A Machine Learning Approach to Predicting Open Access Support in Research Projects

A Machine Learning Approach to Predicting Open Access Support in Research Projects

Hoang-Son Pham[1,2], Evy Neyens[1,2], and Amr Ali-Eldin[1,2,3]

[1] Centre for Research & Development Monitoring (ECOOM-UHasselt), 3500 Hasselt, Belgium,
[2] Data Science Institute, Hasselt University, 3500 Hasselt, Belgium,
[3] Computer engineering and Control systems Department, Faculty of Engineering, Mansoura University, Mansoura 35516, Egypt

Abstract. In recent years, Open Science has emerged as a pivotal element in research and innovation policies at various levels, advocating for immediate and unrestricted access to publicly funded scientific research outcomes. While the benefits of open access to research are recognized, measuring its extent poses a challenge. In this paper, we present a pioneering approach for predicting the level of support for open access within research projects by applying machine learning algorithms. Leveraging various text-mining techniques to create relevant features, our methodology utilizes machine learning models to forecast the extent of open-access support. We incorporate diverse features, including open-access publications by researchers and organizations, funding sources, research disciplines, and interdisciplinarity, to enhance predictive capabilities. The evaluation of our proposed approach involves implementing machine learning models on features extracted from projects available on the FRIS portal. Our findings unveil a significant correlation between the degree of open-access publications of researchers and the degree of open access within the associated projects. Furthermore, the selected models achieved an accuracy range of 76% to 85%, showcasing their effectiveness in precise predictions. The highest accuracy, 85%, was attained with Random Forest Regression, and classical machine learning algorithms outperformed deep learning counterparts.

Keywords: Open Science, Open Access, Indicator Development, Regression model

1 Introduction

In recent years, Open Science has become the cornerstone of research and innovation policies at supranational, national, regional, and institutional levels. The Open Science paradigm advocates for immediate and unrestricted access to all results and outcomes of publicly funded scientific research, including publications, datasets, patents, software, and code [1]. Access to research outcomes, tools, and methods facilitates scientific progress and innovation in various ways. For instance, publicly sharing data, methods, software, and code increases the

transparency of scientific research and fosters trust in research in general. Open access to (raw) datasets ensures scientific integrity and counteracts scientific fraud. It opens up datasets for reuse, allowing for results to be validated and reproduced, thereby increasing the robustness of scientific findings. Moreover, it permits the reuse of data to accelerate scientific progress by reducing duplication and enabling the (re)combination of datasets for new purposes. Additionally, public access to research projects and outcomes promotes collaboration between researchers and facilitates partnerships between research institutions and industry [2].

Although open science has garnered attention from researchers, evaluating or measuring open science remains a challenge. There are several aspects related to open science, such as open data, open source, and open-access articles. In this paper, we specifically focus on developing indicators for open-access articles. Mainly, it aims to propose an approach to predict the level of support for open access within a research project. To the best of our knowledge, this study marks the first endeavor to address this specific predictive challenge. This is an essential task for data management and policymakers. It can be motivated by several factors. For example, understanding the level of support for open access helps efficiently allocate resources. Projects with strong support for open access may benefit from targeted investments to enhance dissemination and accessibility [3]. Additionally, many funding agencies and institutions have policies promoting open access. Predicting the degree of support allows researchers and organizations to assess compliance with these policies and make adjustments if needed [4].

We apply machine learning algorithms to predict the degree of support for open access within a research project. We assume that the degree of support is associated with various information related to the project, such as open-access journal articles authored by researchers or organizations participating in the project, the funding source, and the research disciplines associated with the project. Additionally, interdisciplinary projects often involve researchers from different disciplines, bringing diverse perspectives [5]. Collaboration across disciplines can lead to a richer mix of research outputs and a broader range of open-access materials. Interdisciplinary research may attract attention from a wider audience, potentially increasing the visibility of associated open-access publications. In this study, we employ various text-mining techniques to calculate relevant features, and by utilizing these features, the machine learning model can predict the degree of support for open access within the project.

As a case study, we implemented the proposed approach to projects available on the FRIS portal[4]. The experimental results show that 1) The degree of support is likely to be influenced by open-access articles of authors, and 2) The machine learning models for regression can be applied to predict the degree of supporting open access. The best-achieved accuracy was 85% with Random Forest Regression. The classical machine learning algorithms achieved a better performance in comparison to the deep learning algorithms.

[4] https://researchportal.be/en

The rest of the paper is organized as follows. Following the Introduction section, we provide background literature related to the regression model in Section 2. The proposed approach is outlined in Section 3, followed by the presentation of experimental results in Section 4. Finally, Section 5 concludes the paper.

2 Backgrounds

We employ a regression algorithm to model and indicate the degree of support for open-access articles. Regression analysis is a powerful statistical tool widely used in scientific research to elucidate relationships between variables [6]. The regression model has been widely applied in various applications. For instance, it can be used to measure the journal's contribution to the social attention of research [7], find the correlation between interdisciplinary research and technological impact [8], and discover the correlation between open access and research policy impact [9].

In essence, regression models seek to uncover patterns within data, allowing researchers to make informed predictions or understand the extent of influence one variable may have on another. This method is precious in scenarios where the aim is to quantify and model the connection between a dependent variable and one or more independent variables. Using historical data, regression models discern underlying trends, enabling extrapolation and aiding researchers in uncovering insights into the dynamics of the studied phenomena. As an invaluable asset in the scientific toolkit, regression analysis provides a systematic and quantitative approach to understanding and predicting complex relationships, significantly advancing knowledge across diverse fields. The formula for a multiple linear regression is as follows:

$$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \epsilon,$$

where y is the predicted value of the dependent variable. β_0 is the intercept (y value when all other parameters are set to 0). $\beta_1 X_1$ is the regression coefficient (β_1) of the first independent variable (X_1). Three dots mean that it is the same for however many independent variables. $\beta_n X_n$ is the regression coefficient of the last independent variable. ϵ is the model error (a.k.a. how much variation there is in our estimate of y). The regression finds the line of best-fit line through the data by searching for the regression coefficient that minimizes the model's total error (ϵ).

There are various algorithms for constructing regression models, and the selection of an algorithm depends on factors such as the data's nature, assumptions about the data, and the desired model complexity. Some commonly used regression algorithms include linear regression for multiple independent variables; Random Forest Regression, an ensemble method constructing various decision trees and combining their predictions; Support Vector Regression, extending support vector machines to regression problems by mapping data into a higher-dimensional space; and Deep Learning, where deep learning models, especially

neural networks, are employed for regression tasks, capturing complex patterns in large datasets.

3 Methods

This study aims to predict the degree of support for open access within a research project, calculated by the percentage of open-access articles associated with the project. Typical steps to calculate the percentage of open-access articles include the following:

- Identify relevant articles: Define criteria for selecting articles associated with the research project. Specify the databases, journals, or repositories to be considered.
- Distinguish open-access articles: Develop a method to identify open-access articles. Consider utilizing databases with open-access information or repositories that tag open-access content.
- Calculate percentage: Count the total number of articles associated with the research project. Count the number of open-access articles among them.

The degree of support is expected to be influenced by various project-related factors, such as open-access journal articles authored by project participants (researchers or organizations), the funding source, the associated research disciplines, and the interdisciplinarity of the project. The proposed approach utilizes a regression model to operationalize, leveraging diverse project-related information to predict the value of support for open access.

Suppose a project's metadata includes the following information: researchers, organizations, funding resources, and publications. The challenge is predicting the likelihood of a project supporting open access. To solve this problem, we define a dependent variable, p, as the percentage of open-access articles associated with the project, calculated by the steps described above. This variable needs to be predicted for a new project. As for the independent variables, various features are utilized in this study. They include:

- average percentage open access publications of researchers, denoted by r,
- average percentage open access publications of organizations, denoted by o,
- funding resource, denoted by f,
- the research disciplines related to the project, denoted by d,
- regarding interdisciplinarity, we consider three factors related to interdisciplinarity: diversity of researchers, denoted by DR; diversity of organizations, denoted by DO; and diversity of the project's disciplines, denoted by DD.

The calculation of these features is presented in Subsection 3.1. The regression model for estimating the degree of open-access publications in a project is defined as follows:

$$p_i = \beta_0 + \beta_1.r + \beta_2.o + \beta_3.d + \beta_4.f + \beta_5.DR + \beta_6.DO + \beta_7.DD + \epsilon,$$

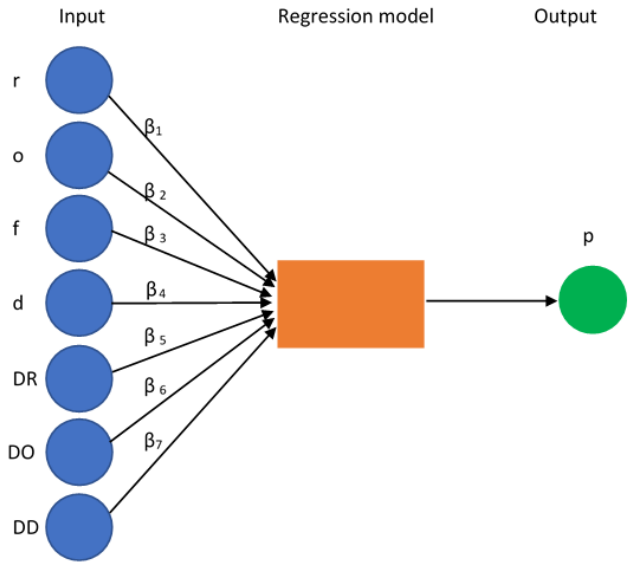

Fig. 1. Illustration of the proposed regression model for degree of support open access prediction.

where p_i indicates the percentage of open access articles published by project i, β_0 is the intercept, and β_i is the coefficient of interest for an independent variable i. The illustration of the proposed regression model is shown in Fig. 1.

A framework for the degree of support open access prediction includes four typical steps: 1) data collection, 2) feature creation, 3) regression model training, and 4) prediction. In the following subsections, we provide details for the second step, which differs from conventional approaches. The other steps are similar to classical machine learning approaches for a regression problem.

3.1 Feature creation

Feature creation algorithm: The feature creation algorithm is designed to generate features for regression models, as illustrated by Algorithm 1. The provided algorithm is designed to create a set of features (denoted as F) based on specific information associated with a given research project identified by its $projectID$. The algorithm consists of several steps:

Step 1: Count Percentage of Open-Access Publications. In this step, it retrieves publications associated with the project (P) and calculates the percentage of open-access publications (p) among the retrieved publications.

Step 2: Calculate the Average Percentage of Open-Access Publications for Researchers and Research Groups. This step retrieves researchers (R) and research groups (O) participating in the project. For each researcher, it finds their publications and calculates the percentage of open-access publications. Then,

Algorithm 1 Feature Creation Algorithm

Input: projectID
Output: F

// **Step 1: Count percentage of open-access publications**
$P \leftarrow$ GetPublications(projectID)
$p \leftarrow$ CalculatePercentageOpenAccess(P)
// **Step 2: Calculate average percentage of open-access publications for researchers and research groups**
$R \leftarrow$ GetResearchers(projectID)
$r \leftarrow$ CalculateAveragePercentageResearchers(R)
$O \leftarrow$ GetResearchGroups(projectID)
$o \leftarrow$ CalculateAveragePercentageGroups(O)
// **Step 3: Get Funder**
$f \leftarrow$ GetFunder(projectID)
// **Step 4: Predict Disciplines**
$d \leftarrow$ PredictDisciplines(projectID)
// **Step 5: Calculate IDR**
$DR, DO, DD \leftarrow$ CalculateIDR(projectID)
// **Output the feature set**
$F \leftarrow \{p, r, o, d, f, DR, DO, DD\}$
return F

the average percentage (r) across all researchers is computed. Similarly, each research group finds and counts the percentage of open-access publications and then calculates the average percentage (o) across all research groups.

Step 3: Get Funder. In this step, it retrieves the funder (f) of the project.

Step 4: Predict Disciplines. This step aims to find research disciplines associated with the project. In particular research information systems, the research disciplines may not be included in the project metadata. To enable the approach to apply to various systems, in this study, we employ a metadata-based framework for discipline prediction proposed in Ref. [10] to predict disciplines associated with the projects. This framework utilizes various text mining techniques to predict research disciplines related to the project by analyzing its description, i.e., a combination of title, keywords, and abstract.

Step 5: Calculate IDR. This step calculates the interdisciplinarity of the project. In particular, three factors are computed: DR, DO, and DD. To calculate these factors, we apply the approach proposed in Ref. [5]. To quantify these factors, it employs various methods, including calculating distance matrices, assessing the distance between researchers, and evaluating the relevance of researchers' expertise to the projects.

Finally, it combines the calculated values (p, r, o, d, f, DR, DO, and DD) into a feature set F and returns F as the output of the algorithm.

Discipline encoding: The *PredictDiscipline* procedure provides a subset of disciplines in a list of disciplines implemented in the information system. In practice, the number of disciplines in the system is often large; for instance,

the FRIS system using VODS [11] encompasses 42 disciplines at the second level and 382 disciplines at the third level. The number of features increases significantly when using one-hot encoding to convert disciplines into numerical data. Additionally, each project typically involves only a few disciplines from a vast pool of possibilities, resulting in many features containing zeros. To address this issue, we employ a discipline encoding technique introduced in reference [10].

The fundamental concept of this technique is as follows: it initially assigns each discipline a unique integer value. Subsequently, for each project, it selects the first n disciplines (e.g., n=2) and replaces them with the corresponding integer values. This approach has been demonstrated to outperform the conventional one-hot encoding method in terms of classification accuracy. For a more in-depth understanding, please refer to the study presented in [10].

Funding sources encoding: The funding sources obtained from project metadata were in string format. To convert them from categorical to numerical data, we utilized a label encoding technique, assigning each funding source a unique integer number.

Data normalization: Additionally, to ensure consistent values across features, we normalize the data using the max-min scale. Consequently, the values of features range from 0 to 1.

3.2 Regression model evaluation

Any machine learning algorithm supporting regression can be employed in this framework. Examples of well-known regression algorithms encompass Linear Regression (LR), Random Forest Regression (RFR), Support Vector Regression (SVR), and deep learning algorithms like Neural Networks (NN), Long Short-Term Memory neural networks (LSTM), Bidirectional Long Short-Term Memory neural networks (BiLSTM), and Convolutional Neural Networks (CNN).

Regression models can be evaluated using the mean squared error (MSE). This score measures the average of the squares of the errors, representing the average squared difference between the estimated values and the actual value.

Additionally, we aim to assess the models in terms of accuracy. In other words, we want to evaluate the model's accuracy in correctly predicting support for open access in a project. We assume a project supports open access if its corresponding p is larger than 0.5. With this assumption, we convert p into binary values, assigning 1s for values larger than 0.5 and 0s for values equal to or smaller than 0.5. This transformation allows us to calculate a confusion matrix necessary for determining accuracy and other measures such as precision, recall, and F1-score.

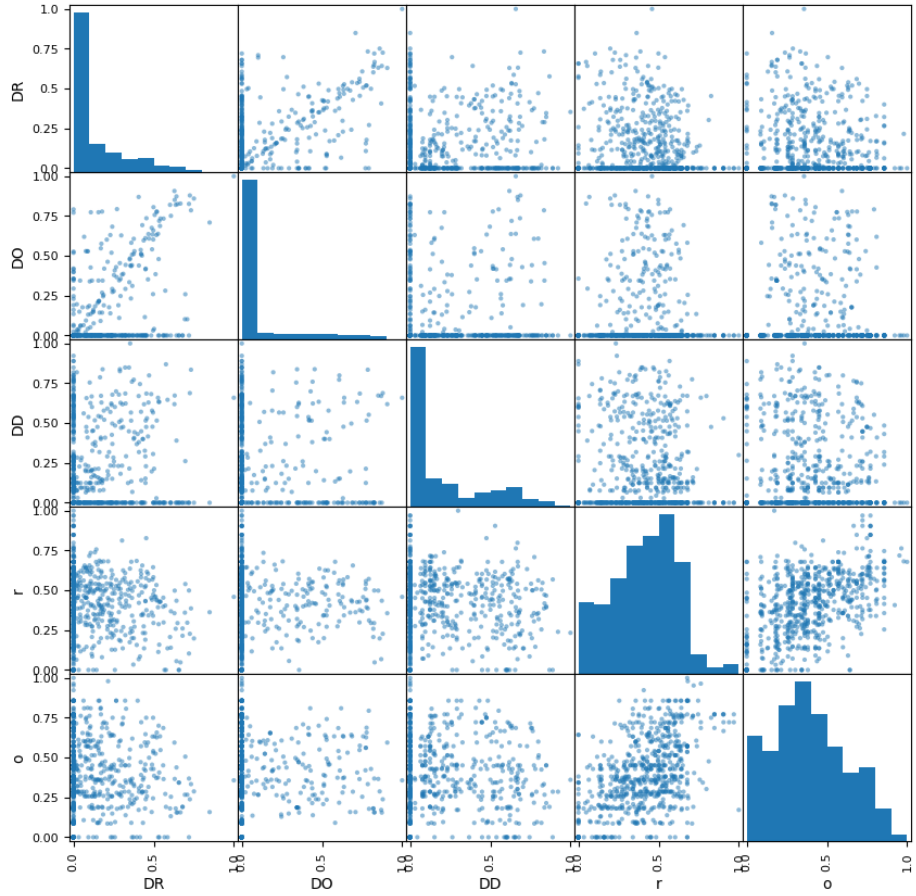

Fig. 2. Correlation of features in training data.

4 Experimental results

4.1 Experimental setup

Project metadata selection: To evaluate the proposed approach, we selected project metadata from the FRIS portal based on specific conditions: 1) the projects must have associated publications and 2) have funding resources. As a result, 703 projects were chosen for analysis. It is essential to note that only projects allowing open access were considered. Projects that do not permit open access, such as those from companies, were excluded from this experiment. Regarding disciplines, the second level of VODS, which encompasses 42 disciplines was chosen.

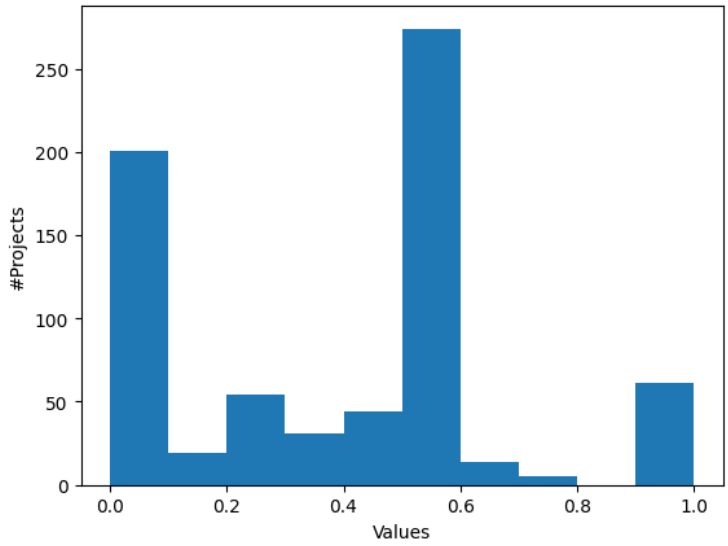

Fig. 3. Distribution of p among projects.

Feature creation: For each project, we initially calculated DR, DO, DD, r, o, p. Due to space constraints and for better readability, we illustrate the correlation of the features DR, DO, DD, r, o in Fig. 2. As can be seen, the correlation between these features is generally weak, except for the correlation between DR and DO, and the correlation between r and o. These correlations indicate a genuine relationship between these features. For instance, DR and DO were calculated based on the disciplines of researchers and organizations where researchers work, while r and o were calculated based on the publications of researchers and organizations.

The distribution of p among the projects is depicted in Fig. 3. As observed, there is a considerable number of projects where p is smaller than 0.5, indicating existing projects with low support of open-access publications.

Regarding disciplines, the *PredictDisciplines* procedure provides two disciplines (denoted d_1, d_2) that were encoded in the features. Similarly, funding resources were transformed into numerical data. As a result of the feature creation process, nine features were created: DR, DO, DD, r, o, p, f, d_1, d_2, in which p was the dependent variable.

Regression models: The purpose of this experiment is to assess whether the created features can effectively predict the target variable p. Determining the optimal hyperparameters to achieve the model's best performance is beyond this experiment's scope. We utilized the regression models mentioned in Subsection 3.2. For models such as LR, RFR, and SVR, default parameters were applied. Concerning deep learning neural networks, three layers were added: an

Table 1. Linear regression model results.

	coef	std err	t	P>‖t‖	[0.025	0.975]
const	0.1026	0.052	1.969	0.049	0	0.205
DR	0.0049	0.069	0.071	0.943	-0.13	0.14
DO	0.0779	0.059	1.315	0.189	-0.038	0.194
DD	0.0243	0.041	0.594	0.553	-0.056	0.104
r	0.5573	0.059	9.422	0	0.441	0.673
o	0.022	0.053	0.416	0.678	-0.082	0.126
f	0.001	0.007	0.154	0.878	-0.012	0.014
d1	0.0004	0.002	0.255	0.799	-0.003	0.003
d2	0.0004	0.001	0.281	0.779	-0.002	0.003

input layer with 64 units, an intermediate layer with 32 units, and an output layer with 1 unit. 'Adam' and 'MSE' were used as the optimizer and loss functions, respectively, for model compilation. To evaluate the performance of the models, we applied cross-validation 10 times. We report the final results based on the average scores obtained from these cross-validation runs.

4.2 Correlation results

We first present the correlation analysis. For this experiment, the LR was implemented to calculate the correlation between the dependent variable p and the independent variables. The result of LR is shown in Table 1. As observed, the regression analysis results indicated that the constant term had a coefficient of 0.1026 with a p-value of 0.049, suggesting statistical significance. However, the coefficients for independent variables were relatively low, all below 0.1, and their associated p-values were greater than 0.05, implying a lack of statistical significance. This indicates that, individually, these variables may not have a substantial impact on the dependent variable (p). Notably, the variable r stands out with a coefficient of 0.55 and a p-value less than 0.05, signifying both statistical significance and a relatively higher impact on the dependent variable. The results suggest that r is a significant predictor of the dependent variable, while the other variables may not contribute significantly in isolation.

4.3 Performance evaluation

The performance of the models is presented in Table 2. As can be seen, the mean squared error (MSE) of the models was all below 0.1, suggesting that the regression models we applied were performing well in terms of prediction accuracy.

The selected models demonstrated an accuracy range from 76% to 85%, indicating their reasonable performance in making accurate predictions on the evaluated dataset. Notably, Random Forest Regression (RFR) achieved the highest accuracy at 85%. In contrast, the accuracies of deep learning models did not

Table 2. Model performance results.

	MSE	Precision	Recall	F1-Score	Accuracy
LR	0.074	0.78	0.80	0.80	0.80
RFR	0.074	0.82	0.84	0.82	0.85
SVR	0.074	0.78	0.80	0.79	0.80
NN	0.096	0.83	0.76	0.76	0.76
LSTM	0.074	0.82	0.81	0.82	0.81
BiLSTM	0.087	0.81	0.75	0.78	0.76
CNN	0.083	0.84	0.76	0.79	0.76

surpass those of classical models. Specifically, the best accuracy was achieved by Long Short-Term Memory (LSTM) at 81%, while other deep learning models such as Neural Network (NN), Bidirectional LSTM (BiLSTM), and Convolutional Neural Network (CNN) exhibited an accuracy of 76%. It's important to highlight that, in this experiment, we implemented straightforward deep-learning models with three layers, and hyperparameter tuning was not a primary focus. This might contribute to the observed lower accuracy, indicating a potential for improvement through more nuanced model architectures and parameter optimization.

The application of regression models to predict the output variable yielded noteworthy insights. Despite the inclusion of various input variables, the results indicate a lack of significant correlation between most variables and the output variable. This suggests that the selected features may not strongly influence the predicted outcome. However, it is notable that the accuracies of the models ranged from 76% to 85%, indicating that the models were relatively successful in making accurate predictions. These findings underscore the complexity of the relationship between input variables and the output variable, highlighting the need for further investigation to better understand the underlying factors influencing the predicted outcome.

5 Conclusion

In this paper, we introduced an innovative approach to predict the degree of support for open access within research projects using machine learning algorithms. To the best of our knowledge, this study represents the first attempt to address this specific problem. Employing various text-mining techniques to select relevant features, we applied machine learning models to predict the degree of open-access support. Relevant features, such as open-access publications of researchers, organizations, funding sources, research disciplines, and interdisciplinarity, were collected and created to enhance the prediction process.

The evaluation of our proposed approach involved implementing machine learning models on features derived from projects available on the FRIS portal. The results revealed a significant correlation between researchers' open-access

publications and the degree of open access within the projects they participated in. Furthermore, the selected models demonstrated an accuracy range from 76% to 85%, indicating their reasonable performance in accurately predicting open access support on the evaluated dataset. The classical machine learning algorithms achieved a better performance than the deep learning algorithms. The best-achieved accuracy was 85% with Random Forest Regression.

It is crucial to note that our experiment focused on straightforward deep-learning models with three layers, and hyperparameter tuning was not a primary emphasis. The observed lower accuracy suggests a potential for improvement through future work that explores more nuanced model architectures and prioritizes hyperparameter optimization. Additionally, to assess the practicality of the proposed method, we intend to implement it on large-scale datasets and compare its performance with that of related state-of-the-art regression approaches.

Acknowledgments

This study was supported by The Expertise Center for Research and Development Monitoring (ECOOM), Flanders, Belgium. The authors would like to thank the reviewers for the feedback and comments that they have made to improve the paper.

References

1. R. T. Thibault, O. B. Amaral, F. Argolo, A. E. Bandrowski, A. R. Davidson, and N. I. Drude, "Open science 2.0: Towards a truly collaborative research ecosystem," *PLOS Biology*, vol. 21, no. 10, pp. 1–22, 10 2023. [Online]. Available: https://doi.org/10.1371/journal.pbio.3002362
2. European Commission. Directorate-General for Research and Innovation. (2021) EOSC Strategic Research and Innovation Agenda 2021-2027. Retrieved from https://eosc.eu/sria-mar/. [Online]. Available: https://eosc.eu/sria-mar/
3. E. Adie, "Do open journals get cited more in government policy?" *Septentrio Conference Series*, no. 4, Sep. 2020.
4. R. H. M. Alkhawtani, T. C. Kwee, and R. M. Kwee, "Citation advantage for open access articles in european radiology," *European Radiology*, vol. 30, no. 1, pp. 482–486, Aug. 2019.
5. H.-S. Pham, B. Vancraeynest, H. Poelmans, S. Vancauwenbergh, and A. Ali-Eldin, "Identifying interdisciplinary research in research projects," *Scientometrics*, vol. 128, no. 10, pp. 5521–5544, Aug. 2023.
6. T. Hastie, R. Tibshirani, and J. Friedman, *Linear Methods for Regression*. Springer New York, Dec. 2008, pp. 43–99.
7. P. Dorta-González, "A multiple linear regression analysis to measure the journal contribution to the social attention of research," *Axioms*, vol. 12, no. 4, p. 337, Mar. 2023.
8. Q. Ke, "Interdisciplinary research and technological impact: evidence from biomedicine," *Scientometrics*, vol. 128, no. 4, pp. 2035–2077, Feb. 2023.

9. Q. Zong, Z. Huang, and J. Huang, "Can open access increase lis research's policy impact? using regression analysis and causal inference," *Scientometrics*, vol. 128, no. 8, pp. 4825–4854, May 2023.
10. H.-S. Pham, H. Poelmans, and A. Ali-Eldin, "A metadata-based approach for research discipline prediction using machine learning techniques and distance metrics," *IEEE Access*, 2023.
11. S. Vancauwenbergh and H. Poelmans, "The creation of the flemish research discipline list, an important step forward in harmonising research information," *Procedia Computer Science*, vol. 146, pp. 265–278, 2019.