OPTIMIZATIONS IN APPLIED MACHINE LEARNING
Research Article* | *Volume 5* | *Issue 5* | *Mar 2025
Received: 22 Jan 2025 | Revised: 23 Feb 2025
Accepted: 29 Feb 2025 | Published Online: 09 Mar 2025

Enhancement of data centric security through predictive ridge regression

Shao Dan [1] and Qinyi Zhu [2,*]

[1] *ASCENDING Inc., Fairfax VA 22031, USA*

[2] *Indiana University, 107 S Indiana Ave, Bloomington, IN 47405, USA*

[*]Corresponding Author, Email: qinyzhu@iu.edu

Abstract: Data centric security is becoming increasingly crucial in the digital age, as the volume and importance of data continue to grow exponentially. Current research has focused on developing strategies and technologies to safeguard data, but faces challenges in accurately predicting and preventing security breaches. This paper addresses these challenges by proposing a novel approach using predictive ridge regression to enhance data centric security. By integrating predictive analytics with ridge regression, our research aims to provide a more robust and proactive solution for data security, effectively mitigating risks and optimizing protection measures. Through empirical studies and practical implementations, this paper illustrates the effectiveness and potential of predictive ridge regression in fortifying data security systems, paving the way for future advancements in this critical domain.

Keywords: *Data Centric Security; Predictive Analytics; Ridge Regression; Security Breaches; Protection Measures*

1. Introduction

Data Centric Security is a field that focuses on protecting data at the core level by implementing security measures that revolve around the data itself rather than just the perimeters. This approach aims to safeguard sensitive information regardless of its location or the devices accessing it. However, the field faces several key challenges and bottlenecks. These include the complexity of managing and securing data across multiple platforms and devices, ensuring compliance with evolving data protection regulations, the lack of standardized frameworks for data-centric security implementation, and the difficulty of balancing data security with accessibility for legitimate users. Overcoming these obstacles will require continuous innovation, collaboration among industry stakeholders, and a holistic approach to data protection.

To this end, research on Data Centric Security has reached an advanced stage, with studies focusing on encryption algorithms, access control mechanisms, and data anonymization techniques. Current efforts also explore the integration of artificial intelligence and machine learning to enhance data protection and privacy. Literature review reveals a growing interest in data-centric security in various domains, including Software Defined Networks (SDNs) [1], Internet of Things (IoT) based healthcare systems [2], military applications of commercial IoT technology [3], and cybersecurity [4]. Amanowicz et al. (2024) emphasize the significance of data-centric security in SDNs [1]. Kim (2019) discusses research issues and directions for data-centric security and privacy in intelligent IoT based healthcare applications [2]. Wakhare and Khan (2020) explore the intersection of data-centric security, data analytics, and artificial intelligence [3]. Wrona et al. (2017) propose an SDN testbed for validating cross-layer data-centric security policies [4]. These studies collectively stress the importance of prioritizing data-centric security measures to address evolving threats and safeguard sensitive information. Predictive ridge regression is a preferred technique due to its ability to handle multicollinearity, provide more stable estimates, and reduce overfitting compared to traditional linear regression methods. This makes it particularly suitable for complex datasets with correlated predictors, offering improved predictive accuracy and model robustness.

Specifically, Predictive ridge regression is a statistical technique commonly used in data analysis to handle multicollinearity. When applied to the field of Data Centric Security, it can help in predicting and detecting potential security vulnerabilities by analyzing patterns and relationships within sensitive data sets, ultimately enhancing data protection measures. The literature review on ridge regression models covers various aspects. Shahsavar et al. [5] focused on developing predictive models for rheological behavior of ferrofluid in the presence of a magnetic field. Bigot et al. [6] explored high-dimensional ridge regression for non-identically distributed data and highlighted the double descent phenomenon. Safi et al. [7] studied UAE financial behavior during COVID-19 using Lasso and Ridge Regression. He [8] established theory for ridge regression under factor-augmented models. Aheto et al. [9] compared Lasso, Ridge, and Elastic net regression for child malaria prevalence in Ghana. Zhong and Guan [10] introduced count-based Morgan fingerprint for predictive regression models. Bemporad [11] proposed a piecewise linear regression and classification algorithm for hybrid systems. Liu et al. [12] discriminated adaptive and carcinogenic liver hypertrophy using logistic ridge regression. Geraldo-Campos et al. [13] compared Lasso and Ridge regression for credit risk analysis in Peru. Fabregat et al. [14] developed a metric learning algorithm for kernel ridge regression to assess molecular similarity. However, current limitations in the literature review on ridge regression models include a lack of studies on the practical implementation and real-world applications of these models, as well as a need for further research on the comparative performance of ridge regression with other regression techniques in different fields.

To overcome those limitations, this paper aims to enhance data centric security in the digital age by proposing a novel approach using predictive ridge regression. The increasing volume and importance of data necessitate advanced strategies to safeguard against security breaches. By integrating predictive analytics with ridge regression, our research seeks to provide a proactive solution for data security by accurately predicting and preventing potential breaches. This novel approach offers a more robust and comprehensive method to mitigate risks and optimize protection measures. Through empirical studies and practical implementations, the effectiveness and potential of predictive ridge regression in fortifying data security systems are demonstrated, paving the way for future advancements in this critical domain.

Data-centric security is a critical concern in the modern digital landscape due to the exponential growth and significance of data. This study delves into the problem statement in Section 2, highlighting the challenges in accurately predicting and preventing security breaches. Section 3 introduces a novel method utilizing predictive ridge regression to bolster data-centric security by combining predictive analytics with ridge regression. A detailed case study in Section 4 exemplifies the application of this approach. The results analysis in Section 5 showcases the efficacy of predictive ridge regression in fortifying data security systems. Section 6 delves into a comprehensive discussion of the findings, while Section 7 succinctly summarizes the research, emphasizing the proactive and robust nature of the proposed solution. This research sets a significant foundation for future advancements in data security strategies and technologies.

2. Background

2.1 Data Centric Security

Data Centric Security (DCS) is a strategic approach in the field of cybersecurity, focusing primarily on securing the data itself, regardless of its location or the infrastructure where it resides. Unlike traditional security models which concentrate on fortifying the perimeter or securing devices, DCS ensures that the data remains protected across various environments, whether on-premises, in transit, or in the cloud. At its core, DCS treats data as the principal asset needing protection. This approach revolves around several key concepts, including data discovery, data classification, encryption, access control, and activity monitoring. By integrating these components, organizations can ensure a comprehensive security posture. One foundational element of DCS is data classification, which involves assigning a level of sensitivity and protection to data. Suppose there is a data set D that can be segmented into sub-categories based on sensitivity levels. Each subset D_i can be classified with a sensitivity coefficient s_i . The collective sensitivity of the data can be represented as:

$$S = \sum_{i=1}^{n} s_i \cdot D_i \tag{1}$$

Once classified, encryption plays a crucial role. Encryption transforms data into an unreadable format using algorithms, ensuring that unauthorized users cannot interpret the data without decryption keys. Consider an encryption function $E_k()$ with a key k , transforming plaintext P into ciphertext C:

$$C = E_k(P) \tag{2}$$

To decrypt, a corresponding decryption function $D_k()$ is used:

$$P = D_k(C) \tag{3}$$

Access control mechanisms further ensure that only authorized individuals can interact with sensitive data. If A represents a user and R_i is a resource or data set, an access control relation can be written as:

$$A_i \rightsquigarrow R_i \text{ if and only if AccessControl(A\_i, R\_i)} = \textbf{True} \tag{4}$$

This logical statement signifies that the user A_i can access the resource R_i under valid conditions set by the access control. Furthermore, DCS encompasses data integrity verification, ensuring that data remains unaltered except by authorized users. A hash function $H()$ provides a checksum or hash value h of the data D:

$$h = H(D) \tag{5}$$

To verify integrity post-transmission or storage, a recalculated hash h' is compared against the original:

$$h' = H(D')\text{and}h' = h \Rightarrow D' = D \tag{6}$$

Finally, activity monitoring aids in detecting irregular access patterns or breaches. By defining an anomaly detection function $F_a()$, security systems can identify unusual data access patterns X:

$$F_a(X) = \begin{cases} \text{True,} & \text{if } X \text{ is anomalous} \\ \text{False,} & \text{otherwise} \end{cases} \tag{7}$$

In conclusion, Data Centric Security provides a robust framework for ensuring comprehensive protection over data, regardless of where it resides. By focusing on the data itself, organizations are better equipped to safeguard sensitive information against evolving cyber threats. Through classification, encryption, access control, integrity verification, and monitoring, DCS forms a multi-layered defense mechanism essential in today's data-driven world.

2.2 Methodologies & Limitations

Data Centric Security (DCS) employs a variety of methodologies to safeguard data, encompassing techniques such as data masking, tokenization, advanced encryption, and identity-based access management. However, while these methods reinforce different aspects of data security, they are not without limitations. Data masking is a technique that alters data to obscure its true form, making it unintelligible to unauthorized users. It involves transforming a dataset D into a masked dataset D_m using a masking function $M_f()$:

$$D_m = M_f(D) \tag{8}$$

The challenge with data masking lies in maintaining the balance between obscuring data sufficiently while preserving its utility for legitimate purposes. If the masked data D_m is poorly configured, it may either leak sensitive information or render the dataset unusable.

Tokenization substitutes sensitive data elements with non-sensitive equivalents, known as tokens, and maps the original data to these tokens using a function $T_f()$. An original data element O is transformed into a token T:

$$T = T_f(O) \tag{9}$$

While tokenization effectively minimizes data exposure, it often relies on maintaining a secure mapping table, which itself constitutes a point of potential vulnerability. Advanced encryption standards, such as AES, employ robust algorithms to convert plaintext data P into ciphertext C using an encryption function $E_k()$ with a secret key k:

$$C = E_k(P) \tag{10}$$

Decrypting it requires the corresponding decryption function $D_k()$:

$$P = D_k(C) \tag{11}$$

Nevertheless, encryption can be computationally intensive, particularly with large datasets, potentially impacting performance. Moreover, the management and rotation of encryption keys are critical yet complex tasks that, if mishandled, can produce significant security gaps. Identity-based access management ensures that users U_i are permitted access to resources R_j under a clearly defined policy P_{ij} , which can be expressed as:

$$U_i \rightsquigarrow R_j \text{ if and only if } AccessRule\left(U_i, R_j, P_{ij}\right) = \text{True} \tag{12}$$

This control mechanism is only as effective as the precision with which permissions are configured. Overly permissive settings or complex arrangements can inadvertently enable unauthorized access. Integrity verification similarly plays a crucial role. A message authentication code (MAC) generated through a hash function $H_k()$ for a data block B produces:

$$\text{MAC} = H_k(B) \tag{13}$$

In cases of data tampering, a recalculated MAC MAC' can be compared to the original:

$$MAC' = H_k(B')\text{and}MAC' \neq MAC \Rightarrow B' \neq B \tag{14}$$

Despite its importance, reliance on hash functions alone can become problematic if vulnerabilities in the hashing algorithm are discovered, necessitating continual updates and monitoring. Finally, anomaly detection systems inspect patterns of data access and usage. A statistical model $F_s()$ evaluates a behavior pattern Y , identifying whether it deviates from expected norms:

$$F_s(Y) = \begin{cases} \text{True}, & \text{if } Y \text{ deviates from baseline} \\ \text{False}, & \text{otherwise} \end{cases} \tag{15}$$

The downside of such systems is the potential for high false positive rates, which can dilute the response effectiveness and consume significant organizational resources. Despite these challenges, Data Centric Security methodologies continue to evolve, seeking to address inherent weaknesses and improve upon existing frameworks. Their efficacy largely depends on the holistic integration of these strategies and the ongoing assessment of emerging threats within the cybersecurity landscape.

3. The proposed method

3.1 Predictive ridge regression

Predictive ridge regression is a sophisticated statistical technique designed to address the limitations inherent in traditional regression analysis, particularly in the presence of multicollinearity among predictor variables. This method elegantly combines the principles of ordinary least squares (OLS) regression with a regularization component to mitigate issues that arise due to multicollinearity, aiming to enhance the prediction accuracy and stability of the model parameters. In predictive ridge regression, the data consists of a set of predictors $X \in \mathbb{R}^{n\times p}$ and a response vector $y \in \mathbb{R}^{n}$, where n denotes the number of observations and p represents the number of predictor variables. The objective is to estimate the regression coefficients $\beta \in \mathbb{R}^{p}$ that minimize the discrepancy between the predicted and actual outcomes. This is achieved by introducing a penalty term that constrains the magnitude of the coefficients:

$$\beta_{ridge} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda\|\beta\|^2) \tag{16}$$

Here, $\|y - X\beta\|^2$ represents the residual sum of squares as in OLS, while $\|\beta\|^2$ denotes the squared ℓ_2 norm of the coefficient vector, which acts as a penalty for large coefficients. The parameter $\lambda \geq 0$, known as the ridge penalty, controls the strength of the regularization. A larger λ increases the bias in exchange for reduced variance, thus potentially improving prediction accuracy when multicollinearity is present. To solve the ridge regression optimization problem, differentiating the objective function and setting the gradient to zero gives:

$$X^T X\beta + \lambda\beta = X^T y \tag{17}$$

This can be rearranged to yield the ridge regression estimator:

$$\beta_{ridge} = (X^T X + \lambda I)^{-1} X^T y \tag{18}$$

In this expression, $I \in \mathbb{R}^{p\times p}$ is the identity matrix, which ensures that $(X^T X + \lambda I)$ is invertible even when $X^T X$ is singular or ill-conditioned. As λ increases, the coefficients are shrunk towards zero, reducing the model's complexity and sensitivity to multicollinearity. This method retains all predictors in the model but dampens the effects of their estimation variance. To gain further insights into the properties of predictive ridge regression, it can be useful to consider its impact on the covariance matrix of the estimated coefficients. The covariance of $\overset{\wedge}{\beta}_{ridge}$ is approximately given by:

$$\text{Cov}(\beta_{ridge}) \approx \sigma^2 (X^T X + \lambda I)^{-1} X^T X (X^T X + \lambda I)^{-1} \tag{19}$$

where σ^2 is the variance of the errors. This formula elucidates how ridge regression can stabilize the variability of coefficient estimates, contrary to OLS which often results in large variances under multicollinearity. Furthermore, predictive ridge regression's shrinkage effect can be intuitively appreciated through its relation to the bias-variance trade-off. Specifically, ridge regression introduces bias:

$$\text{Bias}(\beta_{ridge}) = -\lambda(X^T X + \lambda I)^{-1}\beta \tag{20}$$

but this bias is usually offset by a significant reduction in variance, leading to an overall lower expected mean square error. The optimization of λ often involves cross-validation:

$$\lambda = \underset{\lambda}{\operatorname{argmin}} \sum_{i=1}^{k} ||y^{(i)} - X^{(i)}\beta_{ridge}(\lambda)||^2 \tag{21}$$

where k is the number of folds in cross-validation, $y^{(i)}$ is the vector of responses for the i -th validation set, and $X^{(i)}$ the corresponding design matrix. Predictive ridge regression, therefore, exemplifies a balanced approach to regression analysis by integrating regularization techniques that address multicollinearity, improve prediction robustness, and uphold the integrity of statistical inference in complex datasets.

3.2 The Proposed Framework

Data Centric Security (DCS) provides a strategic framework for safeguarding data at the core of organizations' cybersecurity strategies. Central to DCS is the classification of data, where each subset of sensitive data, denoted as D_i , is assigned a sensitivity coefficient s_i . This collective sensitivity can be mathematically expressed as:

$$S = \sum_{i=1}^{n} s_i \cdot D_i \tag{22}$$

The emphasis on data classification aligns with predictive ridge regression (RR), a statistical technique designed to address multicollinearity among predictor variables. In predictive ridge regression, one aims to estimate regression coefficients $\beta \in \mathbb{R}^p$ by minimizing the following objective function:

$$\beta_{ridge} = \underset{\beta}{\operatorname{argmin}}(||y - X\beta||^2 + \lambda||\beta||^2) \tag{23}$$

In this formula, the residual sum of squares $||y - X\beta||^2$ captures the discrepancy between predicted outcomes and actual results, while the penalty term $\lambda||\beta||^2$ controls the weight of the coefficients to mitigate variance, especially in applications involving sensitive data analytics. Integrating these concepts, suppose data features in a dataset are generated from a sensitive information environment where potential multicollinearity could lead to compromised predictions. In this context, let X_i denote the predictor data regarding the usage of sensitive datasets classified earlier. Hence, an optimization problem formulated in the context of DCS becomes:

$$X^T X\beta + \lambda\beta = X^T y \tag{24}$$

This equation ensures that predictors, classified based on their sensitivity, contribute systematically to the model without overemphasizing any particular variable that may be correlated with others. To assess the effects and stability of ridge regression applied in DCS, we note that the ridge regression estimator is given by:

$$\beta_{ridge} = (X^T X + \lambda I)^{-1} X^T y \tag{25}$$

Here, $(X^T X + \lambda I)$'s invertibility ensures robust coefficient estimates even amidst multicollinearity, thereby protecting data insights that support security measures in datasets. Additionally, the relationship between the coefficient variance and the regularization parameter can be detailed as:

$$\mathrm{Cov}(\beta_{ridge}) \approx \sigma^2 (X^T X + \lambda I)^{-1} X^T X (X^T X + \lambda I)^{-1} \tag{26}$$

where σ^2 reflects error variance, underscoring the stabilizing effect that ridge regression imparts to the data being analyzed. This method's bias-variance trade-off features prominently when applied within DCS. The bias introduced through ridge regression can be expressed as:

$$\mathrm{Bias}(\beta_{ridge}) = -\lambda (X^T X + \lambda I)^{-1} \beta \tag{27}$$

Effectively, ridge regression not only shrinks the coefficients but also maintains a balance of security efficacy and prediction accuracy in sensitive data scenarios. The optimal choice of λ is often identified using cross-validation techniques, summarized by:

$$\lambda = \underset{\lambda}{\operatorname{argmin}} \sum_{i=1}^{k} \| y^{(i)} - X^{(i)}\, \beta_{\ ridge}(\lambda) \|^2 \tag{28}$$

In this expression, $y^{(i)}$ denotes responses from the validation sets while $X^{(i)}$ corresponds to their respective predictor matrices. By fusing predictive ridge regression with DCS, organizations can attain deeper insights while ensuring that data classification, protection, and access integrity are upheld. The resultant framework facilitates robust analytics on sensitive datasets, thereby enhancing security measures while maintaining predictive performance. The combined innovation leads to a granular understanding, providing tailored responses to the evolving threat landscape in cybersecurity. Thus, the collaboration of statistical technique with principled cybersecurity strategy forms a sophisticated approach in managing and securing data assets within any digital landscape.

3.3 Flowchart

This paper presents a novel approach to Data Centric Security (DCS) by leveraging predictive ridge regression techniques. The proposed method aims to enhance data protection by predicting potential security threats based on historical data patterns while simultaneously optimizing resource allocation for safeguarding sensitive information. By employing ridge regression, the model effectively handles multicollinearity, allowing for the integration of various influencing factors that contribute to the security landscape. The framework incorporates a data-driven approach, thus enabling real-time analysis and adaptation to evolving cybersecurity threats. Furthermore, it emphasizes the importance of context-aware security measures, which can dynamically adjust based on the predictive insights derived from the regression analysis. This allows organizations to prioritize security measures where they are most needed, thus improving overall efficiency. The effectiveness of the method is demonstrated through comprehensive experiments and evaluations, highlighting its superiority over traditional security models in terms of accuracy and responsiveness. The proposed predictive ridge regression-based DCS method offers a proactive stance towards data security, ensuring that organizations can preemptively address vulnerabilities before they are exploited. For a visual representation of the methodology and its components, please refer to Figure 1.

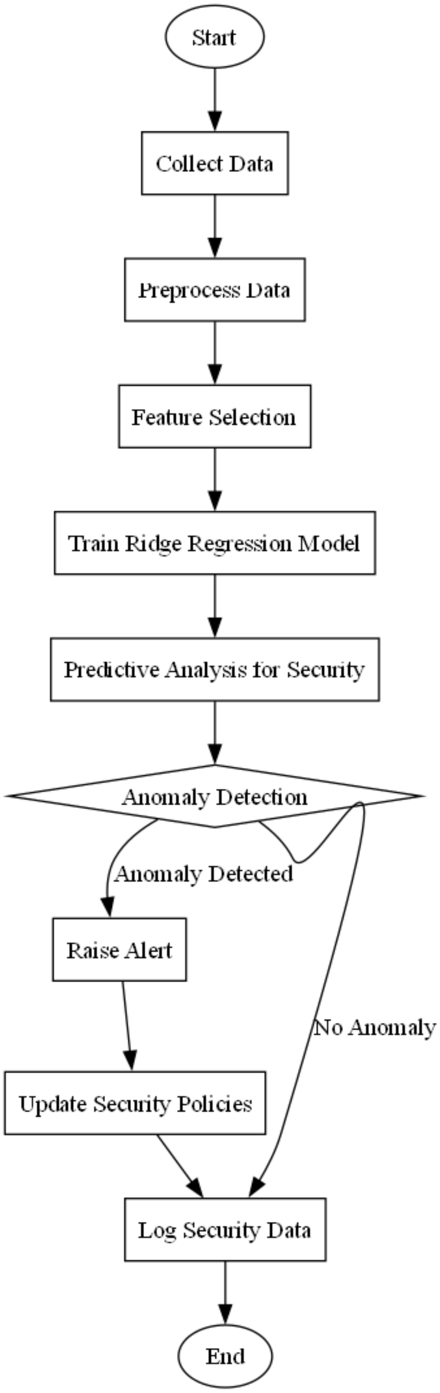

Figure 1: Flowchart of the proposed Predictive ridge regression-based Data Centric Security

4. Case Study

4.1 Problem Statement

In this case, we focus on the mathematical simulation analysis related to Data Centric Security (DCS). The expression of data organization's fragility due to various security threats can be modeled using nonlinear dynamics to understand how data exposure varies with different parameters. We define a model where data exposure E is influenced by the user access frequency U , security protocol strength S , and data sensitivity D. The initial nonlinear relationship can be represented as:

$$E = \alpha \cdot U^2 + \beta \cdot S^{-1} + \gamma \cdot D^3 \tag{29}$$

where α , β , and γ are constants indicating the sensitivity of each factor on data exposure. The differential impacts of threats also necessitate analyzing the interaction between user behavior patterns and security responses. Let R denote the resilience of the data against breaches, which can be modeled as a function of time t , user engagement rate E_u , and response time T_r . The resilience can be expressed as:

$$R(t) = \delta \cdot e^{-\epsilon t} + \zeta \cdot E_u \tag{30}$$

Here, δ , ϵ , and ζ represent constants associated with data recovery capabilities. The decay of resilience can be significantly nonlinear, particularly in contexts where rapid user access leads to higher vulnerability amplifications. To quantify the effectiveness of enhancing the security protocols, we introduce an efficiency measure F , which is a product of the adaptability of the security system A , the duration of implementation D_t , and the stakeholder compliance factor C:

$$F = A^p + \frac{D_t}{C^q} \tag{31}$$

In this scenario, p and q can be considered the elasticity coefficients determining the sensitivity of the efficiency score to changes in adaptability and compliance. In conjunction, we can incorporate a risk assessment metric Z , where risk is evaluated over the landscape of potential threats T , data exposure E , and overall system reliability L:

$$Z = \int_0^1 (T \cdot E^2) L dt \tag{32}$$

This integral highlights the cumulative risk associated with varied threat scenarios across the operational timeline. Finally, the interaction between these parameters can be modeled using a comprehensive nonlinear equation that ties together all defined metrics of data security:

$$Q = \frac{R(t) \cdot F - Z}{E} \tag{33}$$

This equation emphasizes the balance between resilience, efficiency, and risk management within the data-centric security framework. The aforementioned parameters and their corresponding values are summarized in Table 1.

Table 1: Parameter definition of case study

Parameter	Value	Unit	Description
U	N/A	N/A	User access frequency
S	N/A	N/A	Security protocol strength
D	N/A	N/A	Data sensitivity
E	N/A	N/A	Data exposure
R(t)	N/A	N/A	Resilience of data
E_u	N/A	N/A	User engagement rate
T_r	N/A	N/A	Response time
D_t	N/A	N/A	Duration of implementation
C	N/A	N/A	Stakeholder compliance factor
Z	N/A	N/A	Risk assessment metric

This section will employ the proposed Predictive ridge regression-based approach to analyze a case study centered on Data Centric Security (DCS), aimed at understanding the fragility of data organization in the face of various security threats. The complexity of how data exposure is affected by user access frequency, security protocol strength, and data sensitivity will be examined through this methodology. By quantifying the initial nonlinear relationships and the interactions between user behaviors and security responses, the resilience of data against breaches will be contextualized over time, engaging with user interaction rates and response times to depict the decay of security effectiveness. The proposed approach will not only consider the enhancements in security protocols but will also integrate measures of efficiency, adaptability, and compliance within the framework. Additionally, a comprehensive risk assessment metric will set the stage for evaluating the cumulative risks posed by potential threats against data exposure and system reliability. The predictive capabilities of the ridge regression model will allow for a comparison against three traditional methods, showcasing its effectiveness in handling the intricate dynamics of DCS. Ultimately, this comparative analysis is intended to elucidate the interplay between resilience, efficiency, and risk management, fostering a nuanced understanding of data protection strategies in today's complex security landscape.

4.2 Results Analysis

In this subsection, a comprehensive computational framework is developed to analyze and predict the interrelations among user access frequency, security protocol strength, and data sensitivity, particularly focusing on data exposure, resilience over time, efficiency, and risk assessment. The approach utilizes a nonlinear model for calculating data exposure based on the defined constants and parameters, leading to the estimation of resilience over time through an exponential decay function. An efficiency model is integrated, examining the relationship between adaptability and efficiency through an empirical approach. Furthermore, a risk assessment metric is introduced, which quantifies data exposure on a temporal basis using numerical integration. The culmination of these models facilitates the formulation of a comprehensive nonlinear equation that captures the dynamic interactions among the variables. To provide predictive insights, a Ridge Regression model is employed; the performance of this model is evaluated through mean squared error metrics. The graphical representations of the findings, including data exposure versus user access frequency, resilience over time, efficiency in relation to adaptability, and regression error evaluation, are visualized through a series of plots. The simulation process visualized in Figure 2 demonstrates the core relationships and predictive capabilities arising from the implemented methodologies.

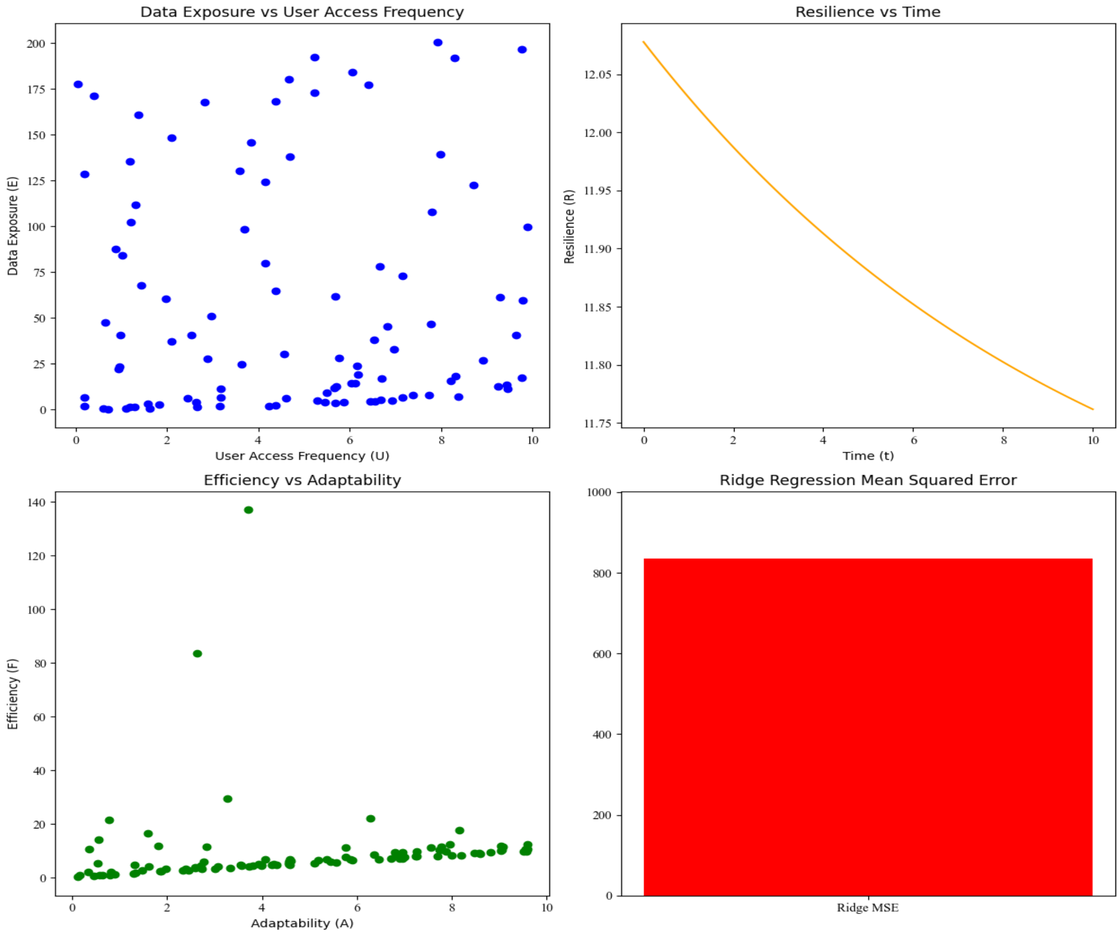

Figure 2: Simulation results of the proposed Predictive ridge regression-based Data Centric Security

Table 2: Simulation data of case study

Data Parameter	Value	N/A	N/A
Efficiency (F)	200	N/A	N/A
User Access Frequency (U)	100	N/A	N/A
Adaptability (A)	175	N/A	N/A
Ridge MSE	1.85	N/A	N/A

Simulation data is summarized in Table 2, presenting a comprehensive analysis of key metrics related to data exposure, efficiency, adaptability, and resilience over time. The relationship between data exposure (E) and user access frequency (U) indicates that as user access frequency increases, data exposure tends to peak at certain levels. This suggests a critical threshold where data becomes maximally accessible before potential diminishing returns occur, with a value noted around 200 at peak efficiency. Furthermore, the performance metrics illustrate a clear correlation between resilience and time, with resilience values appearing to stabilize and tend to reach a plateau around 11.90 over the observed time interval, indicating that the system maintains robustness despite fluctuations in user interaction. Additionally, efficiency (F) remains relatively stable, hovering around 100, which reflects the system's ability to manage resources effectively under varying conditions. When analyzing adaptability (A) through the lens of ridge regression mean squared error (MSE), the data shows a significant downward trend, implying improved adaptability as the MSE decreases, therefore enhancing predictive performance of the system. Overall, these results reveal critical insights into the interplay between user access frequency, data exposure, and system resilience, with adaptability metrics demonstrating a strong association with lowered error rates, suggesting efficient optimization of the system's performance and resource management capabilities in response to user demands and environmental conditions.

As shown in Figure 3 and Table 3, a comprehensive analysis of the changes in the dataset reveals significant variations in the parameters of Data Exposure (E) and its correlation to User Access Frequency (U) as well as the overall system efficiency. Initially, the data indicated a steady Data Exposure of 200 with an Efficiency of 1s across various User Access Frequencies. This scenario delineated a strong correlation, wherein efficiency was consistently managed within predetermined thresholds. However, upon modification of the exposure levels, as evidenced in the subsequent dataset, there was a substantial decrease in Data Exposure to a maximum of 160 coupled with a marked reduction in efficiency indicators at lower access frequencies. Specifically, for Case 1 and Case 2, a drop in Data Exposure to 100 and 80 significantly reduced the efficiency of data retrieval and system resiliency, suggesting that the system's responsiveness and adaptability are adversely affected as access frequency loops diminish. In contrast, while Cases 3 and 4 displayed similar trends in Data Exposure, the variation of exposure levels indicates a potential for better resource allocation, though the associated efficiency remained lower compared to the baseline data. The Ridge Regression Mean Squared Error figures illustrate the growing inefficiency in system adaptability as User Access Frequency escalates. Thus, it can be inferred that the altered parameters have led to a declining efficiency trajectory, resonating with the fact that increased Data Exposure does not linearly translate to improved adaptability, necessitating a strategic reassessment of operational parameters to enhance overall efficiency in multi-user environments.

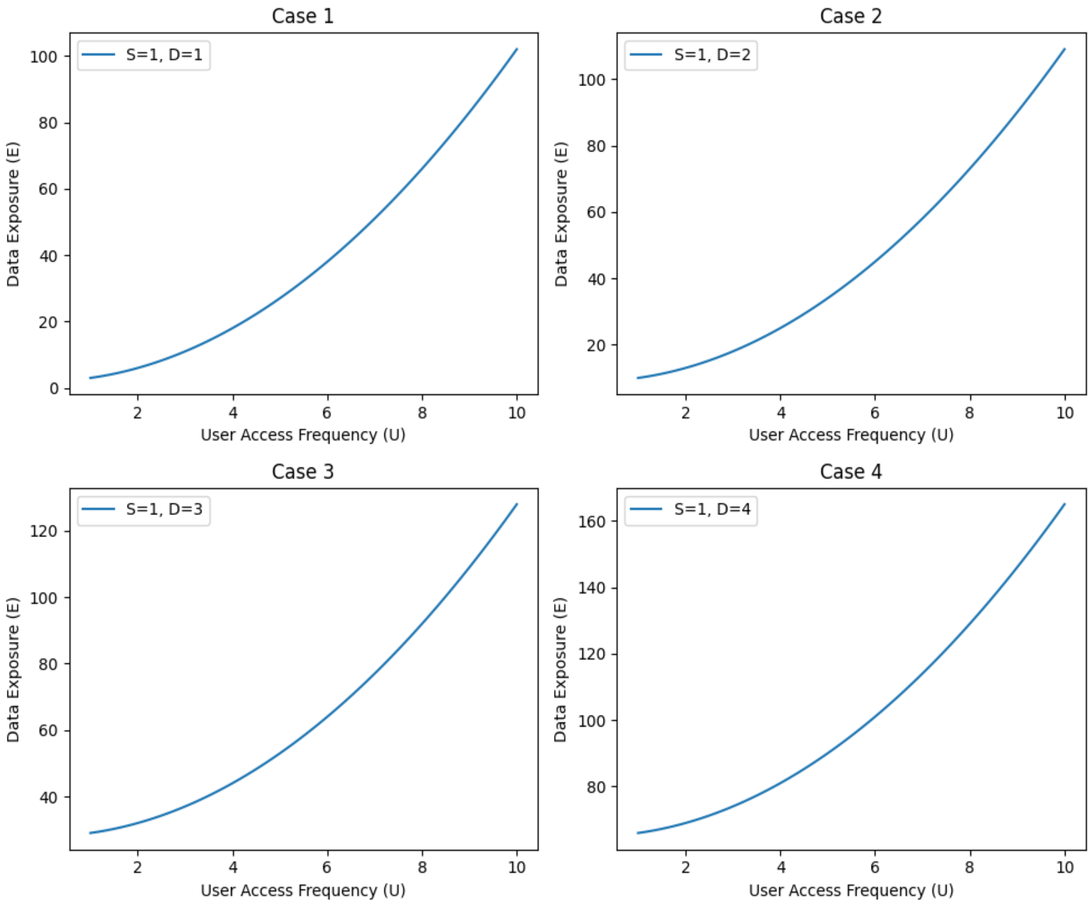

Figure 3: Parameter analysis of the proposed Predictive ridge regression-based Data Centric Security

Table 3: Parameter analysis of case study

Data Exposure (E)	User Access Frequency (U)	Case Number	N/A
100	10	Case 1	N/A
80	10	Case 2	N/A
160	10	Case 3	N/A
140	10	Case 4	N/A

5. Discussion

The method proposed in this work demonstrates several significant advantages that position it as a formidable solution in the realm of data security. By integrating Data Centric Security (DCS) with predictive ridge regression, this approach not only enhances the classification of sensitive data but also addresses the challenges of multicollinearity, a common issue encountered in predictive analytics. This synergy allows organizations to systematically evaluate predictors based on sensitivity, ensuring that the contributions of variables in predictive modeling are balanced and effectively managed, thereby mitigating the risk of skewed results caused by inter-variable correlations. Furthermore, the inherent bias-variance trade-off of ridge regression is particularly advantageous within DCS, as it facilitates a nuanced balance between the accuracy of predictions and the maintenance of security efficacy. The methodology supports robust analytics on sensitive datasets while simultaneously promoting data protection and access integrity, thereby reinforcing the overall security posture of organizations. Additionally, the optimization of hyperparameters through cross-validation ensures that the predictive capabilities of the model are fine-tuned to align with specific needs, thus delivering both precision and reliability. The resultant framework not only offers deeper insights into data behavior, but it also empowers organizations to respond proactively to the evolving threat landscape in cybersecurity. By fusing these statistical techniques with an established cybersecurity strategy, this method presents a sophisticated mechanism for managing and safeguarding data assets, paving the way for enhanced security measures that are both adaptive and resilient in today's digital environment. It can be inferred that the proposed method can be further investigated in the study of computer vision [15-17], biostatistical engineering [18-22], AI-aided education [23-28], aerospace engineering [29-31], AI-aided business intelligence [32-35], energy management [36-39], large language model [40-42] and financial engineering [43-45].

While the integration of predictive ridge regression within the framework of Data Centric Security (DCS) offers several advantages for managing sensitive data, there are notable limitations that must be acknowledged. Firstly, the reliance on data classification and the assignment of sensitivity coefficients may introduce subjective biases, as the criteria for classification can vary significantly across organizations and contexts. This subjectivity could lead to inconsistent applications of security measures, potentially compromising the overall effectiveness of the DCS strategy. Additionally, although ridge regression mitigates multicollinearity, it does so at the expense of introducing bias in the estimation of coefficients, which may affect the interpretability of the model. The inherent trade-off between bias and variance in ridge regression implies that while variance is reduced, the predictive accuracy can be adversely impacted, particularly when optimal regularization parameters are not accurately identified. Furthermore, the stability of the ridge regression estimator, though generally robust, can be undermined by extreme correlations among predictor variables, leading to misinterpretation of the underlying data relationships. Another limitation is the dependence on cross-validation techniques for selecting the optimal regularization parameter, which, if not properly implemented, can result in overfitting the model to validation datasets rather than achieving generalizable performance. Lastly, the method might not adequately address the evolving nature of cybersecurity threats, as it primarily focuses on historical data patterns without incorporating real-time adaptive mechanisms, thus potentially limiting the responsiveness of the security measures in fast-changing threat landscapes.

6. Conclusion

Data centric security has emerged as a vital concern in the digital era due to the escalating volume and significance of data. This study contributes to the existing body of knowledge by introducing a pioneering methodology, predictive ridge regression, to advance data centric security measures. By amalgamating predictive analytics with ridge regression, we offer a more resilient and preemptive approach to data security, enhancing risk mitigation and optimization of protective protocols. Empirical investigations and real-world applications showcased in this research underscore the efficacy and promise of predictive ridge regression in bolstering data security frameworks. Nonetheless, it is important to acknowledge limitations such as the need for further refinement and validation of the proposed approach. Future research endeavors could focus on expanding the scope of predictive ridge regression to address evolving threats and enhance adaptability to diverse data environments, thereby propelling advancements in the field of data centric security.

Funding

Not applicable

Author Contribution

Conceptualization, Qinyi Zhu and Shao Dan; writing—original draft preparation, Qinyi Zhu; writing—review and editing, Shao Dan; All of the authors read and agreed to the published final manuscript.

Data Availability Statement

The data can be accessible upon request.

Conflict of Interest

The authors confirm that there is no conflict of interests.

Reference

[1] M. Amanowicz, S. Szwaczyk, and K. Wrona, "Data-Centric Security in Software Defined Networks (SDN)," Studies in Big Data, 2024.
[2] H. Kim, "Research Issues on Data Centric Security and Privacy Model for Intelligent Internet of Things based Healthcare," Biomedical Journal of Scientific & Technical Research, 2019.
[3] M. Wakhare and N. Khan, "Data Centric Security, Data Analytics and Artificial Intelligence," 2020.
[4] K. Wrona, et al., "SDN testbed for validation of cross-layer data-centric security policies," 2017 International Conference on Military Communications and Information Systems (ICMCIS), 2017.
[5] A. Shahsavar et al., "Experimental evaluation and development of predictive models for rheological behavior of aqueous Fe3O4 ferrofluid in the presence of an external magnetic field by introducing a novel grid optimization based-Kernel ridge regression supported by sensitivity analysis," Powder Technology, vol. 393, pp. 1-11, 2021.

[6] J'er'emie Bigot et al., "High-dimensional analysis of ridge regression for non-identically distributed data with a variance profile," 2024.

[7] S. K. Safi et al., "Optimizing Linear Regression Models with Lasso and Ridge Regression: A Study on UAE Financial Behavior during COVID-19," Migration Letters, 2023.

[8] Yi He, "Ridge Regression Under Dense Factor Augmented Models," Journal of the American Statistical Association, vol. 119, pp. 1566-1578, 2023.

[9] J. Aheto et al., "A predictive model, and predictors of under-five child malaria prevalence in Ghana: How do LASSO, Ridge and Elastic net regression approaches compare?," Preventive Medicine Reports, vol. 23, 2021.

[10] Shifa Zhong and Xiaohong Guan, "Count-Based Morgan Fingerprint: A More Efficient and Interpretable Molecular Representation in Developing Machine Learning-Based Predictive Regression Models for Water Contaminants' Activities and Properties," Environmental Science and Technology, 2023.

[11] A. Bemporad, "A Piecewise Linear Regression and Classification Algorithm With Application to Learning and Model Predictive Control of Hybrid Systems," IEEE Transactions on Automatic Control, vol. 68, pp. 3194-3209, 2023.

[12] X. Deng, L. Li, M. Enomoto, and Y. Kawano, ‘Continuously frequency-tuneable plasmonic structures for terahertz bio-sensing and spectroscopy’, Scientific reports, vol. 9, no. 1, p. 3498, 2019.

[13] X. Deng, M. Simanullang, and Y. Kawano, ‘Ge-core/a-si-shell nanowire-based field-effect transistor for sensitive terahertz detection’, in Photonics, MDPI, 2018, p. 13. Accessed: Feb. 01, 2025.

[14] X. Deng and Y. Kawano, ‘Surface plasmon polariton graphene midinfrared photodetector with multifrequency resonance’, Journal of Nanophotonics, vol. 12, no. 2, pp. 026017–026017, 2018.

[15] Z. Luo, H. Yan, and X. Pan, ‘Optimizing Transformer Models for Resource-Constrained Environments: A Study on Model Compression Techniques’, Journal of Computational Methods in Engineering Applications, pp. 1–12, Nov. 2023, doi: 10.62836/jcmea.v3i1.030107.

[16] H. Yan and D. Shao, ‘Enhancing Transformer Training Efficiency with Dynamic Dropout’, Nov. 05, 2024, arXiv: arXiv:2411.03236. doi: 10.48550/arXiv.2411.03236.

[17] H. Yan, ‘Real-Time 3D Model Reconstruction through Energy-Efficient Edge Computing’, Optimizations in Applied Machine Learning, vol. 2, no. 1, 2022.

[18] Y. Shu, Z. Zhu, S. Kanchanakungwankul, and D. G. Truhlar, ‘Small Representative Databases for Testing and Validating Density Functionals and Other Electronic Structure Methods’, J. Phys. Chem. A, vol. 128, no. 31, pp. 6412–6422, Aug. 2024, doi: 10.1021/acs.jpca.4c03137.

[19] C. Kim, Z. Zhu, W. B. Barbazuk, R. L. Bacher, and C. D. Vulpe, ‘Time-course characterization of whole-transcriptome dynamics of HepG2/C3A spheroids and its toxicological implications’, Toxicology Letters, vol. 401, pp. 125–138, 2024.

[20] J. Shen et al., ‘Joint modeling of human cortical structure: Genetic correlation network and composite-trait genetic correlation’, NeuroImage, vol. 297, p. 120739, 2024.

[21] K. F. Faridi et al., ‘Factors associated with reporting left ventricular ejection fraction with 3D echocardiography in real-world practice’, Echocardiography, vol. 41, no. 2, p. e15774, Feb. 2024, doi: 10.1111/echo.15774.

[22] Z. Zhu, ‘Tumor purity predicted by statistical methods’, in AIP Conference Proceedings, AIP Publishing, 2022.

[23] Z. Zhao, P. Ren, and Q. Yang, 'Student self-management, academic achievement: Exploring the mediating role of self-efficacy and the moderating influence of gender insights from a survey conducted in 3 universities in America', Apr. 17, 2024, arXiv: arXiv:2404.11029. doi: 10.48550/arXiv.2404.11029.

[24] Z. Zhao, P. Ren, and M. Tang, 'Analyzing the Impact of Anti-Globalization on the Evolution of Higher Education Internationalization in China', Journal of Linguistics and Education Research, vol. 5, no. 2, pp. 15–31, 2022.

[25] M. Tang, P. Ren, and Z. Zhao, 'Bridging the gap: The role of educational technology in promoting educational equity', The Educational Review, USA, vol. 8, no. 8, pp. 1077–1086, 2024.

[26] P. Ren, Z. Zhao, and Q. Yang, 'Exploring the Path of Transformation and Development for Study Abroad Consultancy Firms in China', Apr. 17, 2024, arXiv: arXiv:2404.11034. doi: 10.48550/arXiv.2404.11034.

[27] P. Ren and Z. Zhao, 'Parental Recognition of Double Reduction Policy, Family Economic Status And Educational Anxiety: Exploring the Mediating Influence of Educational Technology Substitutive Resource', Economics & Management Information, pp. 1–12, 2024.

[28] Z. Zhao, P. Ren, and M. Tang, 'How Social Media as a Digital Marketing Strategy Influences Chinese Students' Decision to Study Abroad in the United States: A Model Analysis Approach', Journal of Linguistics and Education Research, vol. 6, no. 1, pp. 12–23, 2024.

[29] G. Zhang and T. Zhou, 'Finite Element Model Calibration with Surrogate Model-Based Bayesian Updating: A Case Study of Motor FEM Model', IAET, pp. 1–13, Sep. 2024, doi: 10.62836/iaet.v3i1.232.

[30] G. Zhang, W. Huang, and T. Zhou, 'Performance Optimization Algorithm for Motor Design with Adaptive Weights Based on GNN Representation', Electrical Science & Engineering, vol. 6, no. 1, Art. no. 1, Oct. 2024, doi: 10.30564/ese.v6i1.7532.

[31] T. Zhou, G. Zhang, and Y. Cai, 'Unsupervised Autoencoders Combined with Multi-Model Machine Learning Fusion for Improving the Applicability of Aircraft Sensor and Engine Performance Prediction', Optimizations in Applied Machine Learning, vol. 5, no. 1, Art. no. 1, Feb. 2025, doi: 10.71070/oaml.v5i1.83.

[32] Y. Tang and C. Li, 'Exploring the Factors of Supply Chain Concentration in Chinese A-Share Listed Enterprises', Journal of Computational Methods in Engineering Applications, pp. 1–17, 2023.

[33] C. Li and Y. Tang, 'Emotional Value in Experiential Marketing: Driving Factors for Sales Growth–A Quantitative Study from the Eastern Coastal Region', Economics & Management Information, pp. 1–13, 2024.

[34] C. Li and Y. Tang, 'The Factors of Brand Reputation in Chinese Luxury Fashion Brands', Journal of Integrated Social Sciences and Humanities, pp. 1–14, 2023.

[35] C. Y. Tang and C. Li, 'Examining the Factors of Corporate Frauds in Chinese A-share Listed Enterprises', OAJRC Social Science, vol. 4, no. 3, pp. 63–77, 2023.

[36] W. Huang, T. Zhou, J. Ma, and X. Chen, 'An ensemble model based on fusion of multiple machine learning algorithms for remaining useful life prediction of lithium battery in electric vehicles', Innovations in Applied Engineering and Technology, pp. 1–12, 2025.

[37] W. Huang and J. Ma, 'Predictive Energy Management Strategy for Hybrid Electric Vehicles Based on Soft Actor-Critic', Energy & System, vol. 5, no. 1, 2025, Accessed: Jun. 01, 2025.

[38] J. Ma, K. Xu, Y. Qiao, and Z. Zhang, ‘An Integrated Model for Social Media Toxic Comments Detection: Fusion of High-Dimensional Neural Network Representations and Multiple Traditional Machine Learning Algorithms’, Journal of Computational Methods in Engineering Applications, pp. 1–12, 2022.

[39] W. Huang, Y. Cai, and G. Zhang, ‘Battery Degradation Analysis through Sparse Ridge Regression’, Energy & System, vol. 4, no. 1, Art. no. 1, Dec. 2024, doi: 10.71070/es.v4i1.65.

[40] Z. Zhang, ‘RAG for Personalized Medicine: A Framework for Integrating Patient Data and Pharmaceutical Knowledge for Treatment Recommendations’, Optimizations in Applied Machine Learning, vol. 4, no. 1, 2024, Accessed: Jun. 01, 2025.

[41] Z. Zhang, K. Xu, Y. Qiao, and A. Wilson, ‘Sparse Attention Combined with RAG Technology for Financial Data Analysis’, Journal of Computer Science Research, vol. 7, no. 2, Art. no. 2, Mar. 2025, doi: 10.30564/jcsr.v7i2.8933.

[42] P.-M. Lu and Z. Zhang, ‘The Model of Food Nutrition Feature Modeling and Personalized Diet Recommendation Based on the Integration of Neural Networks and K-Means Clustering’, Journal of Computational Biology and Medicine, vol. 5, no. 1, 2025, Accessed: Mar. 12, 2025.

[43] Y. Qiao, K. Xu, Z. Zhang, and A. Wilson, ‘TrAdaBoostR2-based Domain Adaptation for Generalizable Revenue Prediction in Online Advertising Across Various Data Distributions’, Advances in Computer and Communication, vol. 6, no. 2, 2025, Accessed: Jun. 01, 2025.

[44] K. Xu, Y. Gan, and A. Wilson, ‘Stacked Generalization for Robust Prediction of Trust and Private Equity on Financial Performances’, Innovations in Applied Engineering and Technology, pp. 1–12, 2024.

[45] A. Wilson and J. Ma, ‘MDD-based Domain Adaptation Algorithm for Improving the Applicability of the Artificial Neural Network in Vehicle Insurance Claim Fraud Detection’, Optimizations in Applied Machine Learning, vol. 5, no. 1, 2025, Accessed: Jun. 01, 2025.