**JOURNAL OF SOFT COMPUTING AND DATA MINING**
e-ISSN: 2716-621X

**JSCDM**

Vol. 6 No. 3 (2025) 170-182
https://publisher.uthm.edu.my/ojs/index.php/jscdm

# An Agent-Based Decision Fusion Regression Model for Analyzing Cyberspace Users' Engagement in Informatics Blogs

**Ali Mohammed Saleh Ahmed[1]*, Israa Nazeeh[2], Firas Mohammed Aswad[2]**

[1] *College of Education for Pure Sciences, University of Diyala, Diyala, IRAQ*
[2] *College of Basic Education, University of Diyala, Diyala, IRAQ*

*Corresponding author: dr.alimahmed@uodiyala.edu.iq
DOI: https://doi.org/10.30880/jscdm.2025.06.03.012

**Article Info**

Received: 6 August 2025
Accepted: 13 November 2025
Available online: 30 December 2025

**Keywords**

Information systems, blog informatics, decision fusion, regression, and ensemble modelling

**Abstract**

The Internet provides direct interactive communication technology, such as blogs, which enable users to post comments and share links to maintain interactions with fellow users. In modern information systems, a blog platform allows individuals or groups to write their opinions on a specific topic. Broad informatics blogs provide spaces for people to share ideas and exchange views on various issues. Driven by informatics-based learning methods, these methods are used to detect and analyze users' tendencies in these blog posts. Blog features exhibit a direct correlation with the extent to which bloggers adhere to both national cultural patterns and social and political developments in their respective countries. This paper proposes an agent-based decision fusion regression (ADFR) model for analyzing users' engagement with blog posts in cyberspace. The ADFR model is tested using the BlogFeedback dataset, which has been used in several previous research studies. The model results are compared with three machine learning algorithms: Linear Regression (LR), Decision Tree (c4.5), and Decision Forest (DF) for performing blog post analysis tasks. The $R2$ score reaches 0.0695 with ADFR, while DT shows only 0.0002, DF reveals 0.0104, and LR demonstrates 0.0544 in blog comment prediction. The data analysis indicates that ADFR delivers the highest predictive power and outstanding generalization.

## 1. Introduction

The internet remains a dominant influence in modern human lifestyles. Most users connect their electronic devices to the internet in both personal and professional settings. Most people utilize the boundless nature of the Internet to search for knowledge or share it, making it their primary platform for information. Social media platforms such as Instagram, Facebook, Snapchat, and YouTube operate as modern domains for distributing information that includes regular blog updates. Social media remains an essential part of people's everyday lives [1]. People value blogging content more deeply than other social media platforms because daily reading and writing activities drive the growth of individuals who care about blogging [2]. The internet makes blogs one of the most frequently used online resources for creating virtual space environments [3].

Blogging can be difficult to start due to the specified content, but once the blogger has issues that he or she wants to share and focus on. The blog site will create useful information for the readers. Style and pattern in

blogging can differ for each person and each region or country. The current environment in which the writers live mostly affects how blogging is created [2]. Various elements determine how much people choose to use blogging systems. Research shows that the social, political, and cultural backgrounds of various countries affect their blogging patterns [4]. Blog informatics is a fundamental tool for understanding how users connect with web-based information systems in cyberspace. Organizations gain important user insights about preferences and engagement patterns when systematically analyzing user activity patterns, including posts, comments, and browsing patterns.

The analysis of online community dynamics supports organizations in adjusting content based on user needs, thereby improving overall user satisfaction. Additionally, the mass media operate under government-imposed restrictions in certain countries that silence particular content through media censorship systems. Research shows that evaluating user content helps companies detect behavioral trends for better strategic decisions that enhance user satisfaction rates and maintain user loyalty [1, 5]. Blogging patterns require identification when investigating specific regions and countries. Exploring country-specific blogging patterns enables researchers to identify the present matters that develop within those areas [6, 7]. The blogging site enables citizens in that geographical area to display their thoughts regarding the subject. Some organizations and parties utilize writing and discussion patterns as useful data to understand regional interests and strengths, together with weaknesses. Organizations operating within the region can use blogging trends to develop strategic programs and establish social, political, economic, and cultural pathologies, policies, and plans resulting in appropriate strategies.

This study aims to propose an agent-based decision fusion regression (ADFR) model to identify users' blogging tendency anticipation. The main contribution of this study is developing the ADFR model for improving web-based information system analysis of blog post user engagement. The proposed model combines the decision fusion of Linear Regression (LR), Decision Tree (c4.5), and Decision Forest (DF) with agent-based modeling to process diverse user behavior information for enhanced engagement prediction accuracy. The model is implemented using R Studio, which contains different types of data mining algorithms and tasks that help analyze the blog's data. The ADFR model is tested and evaluated using different algorithms and data mining metrics to make the prediction. The test results show that the model based on regression-based ensemble learning improves the reliability of user interaction patterns to help blogging platforms and content managers achieve better strategy optimization.

This research contributes to the ADFR model, which provides intelligent web analytics with a dependable system for analyzing digital user behaviors on a large scale. The ADFR model uses decision-making agents to create a new approach in ensemble regression learning. The system provides predictive accuracy by nurturing a learning process that adapts its strategic decision parameters through response feedback. The ensemble approach maintains data variability, resistance, and non-linearity robustness through its adaptive weight assignment mechanism. Reinforcement learning within the fusion process generates an adaptive model that constantly improves its performance, which fits nicely in real-world applications of blog post analysis.

The following sections follow this organization. This section reviews the paradigm related to machine learning algorithm use established by other similar research papers that serve as benchmarking references. This research examines both the techniques implemented in previously published papers. Section 3 presents the methodology, which step by step explains how to use the algorithm to perform the regression, along with the dataset and evaluation metrics. Section 4 presents the results obtained from the methodology used and discusses the results. Finally, for Section 5, the project studies are concluded, and future works are added to the research.

## 2. Related Work

In the Related Work section, we review how existing methods of studying cyberspace user engagement in blog posts have utilized agent-based models, decision fusion techniques, and regression-based ensemble learning. In prior studies, data mining and machine learning have been applied to predict user behavior using features like click-through rate, post-interaction, and comment sentiment. For example, many approaches based on traditional content filtering or collaborative filtering have been adopted in engagement prediction, but usually with limited dynamic adaptability to user behavior. The recent evolution of agent-based modeling has allowed for a more granular simulation of user interaction, and decision fusion frameworks more accurately predict user behavior through the integration of separate analytical models. Nevertheless, the scalability, real-time adaptability, and manageability of noisy user-generated data remain a challenge that this study attempts to overcome by applying an agent-based decision fusion regression model for more precise engagement analysis.

Alghobiri [8] performed a classification method evaluation involving Naïve Bayes, C4.5, and Support Vector Machine (SVM) algorithms. Their main consideration was to select data based on their dimensions, along with attribute definitions. The SVM algorithm proved superior to competing types of classification algorithms according to accuracy, precision, and F-Measure evaluation metrics.

Asim et al. [9] analyzed three main machine learning approaches, which included Lazy learning algorithms, Eager learning algorithms, and Ensemble techniques for data classification purposes. The researchers determined

which methods worked best at enhancing data classification accuracy in all available techniques. Data classification using the Random-Forest and Nearest-Neighbor Classifier (IB1) achieves high precision levels in identifying nominal data with 85% accuracy and 84.8% precision.

The C4.5 decision tree classifier serves as the tool of Masetic et al. [10] for malignant website detection. The evaluation measures precision and sensitivity alongside specificity to assess the test results using the ROC curve calculation. A C4.5 decision tree classifier successfully identifies malicious websites with 96.5% accuracy.

The article by Samsudin et al. [11] advocates the use of Random Forest algorithms to classify blogs through the established dataset. The essential objective focuses on achieving an accurate determination of whether bloggers writing blogs maintain a professional status or a seasonal status. According to the experimental results, blog classification using the Random Forest achieved an 11% higher ROC Area than C4.5 and a 6% increase in performance compared to K-NN algorithms. The recall value for Random Forest algorithms exceeds C4.5 and K-NN algorithms by 97%.

The study by Dias and Diasb [12] explored how linear regression, neural network regression, and decision forest regression-based approaches were used to generate accurate monthly ad revenues from blogs by analyzing Google Analytics and Google AdSense statistics. The Decision Forest Regression produced the optimal model because it achieved more than 70% accuracy.

The study by Chen et al. [13] enhanced the recurrent neural network algorithm for analyzing Chinese microblog text. The proposed method applies feature fusion operations between shallow and deep learning approaches to extract features from microblog texts. The recurrent neural network model ensures text sequence analysis via its LSTM implementation to understand internal text elements' correlations. The announced class prediction ability reaches 85.04%, thus surpassing traditional SVM models with shallow learning features by 3.17%. The proposed method showed its effectiveness through the obtained results.

Simaki et al. [14] conducted a study about using various regression methods to develop the age estimation function. Scientists conducted tests on various machine learning programs to determine their capability in carrying out this functionality. The experimental assessment used forty-two text features. According to the obtained results, the combination of the Bagging algorithm using RepTreebase learner achieved the most optimal site user age prediction with a Mean Absolute Error (MAE) value of 5.44 and Root-Mean Square Error (RMSE) of approximately 7.14.

A new DUAPM model, according to Yang et al. [15], enables discovering and modeling microblogging user behavior, which subsequently allows the prediction of activities for CPSS applications to detect spam and fake accounts. The implementation of their method depends on three key characteristics, which include personal information, social relationships, and user interaction. The DUAPM model achieved better prediction accuracy than traditional models, which included logical regression and random forest algorithms during the assessment of 3,621 Sina Weibo users monitored over 20 weeks.

According to Mostafa et al. [16], the diagnosis of Parkinson's disease was achieved through the independent study of three classification methods, including a Decision Tree, along with Naïve Bayes and Neural networks. Choose the most suitable method from among the three identified options. The proposed solution involves evaluating the performance of the three methods in their problem-solving approach. Both Decision Tree and Neural Network provide superior results to Naïve Bayes by reaching 91.63% accuracy and 91.01% accuracy, respectively; thus, the researchers propose using both Decision Tree and Neural Network for datasets with comparable properties.

Woo et al. [17] identified Korean keywords for the identification of outbreaks of influenza from social media info. The researchers followed the steps: selecting initial keywords, preprocessing for keyword time series generation, selecting optimal characteristics to create and validate models using the least absolute shrinkage and selection operator, and then performing Support Vector Machine (SVM) and Random Forest Regression (RFR). Fifteen selected keywords proved most effective for detecting epidemiological influenza occurrences without any concentration preference between Twitter and blog sources. The model suits multiple country settings, linguistic needs, and infectious disease research while working with various social media platforms.

Various machine learning and regression techniques have been demonstrated in the reviewed studies to effectively analyze and predict aspects of blog-related data. According to Alghobiri [8], SVM is the best classification algorithm to use when comparing algorithms on data. In the case of Asim et al. [9], factors affecting the professionalism of blogging, IB1, and Random Forest were top performers, being accurate and precise above 84%. C4.5 decision trees are successfully used by Masetic et al. [10] to detect malicious websites. In the study of Samsudin et al. [11], which analyzed professional bloggers, Random Forest is the best among the others, with 92% ROC, 97% recall, and 88% precision. Using regression models, Dias and Dias [12] predict blog ad revenue, with the Decision Forest Regression coming in at a hair above 70%. In Simaki et al. [14], regression models were assessed to estimate blogger age, which was reported to be the most accurate approach. This work presents how these machine learning models are useful to enhance prediction and decision-making accuracy for blog informatics, including classification and regression approaches.

## 3. Research Methods

The study is intended to propose an automated method to analyze cyberspace users' engagement in informatics blogs by performing regression operations to analyze data mining components. Statistical methods for identifying variable associations function through regression methodology. Among machine learning techniques, the method predicts event outcomes through variable relationships derived from data collection. The application of this technique exists for forecasting purposes as well as time series modeling, climate prediction, and detecting variable causal effects [18]. Through regression analysis, users receive a measurement of the impact of multiple independent factors on dependent variables [19]. The most basic regression approach is linear regression, while multiple regression represents the more complex method [20]. The proposed model for regression operations is an agent-based decision fusion regression (ADFR). ADER utilizes three algorithms known as Linear Regression (LR), Decision Tree (c4.5), and Decision Forest (DF) within its framework. The R Machine Learning instrument served to perform our tests through a ten-fold validation approach that handled training and testing stages.

### 3.1 BlogFeedback Dataset

BlogFeedback Dataset is a large real-world regression dataset, donated to the UCI Machine Learning Repository on May 28, 2014, aimed at modeling and predicting user interaction in terms of the number of future comments on blog posts [21]. It has 60,021 samples, which are individual blog posts, and characteristics are taken out of the textual, structural, and temporal characteristics of online content. The data set contains different statistical measures, including averages, standard deviations, minimums, maximums, and median data about the blog sources and comment activity. Other qualities include the number of comments made before the baseline, the frequency of comments in a few windows before and after the baseline, the change in comment behaviour as time went by, the number of trackbacks, and the temporal gap between the time that the comments were made and the time at which the baseline was made [22]. The characteristics of the dataset make it well-suited for developing predictive models to study how users engage with online information.

In addition to statistical metadata, the data includes a bag-of-words textual representation of 200 features, as well as structural metadata in the number of parent pages and comment statistics on parent pages. It also has binary weekday indicators that capture the publication date as well as the baseline date, and using these models can capture temporal publishing [23]. Notably, the data do not have any missing values; thus, model training is not imputed. Combined with an endless target variable for the number of comments in the next 24 hours, the BlogFeedback dataset is highly dimensional (60 features) and serves as a rich benchmark for testing regression, tree-based, ensemble, neural, and fusion models, including the suggested ADFR. One specimen is an illustration of the dataset structure, including values of the number of tokens in the title, the number of tokens in the content, the number of hyperlinks, the number of pictures, the indicators of the weekday, and the number of targeted comments. Metadata and feature descriptions are also publicly accessible in the UCI Repository [21].

### 3.2 Agent-based Decision Fusion Regression Model

The BlogFeedback dataset is a regression problem that aims to forecast the number of comments on blog posts using extracted features. Regression techniques struggle to identify complex relationships between variables when dealing with statistical and textual data. The Agent-Based Decision Fusion Regression (ADFR) model uses an intelligent agent to fuse together a variety of prediction results of a number of regression models, such as C4.5 Decision Tree, Linear Regression, and Decision Forest. Its agent applies a weighted ensemble method to its fusion tasks, and the dynamic weight allocation is determined based on model performance, with the final decision optimized.

ADFR model design has three core components, including feature extraction and preprocessing, base regression models, and agent-based decision fusion. Step one involves data preprocessing: normalizing numerical data, encoding categorical data, and dropping redundant or correlated variables to ensure maximum model performance. Figure 1 indicates the processing steps of the ADFR model run cycle. The process begins by setting up all three regression models with equal weights, and subsequently, the individual predictions of the model are established. A weighted sum is then used to come up with the fused prediction through these results. It is then followed by the calculation of the Mean Absolute Error (MAE) to measure the accuracy of the prediction, which is then translated into a reward signal to support learning. The algorithm then computes the gradient of the weight updates with respect to the error, enabling the system to adjust each model's contribution via gradient descent. This process of adjustment is repeated across epochs, and the fusion model will improve over time to produce the most successful ensemble output.

J. of Soft Computing and Data Mining Vol. 6 No. 3 (2025) p. 170-182

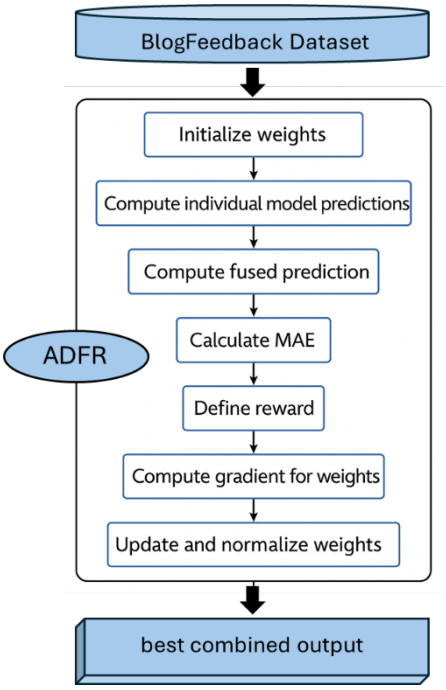


**Fig. 1** *The processing steps of the ADFR model*

The preprocessed BlogFeedback dataset trains the three base models independently through the C4.5 Decision Tree, Linear Regression, and Decision Forest. The predictive models compute the anticipated quantity of comments found in blog posts. After obtaining them, the agent implements a weighted fusion method to combine the multiple predictions from individual models. The last prediction results as *Y*final using the below formula:

$$Y_{final}=w_1 \cdot YC4.5+w_2 \cdot YLR+w_3 \cdot YDF \quad (1)$$

where YC4.5, YLR, and YDF are the three models' individual predictions, and the respective model weights w are optimized based on past performance; the sum of these weights is constrained such that:

$$w1+w2+w3=1 \quad (2)$$

An RL strategy operates within the agent to maximize weight optimization, so the method dynamically adjusts weights according to previous model execution results. Relevant models achieve higher weights through successful predictions, while less accurate models result in lower weights. The equation for updating weights through the reward function appears as follows:

$$R_t=-|Y_{final,t}-Y_{true,t}| \quad (3)$$

where Rt is the reward at time step t, $Y_{final}$ is the fused prediction, and $Y_{true,t}$ is the actual number of comments received. The optimization of the fusion process happens through agent learning which results in continuously decreased prediction errors during the process.

Ensemble approaches enable the beneficial features of different regression models to work together [11]. Strong interpretability through the C4.5 Decision Tree allows it to identify non-linear relations. The linear relationships identified by linear regression provide a base understanding, while decision forests improve overall performance through multiple decision-making paths. The agent achieves better prediction accuracy and reliability by combining model outputs. A brief description of the three ML regression algorithms is as follows:

### 3.2.1 Linear Regression

The statistical method Linear regression (LR) evaluates linear patterns between dependent variables and particular independent variables. Linear regression functions as a basic machine learning tool that belongs to supervised learning methods while implementing regression models [7]. The measurement values for both independent and dependent variables show a linear relationship. The simple linear regression follows the general format of:

$$Y=\beta 0+\beta 1X+\epsilon \quad (4)$$

The computational efficiency, combined with explainable results, makes LR effective, while it lacks the capability to model complex nonlinear patterns [6]. The second type is multiple linear regression, which contains several independent variables. The advantages of this algorithm are: (a) It is easier to implement in various programming languages and more straightforward to understand [24]. (b) Used explicitly for predicting numeric values.

### 3.2.2 Decision Tree

The Decision Tree (DT) type (c4.5) functions as a decision tree algorithm to offer classification among the most powerful and recognized methods. The decision tree classifier produced by this algorithm enables decision-making from given data samples containing univariate or multivariate predictor variables. The data undergo recursive splitting during training to assess how well class separation occurs based on function values [25]. The three main benefits of the C4.5 Classification algorithm include (a) clear comprehension of tree diagram analysis results, (b) simple data sample recovery, and (c) quick computing period. Other classification techniques can process experimental data more easily, producing data that is available to the C4.5 tree. Calculations using this method are found to be faster than those using other classification methods, according to Abdollahzadeh & Gharehchopogh [26]. DT (C4.5) improves ID3 by constructing trees based on entropy and information gain. The method uses recursive partitioning to split data subsets and assign numerical values to terminal nodes during the regression task. The model provides interpretability and the capability to detect non-linear patterns, but requires appropriate pruning techniques to avoid excessive overfitting [10].

### 3.2.3 Decision Forest

Decision Forest (DF) is an ensemble algorithm whereby a classifier is built by assembling a number of independent base classifiers. Decision forests aim to enhance the predictive performance of a single decision tree by training multiple decision trees and pooling their predictions [27]. Decision Forest (DF) is an ensemble learning technique that enhances the accuracy and the robustness of many decision trees. Each tree's two training pipelines include selecting random subsets of data and features, and averaging all the tree's output predictions. The adoption of this strategy will ensure a balance between accuracy, maintenance, and good interpretability, as well as high predictive performance, especially in structured data [11].

Two basic methods can be used to create a decision forest: (1) an ensemble method, such as AdaBoost, which can be applied with any base learning method, such as decision trees, or (2) an ensemble method specifically designed to use decision forests, such as Random Forest [27]. Loef et al. [28] conclude that Random Decision Forests are very efficient and fast in data analysis (classification, regression, clustering, and dimensionality reduction). The technology is compatible with graphics processing units (GPUs) to provide real-time operational results. Random Decision Forest ensemble method is an advancement of the identified features of decision trees, where better generalization properties are utilized [29].

### 3.2.4 Agent-based Decision Fusion

The software agent is critical towards making sure that the decision fusion process is optimized and flexible. The Agent-based Decision Fusion engages in weight adjustment of the models to produce end results that are more accurate than the outcome of the individual prediction models [30, 31]. The learning integration generates improvements to enable the system to self-improve with time. The ADFR model proposed uses the agent acting as a smart mediator to develop an optimal prediction that unites these predictions, LR (M1), DT (M2), and DF (M3). The proposed model has three key functions that are addressed by the agents, and they are: decision combination, adaptive weight management, and minimal error production.

1. Decision Fusion Mechanism: The decision fusion mechanism is the method used to make the agent combine base model predictions by using a weighted sum calculation method. When the predictive model operates with three models (Ms), it predicts the output $Y_{final}$ using the following calculation.

$$Y_{final}=w_1 \cdot YM_1+w_2 \cdot YM_2+w_3 \cdot YM_3 \quad (5)$$

2. Adaptive Weight Optimization: The system is adaptive in changing the weights by employing an adaptive process, which is based on the performance measures of the model. The agent, with its reward functionality, optimizes its weights to reduce prediction error. An MAE computation needs a distinct $Y_{true}$ as the true value.

$$E_t = \frac{1}{N}\sum_{i=1}^{N} |Y_{final,i} - Y_{true,i}| \quad (6)$$

To update weights, the agent applies a reward function:

$$R_t=-E_t \tag{7}$$

where a lower error results in a higher reward, encouraging the agent to prioritize models with better performance.

3. Error Minimization Strategy: During the update of the weights, the agent uses Gradient Descent and modifies every weight *wj* according to its error contribution:

$$w_i^{(t+1)} = w_j^{(t)} - \eta \frac{\partial E_t}{\partial W_j} \tag{8}$$

where $\eta$ is the learning rate controlling weight adjustments.

To calculate the levels of blog comments, Algorithm 1 of the Agent-based Decision Fusion algorithm is optimal to generate true predictions through a combination of multiple regression models, such as DT, LR, and DF. The initial weight distribution of the candidate models is equal to the agent's until the agent finishes a few rounds, after which it adjusts the weights based on the outcomes of the predictions. The output provides final predictions by weighting and combining all models' predictions. MAE is used to measure model accuracy, and the reward system rewards reduced error outputs. Weight updates are a gradient descent method used to adjust model weights based on their effect on minimizing prediction error. The normalized weight values reduce the risk of invalid distribution. The process is repeated with predetermined intervals so that the ensemble model could constantly adapt to the actual changes in data performance. The process produces optimized weights of the model coupled with a state-of-the-art prediction of blog comments operating in a system that does not compromise robustness and adaptability [32, 33].

***Algorithm 1.*** *Agent-based Decision Fusion*

**Input**: A set of regression models $\{M_1, M_2, M_3\}$, feature set $X$ used for prediction, ground truth labels $Y_{true}$, learning rate $\eta$ is a small constant for weight adjustment, and epochs, which represents the number of iterations for weight optimization.

**Output**: Optimized weights, a set of weights $w1$, $w2$, and $w3$ assigned to each model, and final prediction of the fused regression output $Y_{final}$.

**Process**: Initialize weights with equal values to all models, such that $w1 = w2 = w3 = \frac{1}{|M|}$;

**For** each iteration (epoch) t=1 to T, DO:

- Compute individual model predictions: YMi =Mi (X), ∀i∈{1,2,3};
- Compute the fused prediction using weighted sum: $Y_{final} = w_1YM_1+w_2YM_2+w_3YM_3$;
- Calculate MAE as: $Et=\frac{1}{N}\sum_{i=1}^{N}|Y_{final,i} - Y_{true,i}|$;
- Define the reward to guide weight adjustments: $R_t=-E_t$;
- Compute the gradient for weight updates: $\frac{\partial E_t}{\partial W_j} = (Y_{final} - Y_{true})YMi$ , ∀j∈{1,2,3};
- Update model weights using gradient descent: $w_i^{(t+1)} = w_j^{(t)} - \eta \frac{\partial E_t}{\partial W_j}$ ,∀j;
- Normalize weights to ensure their quality: w1+w2+w3=1;

End For;
Return the optimized weights W and final prediction $Y_{final}$;

This algorithm uses an iterative gradient-descent-based update rule to learn optimal weights of each of its multiple regression models. You start with three regression models $\{M_1, M_2, M_3\}$, and weight them equally at the beginning because no model is deemed to be superior initially. In every epoch, the algorithm uses the same input features $X$ to feed all three models and get their respective predictions. The combination of these predictions into one fused output $Y_{final}$ is then achieved by weighted summing up the three model outputs. In order to quantify the quality of this fused prediction, the algorithm computes the MAE between the fused prediction and the ground-truth labels $Y_{true}$; the negative of the MAE is a reward signal that tells how well the system was running. The algorithm then computes the sensitivity of the error in each model weight (i.e., error gradient). This gradient indicates whether the model's weight should be increased or decreased to reduce error in the next epoch. The weights are then optimized with gradient descent: the weights are pushed respectively in the opposite direction of the gradient with a small learning rate $\eta$ in order to prevent unstable changes. The weights are normalized after every update so that they add up to 1 to make sure they can be interpreted as proportions of contributions in the fusion. This cycle repeats itself in all epochs, each time making models that continuously decrease error more influential and weakening models. The algorithm ultimately yields the optimal weights $w_1$, $w_2$, and $w_3$, and the ultimate fused prediction *Y*final, which is an optimal joint performance obtained in all iterations. The agent continuously updates the model weights as a result of an error prediction performance evaluation procedure. The

adaptive decision fusion strategy enhances the impact of effective models on end predictions, resulting in more accurate predictions of blog comments.

## 3.3 Evaluation Metrics

The regression tasks have been evaluated using five metrics: mean absolute error, root mean squared error, relative absolute error, relative squared error, and the coefficient of determination. The following approaches serve to determine the results produced by the provided machine learning algorithms:

1. The Mean Absolute Error (MAE) calculates average values through linear mathematics by treating all difference scores equally. The formula for calculating mean absolute error is shown in Equation 9.

$$\mathrm{MAE} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} |\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i}| \tag{9}$$

2. The mathematical definition of Root Mean Squared Error (RMSE) calculates the square root of the mean square error because the calculation normalizes the error scale against the target scale. The calculation of root-mean-square error follows the expression in Equation 10.

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} (\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i})^2} = \sqrt{\mathrm{MSE}} \tag{10}$$

3. The Relative Absolute Error (RAE) is a measurement unit that varies among models and provides a measuring method. RAE is calculated using the formula below in Equation 11.

$$\mathrm{RAE} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} |\mathrm{p_i} - \mathrm{a_i}|}{\sum_{\mathrm{i}=1}^{\mathrm{n}} |\bar{\mathrm{a}} - \mathrm{a_i}|} \tag{11}$$

4. Model comparison through Relative Squared Error (RSE) becomes possible when working with models having errors that measure different units. Relative Squared Error calculation requires the use of Equation 12.

$$\mathrm{RSE} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{p_i} - \mathrm{a_i})^2}{\sum_{\mathrm{i}=1}^{\mathrm{n}} (\bar{\mathrm{a}} - \mathrm{a_i})^2} \tag{12}$$

5. The coefficient of determination ($R^2$) offers an overview of the explanatory strength of the regression model, derived based on sums-of-squares. In calculating the coefficient of determination, Equation 13 should be used.

$$\mathrm{R}^2 = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} (\hat{\mathrm{y}} - \bar{\mathrm{Y}})^2}{\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{Y} - \bar{\mathrm{Y}})^2} \tag{13}$$

## 4. Results and Discussion

The Adaptive Decision Fusion Regression (ADFR) model will be evaluated based on metric comparisons with the DT (C4.5), LR, and DF models. This would allow the bloggers to behave appropriately and, at the same time, professionally prove the relevance and importance of individual factors. Table 1 shows the performance analysis of Linear Regression (LR), Decision Tree (DT), Decision Forest (DF), and ADFR. Tests performed on the analysis used Mean Absolute Error (MAE) in combination with Root Mean Squared Error (RMSE), along with Relative Absolute Error (RAE), Relative Squared Error (RSE), and the *R*2 score.

**Table 1** *Comparison between the evaluated models*

| Algorithm | LR | DT | DF | ADFR |
|---|---|---|---|---|
| MAE | 0.3866 | 0.1834 | 0.3813 | 0.1714 |
| RMSE | 0.4519 | 0.3876 | 0.4429 | 0.3622 |
| RAE | 0.7873 | 0.4196 | 0.9801 | 0.4535 |
| RSE | 0.9459 | 0.9998 | 0.9895 | 0.9019 |
| **$R^2$** | 0.0544 | 0.0002 | 0.0104 | 0.0695 |

J. of Soft Computing and Data Mining Vol. 6 No. 3 (2025) p. 170-182

In terms of MAE, the ADFR model exhibits the most accurate prediction capability based on its lowest recorded MAE value of 0.1714. Figure 2 shows the reward progress of the agent of the ADFR model for 100 epochs of the training phase, in which the instant reward represents the recorded negative MAE value (Reward = −MAE). The reward reading is negative at the start of training and indicates a rather large prediction error of the first equal-weight fusion. The curve shows an upward trend as epochs advance, indicating that the model's weight-adjustment mechanism effectively reduces the MAE and brings the reward closer to zero. Though there are minor variations that can be seen as a result of noise in the updates of prediction, the general trend is steadily upwards, indicating constant convergence as well as constant learning. The reward will plateau to almost zero in the later epochs, meaning that the fused regression output is close to an improved and more precise state than it was during the initial training phases.

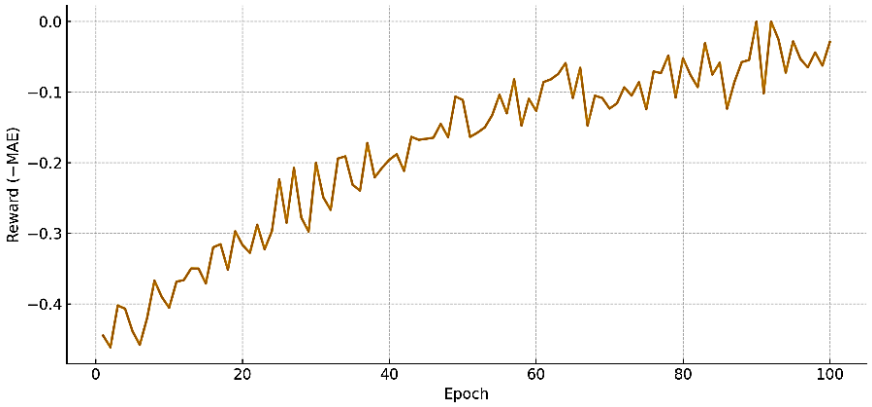


**Fig. 2** *The reward progress of the agent*

The performance of DT ranked second-best at 0.1834, but DF (0.3813) and LR (0.3866) showed notably higher errors. The lower value of MAE indicates that ADFR delivers superior accuracy when estimating blog comments. Consequently, the high error weighting in RMSE calculations indicates that ADFR achieved the lowest value of 0.3622 because it effectively handles extreme deviations in predicted outcomes. Among the compared models, the Decision Tree recorded an RMSE of 0.3876, while Decision Forest followed with 0.4429, and Linear Regression displayed 0.4519, indicating that ADFR brings more effective error reduction. Figure 3 shows a bar chart representation of the obtained results based on the evaluation metrics.

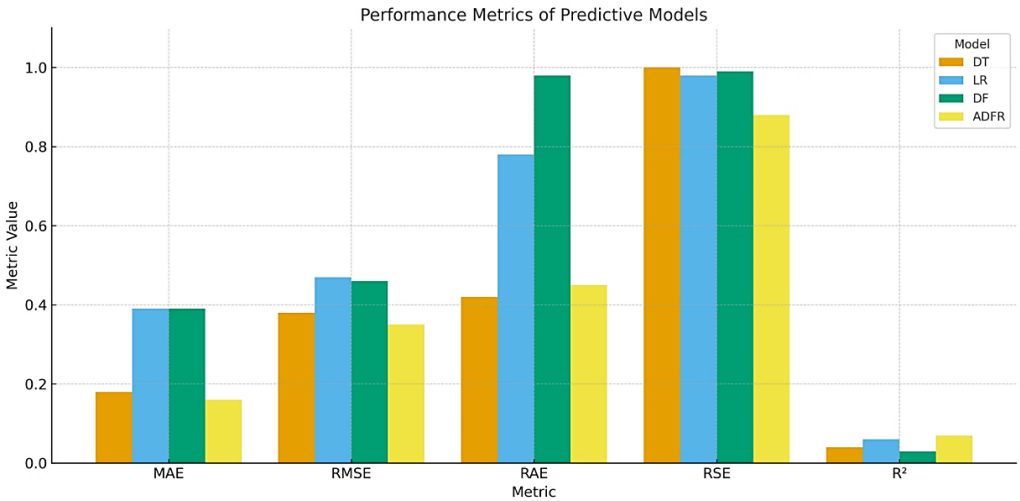


**Fig. 3** *A bar chart of the evaluation metrics*

ADFR also has a lower RAE (0.4535) compared to the Linear Regression (0.7873) and Decision Forest (0.9801) since it offers highly predictable and reliable models of prediction. The ADFR model only missed out on the best RAE performance, yet the other metrics show that this model is superior in generalizing relative to DT.

ADFR has the least value of RSE of 0.9019 as compared to DF of 0.9895 and DT of 0.9998. As much as LR has a lower RSE (0.9459), its performance on other important metrics achieves lower scores than the ADFR.

The *R*2 score (Coefficient of Determination) indicates the extent to which the model demonstrates its ability to describe target variable variation. Based on its calculated RSE value (0.9019), we can determine that the ADFR model outperforms the rest. The *R*2 score reaches 0.0695 with ADFR, while DT shows only 0.0002, DF reveals 0.0104, and LR demonstrates 0.0544 in blog comment prediction. The data analysis demonstrates that ADFR delivers the highest level of predictive power together with outstanding generalization potential. Figure 4 presents a radar (spider) chart that provides a clear visual comparison of all four models across all five metrics in a single circular chart.

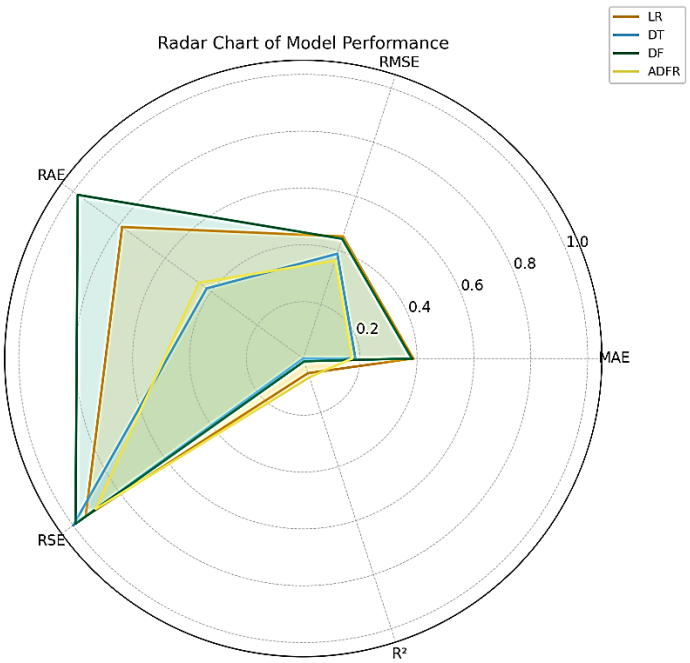


**Fig. 4** *A radar chart of the models' performance*

The radar chart provides a visual comparison of the four models (LR, DT, DF, and ADFR) across the five performance measures reported in the table (MAE, RMSE, RAE, RSE, and R2). The models are displayed as colored polygons that differ in the magnitude of their strengths and weaknesses: models with lower error values (MAE, RMSE, RAE, RSE) are located closer to the center of the corresponding axes, while those with higher values are farther out. As shown in the plot, the best error metrics for DT are achieved with the smallest polygon area, whereas the best error metrics for DF are achieved with the highest error values, resulting in the greatest spread on the chart. ADFR shows mixed results: on the one hand, it performs better across various measures than LR, but on the other hand, it does not achieve the minimum error rates of DT. Nevertheless, the overall results of this research demonstrate that the agent-based decision fusion mechanism in ADFR yields superior regression outcomes compared to competing methods, as assessed by performance evaluation criteria.

## 5. Conclusion

Machine learning regression algorithms employ various techniques to recognize bloggers' writing patterns. These algorithms are implemented based on the unique properties of the dataset they examine. The effectiveness of algorithms depends on several reasons, and one of which is the size of the data to be processed, as some algorithms perform better with small datasets than others. The study uses the BlogFeedback data to evaluate the effectiveness of the Agent-Based Decision Fusion Regression (ADFR) model for text analysis, compared with Linear Regression (LR), Decision Tree (c4.5), and Decision Forest (DF). According to this project, the most appropriate algorithm for small-sample-size datasets is the ADFR, followed by the DT (C4.5) models. The ADFR model performs best across all evaluation metrics for blog comment prediction and is more accurate, stable, and generalizable. The ADFR model is the best option, as it has minimized error rates (MAE, RMSE, RAE, and RSE) and achieved the highest two scores, thus ensuring that the regression task produces the best results on this data. According to our findings, the smart agent-based hybridization of DT, LR, and DF is effective in improving predictive capability. The research

presents a novel theory for analyzing feedback on blog posts using an agent-based ensemble approach. In subsequent studies of the topic, it is recommended to incorporate deep learning models, particularly LSTMs, into the fusion process to achieve the best predictive performance. Future research can be conducted on a larger sample to yield more robust outcomes. The growing nature of online interactions requires more advanced predictive models to cognize users' online engagement.

## Acknowledgement

We would like to express our heartfelt gratitude to the people in the Department of Computer Science, College of Education for Pure Sciences, and the College of Basic Education at the University of Diyala for their support of this research.

## Conflict of Interest

The authors declare that they have no conflicts of interest. The authors certify that the submission is an original work and is not under review at any other publication. All authors have seen and agree with the manuscript's contents, and there is no financial interest to report.

## Author Contribution

*The authors Ali M.S. Ahmed:* ***Conceptualization, Methodology, Software****. Israa N.: Writing –* ***original draft, Visualization, Investigation****. Firas M. Aswad:* ***Supervision, Software, Validation, and Writing****. Israa N:* ***review & editing****. All authors reviewed the results and approved the final version of the manuscript.*

## References

[1] Vankhede, P., & Kumar, S. (2024, February). Predictive Analytics for Website User Behavior Analysis. In 2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS) (pp. 1-6). IEEE. http://doi.org/10.1109/SCEECS61402.2024.10482298

[2] Ebrahimi, M., & Jampour, M. (2022). Identifying Cyberspace Users' Tendency in Blog Writing Using Data Mining Techniques. In Advances in Information Retrieval (pp. 75-86). Springer. http://doi.org/10.1007/978-981-19-2300-5_6

[3] Alsamadani, H. A. (2017). The Effectiveness of Using Online Blogging for Students' Individual and Group Writing. International Education Studies, 11(1), 44. http://doi.org/10.5539/ies.v11n1p44

[4] Gharehchopogh, F. S., Khaze, S. R., & Maleki, I. (2015). A New Approach in Bloggers Classification with Hybrid of K-Nearest Neighbor and Artificial Neural Network Algorithms. Indian Journal of Science and Technology, 8(3), 237. http://doi.org/10.17485/ijst/2015/v8i3/59570

[5] AbuSalim, S. W., Mostafa, S. A., Mustapha, A., Ibrahim, R., & Wahab, M. H. A. (2022). Identifying Cyberspace Users' Tendency in Blog Writing Using Machine Learning Algorithms. In Engineering Mathematics and Computing (pp. 81-92). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-19-2300-5_6

[6] Dalatu, P. I., Fitrianto, A., & Mustapha, A. (2016). A Comparative Study of Linear and Nonlinear Regression Models for Outlier Detection. Recent Advances on Soft Computing and Data Mining, 316–326. http://doi.org/10.1007/978-3-319-51281-5_32

[7] Geetha, M. C. S., Shanthi, I., & Raman, S. (2018). A survey and analysis on regression data mining techniques in agriculture. Int. J. Pure Appl. Math, 118(8), 341-347.

[8] Alghobiri, M. (2018). A comparative analysis of classification algorithms on diverse datasets. Engineering, Technology & Applied Science Research, 8(2), 2790-2795. https://doi.org/10.48084/etasr.1952

[9] Asim, Y., Shahid, A. R., Malik, A. K., & Raza, B. (2018). Significance of machine learning algorithms in professional blogger's classification. Computers & Electrical Engineering, 65, 461-473. https://doi.org/10.1016/j.compeleceng.2017.08.001

[10] Mašetic, Z., Subasi, A., & Azemovic, J. (2016). Malicious web sites detection using C4. 5 decision tree. Southeast Europe Journal of Soft Computing, 5(1).

[11] Samsudin, N. A., Mustapha, A., & Wahab, M. H. A. (2016). Ensemble classification of cyber space users tendency in blog writing using random forest. 2016 12th International Conference on Innovations in Information Technology (IIT). http://doi.org/10.1109/innovations.2016.7880046

[12] Diasa, D. S., & Diasb, N. G. J. (2018). Forecasting monthly ad revenue from blogs using machine learning. In The 3rd International Conference on Advances in Computing and Technology, ICACT.

[13] Chen, Q. H., Guo, Z., Sun, C. H., & Li, W. S. (2017, June). Research on Chinese micro-blog sentiment classification based on recurrent neural network. In Proceedings of 2nd International Conference on Computer Science and Technology, Guilin, China (pp. 859-867). http://doi.org/10.12783/dtcse/cst2017/12594

[14] Simaki, V., Aravantinou, C., Mporas, I., & Megalooikonomou, V. (2015). Automatic estimation of web bloggers' age using regression models. In Speech and Computer: 17th International Conference, SPECOM 2015, Athens, Greece, September 20-24, 2015, Proceedings 17 (pp. 113-120). Springer International Publishing.

[15] Yang, P., Yang, G., Liu, J., Qi, J., Yang, Y., Wang, X., & Wang, T. (2019). DUAPM: An Effective Dynamic Micro-Blogging User Activity Prediction Model towards Cyber-Physical-Social Systems. IEEE Transactions on Industrial Informatics, 1–1. http://doi.org/10.1109/tii.2019.2959791

[16] Mostafa, S. A., Mustapha, A., Khaleefah, S. H., Ahmad, M. S., & Mohammed, M. A. (2018, February). Evaluating the performance of three classification methods in diagnosis of Parkinson's disease. In International Conference on Soft Computing and Data Mining (pp. 43-52). Springer, Cham.

[17] Woo, H., Sung Cho, H., Shim, E., Lee, J. K., Lee, K., Song, G., & Cho, Y. (2017). Identification of Keywords from Twitter and Web Blog Posts to Detect Influenza Epidemics in Korea. Disaster Medicine and Public Health Preparedness, 12(03), 352–359. http://doi.org/10.1017/dmp.2017.84

[18] Moraffah, R., Sheth, P., Karami, M., Bhattacharya, A., Wang, Q., Tahir, A., ... & Liu, H. (2021). Causal inference for time series analysis: Problems, methods and evaluation. Knowledge and Information Systems, 63, 3041-3085. https://doi.org/10.1007/s10115-021-01621-0

[19] Huang, Y., Xu, W., Sukjairungwattana, P., & Yu, Z. (2024). Learners' continuance intention in multimodal language learning education: An innovative multiple linear regression model. Heliyon, 10(6). https://doi.org/10.1016/j.heliyon.2024.e28104

[20] Wang, J., & Zhu, S. (2023). A multi-factor two-stage deep integration model for stock price prediction based on intelligent optimization and feature clustering. Artificial Intelligence Review, 56(7), 7237-7262. https://doi.org/10.1007/s10462-022-10352-9

[21] Buza, K. (2014). BlogFeedback [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C58S3F. Retrieved from https://archive.ics.uci.edu/dataset/304/blogfeedback.

[22] Dua D, Graff, C. UCI Machine Learning Repository, (2019) [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[23] Gharehchopogh, F. S., & Khaze, S. R. (2013). Data mining application for cyber space users tendency in blog writing: a case study. arXiv preprint arXiv:1307.7432. https://doi.org/10.5120/7291-0509

[24] Arif, M. H. (2025). Predicting Oil Prices: A Comparative Study of Machine Learning and Deep Learning Methods. *Bilad Alrafidain Journal for Engineering Science and Technology*, *4*(1), 1-14. https://doi.org/10.56990/bajest/2025.040101

[25] Rahim, R., Zufria, I., Kurniasih, N., Yasin Simargolang, M., Hasibuan, A., Utami Sutiksno, D., … Daengs GS, A. (2018). C4.5 Classification Data Mining for Inventory Control. International Journal of Engineering & Technology, 7(2.3), 68. http://doi.org/10.14419/ijet. v7i2.3.12618

[26] Abdollahzadeh, B., & Gharehchopogh, F. S. (2022). A multi-objective optimization algorithm for feature selection problems. Engineering with Computers, 38(Suppl 3), 1845-1863. https://doi.org/10.1007/s00366-021-01369-9.

[27] Barboza, F., & Altman, E. (2024). Predicting financial distress in Latin American companies: A comparative analysis of logistic regression and random forest models. The North American Journal of Economics and Finance, 72, 102158. https://doi.org/10.1016/j.najef.2024.102158

[28] Loef, B., Wong, A., Janssen, N. A., Strak, M., Hoekstra, J., Picavet, H. S. J., ... & Herber, G. C. M. (2022). Using random forest to identify longitudinal predictors of health in a 30-year cohort study. Scientific Reports, 12(1), 10372. https://doi.org/10.1038/s41598-022-14632-w

[29] Gao, C., Lan, X., Li, N., Yuan, Y., Ding, J., Zhou, Z., Xu, F., & Li, Y. (2023). Large language models empowered agent-based modeling and simulation: A survey and perspectives. arXiv preprint arXiv:2312.11970.

[30] Ma, C., Liang, Y., Yang, X., Wu, H., & Zhang, H. (2024). A privacy-preserving distributed credible evidence fusion algorithm for collective decision-making. arXiv preprint arXiv:2412.02130.

[31] Kayaalp, M., Inan, Y., Koivunen, V., & Sayed, A. H. (2024). Causal influence in federated edge inference. arXiv preprint arXiv:2405.01260.

[32] Zhao, T., Xu, Y., Monfort, M., Choi, W., Baker, C., Zhao, Y., Wang, Y., & Wu, Y. N. (2019). Multi-agent tensor fusion for contextual trajectory prediction. arXiv preprint arXiv:1904.04776.

[33] Weng, J., Xiao, F., & Cao, Z. (2020). Uncertainty modelling in multi-agent information fusion systems. In Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020) (pp. 1494–1502). International Foundation for Autonomous Agents and Multiagent Systems.