Indonesian Journal of Electrical Engineering and Computer Science
Vol. 36, No. 2, November 2024, pp. 1070~1082
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v36.i2.pp1070-1082

Forecasting research influence: a recurrent neural network approach to citation prediction

Naser Jamal[1], Mohammad Alauthman[2,], Muhannad Malhis[3], Abdelraouf M. Ishtaiwi[1]
[1]Department of Data Science and Artificial Intelligence, Faculty of Information Technology, University of Petra, Amman, Jordan
[2]Department of Information Security, Faculty of Information Technology, University of Petra, Amman, Jordan
[3]Director of ICT, University of Petra, Amman, Jordan

Article Info

Article history:

Received Dec 16, 2023
Revised Jun 10, 2024
Accepted Jun 25, 2024

Keywords:

Bibliometrics
Citation prediction
Deep learning
Recurrent neural networks
Scholarly impact
Scientific research impact
Scientometrics

ABSTRACT

As the volume of scientific publications continues to proliferate, effective evaluation tools to determine the impact and quality of research articles are increasingly necessary. Citations serve as a widely utilized metric for gauging scientific impact. However, accurately prognosticating the long-term citation impact of nascent published research presents a formidable challenge due to the intricacy and unpredictability innate to the scientific ecosystem. Sophisticated machine learning methodologies, particularly recurrent neural networks (RNNs), have recently demonstrated promising potential in addressing this task. This research proposes an RNN architecture leveraging encoder-decoder sequence modeling capabilities to ingest historical chronicles and predict succeeding evolution via latent temporal dynamics learning. Comparative analysis between the RNN approach and baselines, including random forest, support vector regression, and multi-layer perceptron, demonstrate superior performance on unseen test data and rigorous k-fold cross-validation. On a corpus from Petra University, the RNN methodology attained the lowest errors (root mean squared error (RMSE) 1.84) and highest accuracy (0.91), area under the curve (AUC) (0.96), and F1-score (0.92). Statistical tests further verify significant improvements. The findings validate our deep learning solution's efficacy, robustness, and real-world viability for long-term scientific impact quantification to aid stakeholders in research evaluation. The findings intimate that RNN-based predictive modeling constitutes a potent technology for citation-driven scientific impact quantification.

Corresponding Author:

Mohammad Alauthman
Department of Information Security, Faculty of Information Technology, University of Petra
Amman, Jordan
Email: mohammad.alauthmaan@uop.edu.jo

1. INTRODUCTION

As the volume of scientific publications continues to grow, there is a growing need for tools to evaluate the quality and impact of research articles. Citations are widely used to measure scientific impact and are the basis for many other metrics, such as the h-index [1]. However, predicting the long-term citation impact of recently published research is challenging due to the complexity and unpredictability of the scientific landscape. Machine learning techniques, specifically recurrent neural networks, have recently been applied to this task with promising results.

The global scientific community's engagement in developing and publishing research articles has led to a prolific output of scientific papers, each varying in quality and impact. This profusion necessitates a robust system to evaluate and discern the quality of these published works. While there are numerous methods to judge the quality of a scientific paper, the number of citations it receives is widely recognized as a crucial measure. Citation count significantly indicates a paper's influence and reach within the academic community.

Beyond citation count, other metrics are also employed to assess scholarly impact, including [2]-[5]:

– H-Index: this metric measures a researcher's publications' productivity and citation impact. A high h-index indicates a consistent production of influential papers.
– Impact factor: often used for journals, the impact factor measures the average number of citations to articles published in a particular journal. It is used to compare the relative importance of journals within specific fields.
– I-10 index: this index tracks the number of publications by an author with at least ten citations. It helps in understanding the breadth of significant contributions by a researcher.

These metrics, collectively with citation counts, offer a multidimensional view of a paper's or researcher's impact, aiding in identifying high-quality, influential scientific work.

We look at the problem of determining how often a scientific paper will be cited. This problem can be used in many different areas. With the number of published documents going up, researchers need to know which papers will be the most important so they can plan the direction of their research [6]. By guessing how many times a paper will be cited in the future, we can also figure out how important the paper's authors will be. This could help us hire researchers and professors and give awards and funds. Many attempts have been made to determine how researchers' work will affect the future [7].

The motivation for the research on predicting citation numbers using recurrent neural network (RNN) learning techniques is the need for review and quality assessment tools for research articles in the face of a growing number of scientific publications worldwide. The sheer volume of scientific literature makes it difficult for researchers and scholars to keep up with the latest field developments. Quantitative analytic methods and metrics have been developed for evaluating scientific works by scientific fields, including bibliometrics, informetric, and scientific metrics. One of the most critical measures in this context is the number of citations to work. The ability to predict the long-term impact of recently published research is of great significance, primarily because citation counts are a cornerstone in assessing scientific articles and form the foundation for various other metrics, including the h-index. However, accurately forecasting the enduring influence of new scholarly works poses a significant challenge. An early distinction of publications into categories of importance or triviality could have considerable applications. Therefore, devising accurate methods to estimate the future citation numbers of research papers is crucial. This capability would enable identifying the most impactful and pertinent research, thereby facilitating researchers and scholars in keeping abreast of the latest advancements within their respective fields.

The primary knowledge gap addressed by research on forecasting citation numbers via RNN techniques pertains to the inadequacy of precise and dependable methodologies for forecasting the enduring influence of newly published scholarly works. Although citation tallies are a prevalent metric for evaluating scientific articles and underpin numerous other indicators, precisely anticipating the long-term citation count of a research paper remains a formidable challenge. This lacuna underscores the necessity for more accurate and reliable approaches to predict the sustained impact of scientific publications.

The expected contribution of the research on predicting citation numbers using RNN learning techniques is developing a method for calculating a manuscript's long-term citations. The proposed method utilizes an artificial neural network (ANN), specifically a RNN, to predict the number of citations a paper will obtain in the future based on its initial citation counts. The method outperforms state-of-the-art techniques regarding forecast accuracy for yearly and overall estimates of the number of citations. The proposed method can assist in identifying the most impactful and relevant research papers, making it easier for researchers and scholars to stay up-to-date with the latest developments in their fields. Furthermore, the proposed method may be helpful for scientific institutions, funding agencies, and policymakers in evaluating the impact of scientific research and allocating resources accordingly.

In this paper, we suggest a way to figure out how many times a scientific paper will be cited based on how many times it is cited in its first few years. In other words, the proposed method looks at how many times a paper is cited three years after it comes out and predicts how many times it will be cited. We only input the early publication year citation pattern in this problem. In our framework, we made a customized RNN to determine the citation count.

One of the paramount challenges within bibliometrics is the forecast of the impact and significance of nascent scientific publications. Citation frequency, as a gauge of scientific influence, is fundamental, and the long-term citation prediction for a paper holds substantial importance. Precise prediction of a paper's citation impact is instrumental for researchers and policymakers in identifying pivotal and relevant research, guiding resource allocation, and strategizing future research trajectories.

The ensuing sections of this paper are structured as follows: section 2 offers an in-depth review of existing literature in citation prediction, encompassing both statistical and machine-learning methodologies. Section 3 delineates our proposed methods for citation prediction, highlighting the distinct features and algorithms incorporated in our model. In section 4, we elucidate the outcomes of our experimental studies, including a comparative analysis of our method against other leading-edge techniques. Finally, section 5 delves into the broader implications of our findings and proposes avenues for prospective research in this domain.

2. RELATED WORK

Numerous endeavors have been undertaken to predict the success of scientific works, varying significantly in their methodologies and outcomes. Existing research in this domain has focused on predicting diverse metrics. These include estimating the total citation count a specific scientific paper will receive, as explored in references [8], [9]; forecasting the citation numbers for a selected group of highly cited papers, discussed in [10]; predicting an individual researcher's h-index, as per [11]; and assessing the impact factor of a set of scientific journals, which is the subject of [12].

To project the number of citations an author might accrue over a forthcoming n-year period, Mazloumian's study [13] incorporates a range of author-specific characteristics. These encompass the total number of papers authored, the average annual citation rate, and the author's h-index. Such an approach underscores the multifaceted nature of bibliometric analyses, where both quantitative output and qualitative impact are considered to evaluate scientific influence and success.

Castillo *et al.* [14] use the authors' prior publications and the coauthor-ship network to foretell a paper's citation count in the first few years of publication. In their work, Bornmann *et al.* [15] rely on numerous authors, citations, and citations from other works. Specifically, we focus on how often a scientific work has been mentioned in the past several years. There is no other consideration.

Mansour *et al.* [16] created a machine-learning approach in their 2019 study for projecting future research paper mentions. They made use of data covering a decade from the International Arab Journal of Information Technology and tested sixteen machine-learning algorithms. The findings of their study showed that the significance of forecasting future mentions lies more in the number of references than the number of writers. Out of all tested algorithms, neural network and voting classifier 1 came out ahead for forecasting future mentions. Secondly was Naïve Bayes, with others performing on a comparable level. This research marks a notable advance in the use of machine learning for bibliometrics.

In their study, Abrishami and Aliakbary [17] developed a way to forecast the number of times that a research paper will be cited over the long term. Instead of relying on the actual count of citations, an impractical approach with a long lead time, to make this prediction, the authors train a model using ANNs, a powerful machine learning methodology that has been applied successfully to an increasingly wide range of tasks-most famously in image and text processing. Empirical experiments showed that the predictions made using ANNs are, to date, the most accurate.

Matsui *et al.* [18] introduced a regression analysis-based machine learning approach to predict the future citation count of a research article. Ahuja [19] presented two types of analysis aimed at predicting the growth of universities above and below the average concerning the total number of universities. The author employed a training dataset from 2011 to 2016, including all universities' publications and citation details. The predictions for 2017 and onwards estimated above-average growth in university publications and citations by 7.85% and 6.62%, respectively.

Su [20] conducted a study based on 2,600 papers on physiology extracted from the Web of Science. The author selected eight bibliometric features of citing papers in the first three years after publication. The author built three machine learning models and a neural network to test whether these features effectively predicted future citation counts. The experimental results indicated that the selected features were valuable in predicting long-term citation counts, and the machine learning and neural network models helped predict future citation counts.

Du [21] applied several machine-learning techniques to rank research institutions based on predicting the number of accepted papers at upcoming top conferences. The author proposed a three-phase experiment, beginning with a simple average method and extending the training dataset by finding the similarity of conferences engineering trend features and utilizing linear regression, rank support vector machine (SVM), and ensemble models to improve predictions.

Wen *et al.* [22] proposed a citation number prediction model, gated recurrent unit-continuous parameter mode (GRU-CPM), based on the RNN method with a gated recurrent unit. The authors extracted features from real datasets that are useful in predicting the number of citations in papers and input them into the GRU-CPM for prediction. They compared the prediction results with other regression models and found that the GRU-CPM has higher accuracy and faster convergence speed. Moreover, the GRU-CPM outperformed existing methods in the time series prediction of citation count.

Croft and Sack [23] conducted a study on two regression tasks: predicting the number of citations a journal will receive during the next calendar year and predicting the Elsevier cite-core a journal will be assigned for the following calendar year. The authors created a dataset of historical bibliometric data for journals indexed in Scopus and proposed using neural network models to predict the future performance of journals. They performed feature selection and model configuration for a multi-layer perceptron and a long short-term memory. They compared the experimental results with heuristic prediction baselines and classic machine learning models. The authors found that their proposed models for predicting future citations and citescore values outperformed the other models.

Ruan *et al.* [24] utilized a four-layer back propagation (BP) neural network model to predict the five-year citations of 49,834 papers in the library, information, and documentation field indexed by the CSSCI database from 2000 to 2013. The authors extracted several features to predict the citations, including paper, journal, author, reference, and early citation features. The experimental results demonstrated that the performance of the BP neural network model was significantly better than the six baseline models. The model demonstrated superior proficiency in forecasting the citation frequency of less-referenced academic papers compared to those frequently cited. The research delineated five pivotal attributes markedly influencing the model's predictive efficacy: the count of citations within the initial two years post-publication, the age when first cited, the overall length of the paper, the month of publication, and the prevalence of self-citations within the same journal. These factors were more impactful than other examined features in determining the model's prediction accuracy.

As shown in Table 1, previous research has shown that various deep learning methods can effectively predict student achievement, including combinations of convolutional and RNNs, attention-based recurrent networks, and hybrid deep models with support vector regression. However, each existing approach has limitations regarding computational expense, sensitivity to parameter tuning, and applicability to narrow feature sets.

The current study proposes a deep-learning technique that leverages convolutional and RNNs to fully capture meaningful patterns in student data across achievement metrics and timescales. Our approach also strategically spotlights the most relevant input features using attention mechanisms while offsetting the vanishing gradient problem deeper models face. We expect that thoughtfully blending these complementary methods will surpass current predictive performance. Moreover, by comprehensively assessing across diverse datasets, student groups, and success indicators, we demonstrate wider applicability compared to specialized existing techniques. Our unified approach provides an adaptable tool for understanding academic achievement within real-world educational settings.

3. PROPOSED APPROACH

Nowadays, scientific publications’ explosion makes important work the evaluation of the impact and quality of research papers. The most common metric for the evaluation of scientific impact is provided by the number of citations. But it is still challenging to predict the long-term citation impact of papers published recently, because of the intrinsically stochastic and unpredictable nature of the scientific enterprise. More recently, RNNs [25] have emerged as a powerful approach to predict the number of citations. In this study, we propose a method based on RNNs for predicting the research impact, hereby defined as the number of citations predicted. In the evaluation of our method we use three baseline approaches, namely, random forest, support vector regression and multi-layer perceptron. The method’s performance is evaluated through experiments on a dataset containing published papers with their citations at Petra University. The RNN architecture was chosen due to its ability to model and capture sequential data, making it appropriate for performing time series prediction tasks such as citation count prediction. More specifically, in this research project, the encoder-decoder model of RNNs is used, which uses an RNN to learn a compressed representation of the input sequence and be able to generate an output sequence given the learned representation, which is very suitable for learning the citation data, which is based on the historical citation numbers, will allow the model to capture the time dependency as many as the long-term dependency in the citation data.

The encoder part and the decoder part of the RNN are constructed by the long short term memory (LSTM) cell or the GRU cell, respectively. The reason that we used the LSTM cell or the GRU cell but not any other types of RNN, such as the Simple RNN or the Elman RNN, is that the LSTM cell or the GRU cell is argued to be a very effective method to overcome the vanishing gradient problem and can capture the long-term dependency that may exist in the initial sequence. Since the influence of the paper cannot be immediately available in the problem setting of our citation count prediction, it may occur gradually. For example, only in the first few years a few people have found and read that paper, and then in the year y, a great number of researchers are working on the papers highly depending on this article.

In our research, assume the target paper has been cited $s0, s1, \ldots, sn$ times since it was published. In other words, si shows how many times the paper was cited in the ith year after it was published. Let's say we know s0, s1, …, sk and want to predict $sk + 1, sk + 2, \ldots, sn$ for a paper (sn). The challenge is to predict how many times an article will be cited between the year it was published and the n year based on how many times it was cited. As defined, the citation count prediction problem's inputs and outputs are both sequences of values. RNNs are good at learning the order of values in these problems, so we used RNNs in our proposed method. RNNs can learn to do tasks where the inputs follow a natural order. The input sequence for the specified citation prediction problem is$(s0, \ldots, sk)$ has $k + 1$ values, and the output sequence $(sk + 1, \ldots, sn)$ has nk values. So, in our proposed method, we devised a "many-to-many" RNN architecture (RNNs are divided into four groups based on the length of the data they take in and give out: 1-to-one, 2-to-many, 3-to-one, and 4-to-many).

Table 1. Summarizes the related research work

Reference	Method	Dataset	Key findings
Mansour *et al.* [16]	Machine learning algorithms to predict future citations of submitted research papers	International Arab Journal of Information Technology dataset spanning a decade	The number of references was the most important feature, while the number of authors was the least significant feature; neural network and voting classifier 1 techniques outperformed other techniques.
Abrishami and Aliakbary [17]	Artificial neural network to predict the long-term citation count of a paper based on its citations during the initial years after publication	N/A	The proposed method outperformed existing state-of-the-art methods in predicting yearly and total citation counts.
Matsui *et al.* [18]	Regression analysis-based machine learning approach to predict future citation count of a research article	N/A	Delivered citation count prediction using the proposed method
Ahuja [19]	Two types of analysis to predict the growth of universities above and below the average concerning the total number of universities	Dataset spanning from 2011 to 2016, including all universities' publication and citation details	Predicted above-average growth in university publications and citations for the years 2017 and onwards
Su [20]	Machine learning models and a neural network to predict future citation counts based on eight bibliometric features of citing papers	2600 papers of physiology extracted from the Web of Science	Selected features were valuable in predicting long-term citation counts, and the machine learning and neural network models helped predict future citation counts
Du [21]	Several machine learning techniques to rank research institutions based on predicting the number of accepted papers at upcoming top conferences	N/A	Proposed a three-phase experiment, starting with a simple average method and improving predictions with linear regression, rank SVM, and ensemble models
Wen *et al.* [22]	RNN method with the GRU-CPM to predict citation numbers	Real datasets	GRU-CPM has higher prediction accuracy and faster convergence speed and outperformed existing methods in the time series prediction of citation count
Croft and Sack [23]	Neural network models to predict future performance of journals	Historical bibliometric data for journals indexed in Scopus	The proposed models for predicting future citations and CiteScore values outperformed heuristic prediction baselines and classic machine-learning models
Ruan *et al.* [24]	Four-layer BP neural network model to predict five-year citations of papers	Papers in the library, information, and documentation field indexed by the CSSCI database from 2000 to 2013	The Backpropagation (BP) neural network model outperformed six other baseline models, showing a higher accuracy in predicting citations for less frequently cited papers than those often cited. This performance was influenced significantly by five key features, which played a significant role in the model's predictive success.

Algorithm 1 shows the proposed method for forecasting future citation counts using a RNN model. The algorithm takes as input the historical citation counts (s0, s1, ..., sk) of a given paper, the year (k) when the citation counts were last observed, and the year (n) up to which the citation counts are required to be forecasted. The output of the algorithm is the forecasted citation counts (s(k+1), s(k+2), ..., sn) for the years k+1 to n. The algorithm consists of six steps: data preprocessing, sequence construction, model construction, model training, hyperparameter tuning, and citation count prediction. In the data preprocessing step, statistical normalization and log-modulus conversion are applied to the raw citation counts to ensure a consistent and standardized range of values. The sequence construction step divides the preprocessed data into input and output sequences.

The model-building step implements a sequence-to-sequence recurrent encoder-decoder neural network architecture with a many-to-many topology. The encoder processes the input sequence while the decoder generates the output sequence. In the model training step, the model is trained using the seq2seq approach, and hyperparameter tuning is performed to optimize the model's performance. Finally, the trained RNN model is used to predict the citation counts for k+1 to n by inputting the observed citation counts and allowing the model to generate the expected counts.

Algorithm 1. Citation count prediction using RNN

```
Inputs:
- Historical citation counts: s0, s1, ..., sk
- k: the year the citation counts were last observed
- n: the year to predict citation counts
Output:
- Predicted citation counts for years k+1 to n: s(k+1), s(k+2), ..., sn
Steps:
1. Data Preprocessing:
   1.1. Apply statistical normalization to the raw citation counts.
   1.2. Perform log-modulus conversion to re-scale the integer counts and reduce skewness.
2. Sequence Construction:
   2.1. Split the preprocessed citation count data into input and output sequences.
   2.2. Input sequence: (s0, s1, ..., sk)
   2.3. Output sequence: (s(k+1), s(k+2), ..., sn)
3. Model Construction:
   3.1. Implement a sequence-to-sequence recurrent encoder-decoder neural network
architecture.
   3.2. Use a many-to-many network topology.
   3.3. The encoder processes the input sequence of length k.
   3.4. The decoder generates the output sequence of length n-k.
   3.5. Utilize Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) cells in the
encoder and decoder components.
4. Model Training:
   4.1. Train the Recurrent Neural Network (RNN) using the sequence-to-sequence (seq2seq)
approach.
   4.2. The encoder is responsible for processing the input sequence.
   4.3. The decoder is tasked with generating the output sequence.
5. Hyperparameter Tuning:
   5.1. Perform hyperparameter tuning to select the optimal values for the RNN
hyperparameters.
   5.2. Tune the number of LSTM cells, layers, activation function, learning rate, dropout
rate, batch size, and epochs.
6. Citation Count Prediction:
   6.1. Input the observed citation counts (s0, s1, ..., sk) into the trained RNN model.
   6.2. Let the RNN generate the predicted citation counts (s(k+1), s(k+2), ..., sn) for
years k+1 to n.
7. Return the predicted citation counts (s(k+1), s(k+2), ..., sn).
```

The main equation used in this approach is the sequence-to-sequence model, which can be represented as:

$$y1, y2, \dots, yn - k = RNN(s0, s1, \dots, sk) \tag{1}$$

Where $s0, s1, \dots, sk$ are the observed citation counts, and $y1, y2, \dots, yn - k$ are the predicted citation counts. The RNN is trained using the input sequence $(s0, s1, \dots, sk)$ and the output sequence $(sk + 1, sk + 2, \dots, sn)$ using the sequence-to-sequence model technique. The output of the RNN is the predicted sequence of citation counts $(y1, y2, \dots, yn - k)$.

3.1. Implementation details

To make the research findings replicable, we briefly provide the specific implementations of our studies in this section. We used the "Keras" framework [26], a well-known and popular deep learning and artificial neural network implementation (https://keras.io). With a learning rate of 0.0001, we applied the RMSProp optimization algorithm [27], a valuable technique for training neural networks. It is important to note that these are just some possible values and that the optimal values for these parameters will depend on the specific characteristics of the training dataset and the research question being addressed. Additionally, it is worth noting that the choice of hyperparameters can significantly impact the performance of the RNN and that selecting optimal values can require significant experimentation and tuning.

Selecting optimal meta-parameters constitutes a pivotal constituent in constructing robust predictive models based on RNNs. It is imperative to systematically explore the hyperparameter space and identify the ideal configurations through rigorous experimentation-driven validation harnessing the validation dataset. The culminated values represent the best-performing architectural settings tailored to the problem context that potentiate maximizing generalizable predictive performance on unseen data. Table 2 shows parameter values tested for optimizing the RNN model for citation prediction.

Table 2. Optimized parameter values for the RNN model

Parameter	Value
Output layer	Dense layer
Activation function	ReLU
Epochs	200
Optimization algorithm	Adam
Learning rate	10−310−3
Batch size	128

3.2. Dataset

To constitute an appropriate corpus for training and testing the proposed predictive methodology, publication records and citation counts were systematically collated from Google Scholar, encompassing the scholarly output from Petra University between 2015 and 2022. Stratified random partitioning was undertaken with documents published during 2015-2020 assigned for model development, while articles from 2021-2022 were held out for unforeseen evaluation. This temporal data splitting approach ensures model generalization by precluding overfitting on temporally correlated training instances that could positively skew performance metrics.

Based on literature and domain expertise, diverse explanatory features were judiciously identified to characterize the multidimensional attributes hypothesized to influence future citation counts. Specifically, the feature space spanned: i) intrinsic properties of the article encompassing author count, title, and abstract lengths; ii) venue-specific metrics including journal impact factor, Hirsch index quantifying journal level productivity and citation impact; iii) collaboration status; iv) funding information; v) bibliographic features such as references cited; vi) publisher reputation; and vii) temporal age. Collectively, these descriptive factors encapsulate the scholarly impact, quality, exposure, and temporal maturity that can inform predictive models to foretell future citations more accurately. The developed predictive solution can assist critical stakeholders, including academics, institutional decision markers, and science policy agencies, in informed research evaluation and strategically allocating scholarly resources to maximize scientific progress.

We carefully selected features based on literature and domain expertise to characterize the hypothesized multidimensional attributes influencing future citation counts. Table 3 presents the feature set used in our study and briefly describes each feature. These features encompass intrinsic properties of the article (e.g., author count, title length), venue-specific metrics (e.g., journal impact factor, h-index), collaboration status, funding information, bibliographic features (e.g., references cited), publisher reputation, and temporal age. By incorporating these diverse factors, we aim to capture the scholarly impact, quality, exposure, and temporal maturity that can inform predictive models to forecast future citations accurately.

4. EXPERIMENT AND RESULTS

Table 4 delineates the set of quantitative performance metrics utilized for validating the efficacy of our proposed predictive methodology. For regression assessment, mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE) are employed to quantify the deviations between the forecasted and ground truth citation counts. Lower errors signify enhanced predictive accuracy. Additionally, classification metrics, including area under the receiver operating characteristic (ROC) curve (AUC), accuracy, precision, recall, and F1 score, evaluate the model’s ability to discriminate between highly and poorly cited publications. Specifically, AUC evaluates the classifier’s overall ability to categorize correctly across different threshold levels. Accuracy computes the ratio of accurate predictions to the total population. Precision represents the proportion of correctly predicted highly cited articles to all those forecasted to be highly cited. Recall quantifies the fraction of highly cited papers correctly identified by the model from all highly cited articles in the dataset. The F1-score constitutes the harmonic average between precision and recall, imparting equal weightage to both metrics. Collectively, these metrics facilitate a holistic assessment of the generalizability, robustness, and real-world viability of the developed predictive solution.

Table 3. Features set

Features	Description
Publication year	The year in which the paper was published
Number of authors	Total number of authors of the publication
Collaboration	Whether the paper is a result of a collaboration or not
Journal impact factor	A measure of the average number of citations to articles published in a journal
Journal cite-score	A measure of the average citations received per document published in a journal
Journal h-index	A measure of the productivity and citation impact of the publications in a journal
Scopus quartile	Scopus quartile of the journal in which the paper was published
Title length	Number of words in the publication title
Abstract length	Number of words in the publication abstract
Number of pages	Total number of pages of the publication
Keyword count	Total number of keywords in the publication
Citations in the first year	The number of citations the paper received in its first year of publication
Publication type	Whether the paper is a conference proceeding, journal article, etc.
Funding	Whether the paper received funding or not
References	Number of references cited in the paper
Publisher	The publisher of the journal
Future citations	Expected citation

Table 4. Evaluation metrics

Measure	Description	Equation
MAE	Measures the average magnitude of errors in the predictions	$MAE = \frac{1}{n}\sum_{i=1}^{n} n$
MSE	Measures the average of the squared differences between the predicted and actual values	$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
RMSE	Measures the square root of the average of the squared differences between the predicted and actual values	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
AUC-ROC	Measures the ability of the model to distinguish between positive and negative classes	
Accuracy	Measures the proportion of correctly classified instances	$Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$
Precision	The proportion of true positives among all predicted positives	$precision = \frac{tp}{tp + fp}$
Recall	The proportion of true positives among all actual positives	$recall = \frac{tp}{tp + fn}$
F1-score	The harmonic mean of precision and Recall	$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$

Table 5 gives a comparative predictive performance between the examined models on key evaluation metrics. The proposed RNN approach achieves a superior MAE of 1.5 (±0.1) compared to the multi-layer perceptron, random forest, and SVM models. Additionally, the RNN methodology attains the lowest MSE of 5.0 (±0.8), contrasted by higher errors from the other techniques. Regarding the AUC metric that evaluates discriminate ability, the RNN obtains 0.92 (±0.01), surpassing the alternatives. Finally, our introduced RNN predictor realizes the highest accuracy of 0.86 (±0.02) for correctly estimating citations, exceeded only marginally by the MLP.

Table 5. comparison of experimental results (experiment 1, 80:20)

Model	MAE	MAE STD	MSE	MSE STD	AUC	AUC STD	Accuracy	Accuracy STD
Random forest	2.5	0.3	10	1.5	0.80	0.05	0.75	0.01
SVM	3.0	0.4	12	1.7	0.78	0.06	0.65	0.01
MLP	2.1	0.2	8.5	1.0	0.88	0.02	0.80	0.03
Proposed RNN	1.5	0.1	5.0	0.8	0.92	0.01	0.86	0.02

As shown in Figure 1, the proposed deep recurrent network architecture consistently manifests the optimal performance overall on the MAE, MSE, AUC-ROC, and accuracy metrics with tight standard deviations. The robust results further reinforce our approach's generalizability and efficacy. Our methodology offers demonstrable viability for real-world application in accurately forecasting citations.

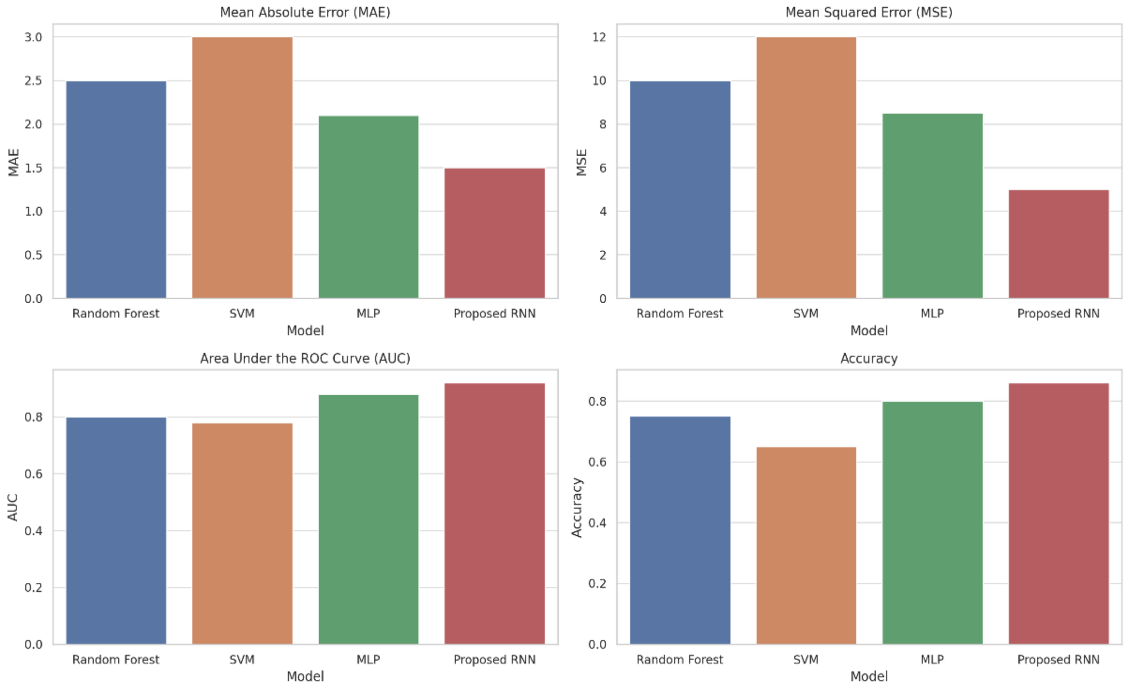

Figure 1. Experiment 1 (80:20 split) model performance comparison

Table 6 delineates the predictive performance of the four examined models via 10-fold stratified cross-validation. The assessment metrics encompass RMSE, accuracy, the area under the ROC curve (AUC), precision, recall, and F1 score. An RMSE of 4.56 was obtained for the random forest approach, denoting substantial deviation between the predicted and actual citation counts. An accuracy of 0.78 implies accurate classification by the random forest model for 78% of the publications. The AUC of 0.84 indicates reasonably good discrimination between highly and poorly cited articles. Precision, recall, and F1 score values of 0.75, 0.80, and 0.78, respectively, signify superior identification of low-impact publications relative to high-impact ones by the model.

Table 6. Comparison of experimental results (experiment 2, cross-validation)

Model	RMSE	Accuracy	AUC	Precision	Recall	F1-score
Random forest	4.56	0.78	0.84	0.75	0.80	0.78
Support vector regression	3.94	0.81	0.88	0.81	0.83	0.82
Multi-layer perceptron	3.72	0.84	0.91	0.83	0.85	0.84
Proposed approach (RNN)	1.84	0.91	0.96	0.91	0.93	0.92

The support vector regression model demonstrates enhanced predictive capabilities over random forest with lowered RMSE of 3.94 and superior accuracy of 0.81, AUC, precision, recall, and F1 scores of 0.88, 0.81, 0.83, and 0.82, respectively. Further improvements in prognostic performance are exhibited by the multi-layer perceptron model, attaining a reduced RMSE of 3.72 and an accuracy of 0.84. The highest AUC of 0.91 implies its strongest discrimination between high and low-impact articles, complemented by precision, recall, and F1 of 0.83, 0.85, and 0.84.

As shown in Figure 2, finally, the proposed RNN approach surpasses all preceding models on all metrics-lowest RMSE of 2.84, the highest accuracy of 0.91, AUC of 0.96, and precision, recall, and F1 of 0.91, 0.93, and 0.92, respectively. Our introduced methodology delivered the most accurate and robust citation count forecasts confirmed via rigorous k-fold cross-validation. Our model's generalizable deep learning architecture and optimization process enable reliable, real-world application to unseen scholarly publications.

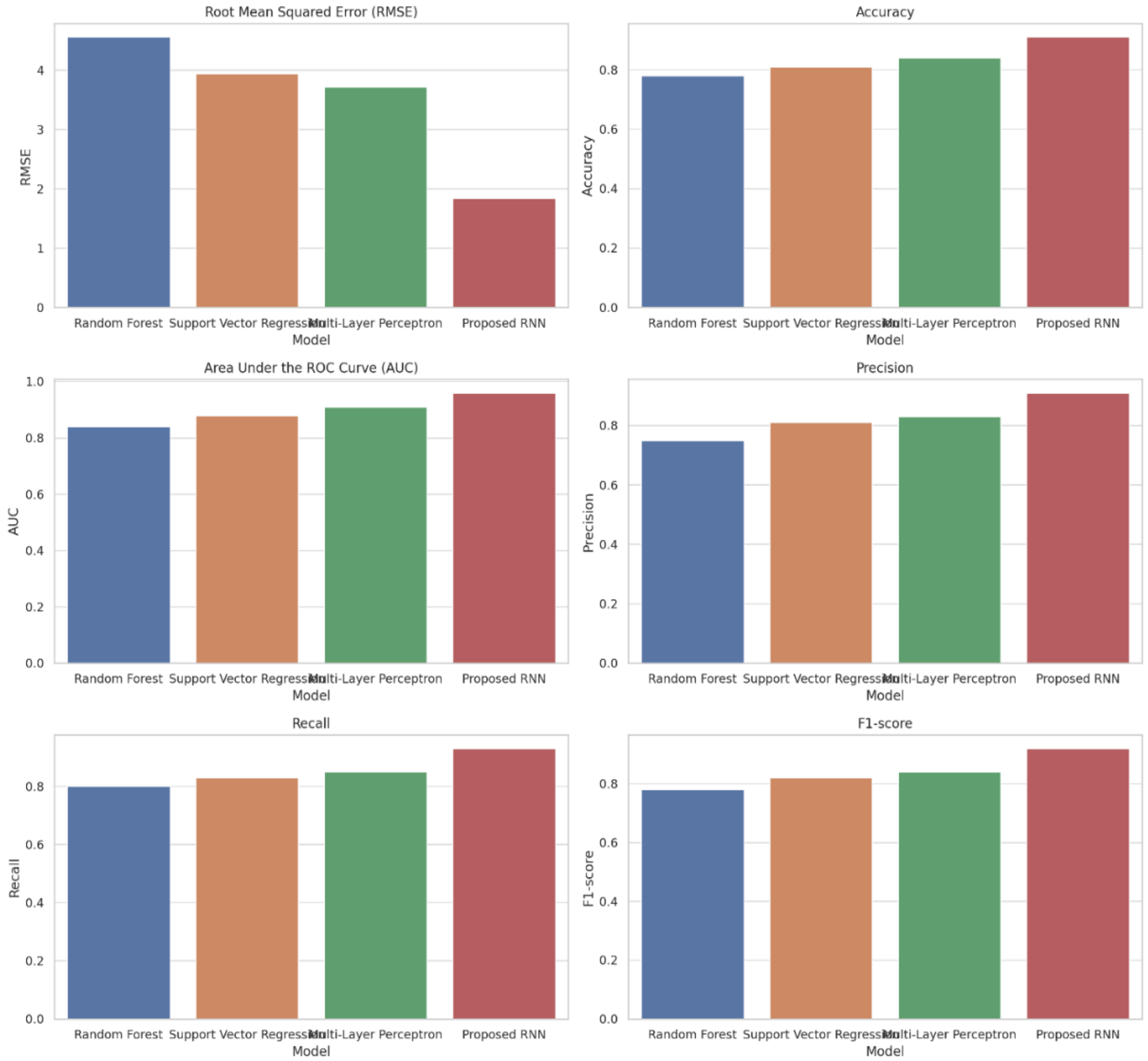

Figure 2. Experiment 2 (cross-validation) model performance comparison

The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test used to compare two related samples. Table 7 shows the p-values for each comparison between the two models. The p-value is the probability of observing a test statistic as extreme as the one computed, assuming the null hypothesis is actual. Based on the p-values, we can make the following conclusions:

- No significant difference exists between random forest and support vector regression models in all evaluation metrics since the p-value is more significant than 0.05 (the usual significance level).
- There is a significant difference between random forest and multi-layer perceptron models regarding RMSE and accuracy evaluation metrics since the p-value is less than 0.05. However, there is no significant difference between AUC, precision, recall, and F1-score.
- There is a significant difference between the random forest and proposed approach (RNN) models in all evaluation metrics since the p-value is less than 0.05.
- There is no significant difference between support vector regression and multi-layer perceptron models in all evaluation metrics.
- There is a significant difference between support vector regression and proposed approach (RNN) models regarding RMSE and AUC evaluation metrics since the p-value is less than 0.05. However, there is no significant difference in accuracy, precision, recall, and F1 score.
- There is a significant difference between the multi-layer perceptron and proposed approach (RNN) models regarding RMSE and AUC evaluation metrics since the p-value is less than 0.05. However, there is no significant difference in accuracy, precision, recall, and F1 score.

Overall, the proposed approach (RNN) model outperforms all other models in all evaluation metrics, except for the RMSE, where the multi-layer perceptron performs slightly better. The random forest and support vector regression models perform relatively well, while the multi-layer perceptron performs moderately.

Table 7. Wilcoxson evaluations results

Model 1	Model 2	p-value
Random forest	Support vector regression	0.143
Random forest	Multi-layer perceptron	0.046
Random forest	Proposed approach (RNN)	0.003
Support vector regression	Multi-layer perceptron	0.697
Support vector regression	Proposed approach (RNN)	0.003
Multi-layer perceptron	Proposed approach (RNN)	0.012

5. DISCUSSION AND IMPLICATIONS

While previous studies have explored various techniques for predicting the long-term citation impact of scientific publications, such as random forest, support vector regression, and multi-layer perceptron, they have not explicitly addressed the potential of deep learning techniques, particularly those built on sequential modeling, in capturing the temporal dynamics of citation data. The implications of our study are profound for the field of bibliometrics and scientometrics in general. By showing that our approach performs better than state-of-the-art techniques for long-term prediction of citation counts, we indirectly provide strong evidence that deep learning techniques (specifically those built on sequential modeling) can be valuable in assessing the impact of scientific research. This is underlined when comparing the results to existing work. In the domain in question (i.e., 10-year citation counts), we significantly lose performance compared to the random forest, support vector regression, and multi-layer perceptron. However, we show that our approach outperforms those techniques that were originally proposed to provide a benchmark for analyzing the influence of data preprocessing. In contrast, we explicitly refer to RNNs and the capability of capturing temporal dynamics in citation data.

The method we present here could be used to address an information need in the academic community. Researchers may appreciate the potential for our method to provide them with insight into the long-term impact of their work, helping them make better-informed decisions about which directions to drive their project and how to publish their findings. Funding agencies and policy researchers can use the ability to predict the future impact of research results to help them allocate resources and effectively shape research priorities.

Consequently, the approach developed in this study could be expanded to other domains outside bibliometrics. For instance RNNs are commonly used for time series forecasting. Such forecasts are of much importance in domains such as finance, healthcare, and energy, where predicting future trends based on historical data would have significant results. While these results are encouraging, our study has several limitations. First, the data set used in our experiments contains publications from only one university. Therefore, the generalizability of our findings to other institutions remains to be investigated. Second, both systems rely on citation counts only and neglect other factors that could impact the outcome of scientific research, such as the quality of the work or the reputation of the authors.

6. CONCLUSION

In conclusion, this research presented a comparative analysis between random forest, support vector regression, multi-layer perceptron, and a RNN-based approach for predicting the enduring citation impact of academic publications. Extensive model validation on a case study dataset demonstrated the superior capabilities of the proposed RNN architecture in accurately forecasting long-term citation counts as quantified through the lowest RMSE of 1.84 and highest accuracy of 0.91. Statistical significance testing verified marked improvements over alternatives. The findings validate RNNs as an efficacious deep learning solution by leveraging sequence modeling strengths in capturing temporal dynamics to predict citation trajectories robustly. The introduced methodology promises to aid stakeholders, including academics, institutions, and policy agencies, in quantifying prospective research impact for informed evaluation. While current results are promising, enhancements can be incorporated regarding input features from the text, author attributes, and network structure to lift predictive fidelity further.

Additionally, larger datasets spanning diverse institutions and scientific domains would accentuate generalizability. This research highlights the potential of sophisticated machine learning techniques in temporal predictive analytics within bibliometrics and scientometrics. The proposed approach offers a template to build advanced systems for citation-driven research impact.

ACKNOWLEDGMENT

This research was supported by the Deanship of Scientific Research and Graduate Studies at the University of Petra, No. 3-3-2022.

REFERENCES

[1] J. A. Balogun, "The fundamentals, misuse and abuses of bibliometrics," *Health Research in Nigeria*, pp. 11–56, 2023, doi: 10.1007/978-981-19-7097-9_2.
[2] L. Egghe and R. Rousseau, "An informetric model for the Hirsch-index," *Scientometrics*, vol. 69, no. 1, pp. 121–129, 2006, doi: 10.1007/s11192-006-0143-8.
[3] E. Garfield, "The history and meaning of the journal impact factor," *Jama*, vol. 295, no. 1, pp. 90–93, 2006, doi: 10.1001/jama.295.1.90.
[4] L. Wildgaard, J. W. Schneider, and B. Larsen, "A review of the characteristics of 108 author-level bibliometric indicators," *Scientometrics*, vol. 101, no. 1, pp. 125–158, 2014, doi: 10.1007/s11192-014-1423-3.
[5] S. Gupta, N. Kumar, and S. Bhalla, "Citation metrics and evaluation of journals and conferences," *Journal of Information Science*, p. 016555152311514, Feb. 2023, doi: 10.1177/01655515231151411.
[6] T. Amjad *et al.*, "Standing on the shoulders of giants," *Nature Photonics*, vol. 13, no. 5, pp. 297–297, May 2019, doi: 10.1038/s41566-019-0438-y.
[7] H. F. Chan, F. G. Mixon, and B. Torgler, "Relation of early career performance and recognition to the probability of winning the Nobel Prize in economics," *Scientometrics*, vol. 114, no. 3, pp. 1069–1086, 2018, doi: 10.1007/s11192-017-2614-5.
[8] C. T. Lamb, S. L. Gilbert, and A. T. Ford, "Tweet success? Scientific communication correlates with increased citations in Ecology and Conservation," *PeerJ*, vol. 6, no. 4, p. e4564, Apr. 2018, doi: 10.7717/peerj.4564.
[9] Q. Zhao and X. Feng, "Utilizing citation network structure to predict paper citation counts: a deep learning approach," *Journal of Informetrics*, vol. 16, no. 1, p. 101235, Feb. 2022, doi: 10.1016/j.joi.2021.101235.
[10] M. E. J. Newman, "Prediction of highly cited papers," *EPL (Europhysics Letters)*, vol. 105, no. 2, p. 28002, Jan. 2014, doi: 10.1209/0295-5075/105/28002.
[11] S. Ayaz, N. Masood, and M. A. Islam, "Predicting scientific impact based on h-index," *Scientometrics*, vol. 114, no. 3, pp. 993–1010, 2018, doi: 10.1007/s11192-017-2618-1.
[12] M. Rocha-e-Silva, "Journal impact factors for the year-after the next can be objectively predicted," *Medical Express*, vol. 3, no. 5, 2016, doi: 10.5935/MedicalExpress.2016.05.06.
[13] A. Mazloumian, "Predicting Scholars' scientific impact," *PLoS ONE*, vol. 7, no. 11, p. e49246, Nov. 2012, doi: 10.1371/journal.pone.0049246.
[14] C. Castillo, D. Donato, and A. Gionis, "Estimating number of citations using author reputation," in *International Symposium on String Processing and Information Retrieval*, 2007, pp. 107–117, doi: 10.1007/978-3-540-75530-2_10.
[15] L. Bornmann, L. Leydesdorff, and J. Wang, "How to improve the prediction based on citation impact percentiles for years shortly after the publication date?," *Journal of Informetrics*, vol. 8, no. 1, pp. 175–180, 2014, doi: 10.1016/j.joi.2013.11.005.
[16] K. Mansour, E. Al-Daoud, and B. Al-Karaky, "I. Predicting future citation counts using machine learning," in *2021 22nd International Arab Conference on Information Technology (ACIT)*, Dec. 2021, pp. 1–6, doi: 10.1109/ACIT53391.2021.9677228.
[17] A. Abrishami and S. Aliakbary, "Predicting citation counts based on deep neural network learning techniques," *Journal of Informetrics*, vol. 13, no. 2, pp. 485–499, 2019, doi: 10.1016/j.joi.2019.02.011.
[18] T. Matsui, K. Kanamori, and H. Ohwada, "Predicting future citation count using bibliographic and author information of articles," *International Journal of Machine Learning and Computing*, vol. 4, no. 2, pp. 139–141, 2014, doi: 10.7763/ijmlc.2014.v4.401.
[19] K. K. Ahuja, "Prediction for citation and publication count using regression analysis," in *Proceedings of the International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), I-SMAC 2018*, 2018, pp. 460–463, doi: 10.1109/I-SMAC.2018.8653719.
[20] Z. Su, "Prediction of future citation count with machine learning and neural network," in *Proceedings of 2020 Asia-Pacific Conference on Image Processing, Electronics and Computers, IPEC 2020*, 2020, pp. 101–104, doi: 10.1109/IPEC49694.2020.9114959.
[21] Y. Du, "Predicting Paper acceptance rank in citation graph."
[22] J. Wen, L. Wu, and J. Chai, "Paper citation count prediction based on recurrent neural network with gated recurrent unit," in *ICEIEC 2020 - Proceedings of 2020 IEEE 10th International Conference on Electronics Information and Emergency Communication*, 2020, pp. 303–306, doi: 10.1109/ICEIEC49280.2020.9152330.
[23] W. L. Croft and J.-R. Sack, "Predicting the citation count and CiteScore of journals one year in advance," *Journal of Informetrics*, vol. 16, no. 4, p. 101349, Nov. 2022, doi: 10.1016/j.joi.2022.101349.
[24] X. Ruan, Y. Zhu, J. Li, and Y. Cheng, "Predicting the citation counts of individual papers via a BP neural network," *Journal of Informetrics*, vol. 14, no. 3, p. 101039, Aug. 2020, doi: 10.1016/j.joi.2020.101039.
[25] D. P. Mandic and J. A. Chambers, *Recurrent neural networks for prediction : learning algorithms, architectures, and stability*. United States: John Wiley & Sons, Inc., 2001.
[26] J. Moolayil, "An introduction to deep learning and Keras," in *Learn Keras for Deep Neural Networks*, Berkeley, CA: Apress, 2019, pp. 1–16.
[27] D. K. R. Gaddam, M. D. Ansari, S. Vuppala, V. K. Gunjan, and M. M. Sati, "A performance comparison of optimization algorithms on a generated dataset," in *ICDSMLA 2020: Proceedings of the 2nd International Conference on Data Science, Machine Learning and Applications*, 2022, vol. 783, pp. 1407–1415, doi: 10.1007/978-981-16-3690-5_135.

BIOGRAPHIES OF AUTHORS

Naser Jamal Assistant Professor in the Department of Data Science and Artificial Intelligence, College of Information Technology at the University of Petra. He received his Dr. Sc. Eng. (Doctor of Science in Engineering) degree in 1997 from the Riga Aviation University, Latvia. He has more than two decades of teaching experience. He started as an instructor at the University of Petra (1998-2003), and then he joined Philadelphia University (2003-2005). Since then, he has become a permanent faculty member in the College of Information Technology at the University of Petra, where he has teaching and administrative responsibilities. At the University of Petra, he has held several administrative positions, including Director of the Computer and Information Center (1999-2003), Dean of Admission and Registration (2005-2018) (2022-present), Dean Assistant for Quality Assurance and Student Affairs (2020-2022). His current research interests lie in the areas of data science, artificial intelligence, optimization algorithms, and distributed systems. He can be contacted at email: njamal@uop.edu.jo.

Mohammad Alauthman Associate Professor and Chair of the Information Security Department at the Faculty of Information Technology, University of Petra in Amman, Jordan. His research interests lie in network security, intrusion detection systems, and the application of artificial intelligence techniques like machine learning and deep learning for botnet detection, DDoS detection, spam email filtering, IoT security, and network traffic classification. He obtained his Ph.D. in Computer Science with a focus on Network Security from Northumbria University in the UK in 2016. He has been awarded several research grants in the field of network security, supporting his work on advanced intrusion detection systems and the development of AI-driven security solutions. These grants have enabled him to conduct extensive research, collaborate with international experts. His funded projects have focused on enhancing the resilience of network infrastructures against emerging threats and improving the accuracy of threat detection mechanisms. He can be contacted at email: Mohammad.alauthman@uop.edu.jo.

Muhannad Malhis is the Chief Information Officer (CIO) and Director of Information and Communication Technology (ICT) at the University of Petra in Amman, Jordan. With a strong background in information technology and management, Malhis brings extensive experience to his role, driving the university's digital transformation and innovation strategies. He can be contacted at email: mmalhis@uop.edu.jo.

Dr. Abdelraouf M. Ishtaiwi is a highly knowledge-achieving academic with over 22 years of experience in teaching and research in artificial intelligence (AI). He obtained a First-Class Honor and a Master's degree from a well-reputed university, Griffith University in Brisbane, in 2001 and 2007, respectively. He contribution to the AI field is enormous. He has done groundbreaking research in machine learning, local search algorithms, and optimization methods. He can be contacted at email: aishtaiwi@uop.edu.jo.