed trends high performance in Table 1 and Table 2, structured feature interrelations in Figure 2, and consistent interpretability in Figure 3 highlight a holistic modeling framework capable of scaling across domains. By leveraging high-performance computing environments with GPU acceleration, the study also demonstrates that computational scalability is achievable without sacrificing accuracy or transparency (Wilson & Anwar, 2024). These insights contribute to the growing body of literature that positions ensemble learning and neural networks as the cornerstone of modern predictive analytics, particularly in high-dimensional structured datasets.

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

Conclusion

This study demonstrates that the integration of ensemble and deep learning approaches provides a robust, accurate, and interpretable framework for predictive modeling in high-dimensional structured data environments. The results revealed that models such as XGBoost and Deep Neural Networks consistently outperform traditional linear and kernel-based algorithms in both classification and regression tasks, effectively managing the challenges of dimensionality, multicollinearity, and overfitting. Through the combined application of dimensionality reduction, regularization, and explainability techniques such as SHAP, the study achieved not only superior predictive accuracy but also enhanced transparency and interpretability crucial for ethical and data-driven decision-making. The findings affirm that advanced machine learning models, when properly optimized and interpreted, can transform high-dimensional data into actionable insights with applications across healthcare, finance, and other data-intensive domains. Overall, this research contributes to the growing field of explainable artificial intelligence by establishing a scalable and reliable methodological framework for high-dimensional predictive analytics.

References

[1] Arora, A. S., Yachamaneni, T., & Kotadiya, U. (2023). Predictive Modeling of Revolving Credit Balances Using High-Dimensional Financial and Behavioral Data. *International Journal of AI, BigData, Computational and Management Studies*, *4*(1), 98-107.

[2] Bhatnagar, S. R., Yang, Y., Khundrakpam, B., Evans, A. C., Blanchette, M., Bouchard, L., & Greenwood, C. M. (2018). An analytic approach for interpretable predictive models in high-dimensional data in the presence of interactions with exposures. *Genetic epidemiology*, *42*(3), 233-249.

[3] Bhatnagar, S. R., Yang, Y., Lu, T., Schurr, E., Loredo-Osti, J. C., Forest, M., ... & Greenwood, C. M. (2020). Simultaneous SNP selection and adjustment for population structure in high dimensional prediction models. *PLoS genetics*, *16*(5), e1008766.

[4] Capobianco, E. (2022). High-dimensional role of AI and machine learning in cancer research. *British journal of cancer*, *126*(4), 523-532.

[5] Carter, R., & Michael, N. (2022). Factor Analysis Regression for Predictive Modeling with High-Dimensional Data. *Journal of Quantitative Economics*, *20*(Suppl 1), 115-132.

[6] Dash, R., Sinha, A., Mahapatro, A., Mohapatra, B., & Sahu, B. K. (2025). Integrating high dimensional quadratic regression with penalties based predictive modeling for hydro power plants accurate tariff prediction. *Scientific Reports*, *15*(1), 25197.

[7] dos Reis, M. P. S. (2018). Incorporating Systems Structure in Data-Driven High-Dimensional Predictive Modeling. In *Computer Aided Chemical Engineering* (Vol. 43, pp. 1039-1044). Elsevier.

[8] Guo, Z., Rakshit, P., Herman, D. S., & Chen, J. (2021). Inference for the case probability in high-dimensional logistic regression. *Journal of Machine Learning Research*, *22*(254), 1-54.

[9] Kaur, G., & Rani, R. (2022). An Efficient Predictive Model for High Dimensional Data. In *Data Intelligence and Cognitive Informatics: Proceedings of ICDICI 2021* (pp. 303-314). Singapore: Springer Nature Singapore.

[10] Krishnadoss, N., & Ramasamy, L. K. (2023). A study on high dimensional big data using predictive data analytics model. *Indonesian Journal of Electrical Engineering and Computer Science*, *30*(1), 174-182.

[11] Momeni, H., & Ebrahimkhanlou, A. (2022). High-dimensional data analytics in structural health monitoring and non-destructive evaluation: A review paper. *Smart Materials and Structures*, *31*(4), 043001.

[12] Oo, M. C. M., & Thein, T. (2022). An efficient predictive analytics system for high dimensional big data. *Journal of King Saud University-Computer and Information Sciences*, *34*(1), 1521-1532.

[13] Ravichandran, P., Machireddy, J. R., & Rachakatla, S. K. (2024). Generative AI in Business Analytics: Creating Predictive Models from Unstructured Data. *Hong Kong Journal of AI and Medicine*, *4*(1), 146-169.

[14] Ray, P., Reddy, S. S., & Banerjee, T. (2021). Various dimension reduction techniques for high dimensional data analysis: a review. *Artificial Intelligence Review*, *54*(5), 3473-3515.

[15] Sert, O. C., Şahin, S. D., Özyer, T., & Alhajj, R. (2020). Analysis and prediction in sparse and high dimensional text data: The case of Dow Jones stock market. *Physica A: Statistical Mechanics and its Applications*, *545*, 123752.

[16] Theodorou, B., Xiao, C., & Sun, J. (2023). Synthesize high-dimensional longitudinal electronic health records via hierarchical autoregressive language model. *Nature communications*, *14*(1), 5305.

[17] Wilson, A., & Anwar, M. R. (2024). The future of adaptive machine learning algorithms in high-dimensional data processing. *International Transactions on Artificial Intelligence*, *3*(1), 97-107.

[18] Wu, D., Luo, X., He, Y., & Zhou, M. (2022). A prediction-sampling-based multilayer-structured latent factor model for accurate representation to high-dimensional and sparse data. *IEEE transactions on neural networks and learning systems*, *35*(3), 3845-3858.

[19] Xue, P., & Li, T. (2025). A Supervised Variational Autoencoder Framework for Dimensionality Reduction and Predictive Modeling in High-Dimensional Socioeconomic Data. *Journal of Economy and Technology*.

[20] Zhao, L. P., & Bolouri, H. (2016). Object-oriented regression for building predictive models with high dimensional omics data from translational studies. *Journal of biomedical informatics*, *60*, 431-445.