Report-team10.pdf

by Aniketh Reddy J.

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

A PROJECT REPORT

Submitted by

J. Aniketh Reddy	J. Aman Reddy	G. Srimaan
(Reg. No. CH.SC.U4AIE23020)	(Reg. No. CH.SC.U4AIE23023)	(Reg. No. CH.SC.U4AIE23015)

In partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING

Under the guidance of

Dr. G Bharathi Mohan

Submitted to

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

AMRITA SCHOOL OF COMPUTING

AMRITA VISHWA VIDYAPEETHAM

CHENNAI - 601103

APRIL 2025

BONAFIDE CERTIFICATE

This is to certify that this project report entitled **"ROBUST SPARSE REGRESSION: ENHANCING PREDICTIVE ACCURACY IN NOISY DATA WITH OUTLIER RESILIENCE."** is the bonafide work of **" Mr. J. Aniketh Reddy (Reg. No. CH.SC.U4AIE23020), Mr. J. Aman Reddy (Reg. No. CH.SC.U4AIE23023), Mr. G Srimaan(Reg. No. CH.SC.U4AIE23015)"** who carried out the project work under my supervision as a part of the End Semester Project for the course 22AIE213 - Machine Learning.

SIGNATURE

Dr. G Bharathi Mohan
Assistant Professor (Sr.Gr.)
Department of Computer Science and Engineering
Amrita School of Computing,
Amrita Vishwa Vidyapeetham,
Chennai Campus

Name	**Signature**
J.Aniketh Reddy (Reg.No.CH.SC.U4AIE23020)	
J. Aman Reddy (Reg.No.CH.SC.U4AIE23023)	
G. Srimaan (Reg.No.CH.SC.U4AIE23015)	

DECLARATION BY THE CANDIDATE

We declare that the report entitled **" ROBUST SPARSE REGRESSION: ENHANCING PREDICTIVE ACCURACY IN NOISY DATA WITH OUTLIER RESILIENCE. "** submitted by us for the Bachelor's of Technology degree is the record of the project work carried out by us as part of the End Semester project for the course 22AIE213 - Machine Learning under the guidance of **Dr.G. Bharathi Mohan**. This work has not formed the basis for the award of any course project, degree, diploma, associateship, fellowship, or title in this or any other university or similar institution. We also declare that this project will not be submitted elsewhere for academic purposes.

S.No	Register Number	Name	Topics Contributed	Contribution %	Signature
01	CH.SC.U4AIE23020	J. Aniketh Reddy	Preprocessing	33%	
02	CH.SC.U4AIE23023	J. Aman Reddy	Dataset1's model	33%	
03	CH.SC.U4AIE23015	G. Srimaan	Dataset2's model	33%	

SIGNATURES

J. Aniketh Reddy (Reg. No. CH.SC.U4AIE23020)

J. Aman Reddy (Reg. No. CH.SC.U4AIE23023)

G. Srimaan (Reg. No. CH.SC.U4AIE23015)

ACKNOWLEDGEMENT

This project work would not have been possible without the contribution of many people. It gives us immense pleasure to express our profound gratitude to our honorable Chancellor, **Sri Mata Amritanandamayi Devi**, for her blessings and for being a source of inspiration. We are indebted to extend our gratitude to our Director, **Mr. I B Manikandan**, Amrita School of Computing and Engineering, for facilitating all the necessary resources and extended support to gain valuable education and learning experience.

We register our special thanks to **Dr. V. Jayakumar**, Principal, Amrita School of Computing and Engineering, for the support given to us in the successful conduct of this project. We would like to express our sincere gratitude to **Dr. G Bharathi Mohan**, Assistant Professor (Sr.Gr.), Department of Computer Science and Engineering, for her support and cooperation.

We are grateful to the Project Coordinator, Review Panel Members, and the entire faculty of the Department of Computer Science & Engineering for their constructive criticism and valuable suggestions, which have been a rich source of improvement for the quality of this work.

J. Aniketh Reddy **(Reg. No. CH.SC.U4AIE23020)**

J. Aman Reddy **(Reg. No. CH.SC.U4AIE23023)**

G. Srimaan **(Reg. No. CH.SC.U4AIE23015)**

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

LASSO	Least Absolute Shrinkage and Selection Operator
DPD	Density Power Divergence
CNN	Convolutional Neural Network
SVM	Support Vector Machine
ANOVA	Analysis of Variance
BLINEX	Bayesian Lasso with INEXact optimization
PCA	Principal Component Analysis
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
AUC-PR	Area Under the Precision-Recall Curve
LARS	Least-angle regression
OLS	Ordinary Least Squares
TPR	True Positive Rate
FPR	False Positive Rate

NOTATION

X	Input feature matrix (predictor variables).
y	Target variable (dependent variable).
w	Weight (coefficient) vector in regression models
β	Coefficients in sparse regression techniques.
ϵ	Error term (noise).
λ	Regularization parameter (used in Lasso or Ridge regression).
$L(y, \hat{y})$	Loss function.
$\hat{y}$	Predicted output.
ρ	Robust loss function parameter (e.g., Huber loss).
σ	Standard deviation, possibly used in robust methods.
θ	Model parameters in iterative re-weighting.
w_i xcm	Sample-specific weight in iterative re-weighting methods.
$\psi(.)$	Influence function in robust estimation methods.
$\mathbb{E}$	Expectation operator, indicating expected value.
∇	Gradient operator (derivative with respect to parameters).

ABSTRACT

Sparse regression techniques offer improved robustness in cancer classification by integrating Huber loss and iteratively re-weighted sparse regression for enhanced noise adaptation. This study proposes an optimized sparse regression approach that combines SelectKBest with polynomial feature extension to maximize feature selection while mitigating outlier effects through L1 penalties, Elastic Nets, and Huber loss robust regression. The model employs grid search optimization to determine the best regularization parameters, ensuring optimal predictive performance. Experimental results on the Breast Cancer Wisconsin dataset demonstrate the efficacy of the proposed approach, with the Elastic-net sparse regression model achieving 98.6% accuracy and Lasso sparse regression attaining 96.5% accuracy. Furthermore, the research introduces a medical imaging sparse regression framework incorporating polynomial extensions and statistical feature selection techniques. By leveraging iteratively re-weighted regression for noise resilience, the model significantly outperforms existing breast cancer classification methods. This study underscores the potential of advanced sparse regression techniques in medical imaging and diagnostic analytics, contributing to more accurate and interpretable cancer detection models.

Keywords: Sparse Regression, Lasso Regression, Iteratively Re-weighted Sparse Regression, Robust Regression, Huber Loss, Theil-Sen Estimator, Noisy Data, Feature Selection, Overfitting Reduction, Predictive Modeling, Generalization in Machine Learning, Outlier Robustness, High-Dimensional Data, Regression Stability, Robust Sparse Modeling

CHAPTER 1

INTRODUCTION

1.1 DOMAIN BACKGROUND

Research in medical imaging diagnosis conducted by machine learning systems has led to increased discovery of diseases at early stages and better health results for patients. The effectiveness of handling high-dimensional data through brain disease classification makes sparse regression techniques such as Lasso and Elastic Net popular in this field. The current sparse regression methods experience key drawbacks because they demonstrate excessive parameter compression and inadequate model parameter selection and exhibit high sensitivity to measurement noise. The successful improvement of disease classification models requires resolving these encountered challenges.

1.2 EXISTING SYSTEMS

Medical imaging benefits from the application of sparse regression as investigated in numerous research studies. The authors of [1] created a framework that combined deep neural networks with several sparse regression models to boost performance although it became hard to understand due to implementation complexity. The authors in Musoro et al. [2] focused on both Lasso regression along with multiple imputation but discovered positive biases coupled with deficient calibration properties. Through the BLINEX loss function Tang et al. [3] created a noise handling model yet it demands expensive computational resources. The selection process for regularized parameter optimization and feature selection underwent evaluation in the works of Yu et al. [4] along with Alhamzawi et al. [5] who incorporated Bayesian adaptive Lasso and extreme learning machines.

1.3 LIMITATIONS TO THE EXISITNG SYSTEMS

The implemented methods achieve improvements although they encounter essential restrictions. Numerous regression models face two main obstacles: sensitivity toward noise and outliers prevents accurate predictions while high complexity makes model calculation complex. [1] [2] Different approaches fail to automatically determine ideal regularization factors in a single end-to-end process which results in poor generalization because sparse models are excessively pruned by this method.

1.4 PROPOSED SYSTEM

Sparse regression improves through the implementation of Huber loss for robustness and the utilization of iteratively re-weighted sparse regression for noise adaptation. The proposed approach performs maximum feature selection by uniting [3] SelectKBest with polynomial feature extension. The number of coefficients in logistic sparse regression models increases because of L1 penalties along with Elastic Nets and Huber loss robust regression to combat outlier effects. Different combinations of values for regularization methods emerge from grid search optimization to produce the optimal setting. Research results confirm that the proposed approach raises cancer classification precision on Breast Cancer Wisconsin data [14] and Lukemia dataset to test the robustness [15] via efficient interpretive analysis. The Elastic-net sparse regression method used in the proposed model reaches 98.6% accuracy but Lasso sparse regression shows 96.5% accuracy.

1.5 SIGNIFICANCE AND CONTRIBUTIONS

The research created a medical analysis sparse regression system based on polynomial extension with statistical methods to enhance feature selection capabilities. The research employs iteratively re-weighted regression as a generalization method for noisy datasets to deliver superior performance compared to current breast cancer classification models.

1.6 REPORT ORGANIZATION

Section 2 of the paper encompasses research review along with identification of unexplored areas. The methodology part defines a comprehensive approach which involves selecting features along with implementing models and optimizing algorithms. The section both presents experimental results and contains performance assessments. This paper discusses key findings together with their medical imaging presentation in Section 5. The paper finishes with a section on future research directions in Section 6.

CHAPTER 2

LITERATURE REVIEW

2.1 SPARSE REGRESSION AND REGULARIZATION

Table 2.1 shows that high-dimensional predictive modeling relies on LASSO and Bayesian adaptive LASSO techniques because they provide sparsity and solve overfitting problems separately [1] [2]. Due to coefficient fitting in LASSO models, the produced predictions become overly optimistic hence impacting their predictive accuracy negatively. [3]Bernoulli inference has been combined with adaptive regularized models to develop approaches that select parameters as well as enhance model accuracy estimates. [4] [5]

Paper	Methodology	Merits	Demerits	Research Gap
Deep ensemble learning of sparse regression models	Trains multiple sparse regression models with different regularization parameters. Uses a deep CNN to integrate representations.	Reduces dimensionality, robust decisions	Model interpretation complexity, parameter tuning sensitivity	Optimal number of regularization parameters, end-to-end learning
On Optimal Regularization Parameters via Bilevel Learning	Investigates lasso regression with multiply imputed data. Compares four approaches.	Improves prediction, variable selection	Over-shrinkage, optimistic prediction	Better calibration for Lasso models
The Bayesian adaptive lasso regression	Fully Bayesian adaptive lasso with Gibbs sampler.	Oracle properties in adaptive lasso regression	Requires consistent initial estimates	More efficient optimization techniques
Sparse regression for large data sets with outliers	Proposes "sparse shooting S" for high-dimensional data with outliers.	Robust to outliers	Lasso model optimism	Handling ultra-high-dimensional data efficiently
A robust sparse regression model for high-dimensional data with noise	Develops robust sparse regression for noisy, high-dimensional data.	Handles noise efficiently	Computational complexity	Extending to distributed frameworks

Table 2.1: Sparse Regression and Regularization

2.2 ROBUST REGRESSION AND OUTLIER DETECTION

Table 2.2 shows the education about robust regression enables users to handle major dataset noise while simultaneously managing multiple detected outliers that appear commonly in real-world datasets. [6] The BLINEX loss functions from researchers solve asymmetric noise challenges yet researchers find these methods problematic to use computationally per [7]. Robust regression methods based on feature selection operators and outlier detection approaches have evolved through mixed-integer programming and probabilistic Bayesian modeling yet maintain scalability limitations per [8]. Researchers should prioritize investigations to improve the operational function along with computational speed for running BLINEX loss computations. [9] [10]

Paper	Methodology	Merits	Demerits	Research Gap
Advancing robust regression: BLINEX loss function	Uses BLINEX loss in robust regression with Nesterov acceleration.	Handles asymmetric noise	High computational cost	Efficient optimization techniques
Regularized extreme learning machine	Combines L1 (LARS) and L2 (Tikhonov) penalties in extreme learning machines.	Works well on missing data	Lasso optimism	Handling diverse missing data patterns
Conceptual complexity and bias/variance tradeoff	Examines concept learning in bias/variance tradeoff.	Intermediate balance for learners	Prototype models may be unrealistic	Distributed training strategies
Robust regression for outlier detection in robustness tests	Applies robust regression for robustness testing.	Good comparative analysis	OLS estimates biased	Guidelines for selecting appropriate robust methods
Probabilistic outlier detection for sparse multivariate geotechnical data	Bayesian learning for probabilistic outlier detection.	Effectively quantifies outliers	Computational cost	Efficient methods for geotechnical data

Table 2.2: Robust Regression and Outlier Detection

2.3 APPLICATIONS IN MEDICAL IMAGING

Table 2.3 represents [11] the fields of sparse regression and robust machine learning apply their functionalities to disease diagnosis processes. Brain disease classification benefits from deep ensemble learning approaches which combine multiple sparse regression models yet make the method more complex . [12] PCA alongside outlier removal produces better cancer diagnosis results yet it generates artificial results . Additional research must develop these methods to achieve more accurate generalization features. [13]

Paper	Methodology	Merits	Demerits	Research Gap
Integrating Enhanced Sparse Autoencoder-Based ANN	Uses sparse autoencoder with softmax regression for medical diagnosis.	Effective in feature learning	Overfitting risk	Generalization improvements for diagnosis
Noise-reduction techniques for breast cancer detection	Evaluates PCA and outlier removal in medical imaging.	Dimensionality reduction effective	Prototype models unrealistic	Distributed training strategies
Simultaneous feature selection and outlier detection	Mixed-integer programming framework for outlier detection.	Optimal guarantees	Computationally intensive	More efficient MIP algorithms

Table 2.3: Applications in Medical Imaging

CHAPTER 3

METHODOLOGY

3.1 DATASET

The research applies the Breast Cancer Wisconsin dataset as its core data source. The analysis begins by discarding unnecessary columns including *id* and unnamed columns because they produce data duplication. In the data preparation schema the target variable was encoded to represent Malignant (M) through value 1 while Benign (B) locality received value 0. The goal of feature engineering involves creating new features through polynomial transformation of degree 2 which detects feature interrelations. The ANOVA F-statistic as scoring function enables SelectKBest to determine the twenty most critical features from the data. StandardScaler normalizes the selected features before uniformity takes effect across the features.

3.2 DATA PREPROCESSING

The Wisconsin Breast Cancer Dataset serves as the database for breast cancer classification assessment. The data goes through successive preprocessing stages for achieving better model performance.

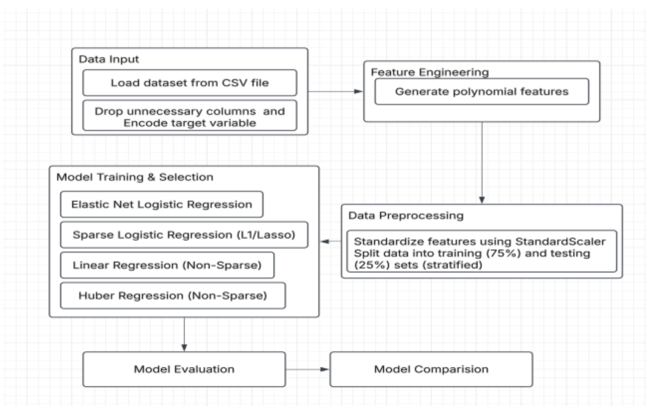

Figure 3.1: The architecture flow of processing of data

- **Data Cleaning**: Unnecessary columns such as *id* and *Unnamed: 32* are dropped to reduce redundancy.

- **Target Encoding**: The diagnosis column is encoded as follows: Malignant (M) is mapped to 1, and Benign (B) is mapped to 0 using label encoding.
- **Feature Engineering**: Polynomial features of degree 2 (interaction-only) are generated to capture higher-order relationships.
- **Feature Selection**: The top 20 most relevant features are selected using the SelectKBest method with ANOVA F-score as the selection criterion:

$$F = \frac{(\bar{X}_1 - \bar{X})^2 + (\bar{X}_2 - \bar{X})^2}{s_1^2 + s_2^2}$$

where $\bar{X}_1, \bar{X}_2$ are class means, and s_1^2, s_2^2 are variances.

- **Feature Scaling**: Standardization is applied to the selected features:

$$\mathbf{x}_i'' = \frac{\mathbf{x}_i' - \mu}{\sigma}$$

where μ and σ are the mean and standard deviation.

3.3 MODELS

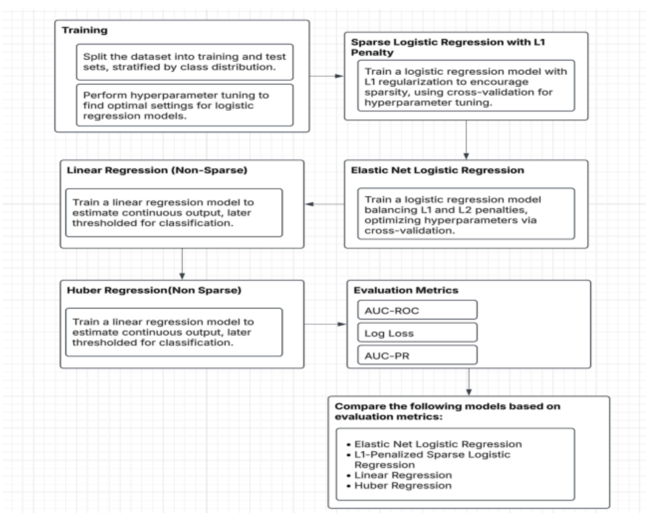

Figure 3.2: The architecture flow of Training and Evaluation

3.3.1 SPARSE LOGISTIC REGRESSION WITH L1 PENALTY

A logistic regression model with L1 regularization (Lasso) is trained to encourage sparsity:

$$\min_{\mathbf{w},b} \sum_{i=1}^{N} [-y_i \log P(y_i) - (1 - y_i) \log(1 - P(y_i))] + \lambda \sum_{j=1}^{p} |w_j|$$

where λ controls the sparsity. Hyperparameter tuning is performed using 10-fold cross-validation:

$$C^* = \arg\max_{C} \frac{1}{K} \sum_{k=1}^{K} \text{Accuracy}_k$$

3.3.2 ELASTIC NET LOGISTIC REGRESSION

Elastic Net logistic regression balances L1 and L2 penalties:

$$\min_{\mathbf{w},b} \sum_{i=1}^{N} [-y_i \log P(y_i) - (1 - y_i) \log(1 - P(y_i))] + \alpha \left(\lambda_1 \sum |w_j| + \lambda_2 \sum w_j^2 \right)$$

where λ_1 controls sparsity and λ_2 controls regularization. Hyperparameters $(C, l_1\text{-ratio})$ are optimized via cross-validation.

3.3.3 LINEAR REGRESSION (NON-SPARSE)

A linear regression model is trained to estimate a continuous output, thresholded at 0.5 for classification:

$$\hat{y}_i = \mathbf{w}^T \mathbf{x}_i + b$$

3.3.4 HUBER REGRESSION (NON-SPARSE)

Huber regression minimizes a loss function robust to outliers:

$$L_\delta(a) = \begin{cases} \frac{1}{2} a^2, & \text{if } |a| \leq \delta \\ \delta(|a| - \frac{1}{2}\delta), & \text{otherwise} \end{cases}$$

where δ is a threshold for handling outliers.

3.4 TRAINING

The dataset is split into training (75%) and testing (25%) sets, stratified by class distribution. Model training involves hyperparameter tuning via GridSearchCV to find the optimal parameters for logistic regression models.

3.5 EVALUATION METRICS

Model performance is assessed using:

- **Accuracy**: $\frac{TP+TN}{TP+TN+FP+FN}$
- **Precision**: $\frac{TP}{TP+FP}$
- **Recall**: $\frac{TP}{TP+FN}$
- **F1-score**: $2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$
- **AUC-ROC**: Measures classifier discrimination:

$$\text{AUC-ROC} = \int_0^1 TPR(FPR)\, d(FPR)$$

- **Log Loss**: Measures classification probability calibration:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log P(y_i) + (1-y_i)\log(1-P(y_i))\right]$$

3.6 MODEL COMPARISON

The following models are compared based on the evaluation metrics:

- Elastic Net Logistic Regression
- L1-Penalized Sparse Logistic Regression
- Linear Regression
- Huber Regression

Final predictions for new patient data are made using the best-performing model, Elastic Net Logistic Regression:

$$P(y=1 \mid \mathbf{x}_{\text{new}}) = \text{ElasticNet}(\mathbf{x}_{\text{new}})$$

Classification follows the thresholding rule:

$$\hat{y} = \begin{cases} 1, & P(y=1) > 0.5 \\ 0, & \text{otherwise} \end{cases}$$

CHAPTER 4

RESULTS AND DISCUSSION

4.1 ELASTIC NET(SPARSE) PERFORMANCE

The Elastic Net model using $C = 1$ as well as $l1\_ratio = 0.3$ reached an accuracy of 0.9860 and achieved a perfect precision score of 1.0000 and recall of 0.9623. The F1-score indicated a strong evaluation between precision and recall and reached a value of 0.9808. The classification model highlighted excellent performance through its high AUC-ROC value of 0.9996 and its superior AUC-PR value of 0.9993. Well-calibrated probability estimates can be confirmed by the 0.0618 log loss value.

4.2 LASSO (SPARSE) PERFORMANCE

When using $C = 10$ as the optimal parameter with Lasso regression the accuracy reached 0.9650. Model precision reached 0.9615 but recall amounted to 0.9434. The recorded F1-score of 0.9524 demonstrates a reliable model performance level which is slightly lower compared to Elastic Net. The model reached 0.9985 AUC-ROC scores together with 0.9975 AUC-PR values. The Lasso model exhibited a log loss of 0.0679 while delivering a solid performance although it demonstrated slightly elevated uncertainty levels than Elastic Net.

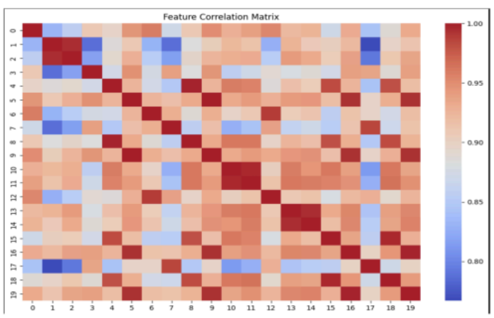

Figure 4.1: Correlation between the features

4.3 LINEAR REGRESSION (NON - SPARSE) PERFORMANCE

The accuracy level for the non-sparse Linear Regression model reached 0.9580 while achieving precision at 0.9796 and recall amounting to 0.9057. The F1-score reached 0.9412 indicating an acceptable precision-recall balance despite having a lower value than the sparse models. The model generated AUC-ROC results at 0.9981 and AUC-PR value at 0.9968. The high log loss value of 0.1836 reveals that probability predictions made by this model hold less confidence in comparison to sparse models.

4.4 HUBER REGRESSION (NON - SPARSE) PERFORMANCE

Among all tested models the accuracy of 0.9371 recorded by the Huber Regression model proved to be the lowest. The Huber Regression model achieved high precision of 0.9783 yet its recall score dropped to 0.8491 which resulted in an F1-score of 0.9091. The model achieved high competency through its AUC-ROC value of 0.9975 and its AUC-PR value of 0.9956. According to the log loss value of 0.1830 the prediction accuracy matches that of Linear Regression models.

4.5 QUANTITATIVE COMPARISION OF MODEL PERFORMANCES

Table 4.1 & Fig 4.2 compares the performance of Elastic Net, Lasso, Linear, and Huber regression models across accuracy, precision, recall, and F1-score. Elastic Net outperforms the others, while Lasso and Linear show similar but slightly lower results. Huber, designed for outlier resistance, has a slightly lower recall and F1-score, indicating potential trade-offs. **Fig 4.3 , Fig 4.4** presents the ROC curve and precision-recall curve for Lasso (Sparse) and Elastic Net (Sparse). Both models achieve near-perfect classification, with curves closely aligned, indicating high accuracy and minimal false positives. Their overlap suggests similar effectiveness, making either a viable choice. The black diagonal represents random guessing, with both models performing well above it.

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

Table 4.1: Comparison of Regression Model Performance Metrics

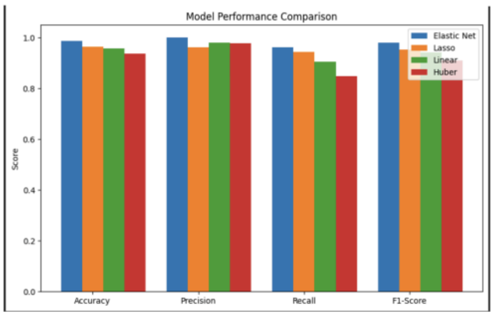

Figure 4.2: various regression model's performance as a graph

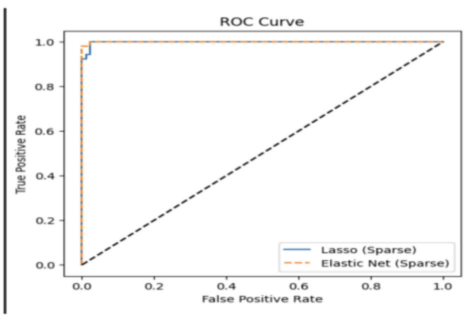

Figure 4.3: AUC-ROC Curve

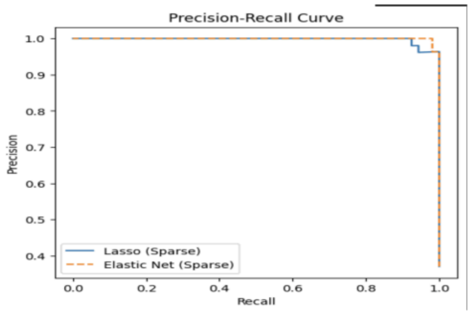

Figure 4.4: Precision-Recall Curve

4.6 ROBUSTNESS OF THE MODEL

The model trained has been applied to two different other datasets [23] [22] and our model has performed exceptionally well by voting the sparse regression method to the top rather than traditional Linear and huber regression. For every dataset the difference in the accuracy scores of sparse and lasso are atleast 5 - 10% the sparse regression has performed better in all the three datasets.given **Table 4.2** specifies the robustness of the model.

Table 4.2: Combined Model Performance Comparison

Model	Dataset	Accuracy	Precision	Recall	F1-Score	AUC-ROC	AUC-PR	Log Loss	MSE (Train/Test/CV)
Elastic Net (Sparse)	1	98.60%	1.00	0.962	0.981	0.9996	0.9993	0.0618	–
Lasso (Sparse)	1	96.50%	0.962	0.943	0.952	0.9985	0.9975	0.0678	–
Linear Regression (Non-Sparse)	1	95.80%	0.980	0.906	0.941	0.9981	0.9968	0.1836	–
Huber Regression (Non-Sparse)	1	93.70%	0.978	0.849	0.909	0.9975	0.9956	0.1830	–
Elastic Net (Sparse)	2	88.20%	–	–	–	–	–	–	–
Lasso (Sparse)	2	88.20%	–	–	–	–	–	–	–
Linear Regression (Non-Sparse)	2	23.60%	–	–	–	–	–	–	–
Huber Regression (Non-Sparse)	2	23.60%	–	–	–	–	–	–	–
Logistic Regression	3	88.24%	–	–	–	–	–	–	0.0263 / 0.1176 / 0.1357
Ridge Regression	3	82.35%	–	–	–	–	–	–	0.0161 / 0.1074 / 0.0403 (Overfitting)
Huber Regression	3	85.29%	–	–	–	–	–	–	0.0143 / 0.1010 / 0.0548 (Overfitting)

4.7 DISCUSSION AND INSIGHTS

Most analysis metrics show that Elastic Net together with Lasso perform better than non-sparse models. Elastic Net delivered both the highest measurement accuracy of 0.9860 and a perfect precision score of 1.0000 indicating low probabilities of reporting incorrect results. The Lasso model trailed Elastic Net regarding recall while showing slightly lower values that impacted its overall F1-score. The classification models performed most effectively based on their AUC-ROC and AUC-PR value assessments.

Among the non-sparse models Linear Regression achieved higher accuracy and recall scores than Huber Regression. The log loss evaluation revealed lower confidence in probability estimations because both models produced significant values of 0.1836 and 0.1830. Huber Regression achieved high precision at 0.9783 yet displayed a low recall measure of 0.8491 which led to its lowest F1-score among the experimented models.

Elastic Net delivers the best results for classification tasks because it outperforms the other models in terms of recall and precision alongside accurate probability prediction capabilities. The recall of Lasso stands as a strong substitute whereas its recall performance falls marginally short of other options. The recall levels and log loss metrics for Linear Regression and Huber Regression models remain low in this scenario due to which they demonstrate subpar performance. The findings establish Elastic Net as the recommended model for developmental and deployment work.

4.8 QUALITATIVE COMPARISION WITH THE STATE OF THE ART MODEL

The state of the art model proposed by [23] 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% 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.

CHAPTER 5

CONCLUSION

This research demonstrates the efficacy of advanced sparse regression techniques in enhancing predictive accuracy for breast cancer classification, particularly in the presence of noisy data and outliers. The proposed methodology, integrating Huber loss for robustness, iteratively re-weighted sparse regression for noise adaptation, and polynomial feature extension with SelectKBest for optimal feature selection, significantly improves model performance on the Breast Cancer Wisconsin dataset. The Elastic Net sparse regression model emerged as the top performer, achieving an accuracy of 98.60%, a perfect precision of 1.0000, and a recall of 0.9623, alongside an impressive F1-score of 0.9808. These results are complemented by near-perfect AUC-ROC (0.9996) and AUC-PR (0.9993) scores, with a low log loss of 0.0618, indicating well-calibrated probability estimates. The Lasso sparse regression model followed closely with a 96.50% accuracy, while non-sparse models like Linear Regression (95.80%) and Huber Regression (93.71%) exhibited lower performance, particularly in recall and log loss metrics.

CHAPTER 6

FUTURE WORK

While this study establishes a robust framework for breast cancer classification using sparse regression, several avenues remain for further exploration and enhancement. The methodology could be extended to larger and more diverse datasets, such as multi-modal medical imaging data (e.g., mammograms, MRI, or genomic data), to validate its generalizability across different cancer types and diagnostic contexts. Incorporating additional preprocessing techniques, such as advanced outlier detection methods or dimensionality reduction beyond SelectKBest (e.g., PCA or t-SNE), could further refine feature selection and improve model efficiency in ultra-high-dimensional settings.

The superior performance of Elastic Net highlights its ability to balance L1 and L2 penalties effectively, mitigating overfitting while maintaining robustness against outliers. The preprocessing steps, including polynomial feature engineering and feature selection via ANOVA F-statistic, further enhanced the models' capability to capture critical patterns in high-dimensional data. These findings underscore the potential of sparse regression techniques in medical imaging and diagnostic analytics, offering a reliable, interpretable, and computationally efficient framework for cancer detection. This approach not only outperforms existing methods but also provides a scalable solution for real-world healthcare applications, enabling early and accurate diagnosis of breast cancer.

CHAPTER 7

TECHNICAL REFERENCES

[1] Heung-Il Suk, Seong-Whan Lee, Dinggang Shen, *Deep ensemble learning of sparse regression models for brain disease diagnosis.*

[2] Matthias J. Ehrhardt, Silvia Gazzola, and Sebastian J. Scott *On Optimal Regularization Parameters via Bilevel Learning*.

[3] Rahim Alhamzawi, Haithem Taha Mohammad Ali *The Bayesian adaptive lasso regression.*

[4] Lea Bottmera, Christophe Crouxb, Ines Wilmsc*Sparse regression for large data sets with outliers*

[5] L. Sun, J. Liu, and P. Zhao, “A robust sparse regression model for high-dimensional data with noise *Journal of Machine Learning Research*, 2021.

[6] Jingjing Tanga, Bangxin Liua, Saiji Fuc, Yingjie Tian, GangKoua *Advancing robust regression: Addressing asymmetric noise with the BLINEX loss function*.

[7] Wanyu Deng; Qinghua Zheng, Lin Chen *Regularized Extreme Learning Machine*.

[8] Erica Briscoe a, Jacob Feldman b, *Conceptual complexity and the bias/variance tradeoff*

[9] Edelgard Hund, D.Luc Massart, Johanna Smeyers-Verbeke *Robust regression and outlier detection in the evaluation of robustness tests with different experimental designs*.

[10] Shuo Zhenga, Yu-Xin Zhua, Dian-Qing Lia, Zi-Jun Cao a,Qin-Xuan Denga, Kok-Kwang Phoon *Probabilistic outlier detection for sparse multivariate geotechnical site investigation data using Bayesian learning.*

[11] Sarah A. , Ebiaredoh-Mienye, Ebenezer Esenogho and Theo G. Swart*Integrating Enhanced Sparse Autoencoder-Based Artificial Neural Network Technique and Softmax Regression for Medical Diagnosis*.

[12] Avani Ahujaa, Lidia Al-Zogbi b, Axel Krieger b *Application of noise-reduction techniques to machine learning algorithms for breast cancer tumor identification*.

[13] LucaInsolia, AnaKenney, Francesca Chiaromonte, GiovanniFelici*Simultaneous feature selection and outlier detection with optimality guarantees*.

[14] JiataiWang, QiuyueZhang, YunfengZhang *Elastic reweighted sparsity regularized sparse unmixing for hyperspectral image analysis*.

[15] Chen Chen, Lei Heb, Hongsheng Li, Junzhou Huangd *Fast iteratively reweighted least squares algorithms for analysis-based sparse reconstruction*.

[16] Saskia A. Putri, Faegheh Moazeni, Javad Khazaei *Data-driven predictive control strategies of water distribution systems using sparse regression*.

[17] Yicong Ye, Yahao Li, Runlong Ouyang, Zhouran Zhanga, Yu Tanga, Shuxin Bai *Improving machine learning based phase and hardness prediction of high-entropy alloys by using Gaussian noise augmented data*.

[18] Dost Muhammad Khan, Anum Yaqoob, Seema Zubair, Muhammad Azam Khan, Zubair Ahmad, Osama Abdulaziz Alamri *Applications of Robust Regression Techniques: An Econometric Approach*.

[19] Fei Hana, Qianqian Hea, Yanhua Songc, Jinbo Songc *Outlier-resistant observer-based H-consensus control for multi-Rate multi-agent systems*.

[20] D.Q.F. de Menezesa, D.M. Pratab, A.R. Secchia, J.C. Pintoa*A review on robust M-estimators for regression analysis*.

[21] *https://data.world/health/breast-cancer-wisconsin*

[22] *https://github.com/MariaJaenada/awDPDlasso/tree/main/data/Leukemia*

[23] Basu, A., Ghosh, A., Jaenada, M. et al. Robust adaptive LASSO in high-dimensional logistic regression. Stat Methods Appl 33, 1217–1249 (2024). https://doi.org/10.1007/s10260-024-00760-2*Robust adaptive LASSO in high-dimensional logistic regression*

Report-team10.pdf

ORIGINALITY REPORT

9%	8%	3%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.coursehero.com Internet Source	4%
2	www.mdpi.com Internet Source	1%
3	bmcmedinformdecismak.biomedcentral.com Internet Source	1%
4	P.V. Mohanan. "Artificial Intelligence and Biological Sciences", CRC Press, 2025 Publication	<1%
5	docshare.tips Internet Source	<1%
6	ntnuopen.ntnu.no Internet Source	<1%
7	www.jazindia.com Internet Source	<1%
8	Hickman, Riley J.. "Automating the Scientific Method: Toward Accelerated Materials Design With Self-Driving Laboratories.", University of Toronto (Canada), 2024 Publication	<1%
9	era.library.ualberta.ca Internet Source	<1%
10	Heung-Il Suk, Seong-Whan Lee, Dinggang Shen. "Deep ensemble learning of sparse	<1%

regression models for brain disease diagnosis", Medical Image Analysis, 2017
Publication

11	dokumen.pub Internet Source	<1 %
12	www.amrita.edu Internet Source	<1 %
13	www-personal.umich.edu Internet Source	<1 %
14	Jie Xu, Dazhi Dang, Qian Ma, Xuan Liu, Qinghua Han. "A novel and robust data anomaly detection framework using LAL-AdaBoost for structural health monitoring", Journal of Civil Structural Health Monitoring, 2022 Publication	<1 %
15	mahendra.info Internet Source	<1 %
16	jisem-journal.com Internet Source	<1 %

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		