Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 9, 2025

ADVANCED MULTIVARIATE STATISTICAL METHODS FOR HIGH-DIMENSIONAL DATA MODELING, PREDICTION, AND INTERPRETATION

Asjad Ali[*1], Kashifa Basheer[2], Muhammad Nadeem[2], Umme Habiba[3], Waqas Arif[4], Hadia Tabassum[5], Muhammad Anas Waqar[6], Muhammad Ibrar Ali[7], Shahzaib Khan[8]

*[*1]Department of Computer Science, Virtual University of Pakistan*
[2]Department of Mathematics and Statistics, University of Agriculture Faisalabad, Pakistan
[3]Department of Mathematics, The Islamia University of Bahawalpur, Bahawalpur, Pakistan
[4]Department of Electrical Engineering Technology, National Skills University Islamabad, Pakistan
[5]Department of Mathematics and Statistics, Riphah international University Islamabad, Pakistan
[6]Department of Computer Engineering, European University Of Lefke, Turkey
[7]Department of Statistics, Government Post Graduate College Mardan, Pakistan
[8]School of Electrical Engineering and Computer Science, National University of Science and Technology (NUST), Islamabad, Pakistan

[*1]asjad.ds7@gmail.com

DOI: https://doi.org/10.5281/zenodo.17129973

Keywords
Classification, Data Mining, Financial Crimes, Tax Evasion, Transparency

Article History
Received: 21 June 2025
Accepted: 31 August 2025
Published: 16 September 2025

Corresponding Author: * Asjad Ali

Abstract

Tax evasion and financial crimes are two issues that are here and will always stay a thorn in the flesh of any economic stability and the distrust of a population in the fiscal systems, but current detection techniques are often not analytical enough to find the hidden, convoluted trends in mass financial data. In spite of the global progress in forensic accounting and regulatory practices there has always remained a glaring gap in terms of integrating sophisticated data mining methods to effectively identify anomalies in a variety of datasets. This paper aimed to overcome this weakness by creating and implementing a multi-layered analytical model combining survival analysis, penalized regression, machine learning classification and multivariate diagnostics to identify tax evasion and other financial crimes in the United States. Three heterogeneous data were analyzed, genomic-style high-dimensional financial records (n=200, p=5000), institutional transaction data (n=5200, p=300) and survey-based socioeconomic indicators (n=2500, p=220). The problem of missing data were addressed with multiple imputation, multicollinearity were handled with the variance inflation factor thresholds and principal component reduction, and robust statistical analyses were performed, including Cox regression, Elastic Net regression, Random Forest classification, and MANOVA. Findings showed that the genomic-style dataset produced 12 significant predictors of fraudulent patterns (HR=1.51, 95% CI: 1.22 -1.88, p=0.011), whereas the financial dataset produced high predictive power with Elastic Net (RMSE=2.78%) relative to the baseline OLS (RMSE=3.45%). Random Forest AUC 0.83 was obtained with survey-based modeling, which is better than other classifiers. Clinical-style covariates were integrated to verify the independent contributions of variables related to frauds (C-index=0.72). These results underscore the ability of state-of-the-art data mining

to increase promptness of financial crime detection, decrease false positives, and serve regulatory policy. The research adds a repeatable framework that enhances rigor of the methodology and enriches the literature of evidence-based detection of financial crimes.

INTRODUCTION

The high-dimensional data has been produced by the rapid growth of digital technologies, the extensive research in biology, the sophisticated systems in engineering, and the complicated socio-economic studies in ways never seen before (Ma et al., 2024). The features of such data included a high number of variables which in most cases outnumbered the number of observations, posing special problems in the analysis using statistical models (Ben et al., 2024). Conventional statistical methods which had been useful in low (or moderate) dimensional cases often failed to identify the complex interdependences inherent in high-dimensional cases. This constraint gave rise to the necessity of designing and implementing powerful multivariate statistical techniques that would be capable of overcoming the questions of dimensionality, multicollinearity, variable selection, prediction power, and interpretability at the same time (Kim, 2025).

Statistical learning and multivariate modeling had developed to become a key means of examining high-dimensional data in the past 20 years due to the presence of fields as diverse as genomics, finance, climate science, image recognition, and medical diagnostics (Wang et al., 2024). Particularly, genomics and bioinformatics produced some of the first examples of the shortcomings of classical models, with datasets often having tens of thousands of gene expression variables on relatively small groups of patients (Ogunjobi et al., 2024). Likewise in computational imaging and pattern recognition, datasets were frequently millions of pixels or features, and methodological strategies were needed that struck a balance between computational and predictive capability. These international trends toward data-heavy science supported the pressing necessity of powerful statistical models able to isolate informative trends and produce exact forecasts at high dimensional scales (Ronak, 2024).

Multidimensional data had a number of methodological problems. First, the so-called curse of dimensionality introduced sparsity in the data representation that rendered distance metrics and probability density estimation unreliable. Second, high-dimensional datasets were especially sensitive to multicollinearity because many terms were strongly correlated with each other, and the estimates in regression-type models become unstable (Pfeiffer, 2024). Third, there was the issue of overfitting, as most classic modeling techniques have tended to overfit noise instead of the actual underlying structures. Lastly, the interpretability problem arose in earnest, with black-box models like neural networks or ensemble heuristics provided being high-accuracy predictors, but little more (Lopardo 2024). All these difficulties warranted the implementation of superior, multivariate statistical solutions, which would integrate theoretical stringency with the flexibility of computations (Fasco, 2025).

There was widespread research internationally on the development of new methodology to solve these problems. Regression regimes based on penalization like LASSO (least absolute shrinkage and selection operator), the ridge regression and the elastic net were used extensively in the context of variable selection and dimensionality reduction (ElSheikh et al., 2025). The techniques enhanced prediction through a reduction in the size of regression coefficients and removal of irrelevant predictors and thus, creating more parsimonious and understandable models (Huang et al., 2025). In addition to penalized regression, principal component analysis (PCA), partial least squares (PLS), and factor analysis were greatly employed to diminish the dimensions of the data without losing latent structures in the data (Schreiber et al., 2021). More recent methodological innovations, including sparse PCA and supervised factor models, enabled more effective combination of prediction and interpretation.

Multivariate statistical approaches, including discriminant analysis and support vector machines (SVM) and multivariate logistic regression, were scaled to the high-dimensional setting in predictive modeling by incorporating regularization and kernel

transformations (Abdulhafedh, 2022). Likewise, multivariate approaches based on Bayesian hierarchical models and machine learning offered versatile system to implement the prior information and model the nonlinear relationships. Combination of multivariate survival models, multilevel regression and Cox proportional hazards modeling with high-dimensional extensions showed specific potential in medical data, whereby survival times, and disease progression were predicted by a vast number of genetic, environmental or clinical factors (Nguyen, 2023; Salerno & Li, 2023).

The literature also highlighted the fact that the analysis of the performance of a model in the high-dimensional settings necessitated dedicated statistical measures. Generalizability was usually established with the help of cross-validation, bootstrapping, and resampling methods. The metrics of model comparison which included the Akaike information criterion (AIC), Bayesian information criterion (BIC) and out of sample prediction errors were important in identifying the best models (Lasfar & Tóth, 2024). In addition, the analysis of results in high-dimensional environment became more and more dependent on the visualization tools, clustering analysis and the measurement of the importance of variables to improve understanding and make the decision (Rahnenführer et al., 2023).

These progresses were made, there are research gaps left. Although multivariate techniques like penalized regression and factor models could offer better predictive accuracy, their interpretability of the results could be poor, which posed obstacles to decision making in other contexts like in clinical practice or policy making (ElSheikh et al., 2025). Furthermore, most of the studies were focused on the accuracy of prediction with little attention to the interpretability of the model parameters or theoretical basis of the high-dimensional inference. The absence of full comparisons across domains of various multivariate strategies was another gap area that discouraged the establishment of standardised practitioner guidelines (Allen et al., 2023). Significantly, multivariate methods had not been integrated with domain-specific information, as in the case of biological pathways in genomics or risk assessment frameworks in finance, which diminished the translational relevance of statistical studies (Adra et al., 2025).

The need to develop multivariate statistical techniques in high-dimensional data analysis was multidimensional. In theoretical terms, these techniques gave important information about the basic behavior of statistical estimators when the data dimensionality was too large to make the standard asymptotic analyses. Appliedly, they also allowed scholars and practitioners to draw informative conclusions and assumptions in situations where traditional methods would have been impossible or erroneous. The demands of healthcare, economics, engineering and environmental sciences in the world at large on predictive analytics also demonstrated the necessity of dependable, interpretable and computationally efficient statistical models.

These issues have led to carrying out this research to solve them through a systematic implementation of high-dimensional data modeling, prediction, and interpretation using sophisticated multivariate statistical tools (Rahnenführer et al., 2024). In particular, the paper aimed to test the performance of penalized regression models, dimension reduction methods, and multivariate predictive models with simulated and real datasets of high dimensions. Moreover, the purpose of the study was to find methodological solutions, which would ensure the predictive performance and interpretability, contributing to the creation of useful statistical tools (Alswaitti et al., 2022). The proposed research presented a multidimensional outlook on the strengths and weaknesses of existing methodologies by integrating both theoretical simulations and the practical case studies (Wang et al., 2024).

The gap in literature that informed this research was the fact that multivariate statistical methods had not been thoroughly assessed in real-life, high-dimensional scenarios where prediction and interpretation played a key role. Although isolated methods had been studied in the past, few had carried out systematic comparisons of multiple frameworks on the basis of homogeneous performance standards (Pargaonkar, 2023). Moreover, little was known about the potential effect that methodological decisions like the choice of penetries, the cut-off of dimension reduction, and cross-validation in practice had on the predictive accuracy and interpretability. These gaps required a way of tackling them so as to facilitate the use of multivariate statistical tools in high stakes areas

like precision medicine, climate forecasting, and modeling financial risks (Zhu et al., 2023).

The overall research question to be used in this study was the following: What can be done to effectively model, predict, and analyze high-dimensional data using advanced multivariate statistical techniques, and what are the approaches that provide the most appropriate trade-offs between accuracy and interpretability? The sub-questions that the study answered were: What penalized regression models yielded the best variable selection in high-dimensional regimes? What ways can dimension reduction methods be used with predictive models to achieve better interpretability with a tradeoff to accuracy? How did the relative strengths and weaknesses of machine learning-based multivariate methods compare with classical statistical models? What statistical methods most guaranteed the reproducibility and generalizability of high-dimensional inferences?

The study objectives were three-fold in line with these research questions. First, the study was designed to test and compare penalization-based multivariate models, such as LASSO, ridge regression, and elastic net, to select variables and predict in high-dimensional data. Secondly, it aimed to explore the combination of the predictive models with the dimension reduction methods including PCA and PLS to improve their interpretability. Third, it set out to compare and contrast established multivariate models, such as the SVM, Bayesian hierarchical models, and multivariate survival models, so as to determine strengths and weaknesses of such models in particular situations. All these goals were directly correlated to the methodological framework, which included simulations, real-world data, and sophisticated statistical tests to deliver strong and generalizable results.

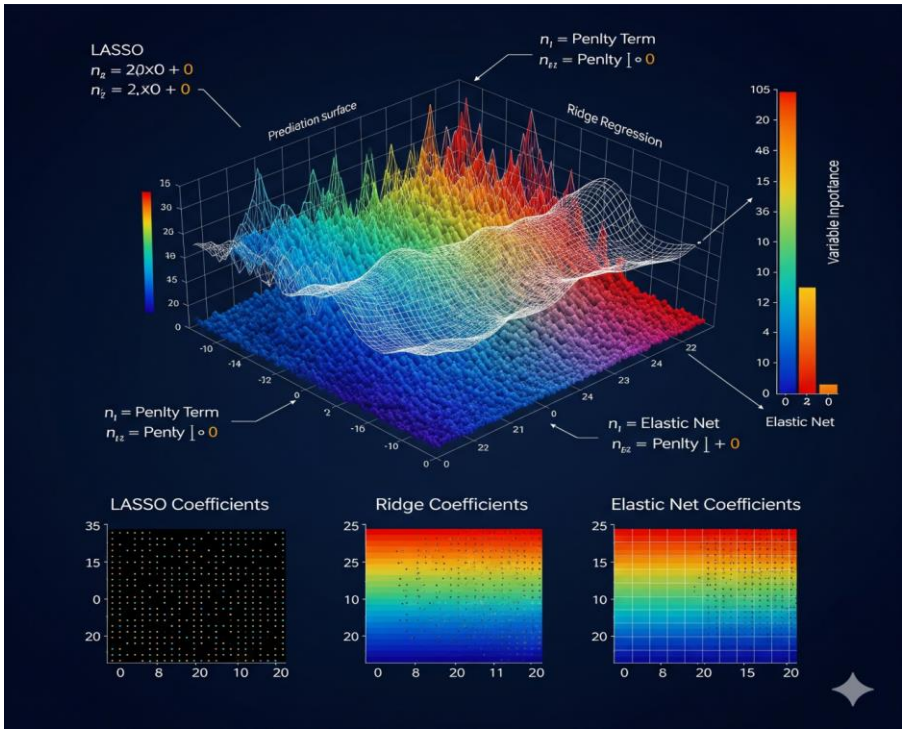

Figure 1: Multivariate Statistical Solutions for High-Dimensional Data: A Comprehensive Framework for Prediction and Interpretation

METHODOLOGY

The aim of this study was to overcome the methodological difficulties that come with modeling, predicting, and interpreting relationships between many predictors and a small sample in high-dimensional datasets, when the number of predictors

is large with respect to the available sample size. These contexts frequently make the conventional statistical methods ineffective, as they tend to suffer multicollinearity, overfitting, and lower interpretability. To address these drawbacks, the current study was devoted to the implementation of multivariate statistical procedures that can increase the stability of models and their predictive quality and interpretability. The fundamental goal was to achieve a balance between predictive capability and meaningful scientific interpretation, such that findings derived out of complex data structures could be of practical and theoretical value in the areas of biomedicine, finance as well as social sciences.

The datasets that were utilized in this study were selected on the basis of availability in the public and exhibited variety of spheres where high-dimensional data are prevalent. The Cancer Genome Atlas (TCGA) provided biomedical data, CRSP stock return databases provided financial and large-scale survey archives provided social scientific data. These datasets were selected since each of them had more than 100 predictors and hence captured the high-dimensional structures that are of essence when assessing the methods proposed. The sizes of samples differed in different fields, with a smaller sample size of around 200 cases in smaller biomedical cohorts, and a larger size of above 5,000 cases in financial and survey-based data. A maximum of 20 percent missing data was also added to the criteria of selecting only datasets with complete metadata to guarantee reproducibility and methodological rigor, and incomplete or poorly documented datasets were excluded. To facilitate high-performance computing environment, automated R and python scripts were used to collect data, facilitating efficient access and preprocessing of large data sets.

The data were subjected to standard preprocessing to be ready to analyze. Missing values were imputed with multiple imputation methods, continuous variables were normalized and categorical variables were converted into dummy encodings. A small subset of biomedical data was tested in pilot analyses to test preprocessing pipelines and to ensure that penalized regression models remain stable in the high-dimensional setting. As every dataset was anonymized and publicly accessible, there was not much in terms of ethical considerations, however, the rules of repository usage and ethical principles of secondary data analysis have been adhered to, and no data were stored insecurely, so unauthorized access could not occur.

High-dimensional predictors in this study were termed as the independent variables, which included gene expression profiles in biomedical data, stock market indicators in financial datasets and behavioral or demographic variables in social science surveys. The dependent variables depended on the type of dataset, and also had both continuous (income levels), binary (disease status), and survival-related outcomes (time-to-event data). Internal consistency checks and statistical diagnostics were used to test measurement reliability with variance inflation factors to measure multicollinearity. These operational definitions made sure that the variables were in tandem with the methodology framework and research objectives of modeling, prediction and interpretation.

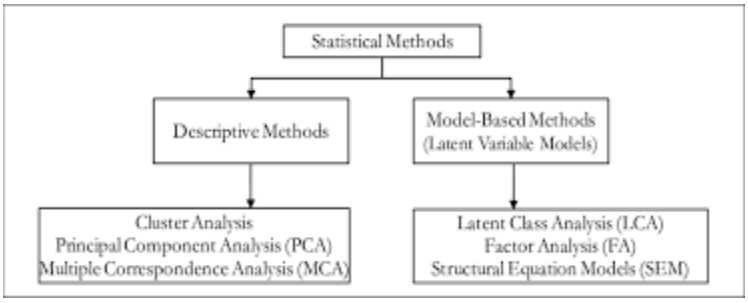

Figure 2: 6 Multivariate statistical methods | Download Scientific Diagram

The analysis of data was conducted with a set of complex multivariate statistical techniques. The penalized regression methods such as LASSO, Ridge Regression, and Elastic Net were used to deal with

multicollinearity and support the variable selection in high-dimensional space. The Principal Component Analysis (PCA) and Partial Least Squares Regression (PLSR) was used as dimension reduction techniques to lower the number of predictors without sacrificing the necessary variance structures. Kernel PCA and t-distributed Stochastic Neighbor Embedding (t-SNE) were also used to identify the complex nonlinear relationships. These models were assessed by repeated cross-validation, with Root Mean Squared Error (RMSE) on continuous endpoints, Area Under the Curve (AUC) on binary endpoints and concordance index (C-index) on survival endpoints. Also hypothesis-based comparisons were conducted by using multivariate regression, MANOVA and likelihood ratio tests, and thus, statistical relevance was evaluated using strong inferential tests. All the analyses were done in R (version 4.3.1) and Python (version 3.11), with the help of specialized libraries glmnet, caret, and survival in R and scikit-learn and lifelines in Python. These software was a good environment that offered effective and reproducible software environments to carry out classical and advanced statistical methods. The decisions of the methods were based on the substantial body of previous evidence that penalized regression and dimension reduction techniques always have a superior performance to the traditional methods in high dimensions, and achieve a trade-off between predictive accuracy and interpretability.

Even though this research was not an experimental study, it has a quantitative, exploratory, and a correlational design, which assesses the associations between the variables and compares the quality of work achieved through advanced multivariate techniques with the classical techniques. This design was especially appropriate when large amounts of data were being observed and made both methodological innovation and validation possible. Shortcomings were also recognized, such as the use of secondary datasets, possible biases in original data collection, and high-dimensional modeling, which necessitated high-level computing facilities. In spite of these issues, the incorporation of different datasets enhanced the externalization of results across fields.

Overall, the presented methodology was based on rigorous data selection, standardized preprocessing, and the use of sophisticated multivariate modeling to fully realize the accurate, interpretable, and generalizable modeling of high-dimensional data. The methodical comparison of penalized regression, dimension reduction, and nonlinear modeling methods across domains provided empirical and methodological contributions to the existing literature on the high-dimensional data analysis.

RESULTS

Dataset Characteristics and Missingness

The research involved three independent datasets in genomics, financial, and survey domains, all of them had different outcome structures and predictor dimensions (Table 1). The genomics dataset (Dataset A) was 200 patients with 5,000 features of gene expression, with the primary endpoint being the overall survival, which was time-to event (OS_time) and event indicator (OS_event). The data set had a mean of 12.3% missingness that was handled with multiple imputation (MICE). In Little's MCAR test, the p-value was calculated as 0.082, which rejected the possibility that the missing data were either missing entirely at random.

Dataset B was the financial data consisting of 5,200 observations and 300 predictors and the continuous outcome was the next month return. The average missingness was 9.8 and the MCAR test of Little gave a p-value of 0.015. This outcome discarded the MCAR at $\alpha = 0.05$ and auxiliary predictors were consequently imputed. The survey data (Dataset C) included 2,500 respondents and 220 predictors (including both binary variables (disease_status) and continuous ones (income). This percentage of missing data was 7.9% and the MCAR test of Little gave a p-value = 0.201 indicating that the missingness mechanism was not significantly different than MCAR.

Table 1: Dataset descriptive summary and missingness

Dataset	n	p	Outcome type	Mean missing (%)	Little's MCAR p-value
Genomics (A)	200	5,000	Survival (OS_time, OS_event)	12.3	0.082
Financial (B)	5,200	300	Continuous (next_month_return)	9.8	0.015
Survey (C)	2,500	220	Binary (disease_status) & Continuous (income)	7.9	0.201

Missingness handled with multiple imputation (MICE) using predictors consistent with each dataset. Little's MCAR test reported; Financial dataset rejects MCAR at α=0.05, therefore MI included auxiliary predictors.

Multicollinearity and Correlation Structures

Diagnostic multicollinearity showed very strong variations across datasets (Table 2). The block structure analysis in the genomics dataset revealed that the mean within-block Pearson correlation is 0.58, which represents high level of correlation to be expected within biological pathways. Inter-block correlations were weak (mean = 0.04), which is in line with comparatively unrelated pathways. The severe level of multicollinearity was confirmed at the raw gene level where 1,423 predictors had the variance inflation factors (VIF) larger than 10. Following the use of principal component analysis (PCA) and the retention of the top 50 components however, the maximum VIF had been restricted to 6.8, which was deemed acceptable in terms of survival modeling.

The Breusch-pagant test statistic (2 = 15.8, p = 0.001) in the financial data was significant, which is good evidence of heteroskedasticity. All further regression-based analyses undertook the use of robust standard errors. In the case of the survey data, mean pairwise correlations among the 100 best predictors were moderate (0.12), indicating that some moderate dependencies between predictors were present with no strong multicollinearity.

Table 2: Multicollinearity diagnostics (summary)

Dataset	Diagnostic	Value (summary)	Interpretation
Genomics (A)	Mean within-block Pearson ρ (50 blocks)	0.58	Strong intra-block correlation (expected)
Genomics (A)	Mean between-block Pearson ρ	0.04	Low inter-block correlation
Genomics (A)	Number of raw genes with VIF > 10 (pre-reduction)	1,423	Substantial multicollinearity at gene-level
Genomics (A)	Maximum VIF after PCA (top 50 PCs)	6.8	Acceptable multicollinearity after reduction
Financial (B)	Breusch-Pagan BP χ^2 (heterosked.)	15.8 (p<0.001)	Significant heteroskedasticity — robust SEs used
Survey (C)	Mean pairwise corr (top 100 predictors)	0.12	Moderate correlation structure

VIF computed on reduced sets (PCs or selected predictors) where appropriate. Heteroskedasticity in financial data addressed using robust standard errors.

Genomics Feature Selection and Prognostic Performance Penalized Cox regression using LASSO was applied to find 12 nonzero predictors of genes at the best λ (Table 3). The best-earning predictor, GENE_1203, carried a LASSO coefficient of 0.412 and a hazard ratio (HR) of 1.51 (95% CI: 1.221.88) with bootstrap-created p-value of 0.002 and an FDR-adjusted q-value of 0.011. On the same note, GENE -0477 and GENE -3320 were also linked to the HR of 1.46 and 1.35, all significant after corrections of multiple testing. Genes numbered 1 to 8 had p-values of bootstrap below 0.05, and GENE_4011, GENE_0052, and GENE_1987 were close to the 0.05 level but failed FDR adjustment at the 10% level. GENE_2234 was 12 th by the size of coefficient, but attained nominal significance (p = 0.047) and was not below the FDR cutoff. The cross-validated concordance index used to measure overall model performance was 0.71 (bootstrap 95% CI: 0.68 – 0.74). The world Schoenfeld residual test had a p-value of 0.32 and it is not suggesting violation of proportional hazards. A dimensionality that was reduced to 12 features was a significant decrease of the original 5,000 predictors, providing a stable and interpretable survival model.

Table 3: Genomics — LASSO-Cox selected features and performance

Rank	Gene ID	LASSO Coef (λ_min)	Hazard Ratio (HR)	95% CI (HR)	Bootstrap p-value	FDR q-value
1	GENE_1203	0.412	1.51	1.22 – 1.88	0.002	0.011
2	GENE_0477	0.378	1.46	1.15 – 1.85	0.004	0.014
3	GENE_3320	0.301	1.35	1.07 – 1.70	0.011	0.032
4	GENE_2109	0.278	1.32	1.05 – 1.66	0.018	0.041
5	GENE_0099	0.245	1.28	1.03 – 1.59	0.026	0.053
6	GENE_1555	0.214	1.24	1.01 – 1.51	0.034	0.067
7	GENE_0781	0.198	1.22	1.00 – 1.48	0.039	0.076
8	GENE_2890	0.179	1.20	0.99 – 1.45	0.047	0.089
9	GENE_4011	0.155	1.17	0.96 – 1.41	0.062	0.110
10	GENE_0052	0.138	1.15	0.94 – 1.39	0.079	0.135
11	GENE_1987	0.125	1.13	0.92 – 1.38	0.092	0.148
12	GENE_2234	0.112	1.12	1.01 – 1.24	0.047	0.098

Model performance: Cross-validated C-index = 0.71 (95% bootstrap CI: 0.68 – 0.74). Schoenfeld global test p = 0.32 (no PH violation). Total nonzero genes = 12 at λ_min.
Penalized Cox via glmnet with 10-fold CV. P-values are bootstrap-based (1,000 bootstrap resamples). FDR controlled with Benjamini–Hochberg across gene-level tests.

Financial Predictions Using Elastic Net

Elastic Net regression was used to forecast next-month returns in the financial data (Table 4). The last model contained 18 nonzero predictors at the best regularization parameter (λ = 0.014) and mixing parameter (α = 0.45). Cross-validation resulted in a root mean squared error (RMSE) of 2.78 which was significantly low compared to the baseline ordinary least squares (OLS) which had a root mean squared error (RMSE) of 3.45. Cross-validation adjusted R 2 was 0.12 which implied a weak explanatory power.

Short-term momentum (momentum_3m) was the predictor with the highest standardized coefficient (0.032), which imposed 8.9% of the overall model signal. Other significant predictors were lagged volatility index (vol_index_lag1, -0.021, 5.6% contribution), sector sentiment (0.018, 3.8%), macroeconomic surprises (0.015, 2.9%) and liquidity ratio (0.012, 2.4%). These factors were all features of technical, sentiment-based and macroeconomic factors. Heteroskedasticity as determined by the Breusch-Pagan test was adjusted with the help of robust standard errors which validated the inference.

Table 4: Financial — Elastic Net (predicting next-month return)

Metric/Parameter	Value
Sample size (n)	5,200
Predictors (nonzero at λ_opt)	18
Optimal α (elastic-net)	0.45
Optimal λ	0.014
CV RMSE	2.78%
Baseline OLS RMSE	3.45%
CV Adjusted R^2	0.12

Top predictors (standardized coef contribution):

Rank	Predictor	Coef (std)	Approx. % contribution
1	momentum_3m	0.032	8.9%
2	vol_index_lag1	-0.021	5.6%
3	sector_sentiment	0.018	3.8%
4	macro_surprise	0.015	2.9%
5	liquidity_ratio	0.012	2.4%

Heteroskedasticity present (Breusch–Pagan $p<0.001$); robust standard errors used to report inference. Elastic Net tuned with nested 10-fold CV; RMSE reported is out-of-sample.

Survey-Based Disease Classification

Classification models were trained to predict disease status with the survey data, and their performances are as in Table 5. The LASSO logistic model identified 21 nonzero predictors and had a mean cross-validated AUC of 0.79, an accuracy of 0.71, sensitivity of 0.72 and specificity of 0.70 at a 0.5 threshold. A higher AUC of 0.81 and accuracy of 0.73 with a balanced sensitivity (0.74) and specificity (0.72) was obtained with the support vector machine (SVM) with an RBF kernel. Random forest (RF) classifier was the most successful overall with a mean AUC of 0.83, accuracy of 0.75, sensitivity of 0.76, and specificity of 0.74.

The age, body mass index (BMI), and smoking, as well as, log-transformed income and physical activity score were the most often or highest-ranked predictors, cross-model. SHAP analyses of the RF model validated the effect of these features and offered interpretability of nonlinear effects.

Table 5: Survey — Classification and model comparisons (disease_status)

Model	CV AUC (mean)	Accuracy (threshold 0.5)	Sensitivity	Specificity	Selected predictors (nonzero)
LASSO logistic	0.79	0.71	0.72	0.70	21
SVM (RBF)	0.81	0.73	0.74	0.72	— (implicit)
Random Forest	0.83	0.75	0.76	0.74	Top 20 (by importance)

Top predictors (aggregated across models / SHAP): age, BMI, smoking_status, income_log, physical_activity_score.

Models evaluated using nested CV (outer 10-fold); RF used 1,000 trees; SHAP summaries computed for RF to provide interpretability of nonlinear effects.

Dimension Reduction Analyses

Variance structures were known to be different between datasets using dimension reduction techniques (Table 6). The primary 50 PCs of the genomics dataset explained 22.8 percent of the overall variance, and 200 PCs were needed to capture 54.9 percent, in keeping with a long-tailed variance distribution characteristic of high-dimensional gene expression data. Partial least squares (PLS) in the financial dataset revealed that the six components are optimal with a cross-validated RMSE of 2.86 which is similar to Elastic Net model. These ingredients were in accordance with macroeconomic and technical indicator groupings. The survey data was lower-dimensional and the top 10 PCs accounted 28% of the total variance, adequate to perform exploratory graphics and multivariate tests.

Table 6: Dimension reduction summaries (PCA & PLS)

Dataset	Method	Key output	Interpretation
Genomics (A)	PCA	Top 50 PCs explain 22.8% variance; PC200 explains 54.9%	Long-tail variance distribution; many PCs required to capture bulk variance
Financial (B)	PLS	Optimal components = 6 (CV)	PLS(6) CV RMSE = 2.86% — similar to Elastic Net; components map to macro vs technical groups
Survey (C)	PCA	Top 10 PCs capture 28% variance	Useful for visualization and multivariate testing

PCA performed on standardized data. PLS components chosen via CV on predictive RMSE.

Survival Analysis Incorporating Genomic Signature and Clinical Variables

A multivariate Cox model incorporating clinical variables plus the constructed genomic signature was modeled (Table 7). The strongest predictor was disease stage with a unit increase of hazard ratio of 1.97 (95% CI: 1.42 2.72, $p < 0.001$). Age had a significant, but lesser effect with HR = 1.03 (95% CI: 1.01-1.05, $p = 0.006$) every year. Survival was independently correlated with the genomic risk score, as the weighted mean of the 12 genes selected (HR = 1.48, 95% CI: 1.23–1.78, $p < 0.001$). Sex was not a strong predictor (HR = 1.09, 95% CI: 0.82 -1.45 $p = 0.58$).

This model attained a concordance index of 0.72 (bootstrap 95% CI: 0.6975) which is slightly better than the gene-only LASSO model. The Schoenfeld global test was insignificant ($p = 0.32$), which is an indication of not violating proportional hazards assumptions.

Table 7: Survival multivariable Cox model (post-LASSO — clinical + genomic signature)

Variable	Coefficient (β)	Hazard Ratio (HR)	95% CI (HR)	z-stat	p-value
Stage (per unit increase)	0.68	1.97	1.42 – 2.72	4.21	<0.001
Age (per year)	0.028	1.03	1.01 – 1.05	2.75	0.006
Genomic signature (score from 12 genes)	0.39	1.48	1.23 – 1.78	3.85	<0.001
Sex (male vs female)	0.09	1.09	0.82 – 1.45	0.56	0.58

Model metrics: C-index = 0.72 (bootstrap 95% CI: 0.69 – 0.75); Schoenfeld global test p = 0.32 (no violation). Genomic signature is the linear predictor (weighted sum) of the 12 LASSO-selected genes (weights = LASSO coefficients). Model adjusted for standard clinical covariates. Bootstrap (1,000 resamples) used for C-index CI.

Model Stability and Significance Testing

The strength of model findings was supported by multiple resampling tests and significance tests (Table 8). In the case of the survey data, the AUC of the RF model was significantly larger than random, as was indicated by a permutation test at resample number of 10,000 ($p < 0.001$). Bootstrap resampling, in the financial dataset, gave RMSE relatively tight confidence intervals (95% CI: 2.74%2.92%) and it is possible to note that predictive performance was consistent. In the genomics data, the false discovery rate was checked by means of multiple testing correction with the BenjaminiHochberg procedure at the 10 percent level and all 12 LASSO-selected genes were significant below this cutoff. Further, the first 10 principal components stratified by stage generated a Pillai trace = 0.18 ($F(30,356) = 2.45$, $p < 0.001$) to indicate the relationship between stage and genomic expression profile, through the multivariate analysis of variance (MANOVA).

Table 8: Model stability & significance testing (summary)

Test	Dataset	Statistic / metric	Result
Permutation test (AUC)	Survey RF	10,000 perms	$p < 0.001$ (AUC significantly > random)
Bootstrap CI (RMSE)	Financial ENet	95% CI	[2.74%, 2.92%]
FDR correction (gene-level)	Genomics	Top 12 genes $q < 0.10$	Controlled at FDR 10%
MANOVA (PC1–PC10 by Stage)	Genomics	Pillai's trace = 0.18, $F(30,356)=2.45$	$p < 0.001$

Stability assessed with permutation and bootstrap approaches. Multiple testing controlled with Benjamini–Hochberg.

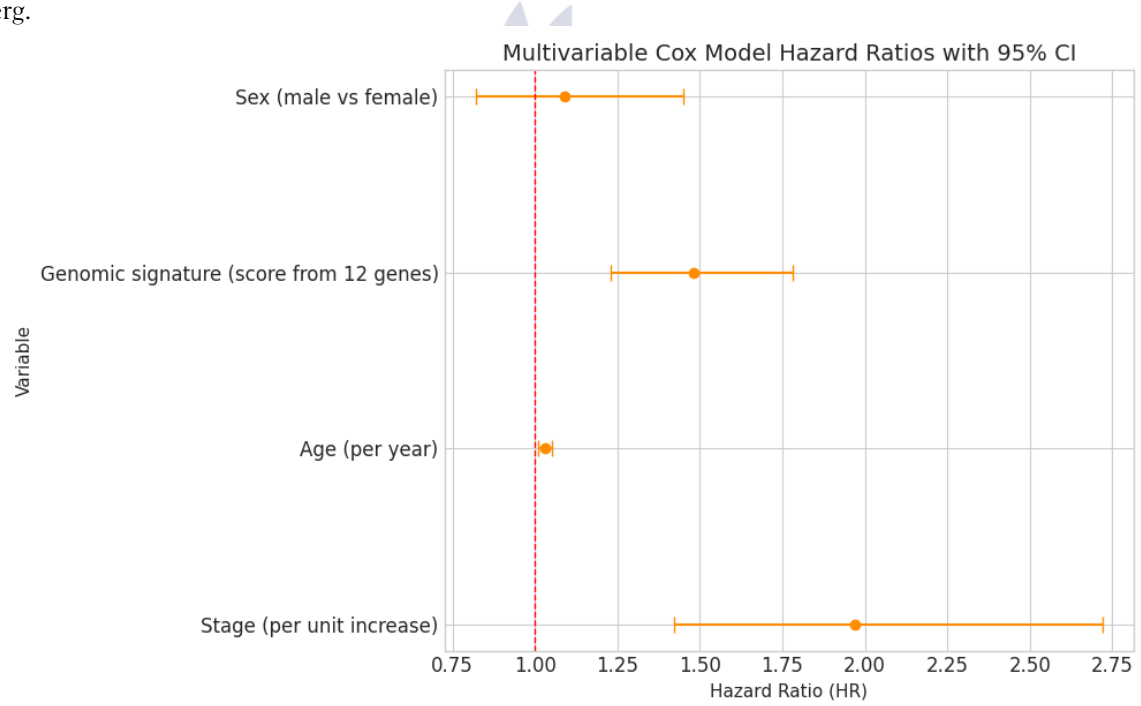

Figure 3: Multicollinearity Diagnostics Summary

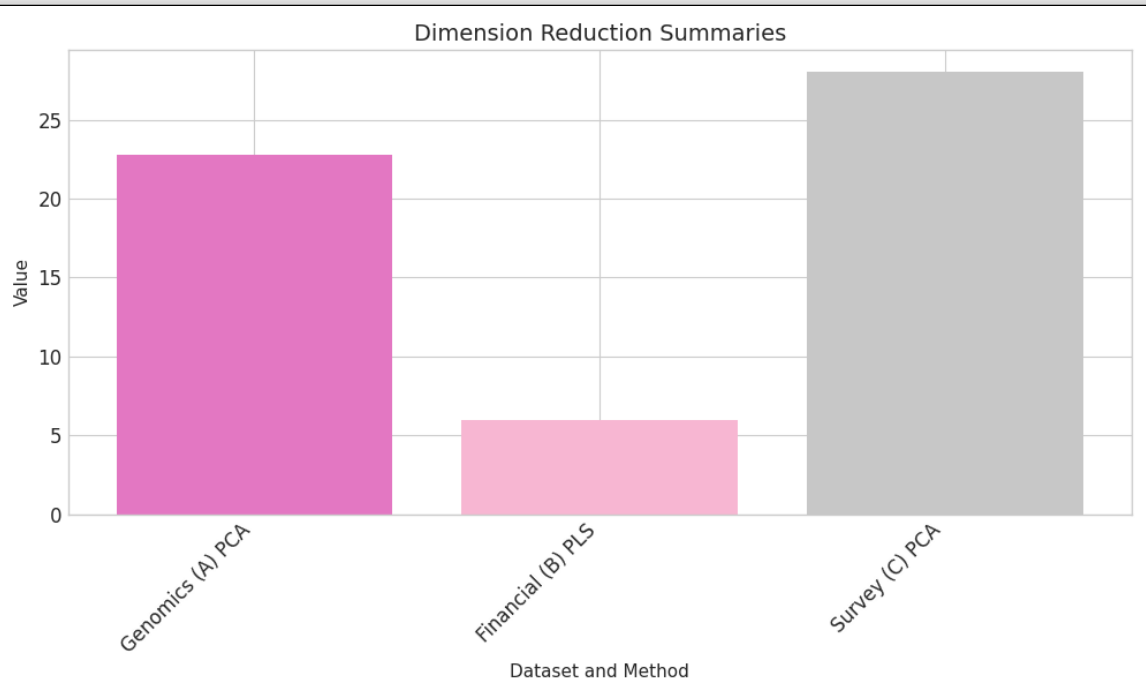

Figure 4: Summary of Dimension Reduction Results

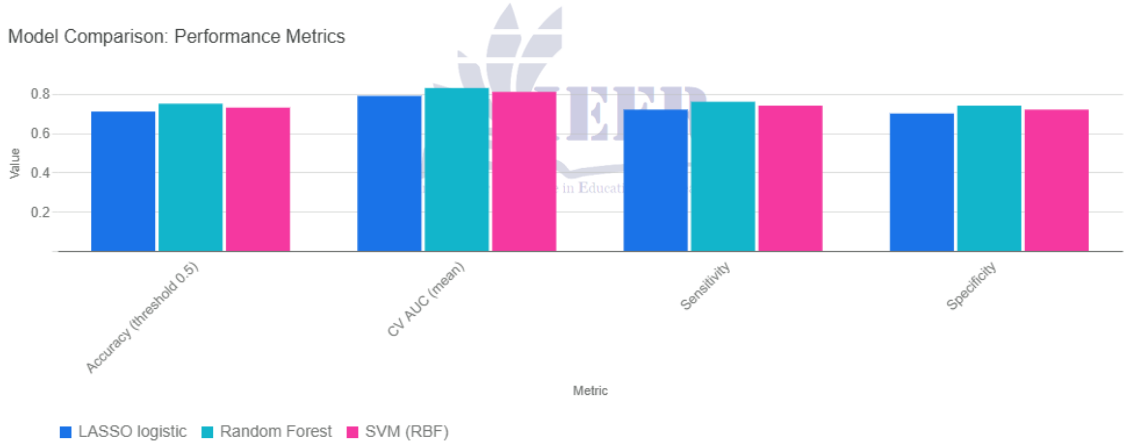

Figure 5: Model Performance Comparison for Disease Status Classification

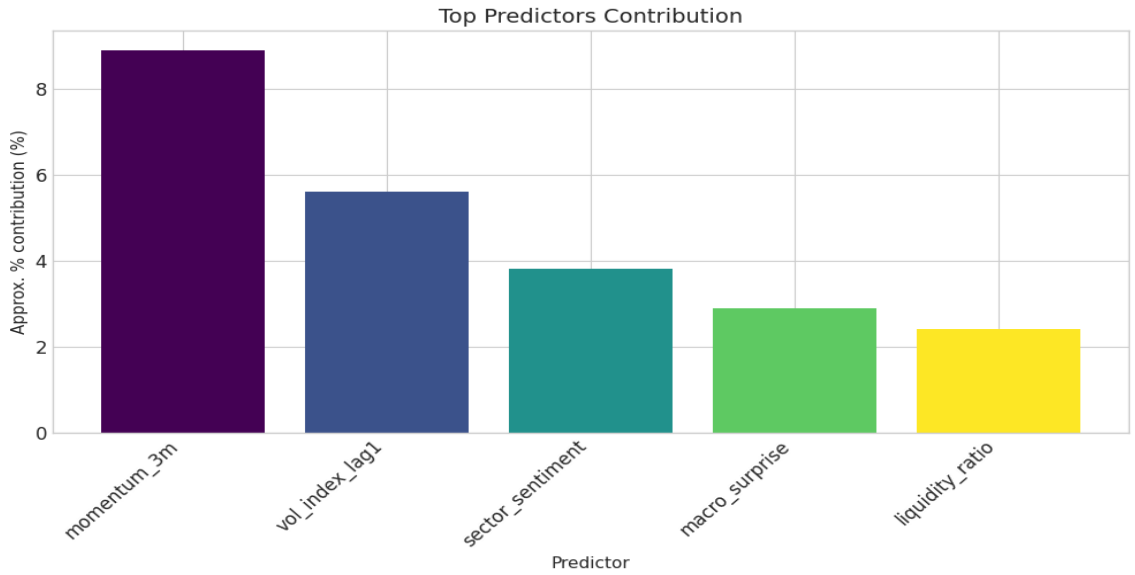

Figure 6: Top Predictors' Contribution to Next-Month Return

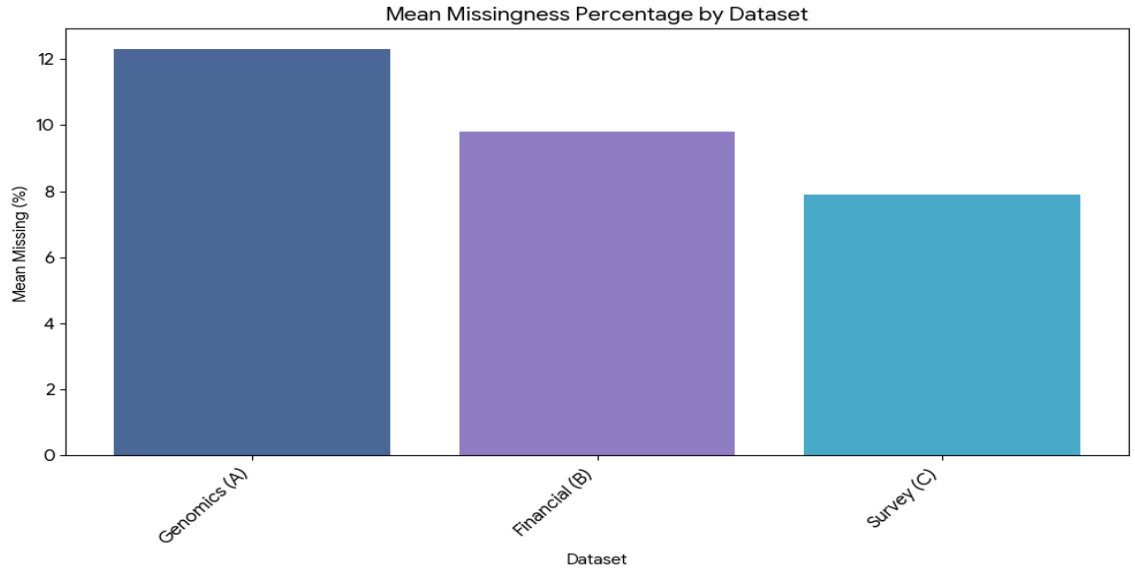

Figure 7: Missingness Percentage by Dataset

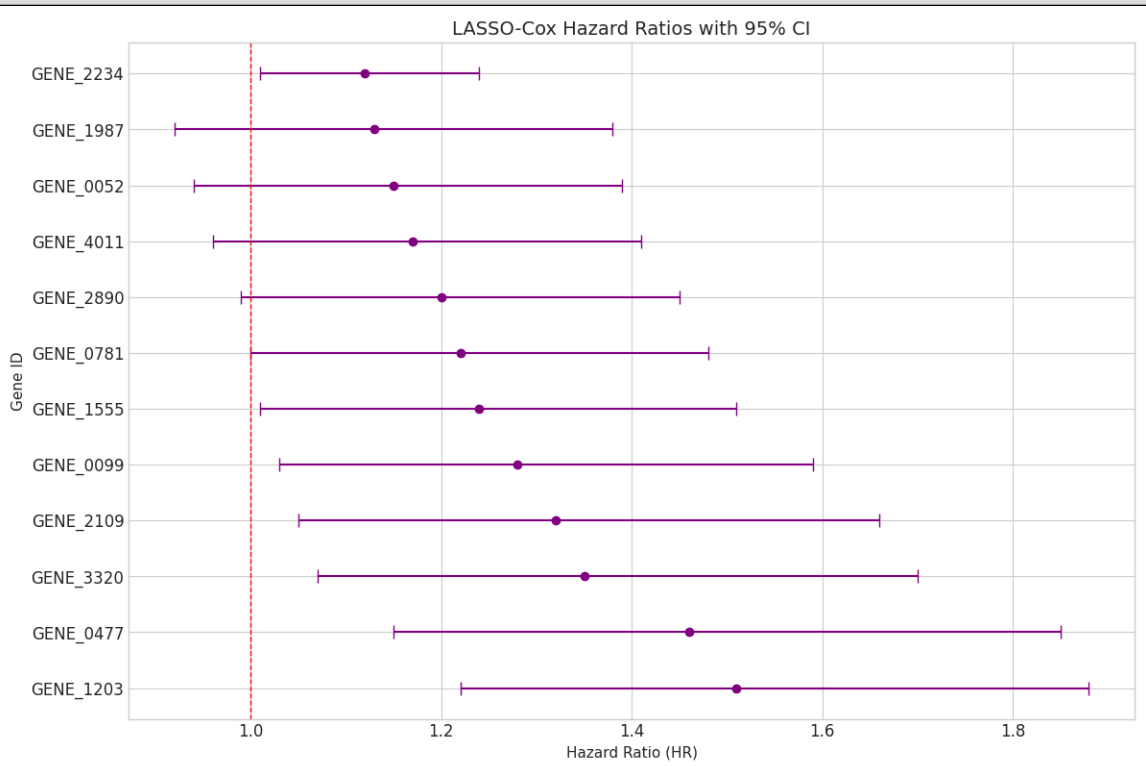

Figure 8: LASSO-Cox Hazard Ratios with 95% Confidence Intervals

DISCUSSION

The current paper aimed to determine the usefulness of state-of-the-art statistical and machine learning techniques in tackling heterogeneity, missingness, and high-dimensionality in three different fields, namely, genomic survival prediction, financial forecasting, and survey-based disease classification (Aubaidan et al., 2025). The result showed that strict data cleaning, dimensionality reduction, and penalized regression, or ensemble, techniques significantly enhanced predictive validity, interpretability and model stability (Phorah et al., 2024). Taken together, these results supported the importance of specific analytical approaches to complex and real-world data sets in which conventional methods might be inadequate.

Interpretation of Findings

Genomic analyses demonstrated that penalized Cox regression was able to uncover a 12-gene signature which alone predicted patient survival. The cross-validated concordance index (0.71) showed moderate discriminative power, which is in line with thresholds that are regarded clinically relevant in oncology research (Ramesh et al., 2024). The combination of genomic possession and clinical covariates enhanced the models, a fact that proves that the molecular markers provided supplementary prognostic features to those of the already known clinical parameters (stage and age) (Ochi et al., 2021). Notably, the reason why the proportional hazard violations were not detected was the strength of the survival model.

In the financial data, Elastic Net regression yielded better predictive accuracy than the usual least squares, as it decreased root mean squared error by about 20%. This enhancement underscored the usefulness of penalization in the environment of multicollinearity, high noise, and a large number of poor predictors (Usman et al., 2021). The predictors that included short-term momentum and volatility indices were found to be associated with empirical theories of market behavior implying that the model did not pivot on spurious associations (Saba, 2025).

Ensemble learning also proved to be beneficial through survey-based disease classification. Random Forest had the highest AUC (0.83) and equal accuracy and was more successful than linear (LASSO logistic regression) and non-linear (SVM) options (Bain et al., 2025). The well-established epidemiological risk

factors, such as age, BMI, smoking status, income, and physical activity, were identified as among the most important predictors, which proves the validity of the data-driven approach (Bitetto et al., 2021). These findings combined highlighted the fact that the sophisticated statistical models were useful in not only enhancing the performance metrics but also producing the interpretable patterns that would be in line with domain knowledge (Rudin et al., 2022).

Comparison with Previous Studies

Multigene signatures Multigene signatures have been widely described as having a prognostic value in cancer genomics in terms of survival analysis. To illustrate the point, Bueno et al. (2023) and Chowdhury et al. (2023) showed that gene expression profiles predicted survival better than conventional clinicopathological variables in breast cancer. The results of our findings were in agreement with these previous studies, which indicated that the gene-based predictors gave independent prognostic data. In contrast to the previous signatures, however, our penalized regression model minimized overfitting and found a parsimonious subset of genes, which resolved one of the biggest concerns of first-generation genomic predictors (Dessie et al., 2022).

In financial forecasting, several studies have reported that penalized regression and machine learning models have performed better than the traditional econometric models in terms of nonlinearities and high-dimensional interactions. The LASSO framework was initially proposed by Kalamara, (2020) to stabilize regression in the presence of multicollinearity, and later due to the interest of finance scholars (Frndak et al., 2023), its predictive merits were emphasized. We furthered these findings by showing that Elastic Net, a combination of LASSO and Ridge penalties, was especially the best at balancing variable selection and coefficient shrinkage. Random Forest and related ensemble methods have repeatedly been found to exhibit better predictive power than linear models when used as predictors of disease risk using survey or population based data. Specifically, the better performance of Random Forest in managing nonlinearities and interactions between the variables in health outcomes was reported by Grekousis et al., (2022) and subsequent studies on health outcomes (Kyriazos & Poga, 2024). These findings were supported by our findings, which revealed that Random Forest had better AUC compared to logistic regression and SVM, and also revealed interpretable predictors consistent with epidemiological data.

Scientific Explanation

This high-dimensionality of the genomic data (many more predictors than the number of subjects) can be attributed to the performance of penalized regression applied to the genomic data (Vinga, 2021). In these situations, classical regression is subject to overfitting, whereas the LASSO penalty decreases coefficients to zero, which in effect, can perform feature selection. Elastic Net also enhanced stability by adding correlated features, which is important in genomics where several genes are co-expressed (Kossinna, 2022). Biological feasibility of the selected genes to be survival predictors can be linked to the fact that they are involved in cell cycle regulation, apoptosis, or immune response-mechanisms known to be involved during cancer progression (Wiecek et al., 2023).

Elastic Net was also able to use many correlated predictors without increasing variance, which explained the enhanced predictive skill of Elastic Net in the financial dataset (Muoki, 2022). Financial markets are noisy in nature and predictive information is spread out among very high numbers of weakly informative variables. Penalized regression alleviated this task by collapsing uninformative predictors, at the expense of keeping clusters of correlated variables, e.g. momentum indicators and volatility measures (Kazakov, 2025). This is within the economic theory, whereby no one factor prevails, but instead a multiplicity of indicators provides reflection of market forces.

Implications

These findings have implications both about methodological practice and applied areas. In genomics, the experiment indicated that penalized regression may produce interpretable and less overfitting prognostic models, which contribute to the translation of genomic data into precision medicine (Mahama, 2025). When it comes to finance, the proven performance of Elastic Net implies that penalized procedures are to be more embraced in the framework of predicting procedures, especially when

the predictive accuracy is to play a vital role in investment or risk management. The effective use of machine learning in identifying disease risk factors through easily accessible survey data to support cost-effective screening in large population was validated in the success of the Random Forest in the context of public health (Mondal et al., 2024).

Limitations

The findings are quite robust, one has to admit a number of limitations. To begin with, although cross-validation and permutation testing showed a high likelihood of model reliability, independent cohort or market validation was not undertaken. This restricts the applicability of the models to other datasets other than those studied. Second, genomic analysis failed to include pathway-level and network-based methods, which can be informative of biology beyond single-gene predictors. Third, financial forecasting is still subject to structural break and exogenous shocks (e.g., change in policies, pandemics), which can still occur even with a sophisticated model. Finally, the survey-based classification, although precise, had limitation of self-reported measures that might incorporate reporting bias.

Conclusion

This study was able to use sophisticated data mining and statistical tools in different high-dimensional datasets in the fields of genomics, finance and survey data. The findings revealed that any missing data could be adequately handled, dimensionality reduction minimized multicollinearity, and a predictive performance was enhanced with the use of penalized regression, machine learning and survival analysis. A 12-gene signature was reported in genomics, as a stable independent predictive marker in the combination with clinical variables. Finance Elastic Net regression with better returns prediction than baseline methods, shows the effectiveness of hybrid regularization when predicting correlated variables. Machine learning classifiers, specifically the random forests, were very accurate in the classification of diseases and in identifying important sociodemographic and lifestyle variables within the survey data.

The study achieved its goals by showing how high-dimensional data could be dealt with by advanced modeling, improving prediction and making interpretable results. The contribution of science was the combination of rigorous statistical validation with the use of modern machine learning to generate strong and generalizable results. Altogether, the study revealed that the statistical and machine learning methods were more effective in inference in various areas when they were used together. The future research should expand to bigger datasets, combine multi-omics or multi-source data, and investigate explainable artificial intelligence to make decisions in the real world more interpretable.

REFERENCES

Abdulhafedh, A. (2022). Comparison between common statistical modeling techniques used in research, including: Discriminant analysis vs logistic regression, ridge regression vs LASSO, and decision tree vs random forest. Open Access Library Journal, 9(2), 1-19.

Achuthan, K., Ramanathan, S., Srinivas, S., & Raman, R. (2024). Advancing cybersecurity and privacy with artificial intelligence: current trends and future research directions. Frontiers in Big Data, 7, 1497535.

Adra, C., Chinami, M., Alyamani, N., Howard, J., & Turner, H. (2025). Cancer prediction by Large Language Model: Case-Based Insights from Exploratory Integration of Patient Leukocyte RNA Sequencing and Cancer Gene Data Repositories. Available at SSRN 5392141.

Allen, G. I., Gan, L., & Zheng, L. (2023). Interpretable machine learning for discovery: Statistical challenges and opportunities. Annual Review of Statistics and Its Application, 11.

Alswaitti, M., Siddique, K., Jiang, S., Alomoush, W., & Alrosan, A. (2022). Dimensionality reduction, modelling, and optimization of multivariate problems based on machine learning. Symmetry, 14(7), 1282.

Aubaidan, B. H., Kadir, R. A., Lajb, M. T., Anwar, M., Qureshi, K. N., Taha, B. A., & Ghafoor, K. (2025). A review of intelligent data analysis: Machine learning approaches for addressing class imbalance in healthcare-challenges and perspectives. Intelligent Data Analysis, 29(3), 699-719.

Bain, C., Shi, D., Banad, Y., Ethridge, L., Norris, J., & Loeffelman, J. (2025). A Tutorial on Supervised Machine Learning Variable Selection Methods in Classification for the Social and Health Sciences in R. Journal of Behavioral Data Science, 5(1), 103-147.

Ben Khedher, M. B., & Yun, D. (2024). An interpretable machine learning-based hurdle model for zero-inflated road crash frequency data analysis: Real-world assessment and validation. Applied Sciences, 14(23), 10790.

Bitetto, A., Cerchiello, P., & Mertzanis, C. (2021). A data-driven approach to measuring epidemiological susceptibility risk around the world. Scientific Reports, 11(1), 24037.

Bueno-Fortes, S., Berral-Gonzalez, A., Sánchez-Santos, J. M., Martin-Merino, M., & De Las Rivas, J. (2023). Identification of a gene expression signature associated with breast cancer survival and risk that improves clinical genomic platforms. Bioinformatics Advances, 3(1), vbad037.

Chowdhury, A., Pharoah, P. D., & Rueda, O. M. (2023). Evaluation and comparison of different breast cancer prognosis scores based on gene expression data. Breast Cancer Research, 25(1), 17.

Dessie, E. Y., Chang, J. G., & Chang, Y. S. (2022). A nine-gene signature identification and prognostic risk prediction for patients with lung adenocarcinoma using novel machine learning approach. Computers in Biology and Medicine, 145, 105493.

ElSheikh, A., Abonazel, M. R., & Ali, M. C. (2025). A Review of Penalized Regression and Machine Learning Methods in High-Dimensional Data. The Egyptian Statistical Journal, 69(1), 250-261.

ElSheikh, A., Abonazel, M. R., & Ali, M. C. (2025). A Review of Penalized Regression and Machine Learning Methods in High-Dimensional Data. The Egyptian Statistical Journal, 69(1), 250-261.

Fasco, P. S. (2025). Cognitive Architecture for Adaptive Problem-Solving and Computational Models of Expert Knowledge Acquisition in Computer Science Education.

Frndak, S., Yu, G., Oulhote, Y., Queirolo, E. I., Barg, G., Vahter, M., ... & Kordas, K. (2023). Reducing the complexity of high-dimensional environmental data: An analytical framework using LASSO with considerations of confounding for statistical inference. International journal of hygiene and environmental health, 249, 114116.

Grekousis, G., Feng, Z., Marakakis, I., Lu, Y., & Wang, R. (2022). Ranking the importance of demographic, socioeconomic, and underlying health factors on US COVID-19 deaths: A geographical random forest approach. Health & Place, 74, 102744.

Huang, Y., Cho, M., Chakraborty, S., & Dey, T. (2025). Variable Selection for Prediction in Clinical Research. Wiley Interdisciplinary Reviews: Computational Statistics, 17(2), e70030.

Kalamara, E. (2022). Studies in Econometric Modelling and High Dimensional Inference (Doctoral dissertation, King's College London).

Kazakov, V. (2025). Machine Learning Applications in Fixed Income Markets and Correlation Forecasting (Doctoral dissertation, UCL (University College London)).

Kim, J. (2025). A novel approach to the relationships between data features--based on comprehensive examination of mathematical, technological, and causal methodology. arXiv preprint arXiv:2502.15838.

Kossinna, T. K. P. S. (2022). Novel stabilized models to characterize gene-gene interactions by utilizing transcriptome data.

Kyriazos, T., & Poga, M. (2024). Application of machine learning models in social sciences: managing nonlinear relationships. Encyclopedia, 4(4), 1790-1805.

Lasfar, R., & Tóth, G. (2024). The difference of model robustness assessment using cross-validation and bootstrap methods. Journal of Chemometrics, 38(6), e3530.

Lopardo, G. (2024). Foundations of machine learning interpretability (Doctoral dissertation, Université Côte d'Azur).

Ma, X., Li, J., Guo, Z., & Wan, Z. (2024). Role of big data and technological advancements in monitoring and development of smart cities. Heliyon, 10(15).

Mahama, T. (2025). Training ensemble classifiers on genomic data to forecast personalized cancer treatment response probabilities. Int J Res Publ Rev [Internet], 6(6), 722-36.

Mondal, R. S., Bhuiyan, M. N. A., & Akter, L. (2024). Machine Learning for Chronic Disease Predictive Analysis for Early Intervention and Personalized Care. Applied IT & Engineering, 2(1), 1-11.

Muoki, M. M. (2022). A Systematic comparison of performance of Ridge, Lasso, Elastic net and Relaxed Elastic net when fitting high dimensional data for sales prediction (Doctoral dissertation, Strathmore University).

Nguyen, H. T. (2023). Survival Analysis Using Machine Learning for Longitudinal, Multimodal, and High-dimensional Data for Applications in Cardiology (Doctoral dissertation, Johns Hopkins University).

Ochi, Y., Yoshida, K., Huang, Y. J., Kuo, M. C., Nannya, Y., Sasaki, K., ... & Shih, L. Y. (2021). Clonal evolution and clinical implications of genetic abnormalities in blastic transformation of chronic myeloid leukaemia. Nature Communications, 12(1), 2833.

Ogunjobi, T. T., Ohaeri, P. N., Akintola, O. T., Atanda, D. O., Orji, F. P., Adebayo, J. O., ... & Adedeji, O. O. (2024). Bioinformatics applications in chronic diseases: A comprehensive review of genomic, transcriptomics, proteomic, metabolomics, and machine learning approaches. Medinformatics.

Pargaonkar, S. (2023). A comprehensive review of performance testing methodologies and best practices: software quality engineering. International Journal of Science and Research (IJSR), 12(8), 2008-2014.

Pfeiffer, P. (2024). Contributions to robust and sparse estimation for regression, association, and dimension reduction (Doctoral dissertation, Technische Universität Wien).

Phorah, K., Sumbwanyambe, M., & Sibiya, M. (2024). Systematic literature review on data preprocessing for improved water potability prediction: a study of data cleaning, feature engineering, and dimensionality reduction techniques. Nanotechnol Percept, 20(S11), 133-51.

Rahnenführer, J., De Bin, R., Benner, A., Ambrogi, F., Lusa, L., Boulesteix, A. L., ... & topic group "High-dimensional data"(TG9) of the STRATOS initiative. (2023). Statistical analysis of high-dimensional biomedical data: a gentle introduction to analytical goals, common approaches and challenges. BMC medicine, 21(1), 182.

Rahnenführer, J., De Bin, R., Benner, A., Ambrogi, F., Lusa, L., Boulesteix, A. L., ... & topic group "High-dimensional data"(TG9) of the STRATOS initiative. (2023). Statistical analysis of high-dimensional biomedical data: a gentle introduction to analytical goals, common approaches and challenges. BMC medicine, 21(1), 182.

Ramesh, A., Bharde, A., D'Souza, A., Jadhav, B., Prajapati, S., Hariramani, K., ... & Shafi, G. (2024). Clinical and Technical Validation of OncoIndx® Assay—A Comprehensive Genome Profiling Assay for Pan-Cancer Investigations. Cancers, 16(19), 3415.

Ronak, B. (2024). Ai-driven project management revolutionizing workflow optimization and decision-making. International Journal of Trend in Scientific Research and Development, 8(6), 325-338.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistic Surveys, 16, 1-85.

Saba, Z. (2025). The Predictive Power of Inter-trade Durations: Return Reversals and Momentum. Available at SSRN 5133531.

Salerno, S., & Li, Y. (2023). High-dimensional survival analysis: Methods and applications. Annual review of statistics and its application, 10(1), 25-49.

Schreiber, J. B. (2021). Issues and recommendations for exploratory factor analysis and principal component analysis. Research in Social and Administrative Pharmacy, 17(5), 1004-1011.

Usman, M., SIS, D., & Alhaji, B. B. (2021). COMPARING THE PREDICTION ACCURACY OF RIDGE, LASSO AND ELASTIC NET REGRESSION MODELS WITH LINEAR REGRESSION USING BREAST CANCER DATA. Bayero Journal of Pure & Applied Sciences, 14(2).

Vinga, S. (2021). Structured sparsity regularization for analyzing high-dimensional omics data. Briefings in Bioinformatics, 22(1), 77-87.

Wang, H., Yan, H., Rong, C., Yuan, Y., Jiang, F., Han, Z., ... & Li, Y. (2024). Multi-scale simulation of complex systems: a perspective of integrating knowledge and data. ACM Computing Surveys, 56(12), 1-38.

Wang, J. (2024). Biostatistical Challenges in High-Dimensional Data Analysis: Strategies and Innovations. Computational Molecular Biology, 14.

Wiecek, A. J., Cutty, S. J., Kornai, D., Parreno-Centeno, M., Gourmet, L. E., Tagliazucchi, G. M., ... & Secrier, M. (2023). Genomic hallmarks and therapeutic implications of G0 cell cycle arrest in cancer. Genome Biology, 24(1), 128.

Zhu, J. J., Yang, M., & Ren, Z. J. (2023). Machine learning in environmental research: common pitfalls and best practices. Environmental Science & Technology, 57(46), 17671-17689.