



Research article



Defining quantitative rules for identifying influential researchers: Insights from mathematics domain

Ghulam Mustafa^a, Abid Rauf^a, Ahmad Sami Al-Shamayleh^b,
Muhammad Tanvir Afzal^c, Ali Waqas^{d,*}, Adnan Akhunzada^e

^a Department of Computer Science, University of Engineering and Technology, Taxila, 47080, Punjab, Pakistan

^b Department of Data Science and Artificial Intelligence, Faculty of Information Technology, Al-Ahliyya Amman University, Amman, 19328, Jordan

^c Department of Computing, Shifa Tameer-e-Millat University, Islamabad, 44000, Pakistan

^d Mathematics Foundation Program, University of Doha for Science and Technology, Doha, 122014, Qatar

^e College of Computing & IT, University of Doha for Science and Technology, Doha, 122014, Qatar

ARTICLE INFO

Dataset link: https://github.com/ghulammustafacomsat/Mathematics_dataset

Keywords:

Decision tree
Rule mining
Author assessment parameters
Multilayer perceptron
Recursive elimination technique

ABSTRACT

In the midst of a vast amount of scientific literature, the need for specific rules arise especially when it comes to deciding which impactful researchers should be nominated. These rules are based on measurable quantities that can easily be applied to a researcher's quantitative data. Various search engines, like Google Scholar, Semantic Scholar, Web of Science etc. Are used for recording metadata such as the researcher's total publications, their citations, h-index etc. However, the scientific community has not yet agreed upon a single set of criteria that a researcher has to meet in order to secure a spot on to the list of impactful researchers. In this study, we have provided a comprehensive set of rules for the scientific community within the field of mathematics, derived from the top five quantitative parameters belonging to each category. Within each categorical grouping, we meticulously selected the five most pivotal parameters. This selection process was guided by an importance score, that was derived after assessing its influence on the model's performance in the classification of data pertaining to both awardees and non awardees. To perform the experiment, we focused on the field of mathematics and used a dataset containing 525 individuals who received awards and 525 individuals who did not receive awards. The rules were developed for each parameter category using the Decision Tree Algorithm, which achieved an average accuracy of 70 to 75 percent for identifying awardees in mathematics domains. Moreover, the highest-ranked parameters belonging to each category were successful in elevating over 50 to 55 percent of the award recipients to positions within the top 100 ranked researchers' list. These findings have the potential to serve as a guidance for individual researchers, who aimed on to making it to the esteemed list of distinguished scientists. Additionally, the scientific community can utilize these rules to sift through the roster of researchers for a subjective evaluation, facilitating the recognition and rewarding of exceptional researchers in the field.

* This Research is funded by Qatar National Library.

* Corresponding author.

E-mail address: ali.waqas@udst.edu.qa (A. Waqas).

<https://doi.org/10.1016/j.heliyon.2024.e30318>

Received 11 September 2023; Received in revised form 22 April 2024; Accepted 23 April 2024

Available online 29 April 2024

2405-8440/© 2024 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Millions of active researchers generate a huge corpus of research articles daily [1–3]. However, the immediate impact and quality of these articles remain uncertain. It takes time for these articles to receive acknowledgment. Thus, creating a delay for award-winning researchers to impact the scientific community based on their research. The evaluation of the researcher's and its impact has created a big fuss within the scientific literature, resulting in various methods being proposed. Different scientific societies have adopted distinct approaches to measure the impact of the researchers. Despite the ongoing discussions, a consensus has yet to be reached within the scientific community regarding a unified method for recognizing outstanding contributions by scientists. Numerous methods for assessing the researchers impact have been explored in the literature [4][5][6]. One approach involves collecting reviews from domain experts. While this method holds credibility due to the expertise of the professionals of the field but it is a labor-intensive manual process. Another common strategy in use involves quantifying the total number of publications or citations. However, these methods are not without their limitations, as researchers might publish in the lower-quality journals [7] or engage in self-citation, which involves citing their own works [8].

The scientific community experienced a significant change when Hirsch proposed the h-index, a measure that looks at both the quality and the quantity quality of a researcher's work [9][10]. Despite the widespread adoption of this index by numerous researchers, subsequent studies have revealed certain constraints and shortcomings. For instance, Diense points out that obtaining more citations for h-core articles does not necessarily improve the h-index score [11]. Another researcher, [12], explained in a survey that less cited articles are ignored by the h-index, even when they have similar citations to the more-cited ones, and so on. To address these issues, researchers have developed many different methods to measure research impacts [13]. A recent study [12] found that in the literature, there are now over 70 proposed measures, such as A-index [14], AR-index [15], M-Quotient [16], k index [17], etc. Whenever a new method is developed, it is usually tested using different sets of hypothetical situations or by data provided. This makes it difficult to determine which measure is most important among all [18].

In light of the introduction of indices designed to overcome the limitations of the h-index, several studies have put forth evaluations of these newly proposed parameters. In 2016, Dienes assessed the h-index, g-index, and complementary h-index to rank researchers in the field of mathematics [11]. In 2018, De et al. conducted an evaluation of the h-index and its variants in civil engineering, including those based on citation intensity and publication age [19]. Schreiber et al. (2019) evaluated the h-index and its variants using a dataset in neurosciences [20]. In 2019, Ain et al. [18] and Ghani et al. [21] systematically evaluated citation intensity-based indices of the h-index, utilizing a comprehensive dataset from mathematics. Moreira et al. (2021) assessed various indices on a comprehensive dataset in civil engineering, aiming to identify the most effective metrics for evaluating author performance [22]. Recently, Mustafa et al. (2023a, 2023b) evaluated Publication, Citation-based metrics, and publication age-based parameters on a mathematics dataset [23], [24]. Ahmed et al. (2023) [3] evaluated author count-based parameters on the same dataset. Usman et al. 2021 [7] employed logistic regression for ranking author assessment parameters, with a limited parameter number; in their second study, they used a neural network for ranking the same author assessment parameters. Similarly, Ahmed et al. 2023 [6] utilized dynamic random forest with brute optimizer to rank author assessment parameters.

In the present circumstances, to the best of our knowledge, there is an absence of well-defined rules that encompass the abundance of parameters at hand. These rules clarify which parameters should be used and which ones should not be. Moreover, there are no criteria defined for a researcher to set a goal that could make him an impactful researcher. This emphasizes the need to establish clear rules that can guide both newcomers in making significant contributions and the scientific community in selecting candidates for research awards [25,26]. Recent studies have attempted to evaluate the impact of parameters (h-index & its variations) on the recipients of research awards [23,24,26]. However, these studies considered only a limited number of parameters. Another study [8] proposed some rules; but they were drawn from a restricted set of parameters.

In this study, we formulated rules based on parameters ranked at the top across various categories. We utilized nearly 64 distinct author assessment parameters, which were categorized into four main groups such as, 1) primitive parameters, 2) publication and citation count-based parameters, 3) publication age-based parameters, & 4) Author Count-based Parameters. Given the extensive number of parameters, our initial step involved ranking using a Multilayer Perceptron classifier in conjunction with the Recursive Elimination technique. Through this process, we ascertained the importance score for each parameter, which allowed us to rank the parameters within each category. Subsequently, we select the top five parameters for rule generation. The mining rules were accomplished using a Decision Tree algorithm from Machine Learning. This study holds the potential to significantly aid individual researchers in enhancing their prospects for winning awards in the right direction. Moreover, it provides the scientific community with a means to select the best researchers by quantifying their impacts within their respective fields. For experimentation, we employed a dataset from the domain of mathematics, which is already used in the existing literature. This dataset comprised 525 award-winning researchers and an equivalent number of non-award-winning researchers.

In this study we have addressed the following research questions.

- Which specific quantitative parameters exert the most substantial influence on researchers who have received awards within each individual category?
- Among the categories of parameters considered, which ones exhibit a higher degree of influence in the process of rule generation, thereby leading to broader coverage of awardees?

The finding from this study hold the potential to serve as a foundation upon which the scientific community can establish criteria for awarding scientific honors. Additionally, these findings can also function as guiding principles for fledgling researchers venturing into the field.

The rest of this paper is organized as follows. In literature Review, we present a concise overview of the literature, the Methodology outlines our approach for ranking and establishing quantitative rules for scientific parameters. The Results, presents the results of our study. In the Conclusion section, we present a summary of this research. Lastly in future Work, we will explore the potential areas for future advancements in this domain.

2. Literature review

In this era, the scientific community needs to define universal criteria for evaluating researchers' performance to ensure equitable and impartial ranking procedures. Researcher impact is assessed through various metrics, including publication count, citation count, h-index, h-index variation, and a combination of approaches. Scholars have nominated for various initiatives within the academic and professional community through subjective evaluations [27–30]. However, these traditional approaches primarily rely on metrics, such as publication and citation counts, which are not universally applicable for global assessments. Critiques of these metrics have arisen because of their limitations. Researchers who produce a high volume of publications are often considered active researchers, but this can be misleading if the publications are in low-impact journals or low impact conferences [31]. Similar to this, citation counts are not always reliable indicators of research influence, as they can be influenced by self-citations or negative citations [32]. In response to these shortcomings, the h-index has gained attention for its simplicity; however, it also has its own limitations. For example, it does not necessarily reflect a researcher's impact when citations increase for h-core papers [33]. Also this is not well suited for new researchers who need time to accumulate citations and increase their h-index. Furthermore, it sometimes favors researchers who are currently inactive [34,35]. To overcome these issues, scientists have proposed numerous alternative parameter to complement the h-index's shortcomings such as A index, k index f index and so on.

With the proposition of new indices, some researchers evaluated (h index & its variants) for finding the impact of these parameters. [13] conducted a study that focused on assessing h-indices. They used award data from mathematical societies, which are considered as a benchmark. In this study, the author identified that the complete h-index outperformed all other indices. In their next study, [26] performed the evaluation of the h index and its variants, which were not discussed in their previous study, and found that the h index outperformed the remaining alternatives. The [36] assessed quantitative parameters within the domain of scientific neuroscience. In their study, they have reported that the hg-index and R-index play a significant role in prominently positioning awardees at the forefront in the researchers list. Similarly, [37] conducted a study to assess quantitative parameters within the mathematics domain. They established correlations between these quantitative parameters, and ranked these parameters based on award-winning researchers. However, they focused on researchers who received awards prior to the introduction of these parameters. As such, the correlation found between awardees and quantitative parameters could be coincidental, as those awardees were not nominated based on indices (h index & its variants). To mitigate this constraint, [7] conducted a study for the assessment of the indices (h index & its variants). This approach is based on the utilization of data from the civil engineering domain. Instead of employing a random selection process, they opted for a targeted approach. They specifically selected individuals who had received awards and those who had not within the same temporal span. This selection was exclusive to researchers who had been honored by prestigious civil engineering scientific societies since 2005. However, their dataset is not sufficiently comprehensive to definitively determine the importance of parameters for award winners. Furthermore, the current literature lacks clearly defined criteria for these parameters, which researchers can use as a foundational benchmark for subjective conduction evaluation. Furthermore, one of the study conducted by [8], generated a rule for scientific community but they chose a top parameter for defining rules from a limited no. of total parameter. Recently, [23,24] conducted two studies, in first they have evaluated publication and citation count based parameter in which they found that normalized h index outperformed all the remaining parameters and in second they have evaluated publication age based parameter in which they found that Ar index outperformed the remaining indices. For experimental purposes, they used a mathematical domain dataset.

After conducting a comprehensive literature review, we studied a number of proposed techniques and conducted evaluation surveys of these parameters. From these studies, our findings indicate that approximately a decade ago, researchers were evaluated based on their publications and citations. However, the approach shifted in the following decade, with a focus on using variations of the h-index without accounting for study limitations or context. When novel methodologies are introduced, they often emerge in unconventional ways or are tested using diverse datasets. Due to this divergence in data sources and validation methods, assessing the individual indices importance is challenging. Thus, it is crucial to thoroughly evaluate the indices when deriving rules that incorporate individual parameters or parameter combinations. This enables the distinction between prolific and nonprolific researchers. To formulate these rules, it is necessary to understand the patterns of award recipients. These rules could potentially elevate non awardees by including them in the pool of awardees.

3. Methodology

After a critical analysis of the literature, we will now discuss rule mining for researchers belonging to the domain of mathematics. We employed a multilayer perceptron classifier with a recursive elimination technique to find the most significant quantitative parameters for each category based on the abundance of parameters. Fig. 1 represents the architecture diagram containing different steps, such as (i) Domain Selection & Dataset Collection, (ii) Calculation of Author Assessment Parameters, (iii) Ranking of Author

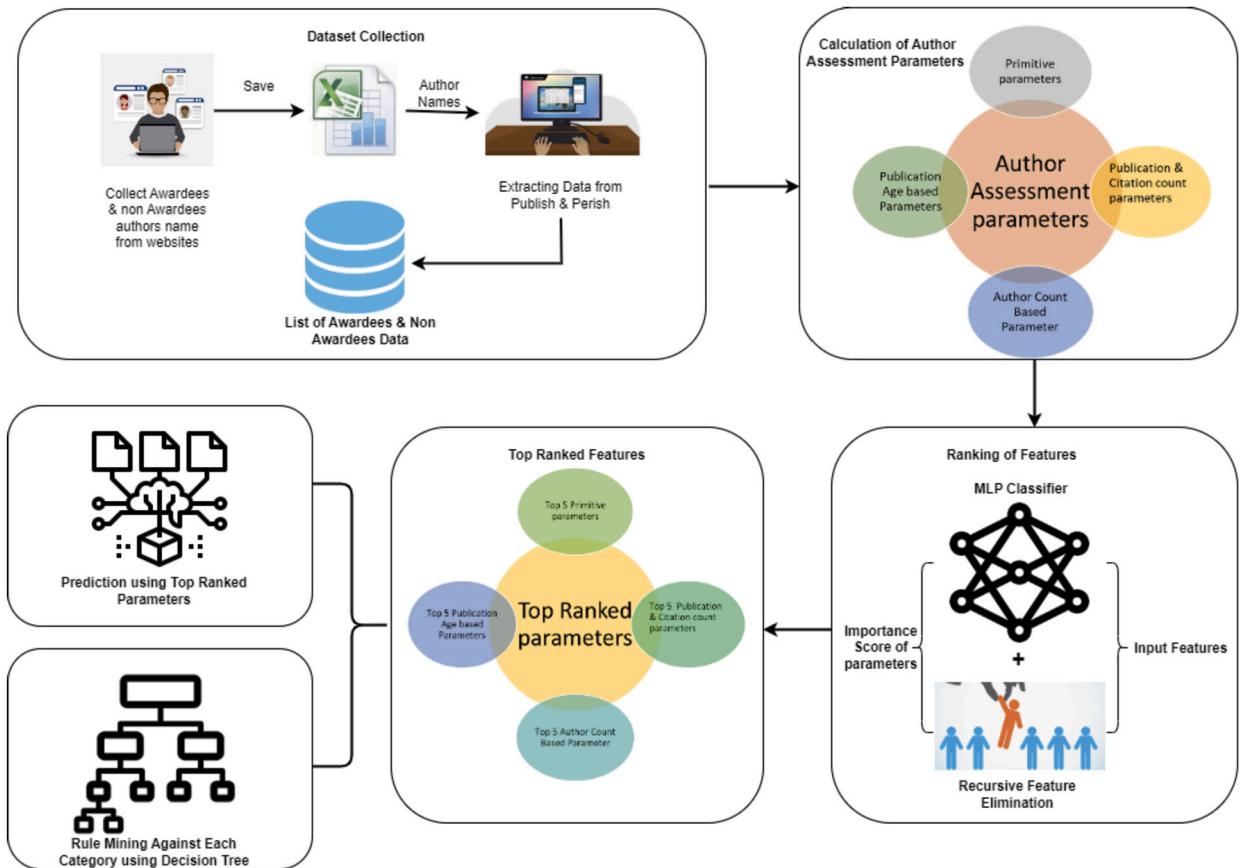


Fig. 1. The architecture diagram represents multiple phases in the proposed system. The dataset collection phase illustrates the process of gathering data, followed by the depiction of parameter categories in the next phase. Subsequently, ranking techniques are represented, and the subsequent phases portray the prediction of awardees based on the top-ranked parameters and the generation of rules for each category.

Assessment Parameters, (iv) prediction of awardees using top-ranked parameters, & (v) rule mining using the top five Parameters against Each Category.

In the following sections, we explain all of the above points.

3.1. Domain selection and dataset collection

For experimentation, a dataset specific to a particular domain was required. In this study, the Mathematics domain was selected as the main focus. That too, due to its extensive history and significant contributions in research. Which makes it the optimal choice for evaluating the proposed methodology. Moreover, within this domain, various scientific societies bestowed annual awards to distinguished researchers, that were based on the impact of their research work. Due to the comprehensiveness of this domain, several studies have already opted for its utilization [18,23,24,38]. Within this domain, we selected the dataset used by [23,24]. This dataset comprised 1050 records, encompassing information about both awardees and non-awardees. Specifically, the dataset contains 525 non-awardees and 525 awardees. For the awardees' data, a list of 30 internationally recognized awards, along with their significant importance (within the mathematical community) were compiled. These awards are esteemed accomplishments for mathematicians and researchers were awarded by renowned mathematical societies such as, the London Mathematical Society (LMS),¹ International Mathematical Union (IMU),² Norwegian Academy of Science and Letters (NASL),³ and the American Mathematical Society (AMS).⁴ The proportion of data related to non-awardees in the dataset was derived from [18] & [21]. However, since the original datasets used by [18] & [21] contained only a limited number of awardees entries, extending until 2013, [23] expanded this dataset by gathering updated information on awardees up to 2023. The detail statistics of dataset is given in Table 1. For the collection of awardees' data, various society websites (LMS, IMU, NASL, and AMS) within the mathematics domain were visited to gather names

¹ <https://www.lms.ac.uk/>.

² <https://www.mathunion.org/>.

³ <https://dnva.no/norwegian-academy-science-and-letters>.

⁴ <https://www.ams.org/home/page>.

Table 1
Dataset Statistics.

Dataset Statistics	
Total Authors	1050 (525 Awardees / 525 Non Awardees)
Citation Count	14,369,007
Publication Count	204,796
Benchmark Dataset Awardees	
Total Awardees	525
Total Awardees belong to AMS	257
Total Awardees belong to IMU	59
Total Awardees belong to LMS	188
Total Awardees belong to NASL	21

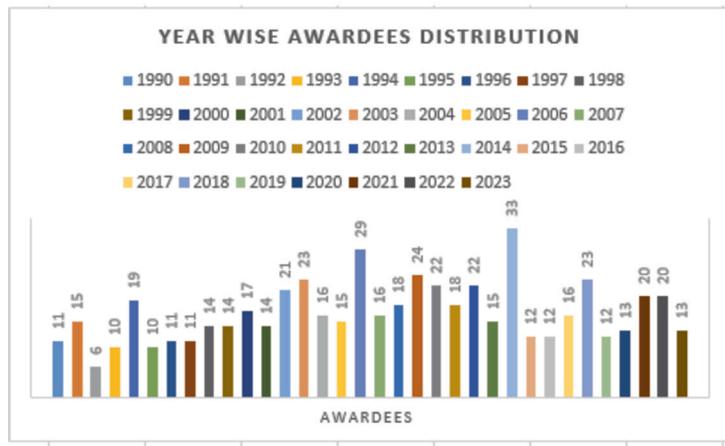


Fig. 2. Present the count of awardees year-wise, with each bar representing the number of awardees for a specific year.

and corresponding years of awards given to researchers over the past three decades. The distribution of awards across years is presented in Fig. 2. To extract data on both awardees and non-awardees, the Publish or Perish⁵ platform was utilized, employing a hold-on strategy that enabled the collection of researchers' data even before their award-receiving years. This tool utilized a sophisticated algorithm to retrieve primary data along with author metadata from Google Scholar. In order to ensure equity within the dataset, data on non-awardees was collected in the same proportion of the number of awardees for each year. For instance, if there were 15 awardees in 1991, data from the 15 non-awardees prior to 1991 were collected using similar techniques. Before conducting any analysis or evaluation, thorough data cleansing from sources such as Google Scholar was essential. This process aims to eliminate incorrect or irrelevant information, commonly referred to as noise, which could compromise the accuracy of the results. The data cleansing comprises several steps, including verification of data accuracy and removal of duplicate entries. In our extensive research dataset, two crucial processes were performed to enhance the quality and relevance of the data. First, a filter is applied to make sure the consistency of each research article within the mathematical field, removing any irrelevant content. This step aided in refining the dataset to concentrate solely on the relevant domain. Subsequently, an author disambiguation process was executed to find and remove duplicate entries resulting from researchers publishing using different variation of names.

3.2. Calculation of author assessment parameters

In this section, we calculate the Sixty-four author assessment parameter belonging to four different categories on the collected dataset. The list of the categories with respective indices is presented in Fig. 3.

Moreover, the formulas of these indices with explanation are presented in Table 3 (Appendix A).

3.3. Ranking of author assessment parameters

In the realm of machine learning, the task of feature ranking holds significant importance. It plays a pivotal role in discerning the most influential features used for a variety of purposes [39]. These purposes encompass dimensionality reduction, enhancing interpretability, mitigating overfitting, optimizing prediction processes, and refining Feature Engineering. The overall process of feature ranking is depicted in Fig. 4. To assess the importance of these features, we employed a multilayer perceptron classifier (MLP) in conjunction with a recursive elimination technique. The MLP classifier is widely recognized in the literature for its versatility,

⁵ <https://harzing.com/resources/publish-or-perish>.

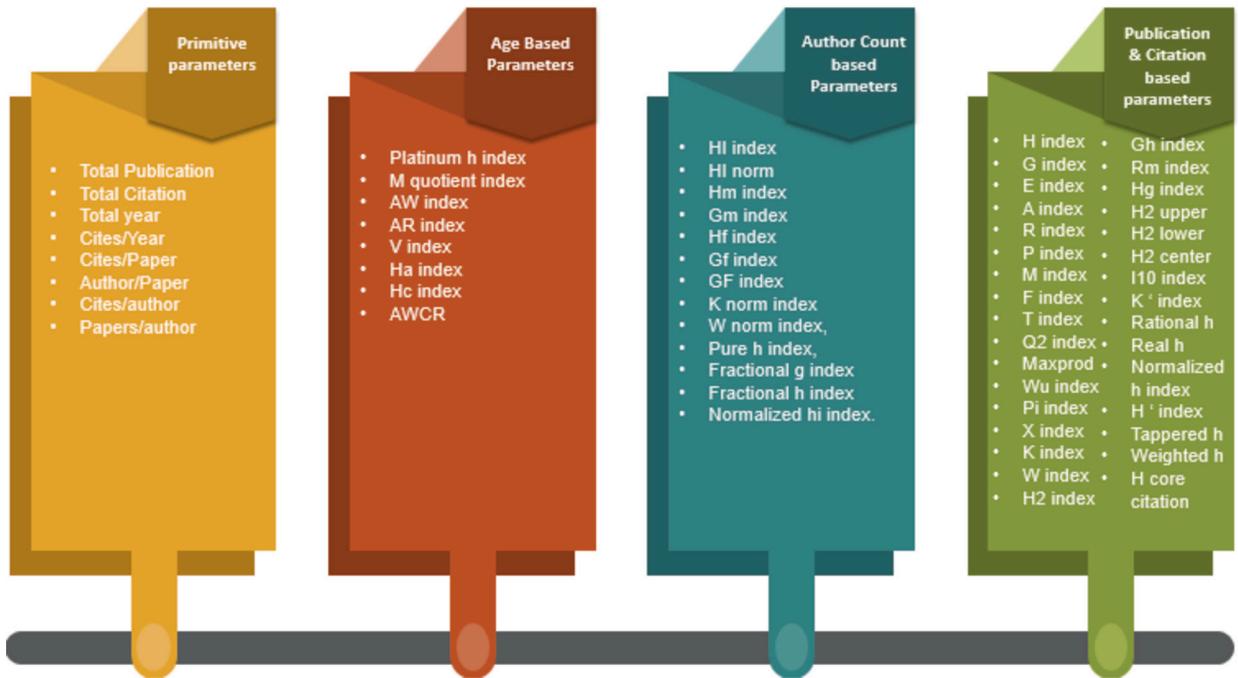


Fig. 3. Parameter Categories, with the names of parameters listed under each respective category.

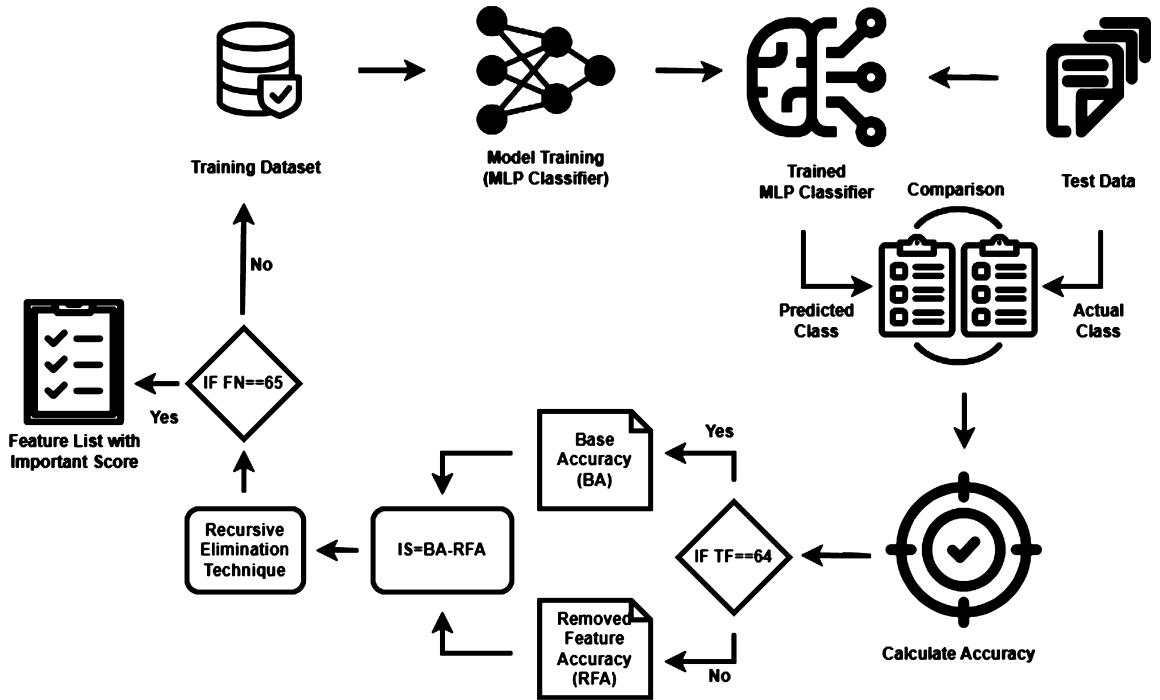


Fig. 4. Feature Ranking process, the dataset was initially divided into a 20:80 ratio. A MLC was trained, and a validation sample was provided to the trained model. The accuracy achieved during this prediction stage was considered the Baseline Accuracy (BA) when Total feature (TF) is equal to sixty four. In the next iteration, the feature removal process began. One parameter was removed from the feature list, and the model was trained on the updated feature set. The new accuracy (RFA) obtained from test sample prediction was then subtracted from the BA, gives Importance Score (IS) against removed feature. The entire procedure was iterated for each parameter in the dataset. Upon iterating through all features, when the Feature Number (FN) reached sixty-five, it signified the completion of iterating through all features. The algorithm generated two lists. The first list contains the names of the features, while the second list contains their corresponding importance scores. Based on these importance scores, the parameters were sorted, providing us with a parameter ranking.

serving as a fundamental algorithm in various problem domains, including classification, prediction, regression, among others [40]. This classifier is constructed as a feed-forward artificial neural network, consisting of multiple hidden layers [41]. In classification-related tasks, the number of features aligns with the number of neurons in the input layer, while the total number of classes used for data classification corresponds to the number of neurons in the output layer. Intermediate layers, situated between the input and output layers, form a fully connected network trained through the backpropagation algorithm. During the forward propagation phase, the network computes the output for each layer using an activation function based on the previous layer's output, combined with the respective weight and bias values presented in Equation (1).

$$X = WA + b \quad (1)$$

Where X represent output matrix, W represent weight matrix and b represent bias vector.

To guarantee that the MLP's output falls within a predefined range, an activation function is applied. This function serves to normalize the output of each layer, ensuring that it conforms to the desired range, as illustrated in Equation (2).

$$C = g(X) \quad (2)$$

Where C represent the activated output matrix.

In this study, we used activation function name as Rectified Linear Unit (ReLU) for the hidden layers and Softmax activation function for the final output layer. The activation function, is defined by Equation (3), acts by setting values less than zero to zero while preserving positive values. This introduces essential nonlinearity into the network, enhancing its capacity to learn complex patterns.

$$f(X) = \max(0, X) \quad (3)$$

In Equation (4) softmax function is defined which is a well-established choice for classification problems. It mitigates the limitations of the sigmoid function and ensures that the probabilities assigned to each class in the output layer sum up to one. By applying Softmax, we can identify the most probable prediction for a given set of inputs.

$$\text{Softmax}(X_i) = \frac{e^{X_i}}{\sum_{i=1}^J e^{X_i}} \quad (4)$$

Where J represent the number of classes and Xi denote the ith output value.

Furthermore, loss function is represented in Equation (5), which is used to quantify the disparity between predicted and actual values within the MLP. It stands as a pivotal metric for evaluating the model's performance. The loss function's computation enables the assessment of the degree of mismatch between the predicted outcomes and the desired target values. Following this assessment, a backpropagation algorithm is employed to iteratively adapt the weights (w) and biases (b) within the neural network. This iterative process is instrumental in optimizing the model's performance by minimizing the loss. Through this fine-tuning of parameters, the network is progressively refined to enhance its ability to make accurate predictions.

$$L(z, \hat{z}) = \frac{1}{n} \sum_{i=1}^n (z_i - \hat{z}_i)^2 \quad (5)$$

Where n denotes number of records and z_i represent the predicted value while (z_i) represent the actual value. Furthermore, whenever we are using deep-learning models, it is of paramount importance to mitigate the risk of overfitting. When a deep neural network possesses an excessive number of layers, it can give rise to issues like gradient vanishing or explosion, severely hampering the model's performance and exacerbating overfitting. To address these challenges, a technique known as batch normalization was introduced by [42]. The primary objective of batch normalization, as outlined in Equations (6), is to counteract gradient explosions or vanishing. This is accomplished by normalizing the output values after each hidden layer, ensuring that they remain within reasonable bounds and do not become excessively large or small. The batch normalization process involves calculating the difference between each output value and the mean value of the entire vector, and subsequently dividing this difference by the standard deviation. In our MLP model, batch normalization was thoughtfully incorporated after each hidden layer to effectively combat overfitting, thereby enhancing the model's generalization capabilities.

$$X_i = \frac{X_i - \text{Mean}_i}{\text{Standard Deviation}_i} \quad (6)$$

Where X_i represent i^{th} hidden layer output matrix and Mean_i represent mean value of X_i .

In this study, 10 hidden layers in MLP Classifier. The ReLU activation function was employed in each hidden layer, which consisted of 10 neurons. The count of hidden layer and neurons was selected based on multiple experiments. To normalize the range of values, batch normalization was applied after each hidden layer. The selected features were input into the neural network by using input layer. Subsequently, model was trained using forward & backward propagation techniques, whereas the output layer employed the Softmax activation function to generate class probabilities. During the prediction phase, a vector is produced containing class probabilities, and argmax function (described in Equation (7)) was used to identify the highest probability value and return its corresponding index.

$$\text{Output} = \max(\text{PredictedVector}) \quad (7)$$

For training our model, we used the Adam optimization algorithm, known for its dynamic adjustment of the learning rate based on recent weight gradients. In our configuration, we set the learning rate to 0.0003, used a batch size of 64, and conducted training for a total of 100 epochs. To prevent overfitting, we thoughtfully implemented the early stopping technique. This technique monitors the training process and halting it when early signs of overfitting become apparent, thereby preserving the best model parameters. In our setup, we configured the early stopping parameter to 40, indicating that if the loss on the validation set failed to decrease for more than 40 consecutive epochs, it signaled that the model had indeed overfit. At this juncture, the training process was promptly terminated, and any changes made to the model's parameters during those epochs were reverted, ensuring the model's generalization capacity was preserved.

Furthermore, with this classification model we have employed Recursive elimination technique with some modification for the calculation of importance score. This technique is mostly used to identify relevant features that contribute significantly to model performance [43]. Furthermore, this technique reduces the dimensionality of the dataset and improves the model interpretability, efficiency, and generalization ability [44]. RET iteratively removes irrelevant or redundant features and focuses on a subset of features that have the most significant impact on the overall model performance. In this study, the first step involves dividing the dataset into a training set and validation set using 20:80 ratios. The training dataset was used to train a multilayer perceptron classifier for classification purposes (discussed earlier in this section). Subsequently, a validation sample was provided to the trained model, and the model predicted the class label for each sample. The accuracy achieved during this prediction stage was considered the baseline accuracy when all the features were included. The next phase of the technique focuses on feature removal. One parameter was removed from the feature list, and the dataset was again divided into training and validation sets. The multilayer perceptron classifier is then trained using the updated feature set. Following training, a test sample was utilized to predict the class label, and the accuracy was noted. The new accuracy obtained was subtracted from the baseline accuracy, yielding a subtraction result that served as the importance score for the removed feature. This process was repeated for each parameter with at least five different epoch phases, yielding an importance score. Equation (8) represents the importance score calculation.

$$F_{IS} = \frac{1}{5} \sum_{i=20}^{100} (BLA_i - WOPA_i) \quad (8)$$

Where F_{IS} represent feature important score, BLA_i represent the baseline accuracy against in i^{th} iteration and $WOPA_i$ represent the without parameter accuracy of the i^{th} iteration.

The entire process was repeated for each parameter in the dataset. After iterating through all the features, the algorithm generates two lists. The first list contains the names of the features, whereas the second list contains the corresponding importance scores. On the basis of this importance score, the parameters were sorted based on their respective scores. This sorting process provides us with parameter ranking.

3.4. Prediction using top ranked parameters

After ranking we have Selected the top five parameters from each category (Primitive parameter, Publication and citation count based parameter, Author count based parameter, Publication age based parameter). Furthermore, by using these parameters we have to check the awardees trend in a ranked list of top 100 researchers records. These results will show the potentiality of top parameters against each category.

3.5. Rules mining using top 5 parameters against each category

In machine learning, the decision tree is widely recognized as a go-to choice for a variety of tasks, including prediction, classification, and rule mining [45,46]. Due to their popularity in this task, we employed decision trees due to its intrinsically interpretable nature and easy to implement.

Moreover, process of building a decision tree commences with the division of the source dataset into two subsets based on a chosen attribute. This division initiates a recursive learning procedure, where each subtree learns from the subsequent subtree. This recursive learning continues until one of two conditions is met: either a subset containing data from only one class is reached, or further subdivision of the data into subtrees ceases to impact the learning process. Decision trees are adept at generalizing, detecting patterns, and categorizing data within a dataset through the application of statistical methods. They are particularly valuable for exploratory tasks because they do not necessitate specific prior knowledge or criteria for construction. Decision trees are renowned for their high accuracy and are considered a standard in the realm of classification tasks.

The process of splitting decision tree nodes is typically guided by attributes. These attributes are either Gini-index or information gain. Due to nature of the dataset, we have employed the Gini-index which is mostly preferable with numeric type data [47]. Moreover, this decision is driven by the fact that information gain is commonly employed to identify the most informative attributes, which can be less suitable in cases which contain numerous different attribute values [48]. In contrast, the Gini-index is a statistical measure used to assess impurities within data [47]. Impurity is quantified as the number of data points that are incorrectly classified when all elements belong to the same class. The Equation (9) represents Gini-formula.

$$GI = 1 - \sum_{j=1}^n (P_j)^2 \quad (9)$$

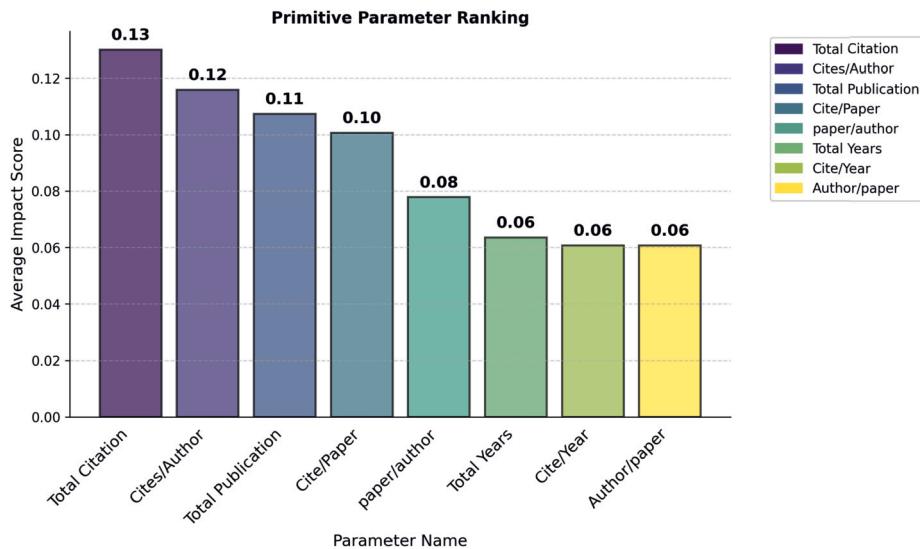


Fig. 5. Primitive Parameter Ranking.

Where P_j represent object probability in the dataset being classified into class j. Moreover, it takes values between 0 and 1, and its interpretation is as follows:

- A Gini index of 0 indicates perfect purity, where all elements in the node belong to the same class.
- A Gini index of 1 implies maximum impurity, meaning the elements are evenly distributed among the classes, and there is no predominant class.

In addition, our dataset encompasses a substantial number of author assessment parameters, resulting in a decision tree that tends to be overly intricate and dense. To alleviate this complexity and eliminate extraneous rules, we employed a post-pruning technique. Post-pruning entails the initial growth of the tree to its full extent, followed by the removal of nodes that contribute negligibly to the overall performance of the tree. This method involves the calculation of a cost parameter for each subtree considering both its accuracy and complexity, quantified by the number of nodes it contains. Subtrees with higher cost values were pruned, resulting in a more streamlined and generalizable tree. This refined tree allows for the extraction of generalized and well-suited rules for each category parameter.

4. Results discussion

Now, in this section we have explained our experimental results obtained from proposed technique by using mathematics dataset.

4.1. Parameters ranking

Moving forward, we have displayed a list of the most effective parameters obtained using RFE technique along MLP. Subsequently, these high-performance parameters are utilized in the crafting rules for the mathematical domain using a decision tree.

4.1.1. Primitive parameters ranking

In the primitive parameters category, we found that the total citation parameter had a higher impact (Importance Score:0.13) on model performance than all the other parameters (see in Fig. 5). The cite/author and total publications take the second and third positions by achieving importance score 0.11 and 0.10 respectively. In order to assess how the total citation count parameter influences the selection of award recipients within the field of mathematics, we conducted a classification of individuals who received awards and those who did not. The total citation individually, achieved 62% percent accuracy in classification of awardees, while the second position cite/author achieved 50% and the third position total publication 47%. The accuracies of the remaining parameters were less than 40. Thus, if you increase your total citations, the chances of your award winner increase.

4.1.2. Author count based parameters ranking

In the Author Count-Based Parameters category, we found that the normalized hi index outperformed all other parameters by creating a huge impact on model performance by achieving an Importance Score of 0.22 (see in Fig. 6). The gf index and hi index take the second and third positions, respectively, by achieving importance score 0.19 and 0.17 respectively. Similar to the above category, for the verification of normalized hi index, we performed classification of awardees and non-awardees. The normalized hi

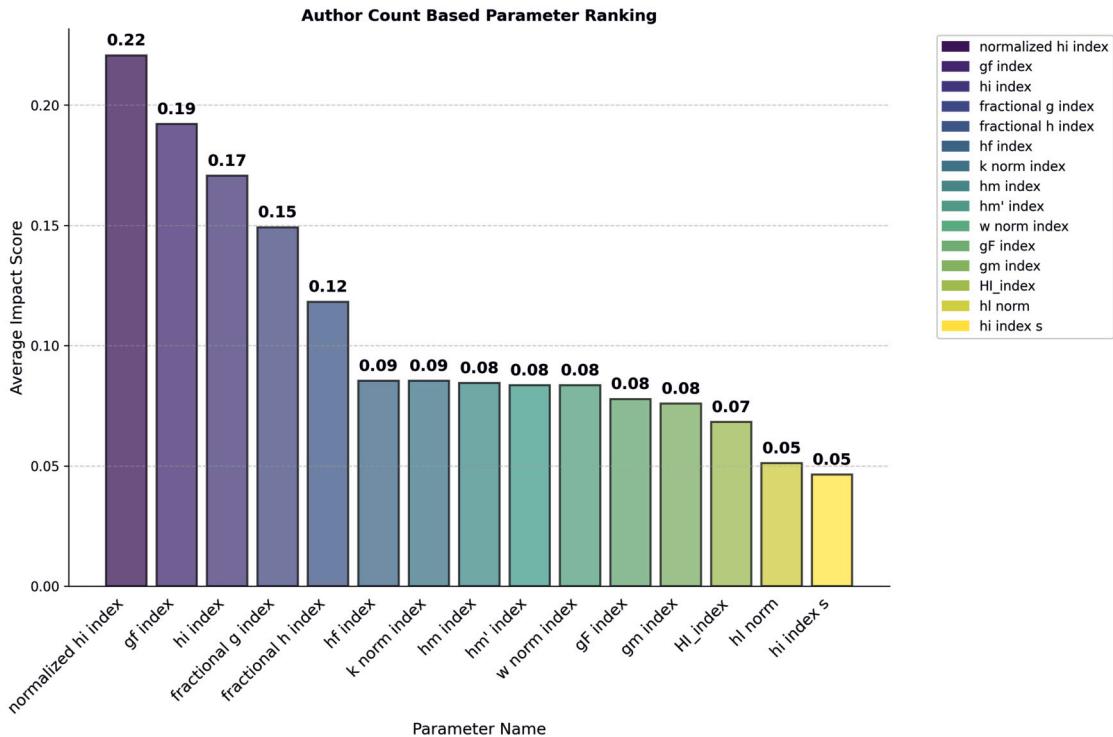


Fig. 6. Author Count based Parameter Ranking.

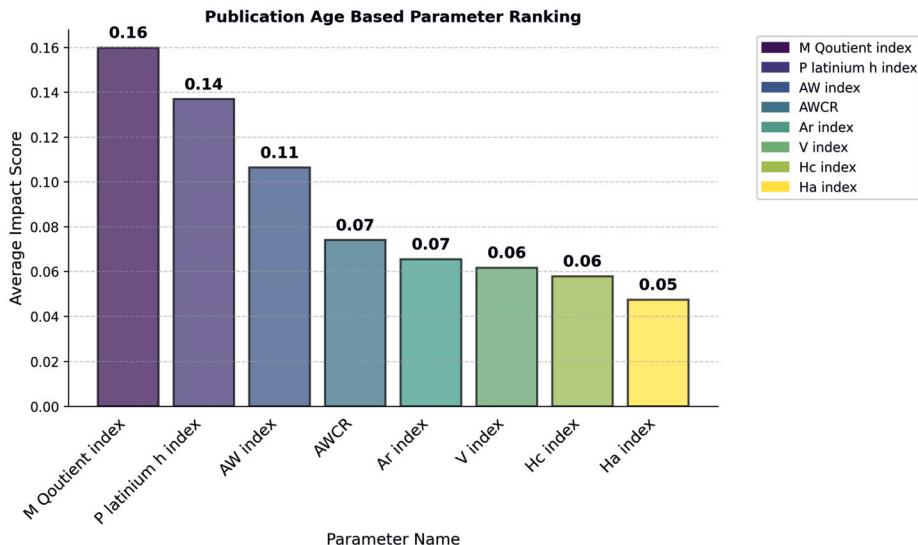


Fig. 7. Publication Age Parameter Ranking.

index achieved 71% accuracy, the gf index achieved 68%, and the hi index achieved 55% accuracy. The remaining parameters in this category achieved an accuracy of less than 45%.

4.1.3. Publication age based parameters

In the Publication Age-based Parameters category, we found that the m-quotient outperforms all the other parameters by impacting model performance by achieving an Importance Score of 0.16 (see in Fig. 7). The platinum h and Aw indices take the second and third positions, respectively, by achieving importance score 0.13 of 0.10 respectively. Similar to the above category, for verification we performed classification of awardees and non-awardees. The m-quotient individually achieved 60% accuracy in classification of awardees, while platinum h index achieved 51%, Aw index achieved 44%, and the remaining one achieved less than 35%.

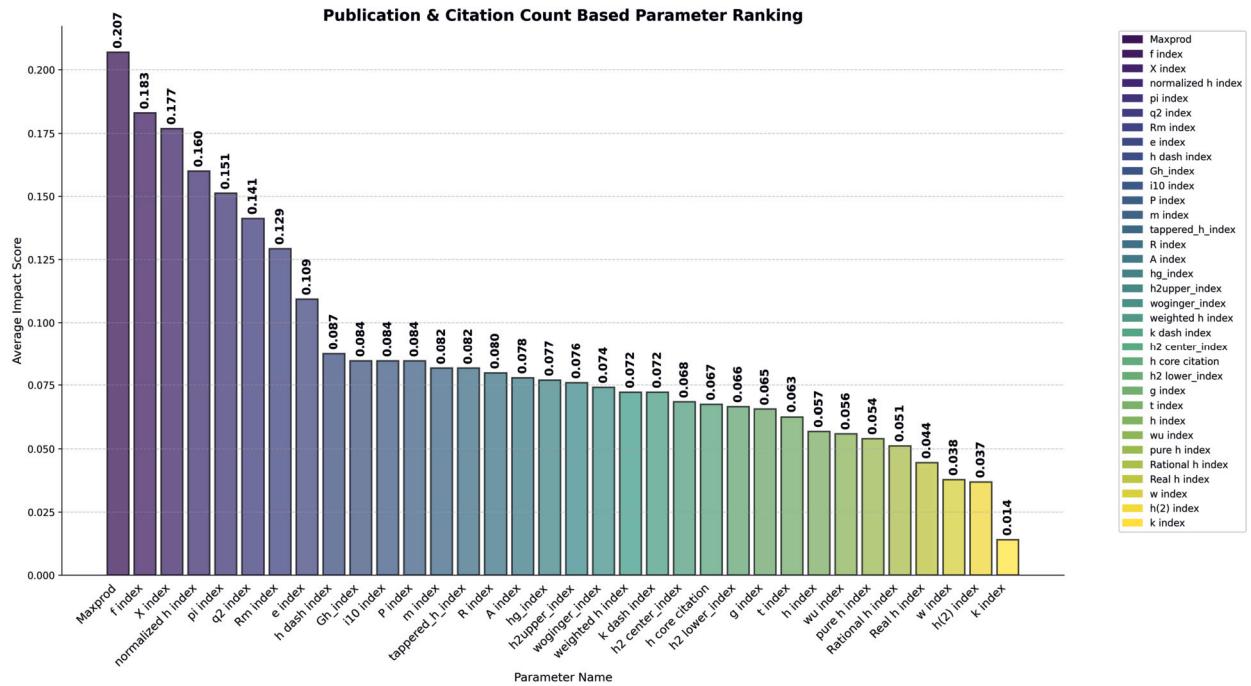


Fig. 8. Publication and Citation Count based Parameter Ranking.

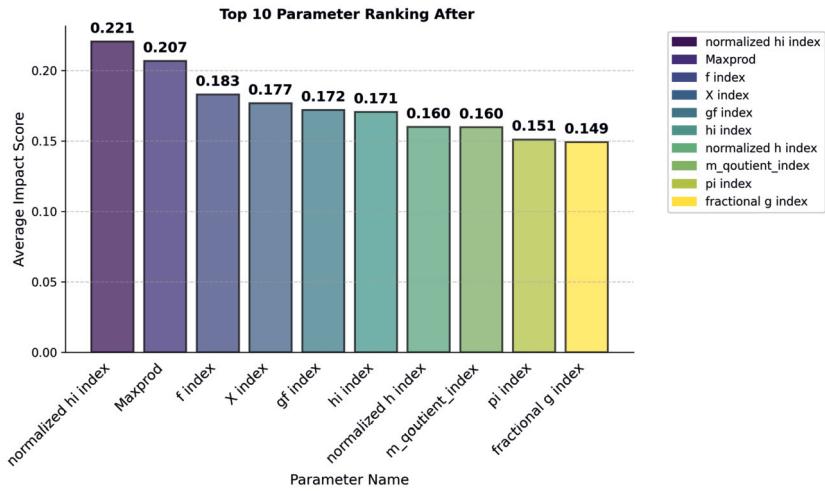


Fig. 9. Combine All Category Parameters.

4.1.4. Publication and citation count based parameter

In the Publication and citation count-based parameter category, we found that Maxprod outperformed all the other parameters by achieving an Importance Score of 0.20 (see in Fig. 8). The f and X indices achieved the second and third positions by achieving importance score 0.18 of 0.17 respectively. Similar to other categories, for verification we performed classification awardees and non-awardees. Maxprod individually achieved 66% accuracy in classification of awardees, while the platinum f index achieved 63%, the Aw index achieved 60%, and the remaining one achieved less than 50%. From this analysis, we concluded that parameters belonging to this category have high performance in classification compared to other categories.

4.1.5. Combination of all parameter

In this section, we combine all the category parameters. When combining all the category parameters, the total count is increased, which is not betterly visualized in one graph; therefore, we only show the top 10 best parameters in the Fig. 9. From this Fig. 9, we have determined that the normalized hi index outperforms all other category parameters. The Maxprod index was placed in the second position and the f index was placed in the third position.

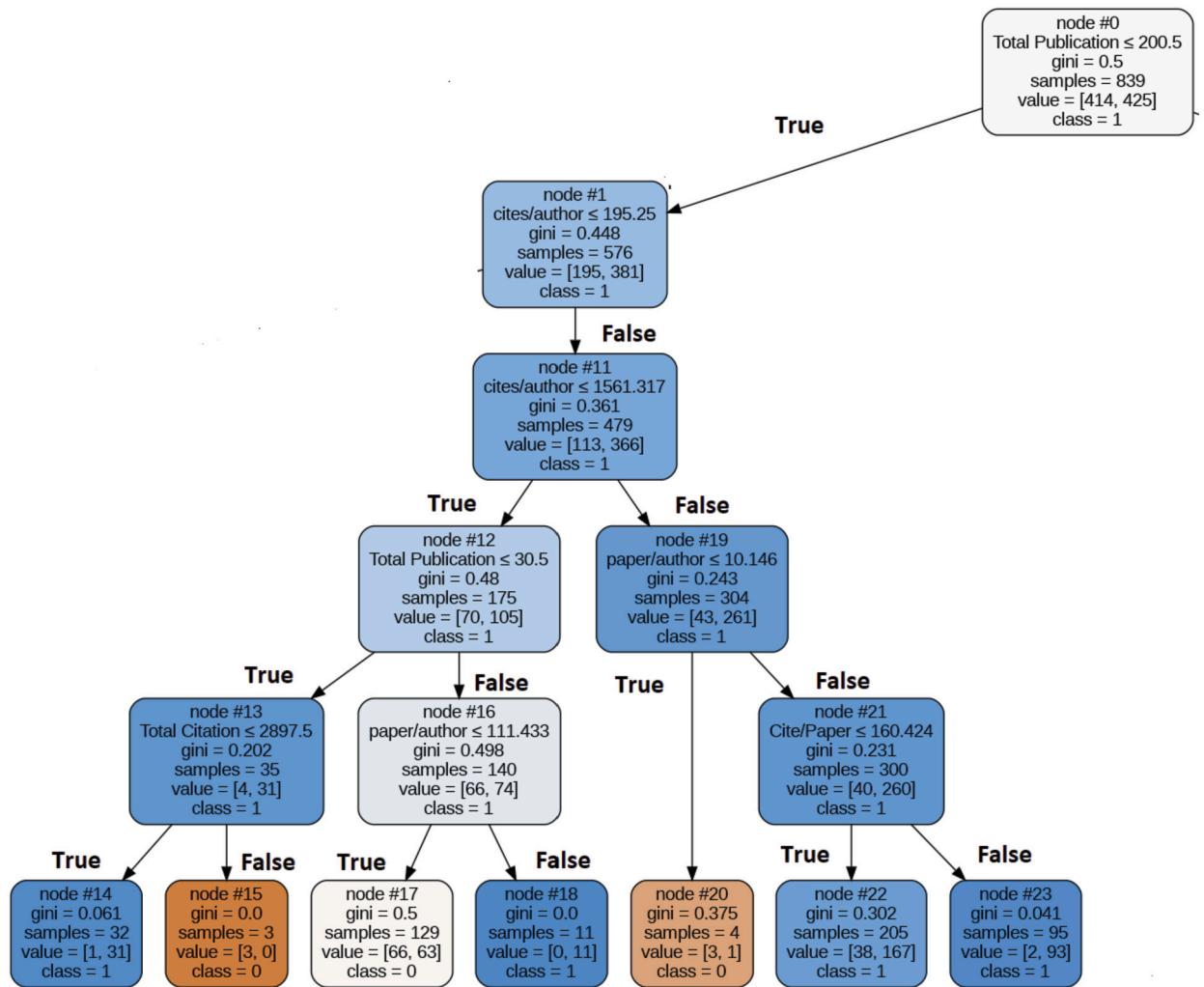


Fig. 10. Decision Tree for Extracted Rule for Primitive Parameter.

5. Rule mining

In this section, we introduce a rule for each category based on the top five parameters within that specific category. Figs. 10 to 14 depict the rules with the highest coverage for each parameter category. These figures were extracted from a complete decision tree against the dataset which contains multiple rules. However, we specifically mined rules that focus only on those aspects we wish to highlight to the scientific community. These highlighted rules hold significant importance in the ranking of the authors.

5.1. Rule for the primitive parameters

In the primitive parameters the top five ranked parameters are: Total Publication, Total Citation cite/paper, cite/author, and paper/author. To extract the prevailing pattern among most awardees, we have employed decision tree with Gini measure (Splitting Purpose). This decision tree attained an accuracy rate of 78% for correctly identifying awardees. Fig. 10 illustrates the decision tree representing the extracted rules for awardees. The top rule, which encompasses 55% of awardees in the primitive category, is presented as follows:

IF Total Publication <= 200 AND Cite/Author > 195.25 AND Paper/Author <= 10 -> Award Recipient

5.2. Rule for author count based parameters

In the author count-based category, the top five ranked parameters were as follows: normalized h index, gf index, hi index, fractional h index, and fractional g index. To uncover the prevailing pattern among most awardees, we have employed decision tree with Gini measure (Splitting Purpose). This decision tree achieved an accuracy rate of 73% for correctly identifying award recipients.

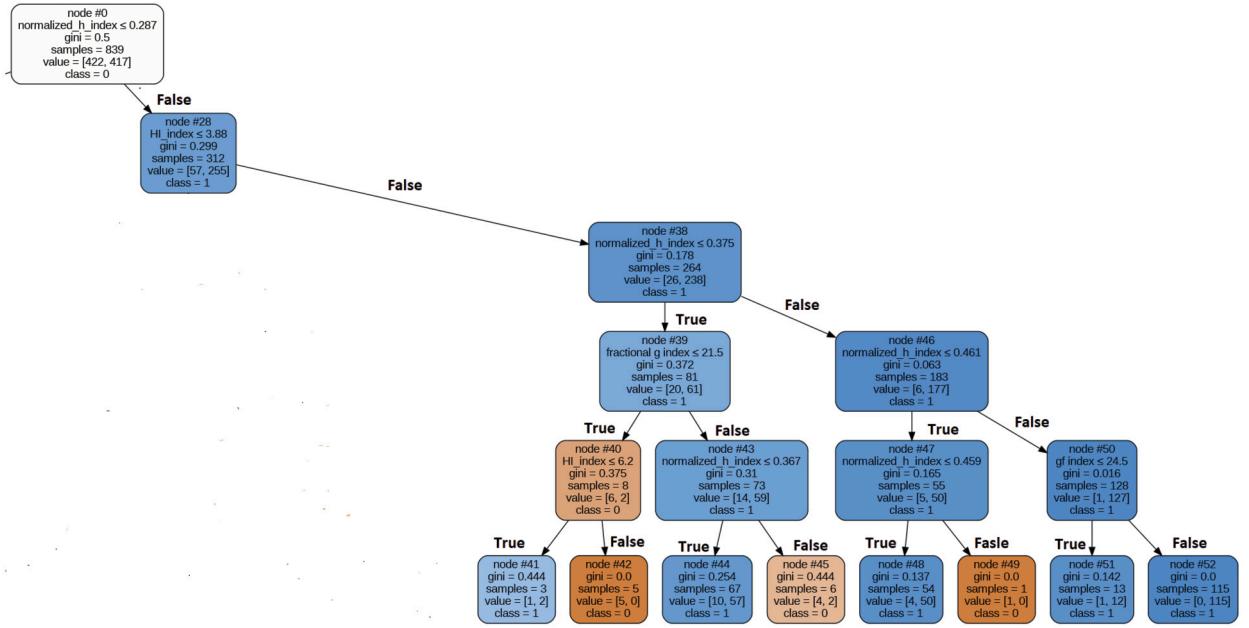


Fig. 11. Decision Tree for Extracted Rule for Author Count Based Category.

Fig. 11 visually represents the decision tree outlining the extracted rules for awardees. The top most rule, which encompasses 56% of awardees within the Author Count based category, is detailed as follows:

IF (0.45 => Normalized h index >= 0.28) And hi index >= 3.88 AND fractional g index > 21.4 -> Award Recipient

5.3. Rule for age based parameters

In the Age-based Category, the top five ranked parameters are as follows: Platinum h index, Aw index, M Quotient index, Ar index, and AWCR. To uncover the prevailing pattern among most awardees, we employed a decision tree with Gini measure (Splitting criteria). This decision tree achieved an accuracy rate of 68% in correctly identifying awardees. Fig. 12 provides a visual representation of the decision tree outlining the extracted rules for the awardees. The top most rule, which covers 41% of the awardees within the Age-based category, is detailed as follows:

IF (216.9 => Platinum h index >= 37.5) AND Aw index >= 208.099 AND M quotient index >= 1.3 -> Award Recipient

5.4. Rule for publication and citation count based parameters

In this Category, the top five ranked parameters were as follows: F index, Maxprod, pi index, X index, and Q2 index. To uncover the prevailing pattern among most awardees, we employed a decision tree with Gini measure (Splitting criteria). This decision tree achieved an accuracy rate of 71% for correctly identifying awardees. Fig. 13 visually represents the decision tree outlining the extracted rules for the awardees. The top most rule, which encompasses 90% of the award recipients within the Publication and Citation Count-based category, is detailed as follows:

IF (9 => X index <= 14.2) AND Maxprod <= 188 AND (189 = < f index <= 102) -> Award Recipient

5.5. Combine all parameter

We have now combined parameters from all categories, and among these parameters, the top five ranked ones are as follows: normalized h index, F index, Maxprod, X index, and Hi index. To uncover the predominant pattern among most awardees, a decision tree using the Gini measure as the splitting criterion was employed. This decision tree achieved an accuracy rate of 78% for correctly identifying award recipients. Fig. 14 provides a visual representation of the decision tree that outlines the extracted rules for the awardees. The top most rule, which covers 53% of the awardees are given below:

IF (0.4 => Normalized h index >= 0.287) AND (6.9 => HI index >= 3.88) AND Maxprod >= 244 -> Award Recipient

5.6. Rules coverage comparison

After creating rules for all categories in this section, we compared the rule coverage between them, as shown in Fig. 15. From the Fig. 15, we have identified that the rule generated using the publication and citation count-based category outperforms all the category rules by achieving 90% coverage, which means that 90% of awardees statistics in the dataset are exactly according to this

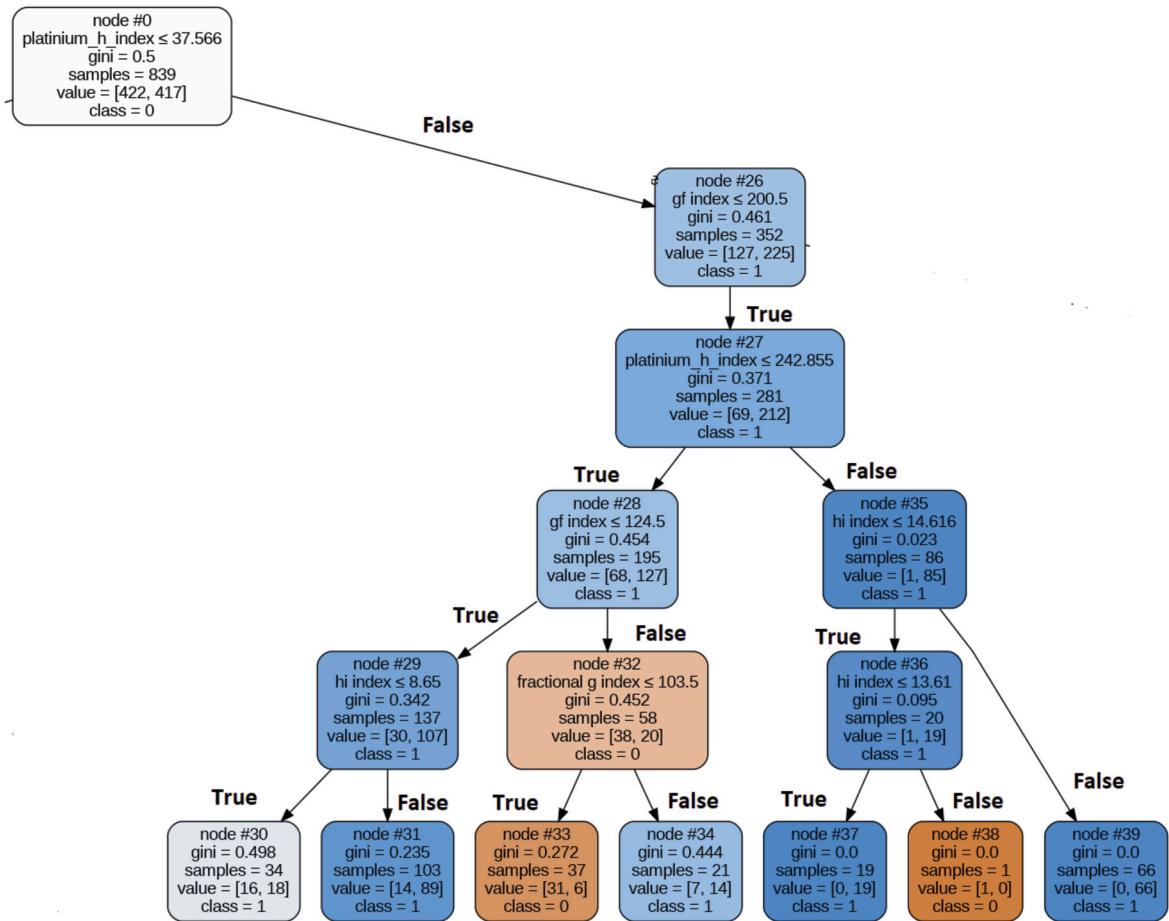


Fig. 12. Decision Tree for Extracted Rule for Age Based Category.

Table 2
Rules Validation Percentage.

Parameter Category	Total Awardees	Percentage
Primitive parameters	53	48%
Author count based parameters	53	52%
Rule for Age based parameters	53	43%
Publication and citation count based parameters	53	78%

rule. From this high coverage, we have determined that you can use the publication and citation count-based parameters in combine manner for ranking authors due to their high performance.

5.7. Rules validation

To validate the generated rules, we selected a sample from the dataset comprising the awardees from the last three years (2023, 2022, and 2021), totaling fifty-three individuals. Subsequently, we applied each search and mapping to every rule across various parameter categories, determining the correct prediction ratio for each rule. The results are presented in Table 2.

Analysis of the results revealed certain recurring patterns, notably that rules based on publication and citation counts outperformed all other category rules, correctly predicting 78 percent of awardees.

6. Awardees trends in top 100 researchers list

In this section, we analyze award trends among the top 100 researchers by condensing the top five parameters from each of the four categories. Primitive parameters, such as Total Publications, Total Citations, Cites per Paper, Papers per Author and Citations per Author, show that total citations significantly contribute to 55% of awardees' rankings in mathematics fields, while Cites per Paper, Papers per Author, and Citations per Author contribute 48%, 45%, and 40%, respectively (see Fig. 16). Author count-based

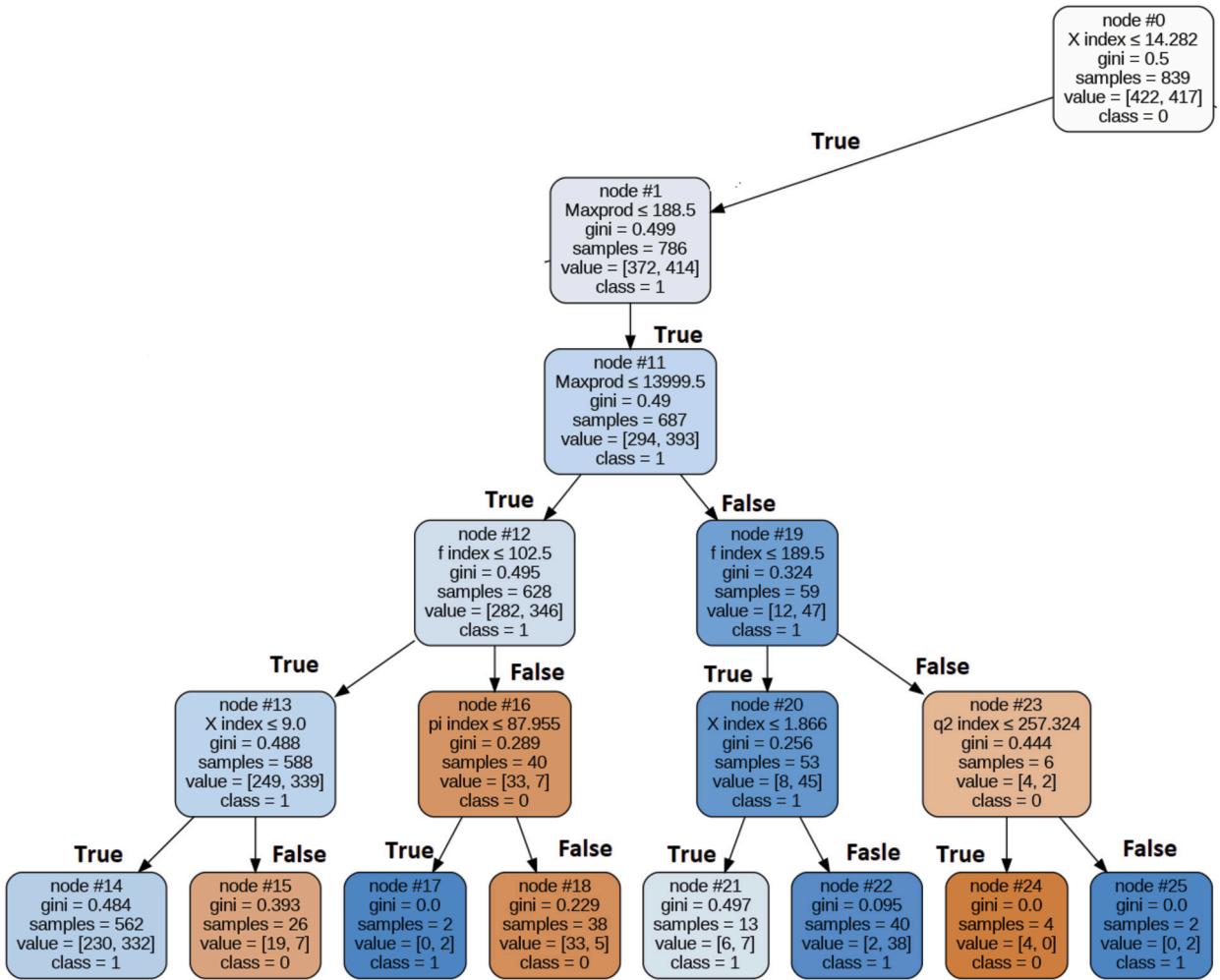


Fig. 13. Decision Tree for Extracted Rule for Publication and Citation Count Based Category.

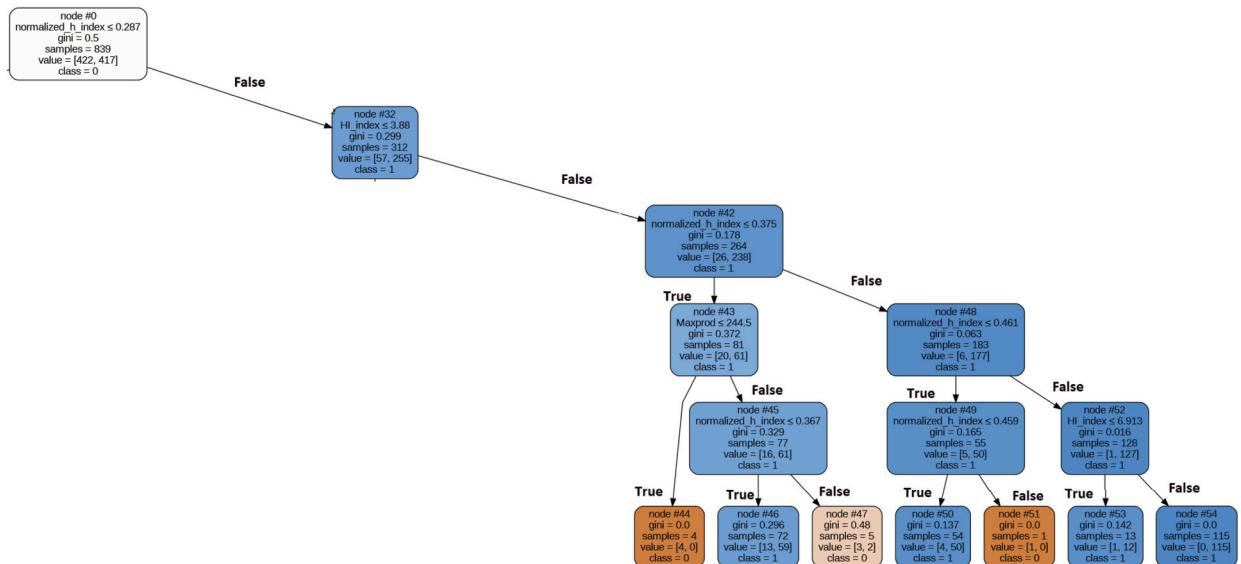


Fig. 14. Decision Tree for Extracted Rule for Combining All Category Parameter.

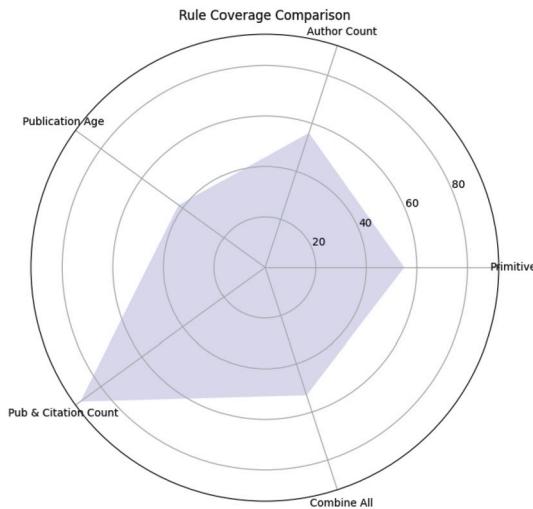


Fig. 15. Rule Coverage Comparison.

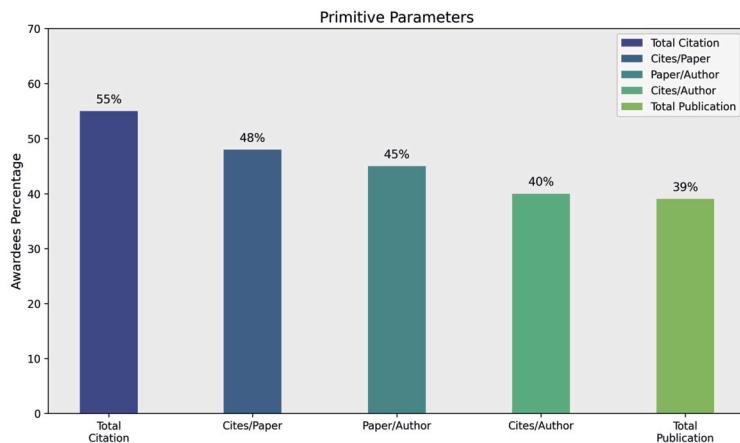


Fig. 16. Awardees Percentage for primitive parameters.

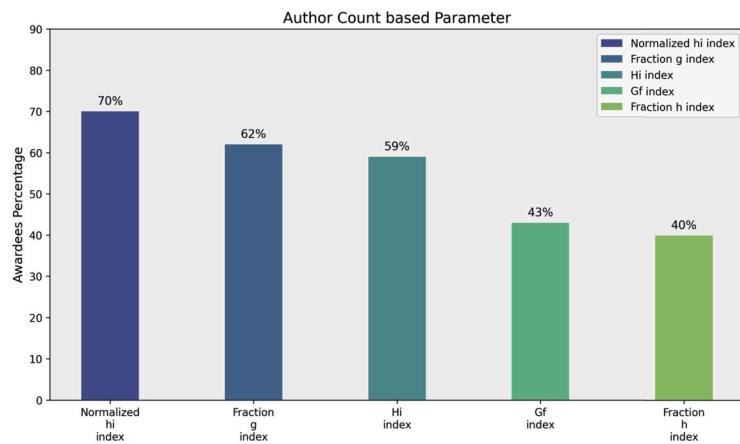


Fig. 17. Awardees Percentage for Author Count based parameters.

parameters, including Normalized h-Index and Fractional g-Index, recognize 70% and 62% of researchers, respectively (see Fig. 17). Publication Age-based parameters, such as Platinum h-Index, demonstrate awards for 70% of researchers, while AWCR identifies 38% (see Fig. 18). Parameters based on Publication and Citation counts, like Q2 Index, Maxprod, and Pi Index, acknowledge 45%,

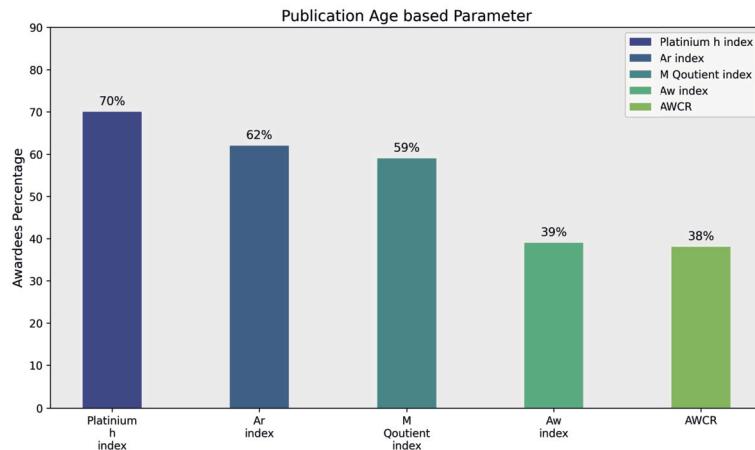


Fig. 18. Awardees Percentage for publication Age based parameters.

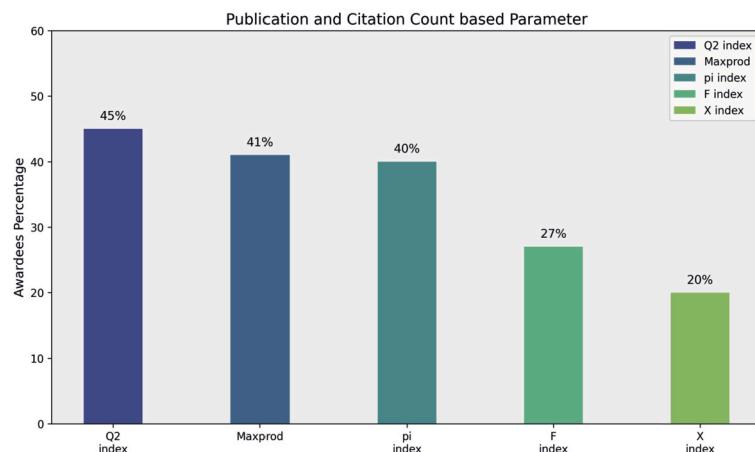


Fig. 19. Awardees Percentage for publication and Citation count based parameters.

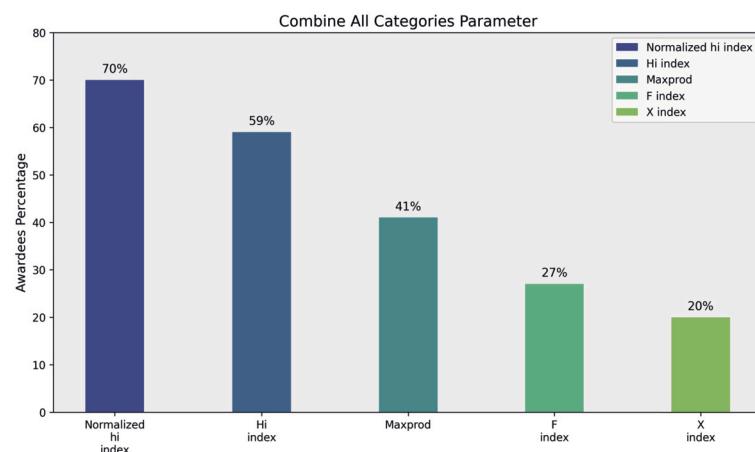


Fig. 20. Awardees Percentage after Combine all parameters.

41%, and 40% of researchers, respectively (see Fig. 19). Combining all parameters reveals that Normalized h-Index has the highest impact, recognizing 70% of awardees, followed by h-Index (59%), Maxprod (41%), F Index (27%), and X Index (20%) (see Fig. 20). Author count-based parameters emerge as dominant in elevating awardees compared to other categories.

7. Conclusion

The assessment of scientists' achievements has become increasingly important over the past two decades. To evaluate authors' contributions, various unique parameters have been proposed in the literature. This study ranks these parameters, which fall into four distinct categories. The findings reveal that within the first category, total citation counts hold significant relevance. In the Author Count-Based category, the gf index stands out as a strong indicator. Meanwhile, the Publication and Citation Count-Based category highlights the importance of the M Quotient parameter, and in the Publication Age-Based category, the Maxprod parameter is closely associated with mathematics awardees. The study also uses machine learning techniques to analyze the quantitative patterns of prolific researchers, which may have contributed to their inclusion on the list of award recipients. While these patterns can increase the likelihood of making it onto the prolific list, inclusion ultimately depends on subjective evaluation, and there are no guarantees. Additionally, the rule generated using the publication and citation count parameters produced more promising results compared to other category parameters. Consequently, the study suggests that combining publication and citation count-based parameters is more effective for author ranking than relying solely on other categories. The findings of this study offer valuable guidance for individual researchers aiming to be recognized as distinguished scientists. Moreover, the scientific community can use these rules to evaluate and identify exceptional researchers, facilitating the recognition and rewarding of outstanding contributions in the field.

8. Future work

In future work, we plan to expand our research from multiple perspectives. Firstly, we aim to evaluate the identified set of parameters across various domains, including Computer Science, Civil Engineering, Neuroscience, and others. This assessment will provide insights into the applicability and effectiveness of our proposed parameters in diverse contexts. Secondly, we intend to introduce new indices that encompass the lifetime achievements of authors. These indices will consider factors like publication count, citation count, and co-author prestige, which are not currently accounted for by existing parameters. Thirdly, we will incorporate Deep Learning techniques to develop a framework for identifying influential researchers. This advanced approach will enhance the accuracy and efficiency of our methodology. Furthermore, beyond the initially chosen parameters, there are additional factors that could contribute to the comprehensiveness of our work. One noteworthy factor is the recently introduced Q-factor, as outlined by (Sinatra et al., 2016). However, determining this diversified parameter goes beyond publications and citations; it also considers aspects such as time, subjective performance in project contribution, dissertation supervision, funding, and more. Our future objective is to construct a dataset that includes subjective factors, thereby broadening the scope of our analysis. Moreover, in upcoming research, we plan to address several inquiries. These include assessing whether the integration of quantitative parameters genuinely enhances the field of Scientometrics, evaluating their practicality for implementation in digital libraries, and exploring the need for new parameters that align with subjective decision-making processes.

Funding

This Research is funded by Qatar National Library.

Ethics declarations

Informed consent was unnecessary for this study as it relied on data sourced from public databases. The research excluded any involvement in animal and human experiments, along with other data pertaining to human privacy.

CRediT authorship contribution statement

Ghulam Mustafa: Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Abid Rauf:** Writing – review & editing, Validation, Supervision, Methodology, Formal analysis, Conceptualization. **Ahmad Sami Al-Shamayleh:** Writing – review & editing, Validation, Resources, Funding acquisition. **Muhammad Tanvir Afzal:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Ali Waqas:** Writing – review & editing, Validation, Resources, Funding acquisition. **Adnan Akhunzada:** Writing – review & editing, Validation, Resources, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The dataset related to this study has been uploaded to a publicly accessible repository. The name of the repository is **ghulam-mustafacomsat** and the link is attached (https://github.com/ghulammustafacomsat/Mathematics_dataset).

Appendix A

Table 3
Indices Calculation Formulas.

Name of Index	Calculation
Total Publication	Total Number of Publications of a Researcher
Total Citation	Total Number of Citations of a Researcher
Total Years	Number of years since the academic's initial publication
Cites per Year	Total citations/years since first paper
Cites per Paper	Total citations/total papers
Author per Paper	Calculate the cumulative count of authors engaged in the publications attributed to the author in question, and subsequently divide this figure by the total number of papers.
Cites per author	Divide the citations for each individual publication by the corresponding number of authors, and then aggregate the resultant citation values. This sum represents the equivalent number of citations for the author in question as if they were the sole author.
Papers per author	Divide each publication by the number of authors and sum the fractional author counts; this is the single-authored equivalent number of papers for the author in question.
H-index	$h = \max(\text{numbers of articles with } >= h \text{ citations})$
G-index	A collection of papers is said to possess a g-index of 'g' if 'g' represents the highest rank where the sum of citations from the top 'g' papers reaches or exceeds g^2 .
Hg-index	$hg - index = \sqrt{h * g}$ Where h represent h index and g represent g index.
A-index	$A - index = \frac{1}{h} \sum_{p=1}^h cit_p$ Where A is the A-index of the scholar, h represents the h-index and cit_p is the citation count of the p^{th} article.
R-index	$R - index = \sqrt{\sum_{p=1}^h Cit_p}$ Where h represents the h-index and cit_p is the citation count of the p^{th} article.
P-index	$P - index = (\frac{C^2}{p})^{\frac{1}{3}}$ The p-index strikes the best balance between total number of citations (C) and the mean citation rate (C/P)
Q ² -index	$Q^2 = \sqrt{h * m}$ Where h represent h index and m represent m index
K-index	$K - index = \frac{C}{\frac{Ch - tail}{Ch - core}}$ Where C represent total citation, p represents of p^{th} article, h-tail represent h-tail article citation and h-core represent h-core citation.
E-index	$e - index = \sum_{p=1}^h Cit_p - h^2$ Where Cit_p represent citation of p^{th} article and h represent h-index
f-index	$f - index = \left(\frac{\max}{f}\right) \frac{1}{\frac{1}{f} \sum_{p=1}^f \frac{1}{Cit_p}} \geq f$ Where Cit_p is the citation count of p^{th} article. The f-index never goes beyond the total number of publications.
T-index	$T - index = \left(\frac{\max}{t}\right) \exp \left[\frac{1}{t} \sum_{k=1}^t \ln(Cit_k) \right] \geq t$ Where Cit_k is the citation count of k^{th} article
Tapered h-index	$H_{T(j)} = \frac{n_j}{2j-1}, n_j \leq j \text{ and } h_{T(j)} = \frac{1}{2j-1} + \sum_{i=j+1}^{n_j} \frac{1}{2j-1}, n_j \geq j$
Wu index	The w-index of an author is determined by considering whether a minimum of 'w' of their articles have acquired '10w' citations each, while the remaining publications have accrued fewer than '10(w+1)' citations each.
Weighted h-index	$R_w(k) = \frac{\sum_{p=1}^k Cit_p}{h}$ Where h is the h-index and Cit_p is the citation count of the p^{th} article. Then, the weighted h-index. Is defined as follows $h_w = \sqrt{\sum_{k=1}^R Cit_k}$ Where Cit_k is the citation count of the k^{th} article and R is the largest rank among all publications such that the k^{th} weighted rank $\leq cit_k$.
h(2)-index	The h(2)-index is the maximum whole number such that the h(2) most cited articles by a scholar have each received at least $(h(2))^2$ citations.
Woeginger index	$w = \left(\frac{\max}{w}\right) (Cit_p \geq w - p + 1) \text{ for all } p \leq w$ where Cit_p is the citation count of p^{th} article and w is the maximum number of publications.
Rm-index	$R_m = \sqrt{\sum_{k=1}^h Cit_k^{\frac{1}{2}}}$ Where Cit_k is the citation count of k^{th} article.
m-index	An m-index is computed as the median quantity of citations garnered by the articles within the h-core.
X-index	$x = \sqrt{\left(\frac{\max}{k}\right) k Cit_k}$ Where Cit_k is the citation count of k^{th} article.
h2 upper-index	$h^2 upper = \frac{\sum_{k=1}^h (Cit_k - h)}{\sum_{k=1}^m Cit_k} * 100 = \frac{e^2}{\sum_{k=1}^m Cit_k} * 100$ Where h is the h-index, Cit_k is the citation count of the k^{th} article, e^2 is the excess citation and m is the total number of articles.
h2 center-index	$h^2 center = \frac{h * h}{\sum_{k=1}^m Cit_k} * 100$ Where h is the h-index and Cit_k is the citation count of the k^{th} article
h2 lower-index	$h^2 lower = \frac{\sum_{k=h+1}^m (Cit_k - h)}{\sum_{k=1}^m Cit_k} * 100$ Where h is the h-index and Cit_k is the citation count of the k^{th} article
h` -index (h dash)	$h' = Rh = \frac{eh}{t}$ where R represents the head-tail ratio of e and t-index

(continued on next page)

Table 3 (continued)

Name of Index	Calculation
Rational h-index	$h_{rat} = h + 1 - \frac{k}{2h+1}$ Where h is the h-index and k is the number of citations required to reach h+1 h-index value.
I10 index	This is a straightforward and direct indexing method that entails tallying the overall count of papers published by a journal, which have attained a minimum of 10 citations each.
Normalized h-index	$normalized\ h\ -\ index = \frac{h}{pub_{count}}$ Where h represents h index and pubcount represent total publication.
Π index	$\Pi\ index = 0.01C(P\Pi)$ The $\Pi - index$ is equal to the 100th of the number of citations, $C(P\Pi)$ to the top square root ($P\Pi$) of the total papers(P).
Gh-index	$Gh^a = \sum_{p=1}^m sing(Cit(pub_b^a) - GH)$ where $sing(x) = 1, x \geq 0$ and $0, x \leq 0$ Where m is the total number of publications of scholar a and GH is the h-index of the scholar. This index is also difficult to compute in comparison to the h-index.
W index	The w-index is characterized by the value 'w', denoting the count of an author's foremost articles that have accumulated no fewer than '10w' citations individually. While the w-index can serve as a valuable metric for assessing a scholar's impact, it might impose a disadvantage on early-career researchers who have recently initiated their work or individuals who haven't yet produced a substantial volume of publications.
Maxprod	The maximum value of $i * c_i$ can be found by evaluating the publication rank of a scholar, where c_i represents the number of citations for the i^{th} most frequently cited paper among all the citations.
H core citation	The h-core citation index exclusively considers publications that have been cited a minimum of 'h' times, disregarding those that haven't reached this threshold.
K dash index	$k' = \frac{Cit_{all} - Pub_{count}}{Cit_t - Cit_h}$ Where Cit_{all} represent total citation, Pub_{count} represent total publication, Cit_t represent total citation of h tail article and Cit_h represent total citation of h core article.
M-Quotient	$M - Quotient = \frac{h - index}{y}$ Where y represent no. of the year the first publication
Hc – index	$hc - index = \alpha * \frac{C(i)}{(Y(now) - Y(i) + 1)}$ Where $Y(now)$ represent current year, $Y(i)$ represent publication year, and $C(i)$ represent paper i citation count.
Aw-index	$hc - index = \frac{C(1)}{1}, \frac{C(2)}{2}, \frac{C(3)}{3}, \dots, \frac{C(n)}{n}$ $Aw - index = \sqrt{\sum_{j=1}^h \frac{Cit_j}{a_m}}$ Where Cit_j citation of the article and a_j represent a^{th} article, m represent total article.
Ar index	$Ar - index = \sqrt{\sum_{j=1}^h \frac{Cit_j}{a_j}}$ Where Cit_j citation of the article and a_j represent a^{th} article, h represent total h core article.
AWCR v-index	This parameter adjusts the citation count based on the length of time that has passed since each publication. $V = \frac{h}{P(y_{this} - y_0)}$ Where h is the h-index, y_{this} is the current year and y_0 is the year of first publication.
Platinum H-index	$Platinum - h = \frac{H}{CL} * \frac{cit_{all}}{Pub_{count}}$ Where H is the h-index, CL is the career length, Cit_{all} is the total citation count and Pub_{count} is the publication count.
Ha index	The ha-index against a dataset is the largest number of papers in the dataset that have obtained at least ha citations per year on average.
HI Index	$h_i = \frac{h}{Avg_a}$ where h represent h index and Avg_a represent average no. of authors in article.
hf index	$\frac{Yh_f}{\phi(Yh_f)} \geq h_f$ where $Y(i)$ represent citation count and $\phi(i)$ represent average no. of authors in article.
gf index	$gf = \sum_{i=1}^k \frac{Y_i}{\phi(i)} \geq g^2 f$ where $Y(i)$ represent citation count and $\phi(i)$ represent average no. of authors in article.
gf index	$gF = (\sum_{i=1}^k \frac{1}{\phi(i)})^2 \leq \sum_{i=1}^k y_i$ where $Y(i)$ represent citation count and $\phi(i)$ represent average no. of authors in article.
Normalized Hi index	normalized hi index = $\frac{h}{pub_{count}}$ Where h represents the h-index and pub_{count} is the total number of articles.
Hm index	$r_{eff}(r) = \sum_{i=1}^r \frac{1}{a(i)} \text{ then } c(r(h_m)) \geq h_m \geq c(r(h_m) + 1)$
k norm index	$k - norm = h - norm + (1 - (h - \frac{norm^2}{\sum_{j=1}^{h-norm} citnorm_j}))$, $\forall h - norm > 1$ and $k - norm = 0$, if $h - norm = 0$
w norm index	$w - norm = h - norm + (1 - (h - \frac{norm^2}{totalcit - norm}))$, $\forall h - norm > 0$ and $w - norm = \frac{totalcit - norm}{1 + totalcit - norm}$, if $h - norm = 0$
gm index	$gm \leq C_{eff}(gm)$ where $C_{eff}(r_{eff})$ and $S_{eff}(r_{eff}) = \sum_{r=1}^{r(r_{eff})} \frac{1}{a(r)} c(r)$
pure h index	$h_p(A) = \frac{h}{\sqrt{E(author)}}$ Where h represent h index and E average no. of author.
fractional h-index	$h_f = max(k \leq \frac{cit(k)}{author(k)})$ Where Cit_k represent citation of article and author(K) represent no. of author in a specific article.
fractional g-index	$g_f = max(\sum_{k=1}^p \frac{cit_k}{Author(k)} \geq p^2)$
hi norm index	The hi-norm is a modified version of the h-index that normalizes citations based on the number of authors per paper.
K index	$Kindex = \frac{c/p}{c(h-tail)/c(h-core)}$ Where C represent total citation, P represent total publication, C(h-tail) represent total citation of h tail article and C(h-core) represent total citation of h core article.
Real h index	$h_r = \frac{(h+1)Cit_h - Cit_{h+1}}{1 - cit_{h+1} + cit_h}$ Where h is the h-index and Cit_h is the citation count of the h^{th} article.

References

- [1] G. Mustafa, M. Usman, L. Yu, M.T. Afzal, M. Sulaiman, A. Shahid, Multi-label classification of research articles using word2vec and identification of similarity threshold, *Sci. Rep.* 11 (1) (2021) 21900, <https://doi.org/10.1038/s41598-021-01460-7>.
- [2] G. Mustafa, A. Rauf, A.S. Al-Shamayleh, M. Sulaiman, W. Alrawagfeh, M.T. Afzal, A. Akhunzada, Optimizing document classification: unleashing the power of genetic algorithms, *IEEE Access* (2023), <https://doi.org/10.1109/ACCESS.2023.3292248>.
- [3] B. Ahmed, W. Li, G. Mustafa, M.T. Afzal, S.Z. Alharthi, A. Akhunzada, Evaluating the effectiveness of author-count based metrics in measuring scientific contributions, *IEEE Access* (2023), <https://doi.org/10.1109/ACCESS.2023.3309416>.
- [4] A. Sidiropoulos, D. Katsaros, Y. Manolopoulos, Generalized hirsch h-index for disclosing latent facts in citation networks, *Scientometrics* 72 (2007) 253–280, <https://doi.org/10.1007/s11192-007-1722-z>.
- [5] R. Wang, M. Lewis, R. Zheng-Powell, J. Julson, M. Smithson, H. Chen, Using the h-index as a factor in the promotion of surgical faculty, *Heliyon* 8 (4) (2022), <https://doi.org/10.1016/j.heliyon.2022.e09319>.
- [6] B. Ahmed, L. Wang, A.S. Al-Shamayleh, M.T. Afzal, G. Mustafa, W. Alrawagfeh, A. Akhunzada, Machine learning approach for effective ranking of researcher assessment parameters, *IEEE Access* 11 (2023) 133294–133312, <https://doi.org/10.1109/ACCESS.2023.3336950>.
- [7] M. Usman, G. Mustafa, M.T. Afzal, Ranking of author assessment parameters using logistic regression, *Scientometrics* 126 (1) (2021) 335–353, <https://doi.org/10.1007/s11192-020-03769-y>.
- [8] A.A. Alshdadi, M. Usman, M.O. Alassafi, M.T. Afzal, R. AlGhamdi, Formulation of rules for the scientific community using deep learning, *Scientometrics* 128 (3) (2023) 1825–1852, <https://doi.org/10.1007/s11192-023-04633-5>.
- [9] J.E. Hirsch, An index to quantify an individual's scientific research output, *Proc. Natl. Acad. Sci.* 102 (46) (2005) 16569–16572, <https://doi.org/10.1073/pnas.0507655102>.
- [10] M. Weinberger, M. Zhitomirsky-Geffet, Modeling a successful citation trajectory structure for scholar's impact evaluation in Israeli Academia, *Heliyon* 9 (5) (2023), <https://doi.org/10.1016/j.heliyon.2023.e15673>.
- [11] K.R. Dienes, Completing h, *J. Informatr.* 9 (2) (2015) 385–397, <https://doi.org/10.1016/j.joi.2015.01.003>.
- [12] A. Bihari, S. Tripathi, A. Deepak, A review on h-index and its alternative indices, *J. Inf. Sci.* 49 (3) (2023) 624–665, <https://doi.org/10.1177/01655515211014478>.
- [13] S. Ayaz, M.T. Afzal, Identification of conversion factor for completing-h index for the field of mathematics, *Scientometrics* 109 (3) (2016) 1511–1524, <https://doi.org/10.1007/s11192-016-2122-z>.
- [14] Q.L. Burrell, On the h-index, the size of the hirsch core and jin's a-index, *J. Informatr.* 1 (2) (2007) 170–177, <https://doi.org/10.1016/j.joi.2007.01.003>.
- [15] B. Jin, L. Liang, R. Rousseau, L. Egghe, The r-and ar-indices: complementing the h-index, *Chin. Sci. Bull.* 52 (6) (2007) 855–863, <https://doi.org/10.1007/s11434-007-0145-9>.
- [16] N.R. Khan, C.J. Thompson, D.R. Taylor, K.S. Gabrick, A.F. Choudhri, F.R. Boop, P. Klimo Jr, Part ii: Should the h-index be modified? An analysis of the m-quotient, contemporary h-index, authorship value, and impact factor, *World Neurosurg.* 80 (6) (2013) 766–774, <https://doi.org/10.1016/j.wneu.2013.07.011>.
- [17] G. Kaptay, The k-index is introduced to replace the h-index to evaluate better the scientific excellence of individuals, *Heliyon* 6 (7) (2020), <https://doi.org/10.1016/j.heliyon.2020.e04415>.
- [18] Q.-u. Ain, H. Riaz, M.T. Afzal, Evaluation of h-index and its citation intensity based variants in the field of mathematics, *Scientometrics* 119 (2019) 187–211, <https://doi.org/10.1007/s11192-019-03009-y>.
- [19] N. De, S.M.A. Nayem, A. Pal, F-index of some graph operations, *Discrete Math. Algorithms Appl.* 8 (02) (2016) 1650025, <https://doi.org/10.1142/S1793830916500257>.
- [20] M. Schreiber, A modification of the h-index: the hm-index accounts for multi-authored manuscripts, *J. Informatr.* 2 (3) (2008) 211–216, <https://doi.org/10.1016/j.joi.2008.05.001>.
- [21] R. Ghani, F. Qayyum, M.T. Afzal, H. Maurer, Comprehensive evaluation of h-index and its extensions in the domain of mathematics, *Scientometrics* 118 (2019) 809–822, <https://doi.org/10.1007/s11192-019-03007-0>.
- [22] C. Moreira, P. Calado, B. Martins, Learning to rank academic experts in the dblp dataset, *Expert Syst.* 32 (4) (2015) 477–493, <https://doi.org/10.1111/exsy.12062>.
- [23] G. Mustafa, A. Rauf, B. Ahmed, M.T. Afzal, A. Akhunzada, S.Z. Alharthi, Comprehensive evaluation of publication and citation metrics for quantifying scholarly influence, *IEEE Access* 11 (2023) 65759–65774, <https://doi.org/10.1109/ACCESS.2023.3290917>.
- [24] G. Mustafa, A. Rauf, A.S. Al-Shamayleh, B. Ahmed, W. Alrawagfeh, M.T. Afzal, A. Akhunzada, Exploring the significance of publication-age-based parameters for evaluating researcher impact, *IEEE Access* (2023), <https://doi.org/10.1109/ACCESS.2023.3304013>.
- [25] R. Crowder, G. Hughes, W. Hall, An agent based approach to finding expertise, in: *International Conference on Practical Aspects of Knowledge Management*, 2002, pp. 179–188.
- [26] M. Raheel, S. Ayaz, M.T. Afzal, Evaluation of h-index, its variants and extensions based on publication age & citation intensity in civil engineering, *Scientometrics* 114 (2018) 1107–1127, <https://doi.org/10.1007/s11192-017-2633-2>.
- [27] Q. Burrell, Hirsch index or hirsch rate? Some thoughts arising from liang's data, *Scientometrics* 73 (1) (2007) 19–28, <https://doi.org/10.1007/s11192-006-1774-5>.
- [28] R. Liang, X. Jiang, Scientific ranking over heterogeneous academic hypernetwork, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, 2016.
- [29] G. Mustafa, M. Usman, M.T. Afzal, A. Shahid, A. Koubaa, A comprehensive evaluation of metadata-based features to classify research paper's topics, *IEEE Access* 9 (2021) 133500–133509, <https://doi.org/10.1109/ACCESS.2021.3115148>.
- [30] G.S. Patience, C.A. Patience, B. Blais, F. Bertrand, Citation analysis of scientific categories, *Heliyon* 3 (5) (2017), <https://doi.org/10.1016/j.heliyon.2017.e00300>.
- [31] R. Pates, P. Candon, K. Stenius, M. Miovský, J. O'Reilly, T. Babor, Publishing addiction science: a guide for the perplexed, <https://doi.org/10.5334/bbd>, 2017.
- [32] G. Mustafa, A. Rauf, M.T. Afzal, L. Egghe, Enhancing author assessment: an advanced modified recursive elimination technique (MRET) for ranking key parameters and conducting statistical analysis of top-ranked parameter, *Int. J. Data Sci. Anal.* (2024) 1–23, <https://doi.org/10.1007/s41060-024-00545-6>.
- [33] F. Ye, R. Rousseau, Probing the h-core: an investigation of the tail core ratio for rank distributions, *Scientometrics* 84 (2) (2010) 431–439, <https://doi.org/10.1007/s11192-009-0099-6>.
- [34] S. Alonso, F.J. Cabrerizo, E. Herrera-Viedma, F. Herrera, h-index: a review focused in its variants, computation and standardization for different scientific fields, *J. Informatr.* 3 (4) (2009) 273–289, <https://doi.org/10.1016/j.joi.2009.04.001>.
- [35] G. Mustafa, A. Rauf, M.T. Afzal, GK index: bridging Gf and K indices for comprehensive author evaluation, *Knowl. Inf. Syst.* (2024) 1–36, <https://doi.org/10.1007/s10115-024-02119-1>.
- [36] M. Ameer, M.T. Afzal, Evaluation of h-index and its qualitative and quantitative variants in neuroscience, *Scientometrics* 121 (2) (2019) 653–673, <https://doi.org/10.1007/s11192-019-03209-6>.
- [37] M. Salman, M.M. Ahmad, M.T. Afzal, Assessment of author ranking indices based on multi-authorship, *Scientometrics* 126 (2021) 4153–4172, <https://doi.org/10.1007/s11192-021-03906-1>.
- [38] S.M.A.H. Shah, A. Ullah, J. Iqbal, S. Bourouis, S.S. Ullah, S. Hussain, M.Q. Khan, Y.A. Shah, G. Mustafa, Classifying and localizing abnormalities in brain mri using channel attention based semi-Bayesian ensemble voting mechanism and convolutional auto-encoder, *IEEE Access* (2023), <https://doi.org/10.1109/ACCESS.2023.3294562>.

- [39] A. Rahangdale, S. Raut, Machine learning methods for ranking, *Int. J. Softw. Eng. Knowl. Eng.* 29 (06) (2019) 729–761, <https://doi.org/10.1142/S021819401930001X>.
- [40] J. Zhang, C. Li, Y. Yin, J. Zhang, M. Grzegorzek, Applications of artificial neural networks in microorganism image analysis: a comprehensive review from conventional multilayer perceptron to popular convolutional neural network and potential visual transformer, *Artif. Intell. Rev.* 56 (2) (2023) 1013–1070, <https://doi.org/10.1007/s10462-022-10192-7>.
- [41] D.H. Shirazi, H. Toosi, Deep multilayer perceptron neural network for the prediction of Iranian dam project delay risks, *J. Constr. Eng. Manage.* 149 (4) (2023) 04023011, <https://doi.org/10.1061/JCEMD4.COENG-12367>.
- [42] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in: *International Conference on Machine Learning*, 2015, pp. 448–456.
- [43] P. Theerthagiri, Predictive analysis of cardiovascular disease using gradient boosting based learning and recursive feature elimination technique, *Intell. Syst. Appl.* 16 (2022) 200121, <https://doi.org/10.1016/j.iswa.2022.200121>.
- [44] P.R. Kannari, N.S. Chowdary, R.L. Biradar, An anomaly-based intrusion detection system using recursive feature elimination technique for improved attack detection, *Theor. Comput. Sci.* 931 (2022) 56–64, <https://doi.org/10.1016/j.tcs.2022.07.030>.
- [45] M. Bilal, H. Israr, M. Shahid, A. Khan, Sentiment classification of roman-urdu opinions using naive bayesian, decision tree and knn classification techniques, *J. King Saud Univ, Comput. Inf. Sci.* 28 (3) (2016) 330–344, <https://doi.org/10.1016/j.jksuci.2015.11.003>.
- [46] A.R. Panhalkar, D.D. Doye, Optimization of decision trees using modified African buffalo algorithm, *J. King Saud Univ, Comput. Inf. Sci.* 34 (8) (2022) 4763–4772, <https://doi.org/10.1016/j.jksuci.2021.01.011>.
- [47] S. Tangirala, Evaluating the impact of Gini index and information gain on classification using decision tree classifier algorithm, *Int. J. Adv. Comput. Sci. Appl.* 11 (2) (2020) 612–619, <https://doi.org/10.14569/IJACSA.2020.0110277>.
- [48] S.A. Fayaz, M. Zaman, M.A. Butt, Performance evaluation of Gini index and information gain criteria on geographical data: an empirical study based on java and python, in: *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2021*, vol. 3, 2022, pp. 249–265.