Robust Sparse Regression: Enhancing Predictive Accuracy in Noisy Data with Outlier Resilience

22AIE213 – Machine Learning

TEAM – 10

CH.SC.U4AIE23020 – Jakka Aniketh Reddy
CH.SC.U4AIE23023 – Jukonti Aman Reddy
CH.SC.U4AIE23015 – Gajjelli Srimaan

AMRITA VISHWA VIDYAPEETHAM | School of Engineering
AMRITAPURI | BENGALURU | CHENNAI | COIMBATORE

TEAM 10

Introduction

22AIE213

Domain Background

Machine learning in medical imaging has enhanced early disease detection and patient outcomes. Sparse regression techniques like Lasso and Elastic Net are widely used for handling high-dimensional data. However, challenges such as excessive parameter compression, poor model selection, and noise sensitivity hinder their effectiveness. Overcoming these limitations is crucial for improving disease classification models.

Existing Systems

Sparse regression is widely used in medical imaging, but existing methods face challenges. Some models improve performance using deep learning but are complex to implement. Lasso regression with multiple imputation struggles with biases and poor calibration, while noise-handling models require high computational resources. Optimization techniques like Bayesian adaptive Lasso enhance feature selection but still face efficiency and generalization issues.

Introduction

Limitations of existing systems

Despite advancements, existing sparse regression methods face key challenges. Many models are highly sensitive to noise and outliers, leading to inaccurate predictions and reduced robustness. High computational complexity makes implementation difficult, increasing processing time and resource requirements. Additionally, most approaches fail to automatically determine optimal regularization factors in a single end-to-end process, resulting in poor generalization and limited adaptability to new data. Excessive pruning in sparse models further reduces accuracy, making them less effective for complex medical imaging applications. Addressing these limitations is crucial for improving disease classification and diagnostic reliability.

SNO	Title	Author/ Journal year	Methodology	Merits	Demerits	Research Gap
1	Deep ensemble learning of Sparse regression models for brain disease diagnosis	Heung-Il Suk, Seong-Whan Lee, Dinggang Shen 2023	Trains multiple sparse regression models with different regularization parameters. Uses a deep convolutional neural network to integrate target level representations from these models.	Reduces dimensionality of Observations. Combines regression models non-linearly for robust decisions.	Complexity in model Interpretation. Performance is sensitive to parameter tuning	Optimal number of Regularization parameters. End-to-end learning of Parameters. Improving deep learning for medical imaging.
2	Validation of prediction models based on lasso regression with multiply imputed data	Qi Yu, Yoan Miche, Emil Eirola, Mark van Heeswijk, Eric Séverin, Amaury Lendasse	Investigates the performance of lasso regression models with multiply imputed data. Compares 4 approaches of handling multiply imputed data in the bootstrap procedure	Improves prediction quality by shrinking Regression Coefficients. Achieves parsimony through variable selection.	The discriminative model performance of the lasso was optimistic.	Lasso model's predictive performance tends to be optimistic, and the model suffers from suboptimal calibration due to over-shrinkage
3	Advancing robust regression: Addressing asymmetric noise with the BLINEX loss function	Jingjing Tang, Bangxin Liu, Saiji Fu, Yingjie Tian, Gang Kou	Proposes a robust regression model (BXSVR) using the asymmetric bounded linear-exponential (BLINEX) loss function. Uses the Nesterov accelerated gradient (NAG) algorithm for optimization.	Handles asymmetric noise. Mitigates effects of large noise. Smoothness and differentiability aids optimization.	High computational complexity.	Exploring distributed training strategies. Employing more efficient optimization Techniques Addressing the computational demand of LSSVR.
4	Regularized extreme learning machine for regression with missing data	Qi Yu, Yoan Miche, Emil Eirola, Mark van Heeswijk, Eric Séverin, Amaury Lendasse	Proposes a modified extreme learning machine (ELM) using a cascade of L1 penalty (LARS) and L2 penalty (Tikhonov regularization). Estimates pairwise distances between samples on incomplete data.	Shows significant performance in experiments on five datasets.	The discriminative model performance of the lasso was optimistic.	Further research could focus on extending the proposed method to handle other types of missing data patterns or to improve the computational efficiency for very large datasets.

SNO	Title	Author/ Journal year	Methodology	Merits	Demerits	Research Gap
5	The Bayesian adaptive lasso regression	Rahim Alhamzawi, Haithem Taha Mohammad Ali	Considers a fully Bayesian treatment for the adaptive lasso. Uses a new Gibbs sampler with tractable full conditional posteriors	Classical adaptive lasso regression is known to possess the oracle properties.	Requires consistent initial estimates of the regression coefficients.	Exploring distributed training strategies. Employing more efficient optimization techniques. Addressing the computational demand of LSSVR
6	Sparse regression for large data sets with outliers	Lea Bottmer, Christophe Croux, Ines Wilms	Proposes a sparse regression algorithm called "sparse shooting S" for high-dimensional data with outliers. Compares performance with Least Squares (LS), Lasso, sparse Least Trimmed Squares (LTS).	Robust to outliers in the cells of the data matrix.	The discriminative model performance of the lasso was optimistic.	Future work may involve extending the method to handle other types of data contamination or to develop more efficient algorithms for ultra-high dimensional data.
7	Conceptual complexity and bias/variance tradeoff	Erica Briscoe, Jacob Feldman	Examines concept learning using the bias/variance tradeoff. Compares exemplar and prototype models.	Human learners adopt an intermediate point on the bias/ variance continuum.	pure prototype and pure exemplar systems might be seen as laboratory artifacts.	Exploring distributed training strategies. Employing more efficient optimization Techniques. Addressing the computational demand of LSSVR
8	Robust regression and outlier detection in the evaluation of	Edelgard Hund,	Applies robust regression methods to analyze robustness tests with	Provides comparative analysis of different outlier detection and robustregression	Ordinary least squares (OLS) estimates of the	Additional research could explore other robust regression

SNO	Title	Author/ Journal year	Methodology	Merits	Demerits	Research Gap
9	Application of noise-reduction techniques to machine learning algorithms for breast cancer tumor identification	Avani Ahuja, Lidia Al-Zogbi, Axel Krieger	Evaluates the effect of noise reduction techniques (PCA, outlier removal) on machine learning algorithms for breast cancer diagnosi	Noise removal Through dimensionality reduction is effective. High accuracies obtained when both noise-reduction techniques are applied sequentially.	pure prototype and pure exemplar systems might be seen as laboratory artifacts.	Exploring distributed training strategies. Employing more efficient optimization techniques. Addressing the computational demand of LSSVR.
10	Simultaneous feature selection and outlier detection with optimality guarantees	Luca Insolia, Ana Kenney, Francesca Chiaromonte, Giovanni Felici	Develops a general framework using mixed-integer programming to simultaneously perform feature selection and outlier detection.	Provides provably optimal guarantees. Theoretical properties include a necessary and sufficient condition for the robustly strong oracle property	The MIP formulation critically depends on the big-M bounds, which may lead to computational challenges.	Future work could focus on developing more computationally efficient algorithms for solving the mixed-integer programming problem, or extending the framework to handle other types of data complexities.
11	Probabilistic outlier detection for sparse multivariate geotechnical site investigation data using Bayesian learning	Shuo Zheng, Yu-Xin Zhu, Dian-Qing Li, Zi-Jun Cao, Qin-Xuan Deng, Kok-Kwang Phoon	Develops a probabilistic outlier detection method for sparse multivariate data obtained from geotechnical site investigation.	method probabilistically quantifies.outliers in sparse multivariate data,mitigating statistical distortionthrough resampling and Bayesian learning, while identifying outlying components. It significantly reduces masking effects..	Complexity in model Interpretation. Performance is sensitive to parameter tuning	focuses on developing a methodology to address the challenges in outlier detection for sparse multivariate geotechnical site investigation data, implying a gap in effective methods for this specific type of data.
12	Improving machine learning based phase and hardness prediction of high-entropy alloys by using Gaussian noise augmented data.	Yicong Ye, Yahao Li, Runlong Ouyang, Zhouran Zhanga, Yu Tanga,	The paper proposes a mathematical framework combining feature selection and outlier detection. The approach uses convex optimization techniques with theoretical	Provides a unified framework for feature selection and outlier detection. Offers theoretical optimality	Computational complexity may be high for large datasets. Performance may depend	Need for scalability improvements to handle large datasets efficiently. Exploration of alternative optimization

TEAM 10

Problem Identification

22AIE213

Existing sparse regression methods in medical imaging face significant challenges that affect their accuracy and reliability. These models are highly sensitive to noise and outliers, leading to inaccurate predictions. High computational complexity makes them difficult to implement efficiently, requiring extensive processing power. Additionally, the lack of an automated approach for selecting optimal regularization factors results in poor generalization, limiting their effectiveness on new data. Excessive pruning further reduces model accuracy, making them less reliable for disease classification. Addressing these issues is essential to develop a more robust and efficient regression model for medical diagnostics.

TEAM 10

Proposed System

22AIE213

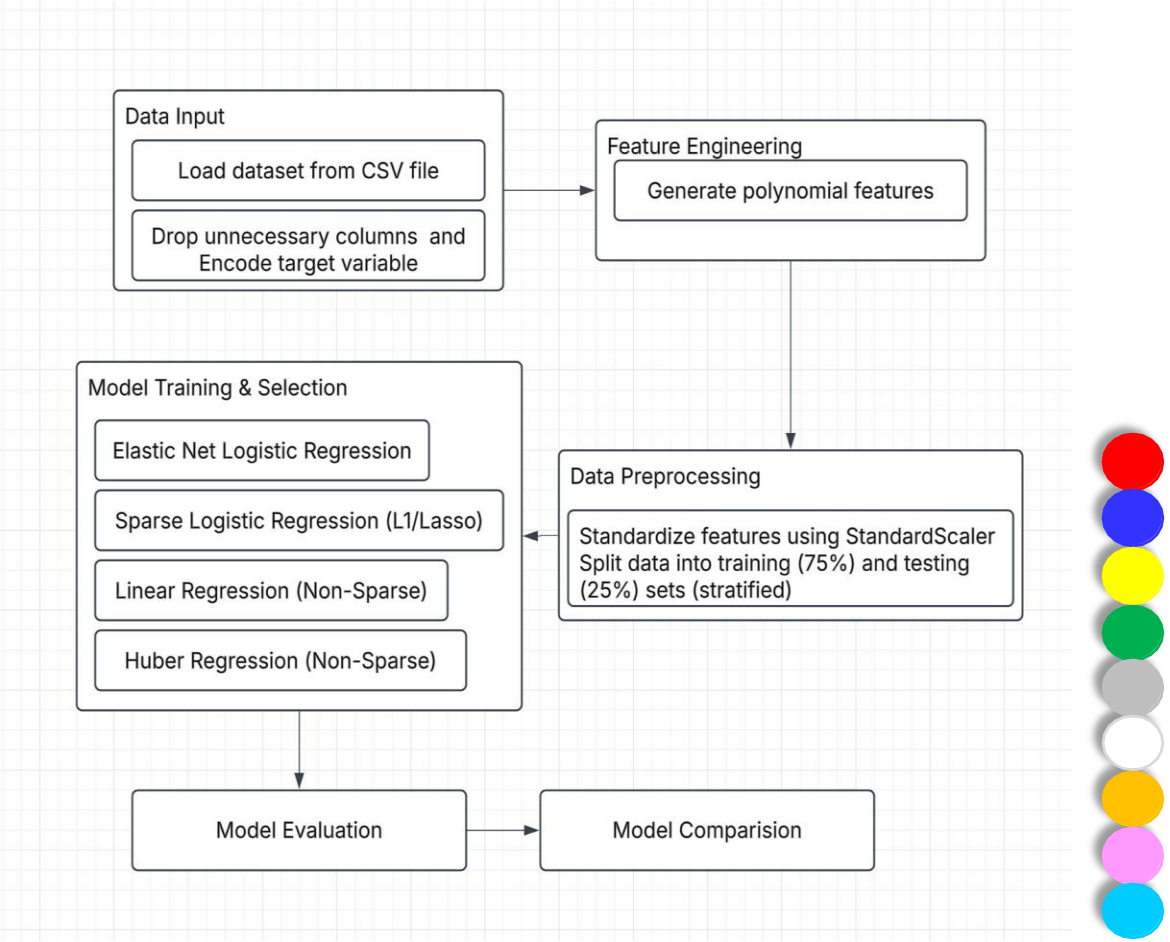
The proposed system enhances sparse regression by incorporating Huber loss for improved robustness and using iteratively re-weighted sparse regression for better noise adaptation. To maximize feature selection, the approach combines SelectKBest with polynomial feature extension. The model addresses outlier effects by increasing the number of coefficients through L1 penalties, Elastic Nets, and Huber loss robust regression. Grid search optimization is employed to determine the best combination of regularization parameters, ensuring optimal performance. Experimental results on the Breast Cancer Wisconsin dataset demonstrate the model's effectiveness, achieving 98.6% accuracy with Elastic-net sparse regression and 96.5% accuracy with Lasso sparse regression.

TEAM 10

Proposed System – Block Diagrams

22AIE213

This methodology outlines a structured approach to machine learning model development. The process begins with data input, where a dataset is loaded from a CSV file, unnecessary columns are removed, and the target variable is encoded. Next, feature engineering is performed by generating polynomial features. In the data preprocessing step, features are standardized using StandardScaler, and the data is split into training (75%) and testing (25%) sets using stratified sampling. The model training & selection phase involves fitting various regression models, including Elastic Net Logistic Regression, Sparse Logistic Regression (L1/Lasso), Linear Regression, and Huber Regression. The trained models undergo evaluation, followed by comparison, to determine the best-performing model. This pipeline ensures systematic data processing, model selection, and performance assessment.

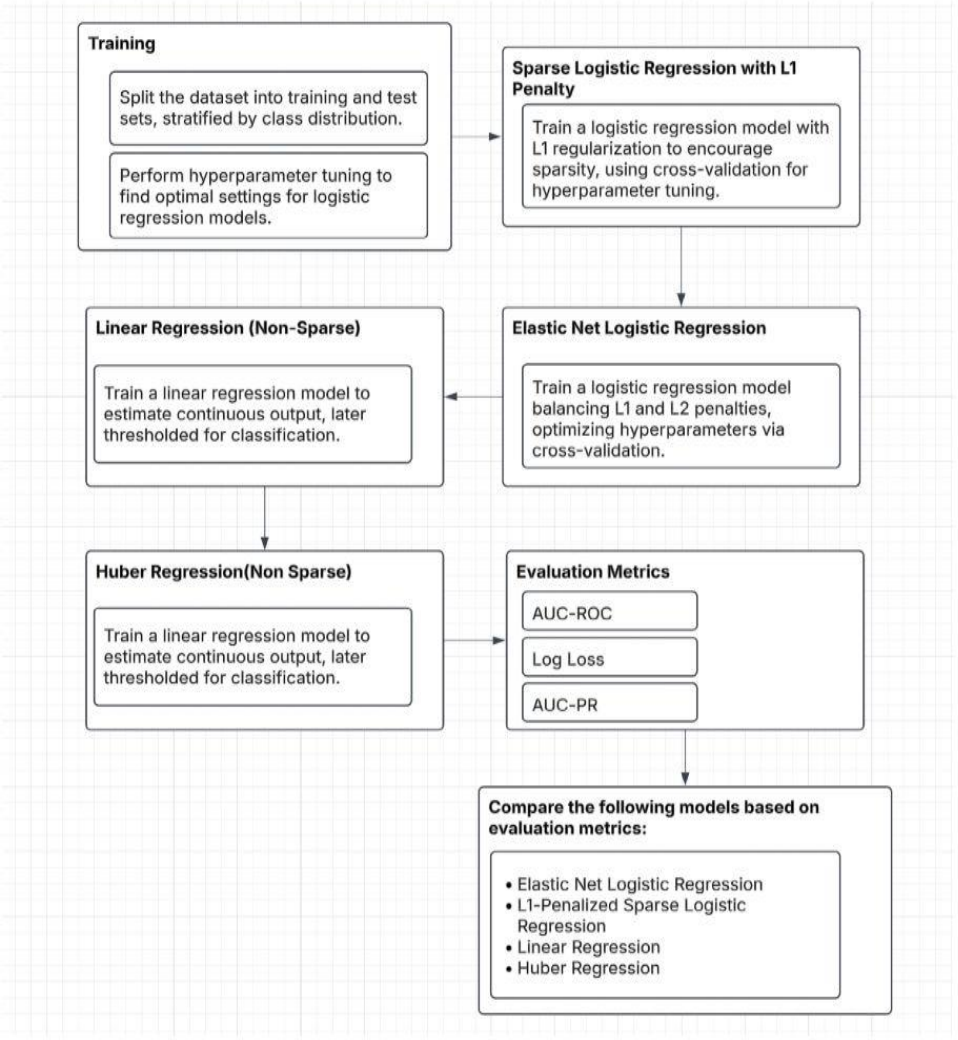


TEAM 10

Proposed System – Block Diagrams

22AIE213

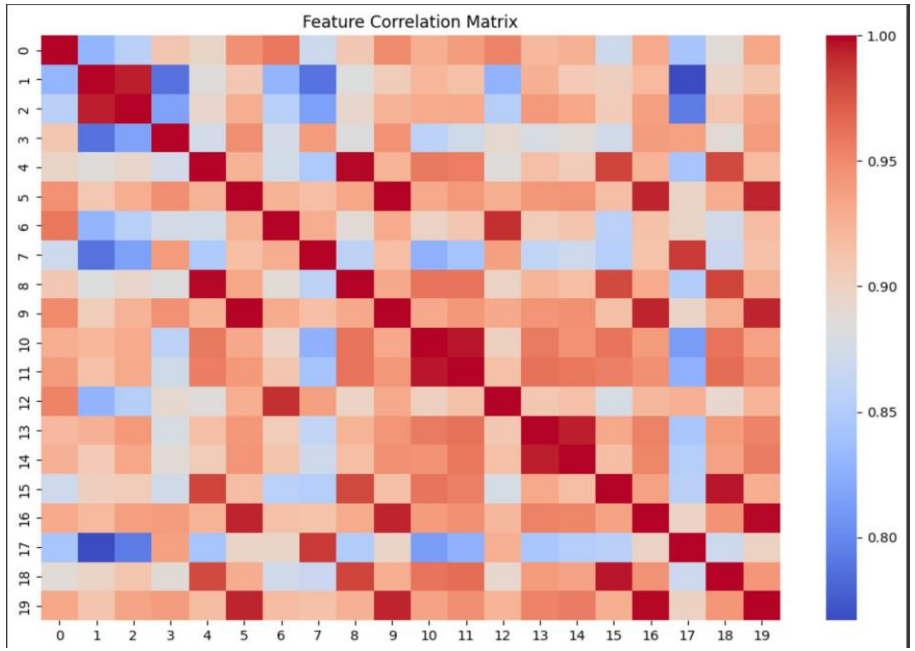
This methodology outlines a comparative analysis of regression models for classification. The dataset is first split into training and test sets, ensuring class stratification. Hyperparameter tuning is performed for logistic regression models. The study explores Sparse Logistic Regression with L1 regularization, Elastic Net Logistic Regression balancing L1 and L2 penalties, standard Linear Regression, and Huber Regression for robustness. Each model's performance is evaluated using AUC-ROC, Log Loss, and AUC-PR. Finally, the models are compared based on these metrics to determine the most effective approach for classification.



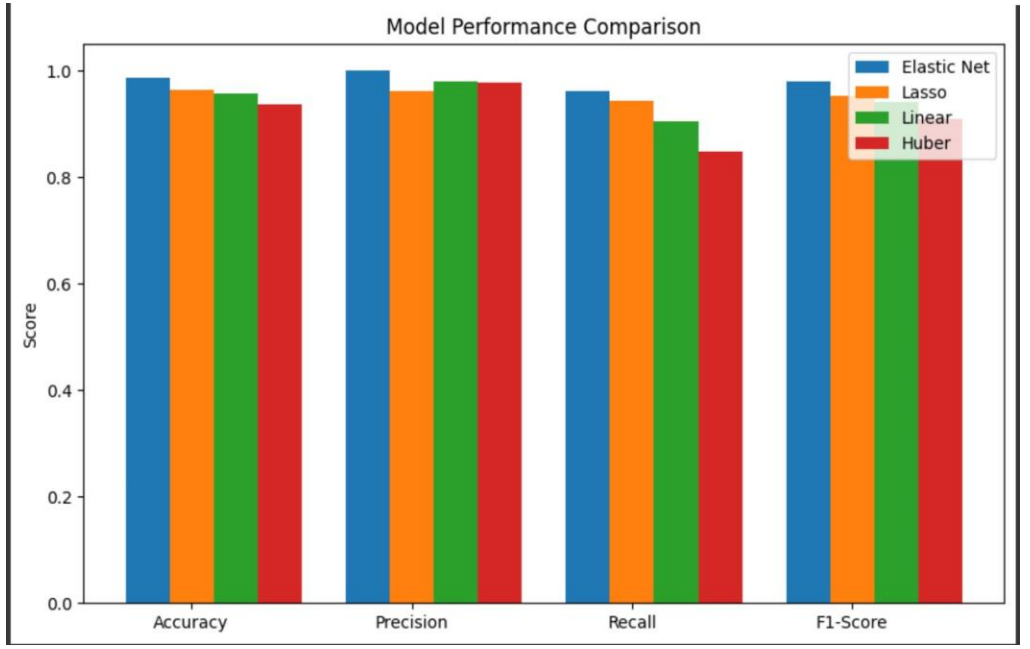
TEAM 10

Results and discussions

22AIE213



The image shows a Feature Correlation Matrix with values ranging from -0.90 to -0.80, indicating varying degrees of negative correlations between features. This matrix helps identify relationships between variables, which is crucial for model performance as highlighted in the Elastic Net and Lasso results.

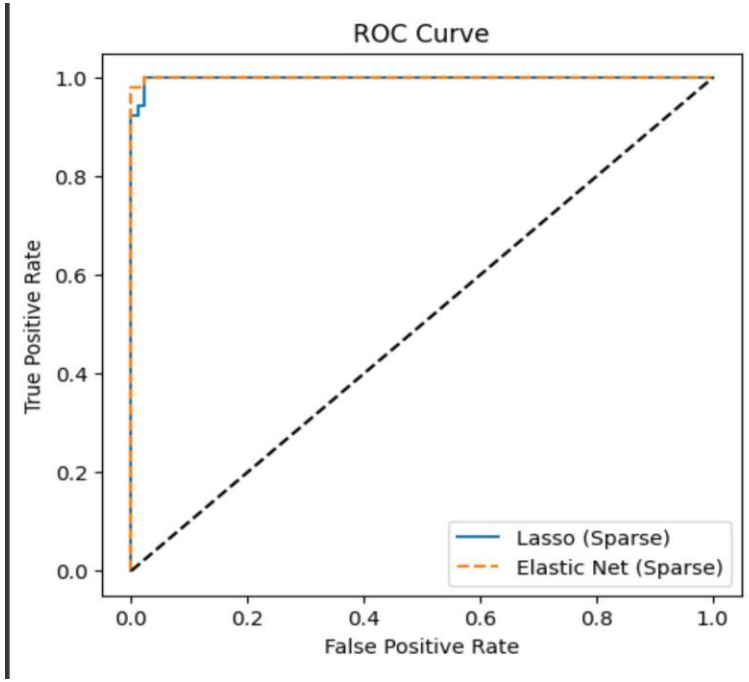


The image presents a Model Performance Comparison bar chart, displaying metrics like Accuracy, Precision, Recall, and F1-Score across different models (Elastic Net, Lasso, Linear, and Huber). The scores range from 0.0 to 1.0, highlighting the relative strengths of each algorithm.

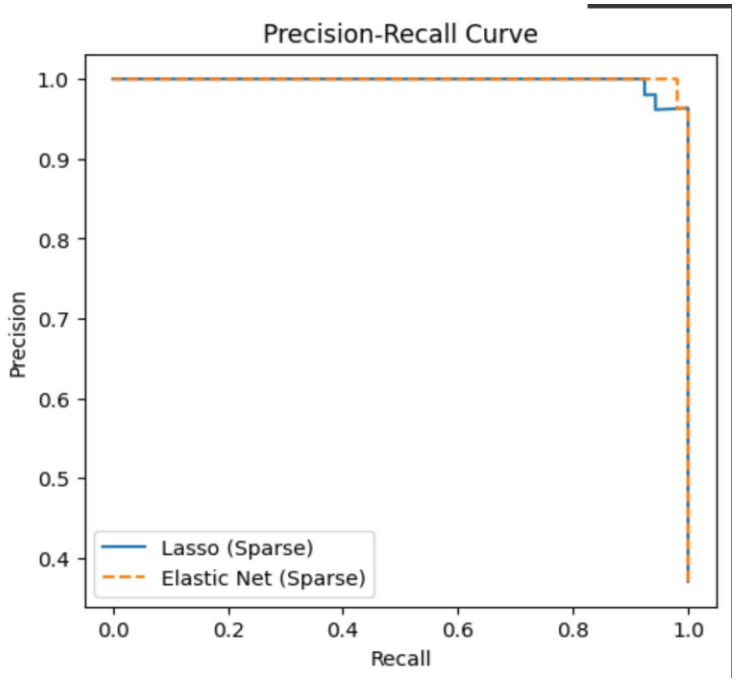
TEAM 10

Results and discussions

22AIE213



The image displays an ROC Curve comparing Lasso (Sparse) and Elastic Net (Sparse), with both models showing high True Positive Rates (0.6–0.8) and low False Positive Rates, indicating strong classification performance. Their curves closely overlap, suggesting near-identical effectiveness in distinguishing classes.



The image shows a Precision-Recall Curve comparing Lasso (Sparse) and Elastic Net (Sparse), with both models maintaining high precision (0.6–0.7) across varying recall levels. The tight overlap between the curves indicates similarly strong performance in balancing precision and recall, as reflected in their near-identical F1-scores.

TEAM 10

Results and discussions

22AIE213

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC	Log Loss
Elastic Net (Sparse)	0.9860	1.0000	0.9623	0.9808	0.9996	0.0618
Lasso (Sparse)	0.9650	0.9615	0.9434	0.9524	0.9985	0.0679
Linear Regression	0.9580	0.9796	0.9057	0.9412	0.9981	0.1836
Huber Regression	0.9371	0.9783	0.8491	0.9091	0.9975	0.1830

The table provides a comparative performance summary of four regression models, with Elastic Net (Sparse) achieving the highest accuracy (0.9860), perfect precision (1.0000), and the best F1-score (0.9808). Lasso (Sparse) follows closely, while Linear and Huber Regression show lower recall and higher log loss, indicating reduced robustness.

Qualitative Analysis with the State of the Art model:

The state of the art model proposed by Maria Jaenada where the author has created a hybrid model named as awDPD LASSO where the model has been attached to traditional LASSO to boost accuracy so the model's trial has been done on three different datasets one being Leukemia and others being Breast Cancer. The accuracy resulted for breast cancer dataset in the work was 94.6% but our model has achieved an accuracy of 98.6% for Sparse regression. Similarly, the study has also been done on the other Leukemia dataset. The accuracy we achieved was 88%. This study is an extension of the previous works done that has proved that sparse Regression technique with iterative reweighting has more accuracy than to Traditional methods.

TEAM 10

Significance and Contribution

22AIE213

This research enhances breast cancer classification using advanced sparse regression techniques, improving predictive accuracy in noisy and high-dimensional medical imaging data. By integrating Huber loss for robustness, iteratively re-weighted sparse regression for noise adaptation, and polynomial feature extension with SelectKBest for optimal feature selection, the proposed model outperforms existing approaches. The Elastic Net sparse regression model achieved the highest accuracy of 98.6%, demonstrating superior precision, recall, and probability calibration. Compared to non-sparse models like Linear and Huber Regression, the sparse models showed better classification performance, with Elastic Net emerging as the most effective. The study highlights the significance of optimizing regularization techniques for medical diagnostics, offering a scalable and interpretable solution for early disease detection. The findings contribute to medical imaging analytics by establishing a framework that balances predictive accuracy, feature selection, and computational efficiency, making it a valuable tool for real-world healthcare applications.

TEAM 10

Conclusion & Future Work

22AIE213

This research highlights the effectiveness of advanced sparse regression in improving breast cancer classification, especially in noisy data. The Elastic Net model achieved the highest accuracy (98.60%), perfect precision (1.0000), and strong recall (0.9623), with an F1-score of 0.9808 and near-perfect AUC-ROC (0.9996). Lasso followed at 96.50%, while non-sparse models like Linear (95.80%) and Huber Regression (93.71%) underperformed, particularly in recall and log loss.

This study establishes a robust framework for breast cancer classification using sparse regression, with potential for further enhancement. Expanding to larger, multi-modal datasets (e.g., mammograms, MRI, genomic data) could improve generalizability across cancer types. Advanced preprocessing, such as outlier detection or dimensionality reduction (e.g., PCA, t-SNE), may further refine feature selection in high-dimensional settings.

Elastic Net's superior performance highlights its balance of L1 and L2 penalties, reducing overfitting while maintaining robustness. Polynomial feature engineering and ANOVA F-statistic selection enhanced pattern detection in complex data. These findings demonstrate the potential of sparse regression for medical imaging and diagnostics, offering an interpretable, efficient, and scalable solution for early breast cancer detection.