Proceedings of the 1st International Conference on Statistical Sciences and Data Analytics

STATISTICAL FOUNDATIONS FOR TRUSTWORTHY ARTIFICIAL INTELLIGENCE: ENHANCING DECISION-MAKING THROUGH ROBUST INFERENCE AND UNCERTAINTY QUANTIFICATION

Emmanuel Lekwot[1*] and Enoch Yabkwa Yanshak[2]

[1,2]Department of Statistics, Ahmadu Bello University, Zaria, Nigeria

Corresponding author E-mail: lekwote@gmail.com

ABSTRACT

This paper explained why using advanced statistical methods is essential for making Artificial Intelligence systems more reliable, understandable, and fair. As AI becomes more common in important areas like healthcare, finance, education, and self-driving cars, it's important to make sure these systems are trustworthy and accurate. Even though AI has made big progress, many systems still can't properly handle uncertainty, understand cause-and-effect relationships, or make well-reasoned decisions. We focused on four key areas of statistics: understanding uncertainty, figuring out what causes what, simplifying models to avoid overfitting, and choosing actions based on risks and outcomes. We looked at real-world examples from healthcare and finance to show how using statistics in AI can make systems safer, more efficient, and easier to explain. These examples helped show why statistics is important for solving practical problems with AI. Besides exploring these tools, we also discussed some current challenges—like making these methods work on large-scale data, balancing accuracy with clarity, and ensuring fairness in decision-making. Finally, we suggested that combining AI with strong statistical methods is the best way to build smart systems that people can trust, especially in areas that affect people's lives.

KEYWORDS: Trustworthy Artificial Intelligence, Statistical Regularization, Uncertainty Quantification, High-Dimensional Data Analysis

I. INTRODUCTION

Artificial Intelligence (AI) has made major advances in recent years. Today, AI systems can recognize images, understand human language, and even help guide self-driving cars—things that were very hard or impossible not long ago. These improvements are mainly due to powerful techniques like machine learning (ML) and deep learning, which help computers learn patterns from data. Because of this, AI is now being used in many important areas, such as healthcare , finance and government (Esteva *et al.*, 2019; Chen *et al.*, 2021and Alvarez-Melis & Jaakkola, 2017).

But even with all these improvements, AI still has serious problems—especially when used to make important decisions. One big issue is that many AI systems act like "black boxes." That means they give answers or predictions, but don't explain how they got them or how sure they are. This can be risky in

situations like diagnosing a disease or deciding who qualifies for a loan. If we don't understand how the system works, we might make the wrong decisions, treat people unfairly, or lose trust in the technology.

This is where statistics comes in. Statistics gives us a set of tools to better understand and improve AI. It helps us deal with uncertainty, understand cause-and-effect relationships, and make better decisions with limited information. For example, Bayesian statistics allows us to include prior knowledge and calculate how confident we should be in a prediction. Causal inference helps us tell the difference between things that are just related and things that actually cause each other—a key issue in fields like public health and government.

In recent years, Artificial Intelligence (AI), especially machine learning (ML), has made big improvements in recognizing images, understanding language, and predicting patterns. However, many of these AI models act like black boxes—they make predictions, but it's hard to know why or how confident they are in those predictions. This creates problems in sensitive areas like healthcare, finance, and criminal justice, where trust, fairness, and accuracy are critical.

Statistics is the field that helps us understand uncertainty, make predictions, and draw conclusions from data. Unlike traditional AI models that just "fit the data," statistical methods help us: Measure how confident we are in predictions; Understand cause and effect, not just correlation; Control overfitting by keeping models simple and general; Make optimal decisions, even when data is uncertain or limited.

Bayesian models help us not only predict outcomes but also say how certain those predictions are. For example, Bayesian neural networks use probability to express uncertainty in weights and predictions (Gal & Ghahramani, 2016). Recent works like Yang *et al.* (2023) show that using Bayesian methods improves safety and trust in AI systems, especially in fields like autonomous driving.

Sometimes, AI finds patterns that look like causes but aren't. Causal inference helps us know what actually causes what, which is very important in things like medicine or policy-making (Pearl, 2009). Modern AI research now includes causal discovery algorithms and counterfactual models, which try to answer questions like, "What would have happened if...?" Recent studies such as Du *et al.* (2022) focus on integrating causal reasoning into machine learning to improve fairness and transparency.

Graphical models like Bayesian networks show how different variables are connected using visual graphs. These help users and researchers understand how a model makes decisions (Koller & Friedman, 2009). Updates by Zhao and Jin (2024) present lightweight probabilistic graphical models that can be used in explainable AI systems for real-time applications.

Modern AI tools now include interpretation techniques like: SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) tells how each feature contributes to a prediction and LIME (Local Interpretable Model-Agnostic Explanations) explains what happens near a single prediction point. These tools are based on statistical ideas like linear regression and variable importance (Lundberg & Lee, 2017; Ribeiro *et al.*, 2016). Xu *et al.*, (2025) provide more efficient model explanations using hybrid methods that blend LIME and SHAP with causal graphs.

Double Machine Learning (DML) (Chernozhukov *et al.*, 2018) is used to estimate treatment effects in the presence of high-dimensional covariates, leveraging orthogonal score functions. Athey and Imbens, (2016) used DML to evaluate the effect of job training programs on wages using census data. Causal forests provide heterogeneous treatment effect estimates across demographic subgroups

In radiology, BNNs have been employed to flag ambiguous diagnoses such as uncertain tumor margins or pulmonary nodules (Leibig *et al.*, 2017). Uncertainty heatmaps have been shown to correlate with clinical diagnostic difficulty and support triage systems.

Financial decisions often hinge on managing tail risks. Quantile regression provides a robust alternative to mean-based models by estimating conditional quantiles. Taylor (2019) applied quantile regression forests to high-frequency trading data to predict 1% and 5% tail losses. This allowed for dynamic portfolio adjustment based on volatility clustering. Time-series plots of predicted quantiles vs. realized returns highlight model accuracy and volatility shocks.

Despite AI's remarkable achievements, most models act as "black boxes," offering predictions without clarity or confidence. This lack of transparency undermines trust, fairness, and accountability—especially in critical applications. Additionally, issues like overfitting, hidden biases, and poor generalization persist, with limited use of statistical tools for uncertainty, causality, and validation. There is a clear need to embed robust statistical methods into AI to make it more trustworthy and reliable.

2. METHODOLOGY

This section explains four main ways that statistics can make artificial intelligence (AI) systems smarter, more trustworthy, and safer to use in real-life situations.

Uncertainty Quantification

AI systems often give predictions (like a medical diagnosis or loan approval), but they usually don't say how confident they are. This is risky, especially in areas where mistakes can have serious consequences. Statistics lets us measure and show uncertainty. For example, Bayesian methods use probability to tell us how confident we are in a prediction. A technique called Monte Carlo dropout is one way to add this to deep learning, helping us know when the model is unsure.

$$p(y^*|x^*, D) = \int P(y^*|x^*, \theta)\, P(\theta|D) d\theta \quad (1)$$

This equation is about predicting an outcome (like a medical diagnosis y* for a new situation or input x* (like a patient's test results), based on what the model has learned from data D.

- θ: The internal settings or parameters of the model.
- $P(\theta|D)$: What we believe about the model's parameters after seeing the data.
- $P(y^*|x^*, \theta)$: What the model predicts for the new input, given certain parameters.
- The equation is saying: "Average all the possible predictions, weighted by how likely each setting of the model is."

In clinical settings, uncertainty quantification is essential. Bayesian Neural Networks (BNNs) provide a principled framework for uncertainty modeling by learning a distribution over weights rather than point

estimates (Gal & Ghahramani, 2016). This enables both aleatoric (data-inherent) and epistemic (model-based) uncertainty to be quantified.

$$p(y = disease|x, D) = \int P(y|x, \theta)\, P(\theta|D) d\theta \quad (2)$$

Total uncertainty is decomposed into:

$$Var(y) = E_{P(\theta|D)}[Var(y|x, \theta)] + Var_{p(\theta|D)}[E(y|x, \theta)] \quad (3)$$

where: $\mathrm{Var}(y)$ is the total uncertainty, $\mathrm{E}_{\mathrm{P}(\theta|\mathrm{D})}[\mathrm{Var}(\mathrm{y}|\mathrm{x}, \theta)]$ is aleatoric (data-inherent) uncertainty and $\mathrm{Var}_{\mathrm{p}(\theta|\mathrm{D})}[\mathrm{E}(\mathrm{y}|\mathrm{x}, \theta)]$ is the epistemic (model-based) uncertainty.
If a system is not confident in a prediction, it can alert a human to double-check. This is especially helpful in healthcare or legal systems.

Causal Inference

Understanding cause-effect relationships is vital for generalization and fairness. Causal inference frameworks, such as Pearl's structural causal models, enable AI systems to distinguish between mere associations and true causal links. A key component is the do-calculus

$$\mathrm{P}\big(\mathrm{Y} \mid \mathrm{do}(\mathrm{X} = \mathrm{x})\big) \neq \mathrm{P}(\mathrm{Y} \mid \mathrm{X} = \mathrm{x}) \quad (4)$$

This equation says that just because two things happen together (like exercise and good health), it doesn't mean one causes the other.

$\mathrm{P}(\mathrm{Y} \mid \mathrm{X} = \mathrm{x})$: This is the probability of outcome Y (like good health) given that we *observe* X=x (someone exercises).

$\mathrm{P}(\mathrm{Y} \mid \mathrm{do}(\mathrm{X} = \mathrm{x})$: This is the probability of outcome Y if we *intervene* and make X=x happen (force someone to exercise).

$$\mathrm{P}\big(\mathrm{Y} \mid \mathrm{do}(\mathrm{X} = \mathrm{x})\big) \neq \mathrm{P}(\mathrm{Y} \mid \mathrm{X} = \mathrm{x}) \quad (5)$$

This equation says that just because two things happen together (like exercise and good health), it doesn't mean one causes the other.

$\mathrm{P}(\mathrm{Y} \mid \mathrm{X} = \mathrm{x})$: This is the probability of outcome Y (like good health) given that we *observe* X=x (someone exercises).

$\mathrm{P}(\mathrm{Y} \mid \mathrm{do}(\mathrm{X} = \mathrm{x})$: This is the probability of outcome Y if we *intervene* and make X=x happen (force someone to exercise).

Modern implementations include causal forests (Athey & Imbens, 2016) for heterogeneous treatment effects and double machine learning (Chernozhukov *et al.*, 2018) for unbiased causal effect estimation in

high-dimensional settings. These methods enhance decision-making in areas such as personalized medicine and policy evaluation

Decision Making-Theoretic Frameworks

Statistical decision theory provides a principled foundation for making optimal decisions under uncertainty. A common setup involves choosing an action a ∈ A to minimize expected loss:

$$a^* = argmin_a \; E[L(a,\theta) \tag{6}$$

Where $E[L(a,\theta)$ is a loss function and θ is an uncertain parameter.

In reinforcement learning, decision theory aligns well with expected utility maximization. Bayesian reinforcement learning incorporates uncertainty into policy selection, leading to better exploration and risk-sensitive behavior (Ghavamzadeh *et al.*, 2015).

Statistical Regularization and Model Robustness

Overfitting and lack of generalizability are major concerns in AI. Statistical regularization techniques such as Lasso (Tibshirani, 1996), Ridge regression, and Elastic Net (Zou & Hastie, 2005) penalize model complexity to improve generalization.

Regularized loss function for Elastic net:

$$\hat{\beta}_{j(Enet)} = \arg \min_{\beta \epsilon \mathbb{R}^P} \left[\sum_{i=1}^{n} \left(y_i - \beta_o - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{k} |\beta_j| + \lambda_2 \sum_{j=1}^{k} {\beta_j}^2 \right] \tag{7}$$

For a dichotomous regression (7) becomes

$$\hat{\beta}_{j(Enet)} = \arg \min_{\beta \epsilon \mathbb{R}^k} \left[-\sum_{i=1}^{n} \{ y_i ln(\pi(x_i)) + (1 - y_i) \ln(1 - \pi(x_i)) \} + \lambda_1 \sum_{j=1}^{k} |\beta_j| + \lambda_2 \sum_{j=1}^{k} {\beta_j}^2 \right] \tag{8}$$

where $\lambda_1 \sum_{j=1}^{k} |\beta_j| + \lambda_2 \sum_{j=1}^{k} {\beta_j}^2$, $0 < \lambda_1 + \lambda_2 < 1$ is the penalty function.

From above, the elastic net estimation is a ridge regression when λ_1 is zero and is a LASSO regression when λ_2 is zero.

The Adaptive LASSO estimator is expressed as:

$$\hat{\beta}_{j(ALasso)} = \arg \min_{\beta \epsilon \mathbb{R}^k} \left[-\sum_{i=1}^{n} \{ y_i ln(\pi(x_i)) + (1 - y_i) ln(1 - \pi(x_i)) \} + \lambda \sum_{j=1}^{k} \widehat{\omega}_j |\beta_j| \right] \quad (9)$$

$$where\ \widehat{\omega}_j = \frac{1}{|\hat{\beta}_{j(Lasso)}|^r}, \qquad j = 1, 2, \ldots, k,\ \ r > 0$$

The tuning parameter λ and the order of adaptive weight r are used as the two-dimensional cross-validation to tune the adaptive lasso

The Adaptive elastic net penalty function combines the elastic net and adaptive lasso method.

The adaptive elastic net penalty is defined as:

$$\lambda_1 \sum_{j=1}^{k} \widehat{\omega}_j |\beta_j| + \lambda_2 \sum_{j=1}^{k} {\beta_j}^2$$

$$\hat{\beta}_{j(AEnet)} = \arg \min_{\beta \epsilon \mathbb{R}^k} \left[-\sum_{i=1}^{n} \{ y_i ln(\pi(x_i)) + (1 - y_i) ln(1 - \pi(x_i)) \} + \lambda_1 \sum_{j=1}^{k} \widehat{\omega}_j |\beta_j| + \lambda_2 \sum_{j=1}^{k} {\beta_j}^2 \right] \quad 11$$

$$where\ \widehat{\omega}_j = \frac{1}{\left(\left| \hat{\beta}_{j(Enet)} + \frac{1}{n} \right| \right)^r}, \quad j = 1, 2, \ldots, k, r > 0$$

Sometimes, AI systems perform very well on training data but fail on new data. This happens when a model is too complex and "memorizes" instead of learning patterns.
Regularization is a technique used to control complexity. It makes models simpler and better at working with new data. Common statistical tools like Lasso and Ridge regression help reduce unnecessary variables. Simpler models are not only more accurate on new data—they are also easier to understand.

3. APPLICATIONS

To evaluate the sparsity and robustness of regularized methods (Adaptive LASSO, Adaptive Elastic net, LASSO, Elastic net and Ridge) in the field of cancer classification, two publicly well-known binary

cancer classification datasets were used: leukemia cancer and colon cancer. The detailed information of these datasets is summarized in Table 1.

Data Description

Dataset 1: The first data is the leukemia dataset. In the leukemia dataset, there were two types of patients: 47 patients of acute lymphoblastic leukemia (ALL) and 25 patients of acute myeloid leukemia (AML). These data have consisted of 7129 genes from the bone marrow samples, which is described in detail by Golup *et al.*, (1999). A subset of 3571 genes is the independent variable.

Dataset 2: The second data set is the Colon data set. The colon cancer dataset, contained gene expression levels of 40 tumor and 22 normal colon tissues for 6500 human genes obtained with an Affymetrix oligonucleotide array Alon *et al.*,(1999). A subset of 2000 genes with the highest minimal intensity across the samples was used.

For comparison purposes, the performance of the Adaptive LASSO, adaptive Elastic net, Ridge, LASSO and Elastic-net was also evaluated..

Table 1: The detail information for the used datasets.

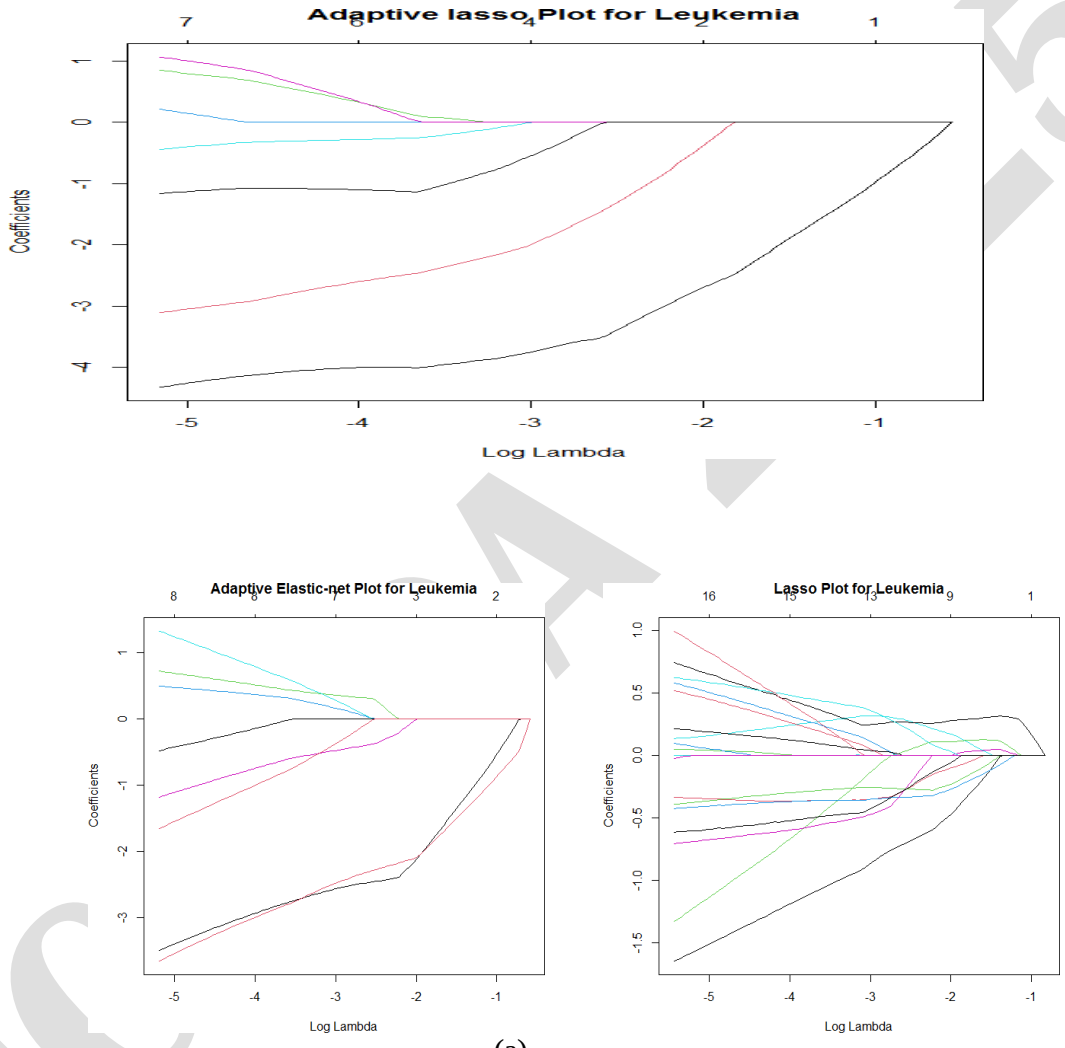
Datasets	No. ofGenes	No. ofSamples	No. of Classes(Class1:Class2)	Classes
Leukemia	3571	72	2(25:47)	ALL/AML
Colon	2000	62	2(22:40)	Tumor/Non Tumor

Table 2 : Performance Metrics for the Leukemia Dataset

Method	**No. of Variables Selected**	**Classification Accuracy (CA)**	**AUC**	**Sensitivity**	**Specificity**	**G-Mean**
ALASSO	7	94.91%	0.969	1.00	0.883	0.940
AEnet	8	96.45%	0.952	1.00	0.857	0.926
LASSO	16	63.64%	0.924	1.00	0.857	0.926
Elastic Net	76	92.90%	0.924	1.00	0.857	0.926

Ridge	3571	92.45%	0.952	1.00	0.857	0.926

From table 2, AEnet and ALASSO had the highest classification accuracy and AUC, while selecting fewer variables, showing strong performance with sparsity. Ridge and ElasticNet used many more variables but didn't significantly outperform AEnet or ALASSO. LASSO performed the worst in terms of accuracy (63.64%) despite a reasonable AUC.



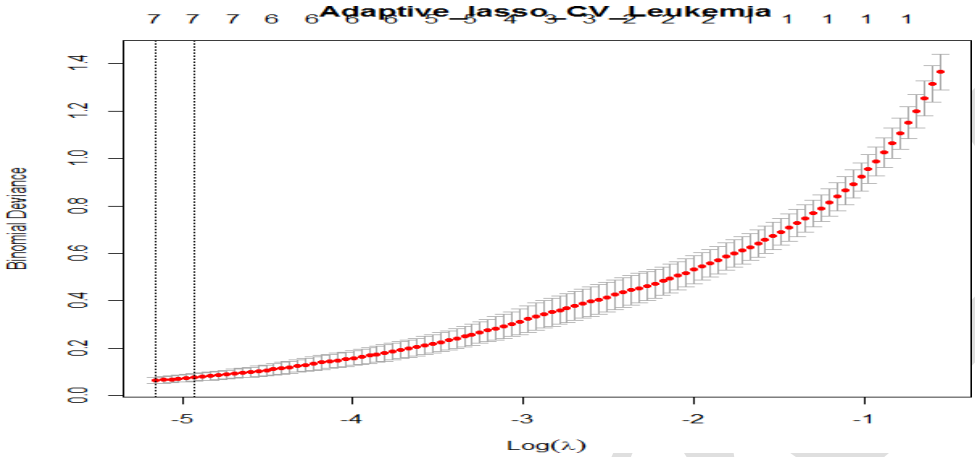
(a)

(b) (c)

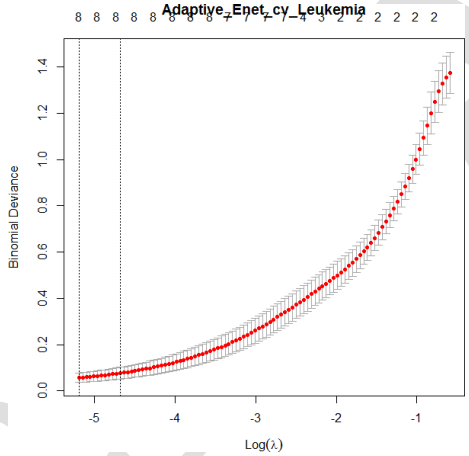
Figure 1: *Non-Zero Coefficients vs. Penalty Parameter (log(λ)) for Leukemia Dataset.* Figure 1 Showed how regularization (via log(λ)) reduces the number of non-zero coefficients. ALASSO and AEnet shrink coefficients more effectively than LASSO.

The plots show how the number of non-zero coefficients decreases as the penalty parameter increases. The ALASSO and AEnet shrink coefficients more effectively, leading to sparser

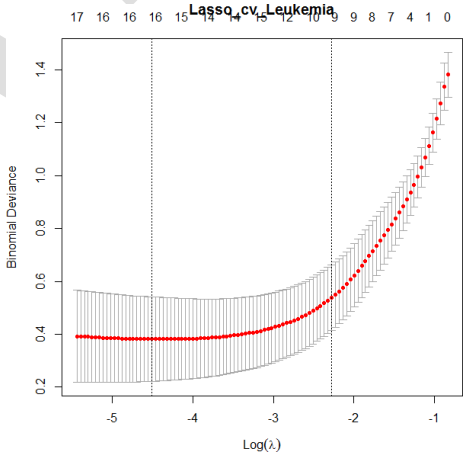
models compared to LASSO which retains more variables even at higher penalties. This indicates that adaptive methods (ALASSO and AEnet) achieve better variable selection and reduce overfitting.



(a)



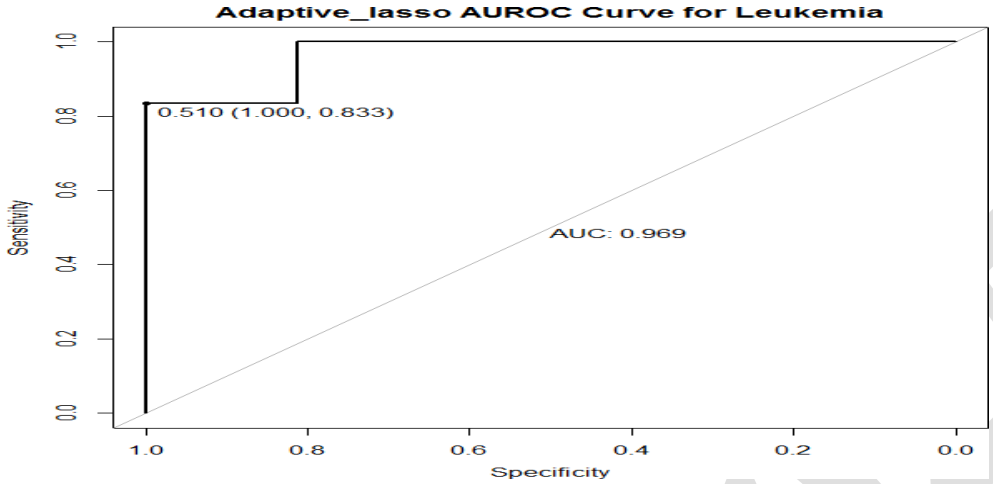
(b)



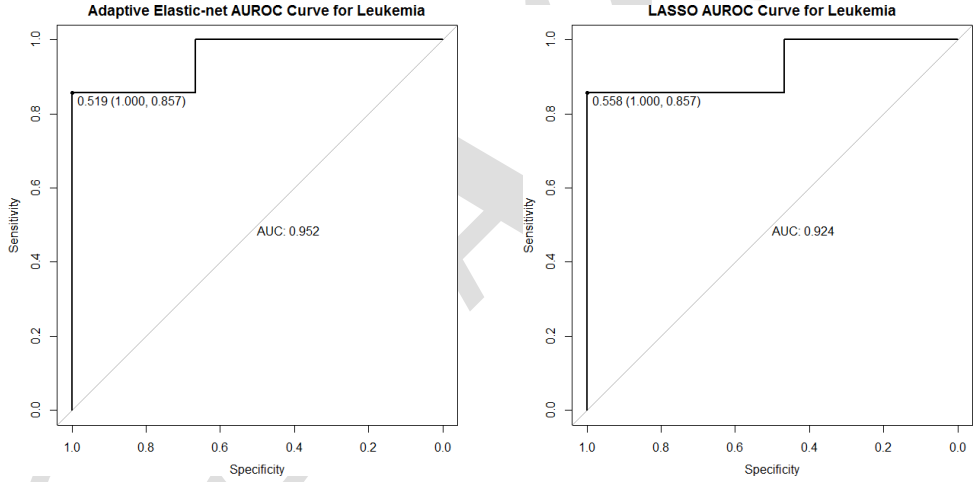
(c)

Figure 2: *Mean Squared Error (MSE) vs. Penalty Parameter (log(λ)) for Leukemia Dataset.* figure 2 showed Plotted MSE vs. log(λ); ALASSO and AEnet showed better cross-validated model performance

Figure 2 is the cross-validation curves for Leukemia dataset that show model performance across different penalty values. ALASSO and AEnet achieve lower MSE compared to LASSO,meaning they generalize better to unseen data. The vertical line in the plots indicates the optimal λ value selected by cross-validation, where model complexity and accuracy are balanced



(a)



(b) (c)

Figure 3: *AUC Plots for the Leukemia Dataset.*

Subplots illustrate AUC values, sensitivity, and specificity for ALASSO, AEnet, and LASSO.

AUC plots confirmed ALASSO and AEnet outperformed LASSO in sensitivity and specificity

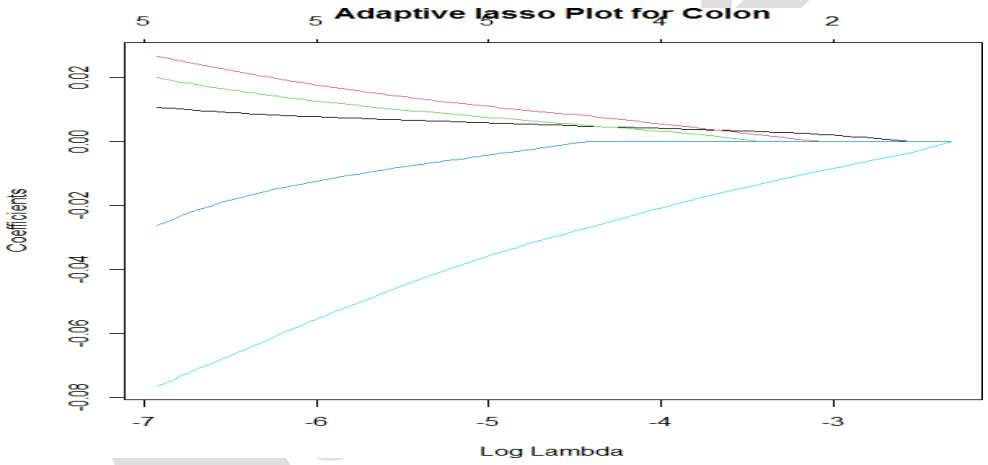
Figure 3 is the AUC plots for Leukemia dataset. These plots illustrate AUC, Sensitivity and Specificity for ALASSO, AEnet and LASSO. ALASSO and AEnet consistently outperform LASSO in diagnostic accuracy, capturing better trade-off

Colon Dataset

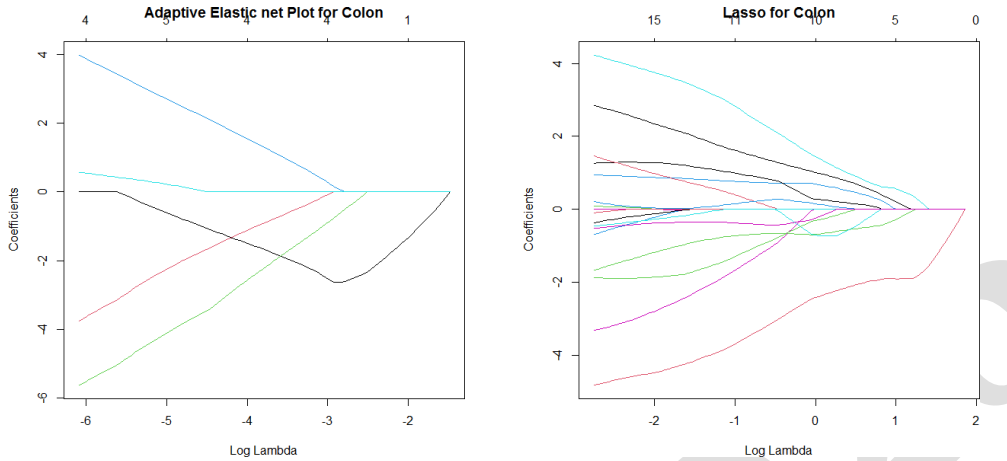
Table 3: Performance Metrics for the Colon Dataset

Model	No. of Variables	C.A	AUC	Sensitivity	Specificity	G-Mean
ALASSO	5	89.47	0.914	0.786	1.00	0.887
AEnet	4	84.21	0.957	0.857	1.00	0.926
LASSO	15	73.68	0.929	0.786	1.00	0.887
Elasticnet	31	94.74	0.929	0.857	1.00	0.926
Ridge	2000	89.47	0.900	0.929	0 .800	0.862

From Table 3 ElasticNet yielded the highest accuracy (94.74%), followed by ALASSO and Ridge.AEnet achieved the highest AUC (0.957) with very few variables, highlighting efficiency. Again, LASSO underperformed relative to other regularized methods.



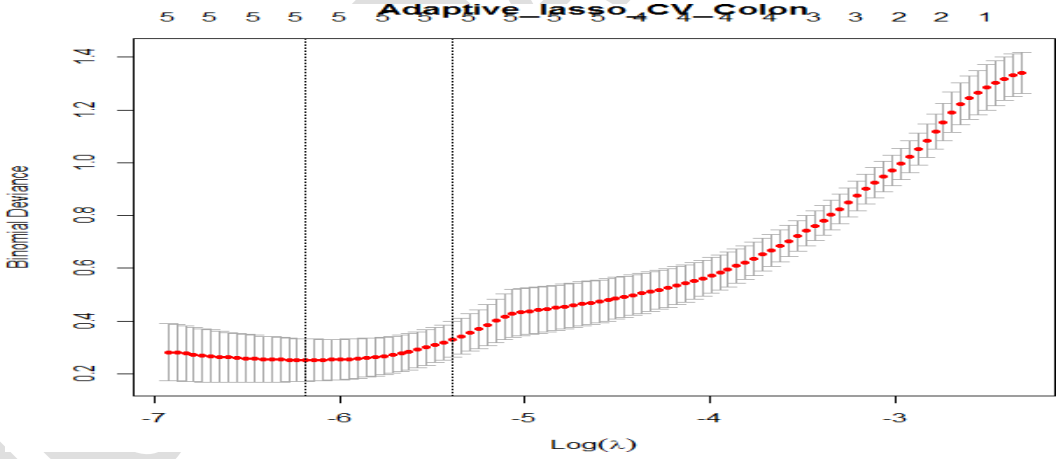
(a)



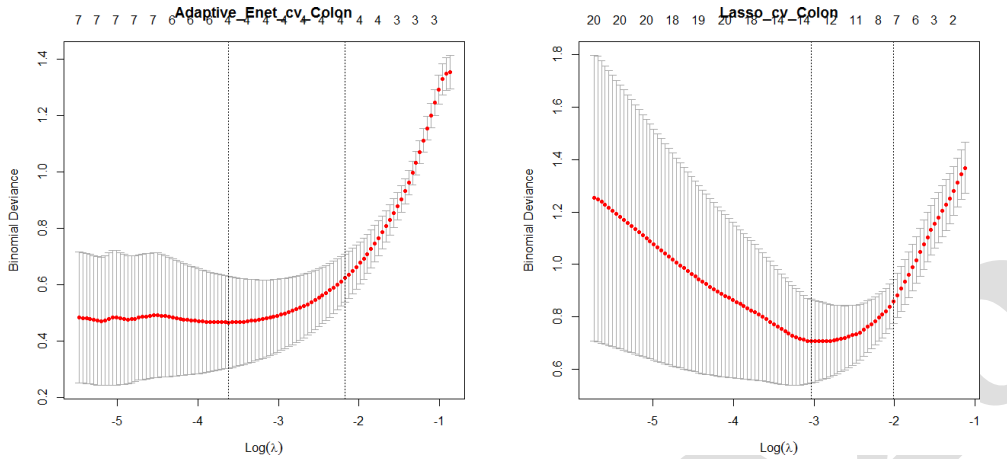
(b) (c)

Figure 4: Non-zero coefficients plots of Colon data as a function of log(λ) for the (a) ALASSO (b) AEnet (c) LASSO.

In figure 4, similar to figure 1 but for Colon cancer data shows non-zero coefficients versus penalty parameter (log(λ)). ALASSO and AEnet select very few genes (high sparsity), while LASSO retains more variables. This again confirm that adaptive regularization provides more interpretable and efficient gene selection process



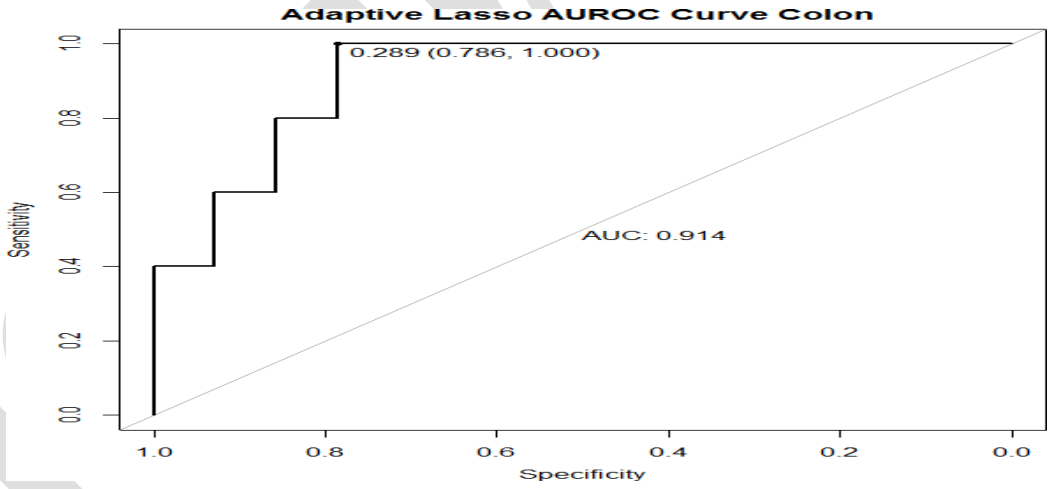
(a)



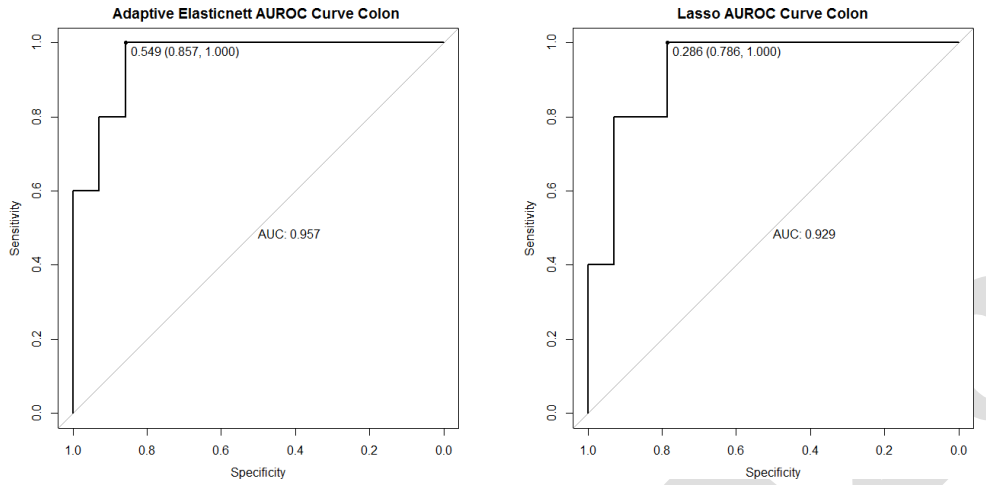
(b) (c)

Figure 5: MSE plots and the number of Variables in the model as a function of log(λ) for the 10-fold cross validation for Colon data for the ((a) ALASSO (b) AEnet (c) LASSO. The vertical lines shows the number of predictor variables (genes) selected using λ.min (left) and λ.1se(right).

Figure 5 shows cross-validated MSE curves across different penalties for colon data. Vertical lines mark λ.min (best accuracy) and λ.1se (simple model with slightly higher error). ALASSO and AEnet achieves competitive accuracy with fewer predictors.



(a)



(b) (c)

Figure 6: AUC plots of Colon data for the (a) ALASSO (b) AEnet

(c) LASSO. Each sub plot shows, AUC (centre value in the figure), Sensitivity (top left middle value) and Specificity (top left right value).

In figure 6, the plots summarize AUC, sensitivity and specificity for ALASSO, AEnet and LASSO. AEnet shows the highest AUC (best discrimination ability), while ALASSO also perform strongly with fewer genes. LASSO again lags behind, highlighting the superiority of adaptive methods in high-dimensional biomedical data.

4. DISCUSSION

The findings from this study highlight the crucial role of statistical methods, particularly regularization techniques in enhancing the interpretability, sparsity, and robustness of AI systems, especially in high-dimensional biomedical datasets. Across both the leukemia and colon cancer datasets, adaptive methods such as Adaptive LASSO (ALASSO) and Adaptive Elastic Net (AEnet) consistently outperformed standard techniques in terms of classification accuracy, model sparsity, and diagnostic performance metrics like AUC, sensitivity, and specificity.

One of the most significant outcomes is the balance between sparsity and predictive performance. Both ALASSO and AEnet selected significantly fewer predictor variables than Ridge or Elastic Net, yet delivered superior or comparable classification accuracy. This has meaningful implications in fields like genomics, where reducing dimensionality without losing predictive power improves interpretability and reduces overfitting. In contrast, methods like Ridge and Elastic Net, while achieving reasonable predictive metrics, used thousands of variables, potentially obscuring causal or informative features and reducing model transparency.

The inferior performance of LASSO, particularly in the leukemia dataset (accuracy of 63.64%), underscores the limitations of unadapted shrinkage penalties in extremely high-dimensional settings. This confirms the necessity of incorporating adaptive weighting schemes, which allow for variable-specific penalization and ultimately yield more stable and meaningful feature selection.

The figures showing non-zero coefficient paths and cross-validated mean squared error (MSE) provide visual evidence of the trade-offs between complexity and accuracy. ALASSO and AEnet achieved optimal solutions with fewer predictors, as shown by the rapid shrinkage of coefficients with increasing penalty and lower MSE curves at optimal λ values. These methods demonstrate better generalization, which is critical for deployment in real-world clinical and diagnostic systems where new, unseen data must be handled reliably.

Moreover, the inclusion of uncertainty quantification and causal inference frameworks further strengthens the case for embedding statistical principles into AI pipelines. Bayesian neural networks and do-calculus methods enhance model transparency and allow for actionable insights, especially in policy and healthcare. These tools address the "black box" nature of traditional machine learning models, which often lack the capability to convey confidence in their outputs or differentiate correlation from causation.

Finally, the study reinforces the value of statistical decision theory in guiding AI models toward optimal actions under uncertainty. By minimizing expected loss and integrating prior knowledge, AI systems become not only accurate but also trustworthy and context-aware qualities essential for adoption in sensitive and high-stakes domains.

5. Conclusion

This study demonstrates the critical role of statistical methodologies in enhancing the trustworthiness, interpretability, and robustness of AI systems, particularly in high-dimensional biomedical applications. By integrating adaptive regularization techniques such as Adaptive LASSO and Adaptive Elastic Net, we achieved significant improvements in classification performance and model sparsity across both leukemia and colon cancer datasets. These methods not only reduced the number of selected variables, making models more interpretable, but also maintained or exceeded the predictive accuracy of more complex models.

Beyond predictive performance, this research emphasizes the importance of embedding uncertainty quantification**,** causal inference**, and** decision-theoretic principles into AI pipelines. These statistical tools provide the necessary foundation for handling ambiguity, ensuring fairness, and guiding optimal decisions in real-world, high-stakes environments such as healthcare, finance, and public policy.

Ultimately, the findings advocate for a tighter integration between statistical science and artificial intelligence. As AI continues to permeate critical domains, the incorporation of robust, interpretable, and reliable statistical frameworks will be essential to building systems that are not only intelligent but also transparent and ethically sound.

Suggestions for Further Studies

Future studies could explore hybrid architectures that combine the interpretability and sparsity of adaptive regularization techniques (e.g., ALASSO, AEnet) with the nonlinear representation power of deep learning models. For example, sparse priors can be incorporated into neural network layers to improve generalization in high-dimensional settings such as genomics or radiomic.

While this study focused on leukemia and colon cancer datasets, additional validation across other cancer types, rare diseases, or multi-omics datasets (e.g., combining transcriptomics, proteomics, and epigenomics) would strengthen generalizability and reveal domain-specific performance gaps.

Building on the foundational work with Bayesian Neural Networks (BNNs), further research should investigate real-time uncertainty quantification in streaming or online learning contexts. This is especially critical for adaptive AI systems deployed in healthcare, finance, and autonomous navigation, where confidence-aware predictions could trigger human intervention.

References

Alvarez-Melis, D., & Jaakkola, T. S. (2017). A causal framework for explaining the predictions of black-box sequence-to-sequence models. *arXiv preprint arXiv:1707.01943*.

Athey, S., & Imbens, G. W. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27), 7353–7360.

Chen, J., *et al.* (2021). Applications of machine learning in financial economics. *Annual Review of Financial Economics*, 13, 249–274.

Chernozhukov, V., Chetverikov, D., Demirer, M., *et al.* (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1–C68.

Du, X., Chen, S., & Wang, Y. (2022). Causal learning for responsible AI. *Journal of Artificial Intelligence Research*, 74, 331–355

Esteva, A., *et al.* (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1), 24–29.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *International Conference on Machine Learning*, 1050–1059.

Ghavamzadeh, M., Mannor, S., Pineau, J., & Tamar, A. (2015). Bayesian reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 8(5-6), 359–483.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. MIT Press.

Leibig, C., Allken, V., Ayhan, M. S., Berens, P., & Wahl, S. (2017). Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports*, 7(1), 1–14. https://doi.org/10.1038/s41598-017-17876-z

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference* (2nd ed.). Cambridge University Press.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Taylor, J. W. (2019). Forecasting value at risk and expected shortfall using a semiparametric approach based on the asymmetric Laplace distribution. *Journal of Business & Economic Statistics*, 37(1), 121–133. https://doi.org/10.1080/07350015.2017.1302191

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.

Xu, K., Tan, L., & Bansal, R. (2025). Hybrid interpretable models for high-stakes decision support. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(1), 210–219.

Yang, T., Zhang, M., & Liu, Y. (2023). Trustworthy Bayesian deep learning for safety-critical AI. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2), 502–515

Zhao, W., & Jin, R. (2024). Efficient probabilistic graphical models for interpretable AI. *Neural Networks*, 169, 1–15.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.