Evolutionary Intelligence (2026) 19:14
https://doi.org/10.1007/s12065-025-01125-8

RESEARCH PAPER

Interpretable genetic programming with SHAP-guided multi-objective optimization for scientific impact modeling

Ghulam Mustafa[1] · Muhammad Saeed Khattak[1] · Sidra Ishfaq[1] · Muhammad Tanvir Afzal[1] · Qamar Mahmood[1] · Yasir Noman Khalid[2]

Received: 10 June 2025 / Revised: 22 September 2025 / Accepted: 28 November 2025

Abstract

In academic evaluation, identifying high-impact researchers using bibliometric indicators requires models that are not only accurate but also interpretable. This study introduces a novel SHAP-guided Genetic Programming (SHAP-GP) framework that evolves symbolic ranking expressions guided by accuracy, complexity, and SHAP-based interpretability. Using curated datasets spanning four academic domains Mathematics, Civil Engineering, Computer Science, and Neuroscience our multi-objective optimization approach simultaneously maximizes classification performance, minimizes symbolic complexity, and improves explanation stability through SHAP metrics. We encode 64 bibliometric indicators as terminal nodes and evolve closed-form expressions that rely on as few as three to five features. A surrogate model is trained for each symbolic candidate to compute SHAP values, enabling quantification of feature Compactness and Stability. Comparative evaluations against Decision Trees, Explainable Boosting Machines (EBM), and Symbolic Regressors (SR) confirm the framework's ability to produce domain-aligned, interpretable expressions with competitive F1-scores. In addition, we benchmark against three established GP baselines such as a standard accuracy-only GP, a parsimony-penalized GP, and a multi-objective GP without SHAP demonstrating that SHAP-GP consistently matches or outperforms these baselines while producing more compact and stable symbolic rules. Paired t-tests and Wilcoxon signed-rank tests validate the statistical significance of observed improvements. The proposed method offers a transparent and generalizable alternative to black-box classifiers for researcher profiling. By integrating interpretable AI into evolutionary design, the SHAP-GP framework advances decision transparency in research policy, funding allocation, and academic recognition.

Keywords Machine learning · Artificial intelligence · Genetic programming · Author assessment parameter · Shap-GP · Ranking parameters · Scientometrics

Ghulam Mustafa, Muhammad Saeed Khattak, Sidra Ishfaq, Muhammad Tanvir Afzal, Qamar Mahmood, and Yasir Noman Khalid contributed equally to this work.

✉ Ghulam Mustafa
ghulam.mustafa.ssc@stmu.edu.pk

Muhammad Saeed Khattak
saeed.ssc@stmu.edu.pk

Sidra Ishfaq
sidra_ishfaq.ssc@stmu.edu.pk

Muhammad Tanvir Afzal
dean.foc@stmu.edu.pk

Qamar Mahmood
qamar.ssc@stmu.edu.pk

Yasir Noman Khalid
yasir.noman.khalid@hitecuni.edu.pk

[1] Department of Computing, Shifa Tameer-e-Millat University, Islamabad 44000, Pakistan

[2] Department of Computing, HITEC, Taxila 24000, Pakistan

Published online: 27 December 2025

1 Introduction

Millions of researchers contribute to the ever expanding corpus of scientific literature each year [1–3]. However, the immediate impact and quality of their contributions often remain opaque due to the time required for scholarly recognition. This delay can hinder the timely acknowledgment of deserving individuals, potentially limiting their influence within the broader scientific community. The evaluation of researchers has long been a subject of debate, with various scientific societies adopting diverse criteria for measuring research impact. Despite these efforts, a universally accepted framework for identifying outstanding researchers has yet to emerge.

Numerous methodologies for assessing research impact have been proposed [4–7]. Traditional approaches such as expert reviews, while credible, are resource-intensive and time-consuming. In contrast, quantitative metrics–including total publications and citation counts–are more scalable but present their own limitations. For instance, some researchers may prioritize quantity over quality by publishing in low-impact venues [8, 9], or engage in self-citation to inflate citation metrics artificially [10, 11].

The introduction of the h-index by Hirsch marked a pivotal advancement by integrating both quantity and quality of output into a single metric [12–14]. Nonetheless, the h-index has inherent shortcomings. For example, increases in citations within the h-core do not necessarily improve the index [15], and papers with similar citation counts may fall outside its scope [16, 17]. To address these issues, numerous alternative metrics have been introduced [12, 18, 19], including the A-index [20], AR-index [21], M-Quotient [22], and k-index [23]. However, many of these metrics have been validated only on limited or hypothetical datasets, raising questions about their generalizability [24].

Evaluations of these alternative indices have produced mixed findings. For instance, Dienes compared the h-index with the g-index and complementary h-index in mathematics [15, 25], while De et al. focused on h-index variants in civil engineering, emphasizing citation intensity and publication age [26]. Schreiber assessed h-index variants using neuroscience data [27], while Ain et al. [24] and Ghani et al. [28] examined citation intensity-based indices in mathematics. Moreira et al. explored performance metrics for civil engineering researchers [29]. More recent studies by Mustafa et al. [30, 31] and Ahmed et al. [2] have further extended this line of research, analyzing parameters such as author count and publication age. Ahmed et al. also introduced dynamic random forests with brute-force optimizers for parameter ranking [6]. Despite these contributions, the field still lacks a transparent and actionable set of rules for systematically evaluating researchers[32–34].

In our previous work [5], interpretable decision rules were mined from bibliometric data using decision trees trained on the most influential indicators selected via Multilayer Perceptron (MLP) and Recursive Feature Elimination (RFE). While this framework produced promising classification performance and rule-based explainability, decision trees are constrained by greedy splitting strategies, limited symbolic expressiveness, and rigid thresholding. Moreover, they do not natively support multi-objective optimization, nor do they quantify the consistency of feature attributions across samples.

To overcome these limitations, we propose an interpretable framework that combines Genetic Programming (GP) with SHAP-guided multi-objective optimization for scientific impact modeling. GP evolves symbolic expressions that directly represent researcher evaluation rules, while SHAP (SHapley Additive exPlanations) [35] provides a principled method for assessing the stability and transparency of these evolved expressions.. By embedding SHAP-based criteria such as attribution sparsity and consistency into the fitness function, we guide GP toward solutions that are not only accurate but also intelligible and trustworthy.

In this study, we utilize a dataset of researchers across four domains such as Mathematics, Civil Engineering, Computer Science, and Neuroscience comprising equal numbers of awardees and non-awardees. A total of 64 bibliometric indicators are used, grouped into primitive, citation-based, authorship-based, and age-adjusted categories. These features serve as the terminal nodes in the GP engine. The symbolic evolution is driven by a multi-objective fitness function that optimizes classification performance (via F1-score), symbolic complexity (tree size), and SHAP-based interpretability (attribution compactness and stability).

Our experimental results show that the proposed framework outperforms traditional decision tree baselines by evolving algebraic expressions that are both compact and highly discriminative. For instance, in the Mathematics domain, a symbolic model with only 7 nodes achieved an F1-score of 0.7892. Other domains, such as Civil Engineering and Neuroscience, also yielded interpretable expressions with F1-scores above 0.65, demonstrating the generalizability of our approach across varied citation cultures. Furthermore, SHAP-based analysis confirmed that the most influential features in each model align with established domain heuristics, supporting the transparency and trustworthiness of the results.

The main contributions of this study are as follows: (1) We introduce the first SHAP-guided, multi-objective genetic programming framework tailored for researcher impact modeling using bibliometric data. (2) We design a fitness function that integrates SHAP-based interpretability measures to evolve symbolic expressions that are both accurate

and explainable. (3) We demonstrate, through empirical validation, that our GP-based models not only outperform traditional decision tree, EBM, and symbolic regression baselines, but also achieve competitive or superior results compared to established GP variants. (4) We make available the source code and benchmark dataset to facilitate reproducibility and future research. (5) We conduct an ablation study isolating the SHAP objective, showing that explicitly optimizing for SHAP-based stability yields more consistent and compact symbolic rules without sacrificing accuracy.

The remainder of this paper is structured as follows: Sect. 2 reviews related literature; Sect. 3 presents the proposed methodology; Sect. 4 outlines the experimental setup and results; and Sect. 5 concludes with key findings and future directions.

2 Literature

In the current scientific landscape, establishing standardized criteria for evaluating researcher performance is essential to ensure fair and unbiased recognition systems. Commonly used metrics include total publication counts, citation volumes, the h-index, and its numerous derivatives. While widely adopted, these indicators often serve as proxies for academic excellence and are used in tenure, funding, and award decisions. However, the reliability and comprehensiveness of such bibliometric indicators have been questioned, especially in contexts requiring equitable and domain-sensitive assessments.

Traditional methods of researcher evaluation such as peer review and expert panels carry the advantage of domain insight but are resource intensive and potentially biased [36–39]. In contrast, quantitative bibliometric metrics offer scalability but are susceptible to gaming and distortion. For example, high publication counts may reflect quantity over quality, especially when derived from low-impact venues [22]. Citation-based measures can be inflated through self-citations or driven by negative references rather than genuine impact [40].

Among these, the h-index has emerged as a widely accepted yet imperfect metric. It attempts to capture both productivity and citation impact in a single value. However, its limitations are well-documented: it does not increase with additional citations to already influential papers, and it inherently favors senior researchers with longer academic careers [41–43]. To mitigate these weaknesses, several alternative indices have been proposed, including the A-index, AR-index, M-Quotient, k-index, and f-index. These aim to normalize for factors such as publication age, author count, and collaboration density.

Recent studies have evaluated these variants across multiple academic disciplines. For example, Ayaz et al. [48] compared h-index variants using award data from mathematical societies and found the original index to be robust in that context. Raheel et al. [44] extended this by validating additional metrics. Ameer et al. [45] identified the hg-index and R-index as particularly useful in neuroscience, while Ain et al. [46] explored the predictive value of various indicators in mathematics. However, these studies often relied on award data from periods that predate the development of modern metrics, introducing potential historical bias.

To address this, Usman et al. [8] evaluated civil engineering researchers based on contemporaneous award data, reducing temporal confounding. Still, their dataset lacked sufficient scale to support generalizable conclusions. Alshdadi et al. [10] proposed rule-based guidelines for scientific evaluation, though their feature set was narrow. More recently, domain-specific studies by Mustafa et al. [30, 31] investigated the predictive power of citation and age-weighted metrics, identifying the normalized h-index and AR-index as top performers within mathematics.

While these studies represent important advances, most have focused on single metrics or limited combinations. Crucially, few approaches integrate multiple indicator types in a holistic, interpretable framework. Even fewer attempt to evolve symbolic rules automatically, nor do they explicitly account for trade-offs between accuracy, model simplicity, and explainability.

A critical gap in the literature is the lack of evolutionary, interpretable models that can simultaneously optimize performance, complexity, and transparency in researcher ranking. Existing models are either black-box in nature such as neural networks or limited to greedy, non-generalizable rule learners like decision trees. Additionally, no prior work has incorporated SHAP (SHapley Additive Explanations) to evaluate or guide the interpretability of evolved ranking rules. To the best of our knowledge, while SHAP has been widely adopted in tree-based and differentiable models [35, 47], its integration into symbolic GP workflows particularly through surrogate modeling remains limited.

This paper addresses these gaps by introducing a SHAP-guided multi-objective genetic programming framework that evolves symbolic rules for identifying high-impact researchers. Our approach builds on past parameter ranking studies but advances the field by combining bibliometric diversity, symbolic interpretability, and explainable AI within a single, transparent system.

3 Methodology

Building upon insights from the literature and limitations identified in previous rule-based researcher recognition systems, this study introduces a novel methodology that integrates symbolic learning, multi-objective optimization, and explainable AI. The aim is to evolve interpretable ranking rules that identify high-impact researchers using a transparent and generalizable framework.

Unlike prior approaches that rely on decision trees for rule mining, our method utilizes Genetic Programming (GP) to evolve symbolic expressions. These expressions are evaluated not only for classification performance but also for model simplicity and interpretability using SHAP (SHapley Additive Explanations). This results in evolved rules that are compact, accurate, and explainable at both global and local levels.

The proposed methodology, visualized in Fig. 1, follows these sequential stages:

(i) Selection of the domain and collection of dataset,
(ii) Computation of author assessment parameters,
(iii) Ranking of features using Multilayer Perceptron (MLP) and Recursive Feature Elimination (RFE),
(iv) Initialization of a Genetic Programming system using top-ranked features,
(v) SHAP-guided multi-objective evolution of interpretable symbolic rules,
(vi) Evaluation and selection of evolved models based on accuracy, complexity, and SHAP-based interpretability.

The subsequent subsections detail each component of this methodology, including dataset construction, parameter computation, feature ranking, genetic program structure, SHAP integration, and model evaluation.

3.1 Domain selection

This study evaluates the effectiveness of SHAP-guided genetic programming in modeling researcher impact across four diverse academic disciplines: Computer Science, Civil Engineering, Neuroscience, and Mathematics. These domains were selected to ensure coverage of both computational and applied sciences, allowing the framework to be evaluated for both generalizability and discipline-specific behavior.

- Computer Science (CS): A dynamic and fast-evolving field known for its volume of publications and diverse collaboration structures. CS was chosen due to its strong representation in bibliometric analysis and frequent inclusion in prior researcher ranking studies.
- Civil Engineering (CE): A well-established and foundational discipline with a rich history of scholarly contributions. Despite its importance, Civil Engineering lacks standardized, data-driven frameworks for identifying high-impact contributors, making it a prime candidate for validation.
- Mathematics (Math): A core theoretical domain with strong interdisciplinary ties. Its relatively slower citation lifecycle provides contrast to high-volume fields like CS.
- Neuroscience (Neuro): A biomedical field with rapid growth and high-impact research focused on human health, allowing the framework to be tested on datasets that prioritize real-world impact.

These domains were selected based on both diversity in publication and citation patterns and availability of curated awardee data from reputable sources.

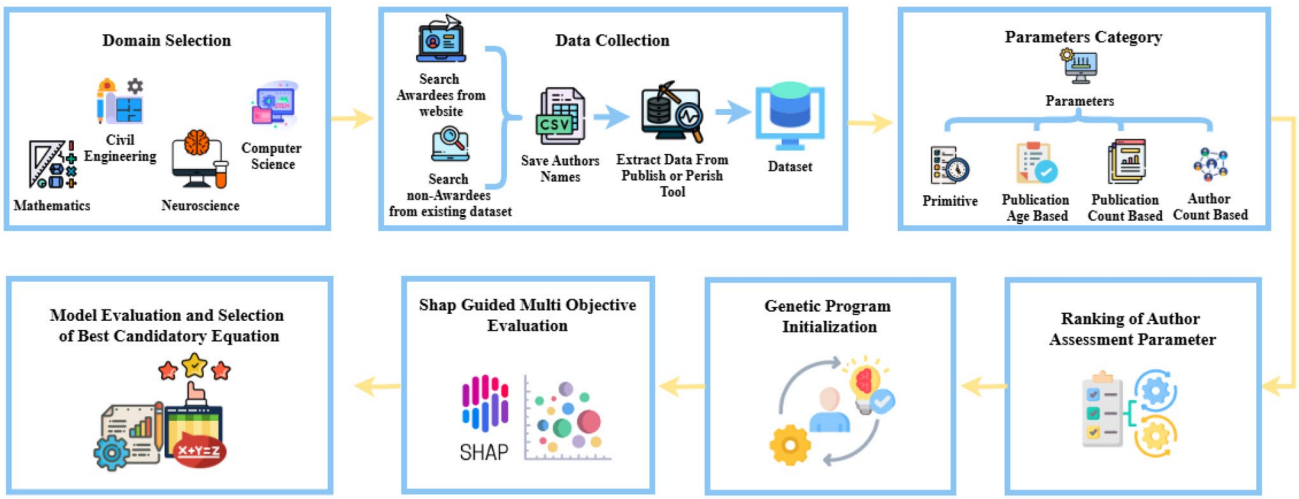

Fig. 1 Architecture of the proposed SHAP-Guided GP methodology

3.2 Dataset collection

To create a balanced evaluation dataset for each selected domain, we compiled bibliometric profiles of recognized **awardees** and their **non-awardee** peers. Data sources included scientific society websites, institutional award records, and prior benchmark studies. Key details include:

- Computer Science: Awardees were sourced from ACM and IEEE (e.g., Turing Award, IEEE Fellow), while non-awardees were collected using the Artminor dataset [48].
- Civil Engineering: Awardees were identified from professional societies such as ACI, ASCE, CSCE, and ICE. Non-awardees were drawn from Usman et al. [49].
- Neuroscience: Awardees were collected from societies like SFN, FENS, CNS, and ANS, with non-awardees derived from Ameer & Afzal [50].
- Mathematics: Awardees were retrieved from IMU, AMS, and LMS databases, with corresponding non-awardees sourced from Mustafa et al. [15].

Researcher profiles were extracted using the *Publish or Perish* tool, targeting Google Scholar-based citation metadata. A 'hold-on' policy was implemented to ensure non-awardees were matched by publication periods relative to award years, ensuring temporal fairness. An equal number of awardees and non-awardees were collected per domain. Data cleaning included such as, Removal of duplicates and irrelevant profiles, Disambiguation of author names across different spellings, Consistency filtering for publication data (e.g., impact, venue, age). Table 1 summarizes the final dataset statistics.

Table 1 Dataset statistics

Mathematics	
Authors count	1050
Awardees count	525
Non-awardees count	525
Publications count	204,896
Citations count	14,370,007
Civil Engineering	
Authors count	1180
Awardees count	590
Non-awardees count	590
Publications count	214,672
Citations count	24,061,210
Neuroscience	
Authors count	1060
Awardees count	530
Non-awardees count	530
Publications count	166,871
Citations count	25,855,493
Computer Science	
Authors count	1200
Awardees count	600
Non-awardees count	600
Publications count	171,388
Citations count	32,801,476

3.3 Calculation of author assessment parameters

This study utilizes a comprehensive set of 64 author assessment parameters, systematically computed from the curated datasets across four academic domains. These parameters are grouped into four distinct categories designed to capture the multifaceted nature of scholarly performance. Each category represents a unique analytical perspective, facilitating robust and interpretable rule evolution.

3.3.1 Primitive parameters

These are foundational metrics reflecting basic productivity and collaboration dynamics:

- Total Publications, Total Citations, Total Years
- Cites per Year, Cites per Paper
- Authors per Paper, Cites per Author, Papers per Author

3.3.2 Publication and citation count-based parameters

Advanced indices capturing citation depth and distribution:

- H-index, G-index, E-index, A-index, R-index, M-index, F-index, T-index
- Tapered H-index, Maxprod, Wu-index, Weighted H-index, I10-index
- H2-index (upper, center, lower), HG-index, Rational H-index, Real H-index, Normalized H-index, H-core Citation
- Pi-index, P-index, K-index, W-index, Q2-index, H-dash index, Woginger-index, GH-index, RM-index, X-index, K-dash index

3.3.3 Author count-based parameters

Metrics adjusted for co-authorship, reflecting individual contribution:

- HI-index, Hm-index, HI-normalized, Gm-index, Hf-index, Gf-index, GF-index
- Pure H-index, Fractional H-index, Fractional G-index, K-norm index, W-norm index, Normalized HI-index

3.3.4 Age-based parameters

Indices accounting for citation accumulation over time:

- Platinum H-index, AR-index, AW-index, M-Quotient
- Hc-index (Contemporary H), Ha-index, V-index, AWCR

3.4 Ranking of author assessment parameters

Effective symbolic modeling, such as Genetic Programming (GP), depends critically on identifying and utilizing the most informative features. With 64 bibliometric indicators available, it is neither computationally efficient nor interpretable to evolve models using all variables. To address this, we adopt a supervised feature ranking approach using a Multilayer Perceptron (MLP)classifier combined with Recursive Feature Elimination (RFE) to isolate the most impactful parameters.

This ranking strategy not only streamlines the GP's search space but also enhances the quality of evolved rules by focusing on features with strong discriminative power. As illustrated in Fig. 2, we begin by splitting the dataset into training and validation subsets in an 80:20 ratio. An MLP classifier is trained on the full feature set to establish a baseline accuracy (BA). In each iteration, one feature is removed from the dataset, and the model is retrained. The resulting accuracy (RFA) is recorded, and the difference from the baseline is used to compute the *importance score (IS)* for the excluded feature. This process is repeated for all features, yielding a ranked list based on their contribution to classification performance.

To understand the internal mechanics of the MLP, consider the forward propagation step, where input features are transformed linearly:

$$X = WA + b \tag{1}$$

Here, A represents the input feature vector, W is the learned weight matrix, and b is the bias. The result X is then passed through an activation function. For hidden layers, we use the Rectified Linear Unit (ReLU):

$$f(X) = \max(0, X) \tag{2}$$

The output layer applies a Softmax activation to convert logits into a probability distribution across the output classes (awardee vs. non-awardee):

$$\text{Softmax}(X_i) = \frac{e^{X_i}}{\sum_{j=1}^{J} e^{X_j}} \tag{3}$$

To guide learning, the model minimizes the Mean Squared Error (MSE) between true labels z_i and predictions $\hat{z}_i$:

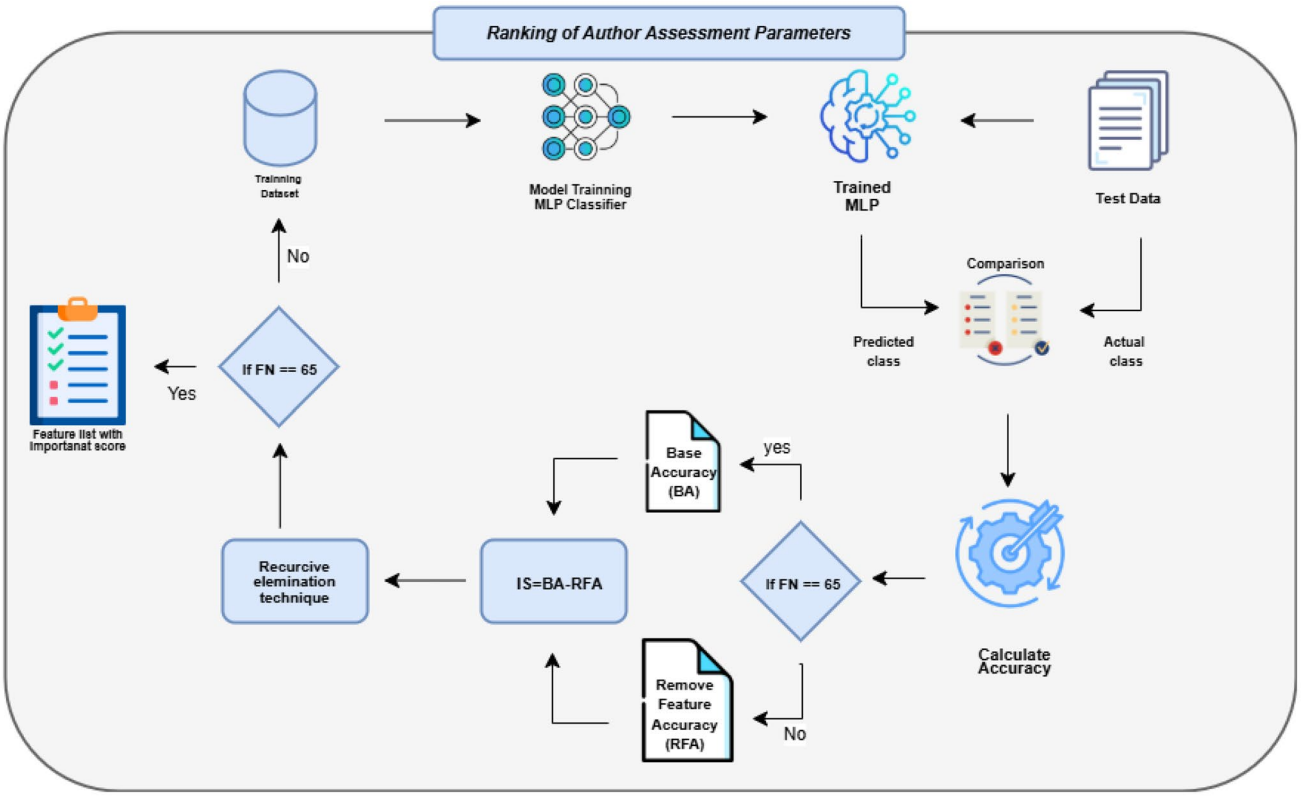

Fig. 2 Feature ranking process: The dataset is split into training and validation sets (80:20). A Multilayer Perceptron (MLP) is trained on the training set, and baseline accuracy (BA) is computed using all 64 features. In successive iterations, one feature is removed, the model is retrained, and the resulting accuracy (RFA) is recorded. The importance score (IS) is calculated as the difference between BA and RFA. This process continues until all features are evaluated, resulting in a ranked list based on IS

$$L(z, \hat{z}) = \frac{1}{n} \sum_{i=1}^{n} (z_i - \hat{z}_i)^2 \tag{4}$$

For training stability and faster convergence, input data is standardized using batch normalization:

$$X_i = \frac{X_i - \text{Mean}_i}{\text{Standard Deviation}_i} \tag{5}$$

The MLP consists of 10 hidden layers with 10 neurons each. It is optimized using the Adam algorithm with a learning rate of 0.0003 and a batch size of 64. Early stopping is applied after 40 epochs without validation improvement to prevent overfitting.

Once all features are evaluated, the top-ranked subset typically the most informative 15–20 parameters is selected. These features are then passed as terminal nodes to the Genetic Programming engine. This targeted selection enhances interpretability and ensures that evolved symbolic rules are both computationally efficient and domain-relevant.

3.5 Initialization of genetic programming

Genetic Programming (GP) is an evolutionary algorithm that automatically discovers symbolic expressions to solve a given task in this case, distinguishing between awardees and non-awardees using bibliometric parameters. Unlike traditional machine learning models, GP evolves interpretable mathematical formulas represented as tree structures, where each node corresponds to either a function (e.g., arithmetic operator) or a terminal (e.g., a selected bibliometric feature).

GP is particularly well-suited for this task because it produces interpretable, closed-form mathematical expressions, which are essential for high-stakes academic decisions. Unlike neural networks or ensemble models, GP expressions can be directly examined, understood, and audited by domain experts making them ideal for transparent researcher evaluation frameworks.

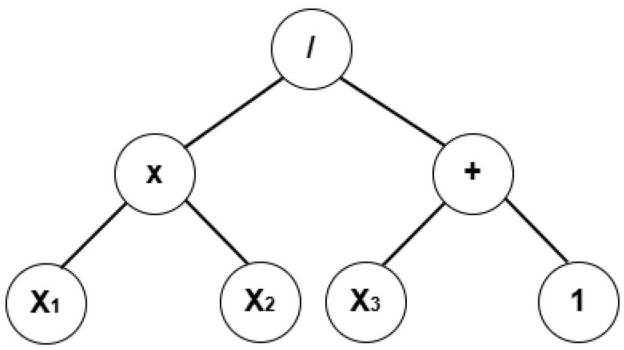

Fig. 3 Example of a GP individual represented as a tree structure. Internal nodes are mathematical operations (function set), while leaf nodes are selected bibliometric features (terminal set). The expression encodes a symbolic rule for classifying researchers

In our framework, the terminal set is constructed using the top-ranked features identified through MLP and RFE-based feature ranking. These include the most informative bibliometric indicators across categories such as citation counts, author-normalized indices, and age-adjusted metrics. The function set includes basic arithmetic operations as well as simple nonlinear transformations to maintain expression interpretability:

$$\text{Function set} = \{+, -, \times, \div, \log, \sqrt{\cdot}\}$$

To prevent runtime errors during expression evaluation, protected versions of sensitive functions (e.g., division by zero, log of negative values) are employed throughout the evolutionary process.

Each candidate solution in GP is represented as a syntax tree composed of function nodes (e.g., arithmetic operations) and terminal nodes (e.g., input variables). For example, the following symbolic expression:

$$f(x) = \frac{x_1 \times x_2}{x_3 + 1}$$

corresponds to the tree structure shown in Fig. 3. In this expression, the internal nodes represent division, multiplication, and addition operations, while the leaf nodes x_1, x_2, x_3, and the constant 1 represent selected input features and literals. This tree-structured representation enables the evolution of transparent and human-readable rules for classifying researchers.

The initial population is generated using the ramped half-and-half method, which combines two strategies for tree generation: the "full" method, where trees are generated to their maximum depth, and the "grow" method, which allows variable depth and randomness in function-terminal placement. This hybrid approach ensures structural diversity in early generations and helps avoid premature convergence.

The population size, tree depth, and initial parameters are empirically set to balance expressiveness and computational efficiency. These symbolic individuals then undergo fitness evaluation in the next stage, where multi-objective optimization guides their evolution. An illustrative example of such a GP individual is shown in Fig. 3, where a symbolic expression is represented as a tree. Internal nodes denote arithmetic operations, while the terminal nodes are selected bibliometric indicators. This hierarchical structure enables the evolution of interpretable mathematical rules for researcher classification.

3.6 SHAP-guided multi-objective evolution

After initializing the population of symbolic expressions, the Genetic Programming (GP) framework enters its core evolutionary loop. Our approach enhances standard GP by embedding a

multi-objective fitness mechanism that evaluates not only prediction accuracy but also model simplicity and interpretability.

We decompose the evolution process into six distinct steps, as illustrated in Fig. 4.

Step 1: Population initialization

The evolutionary process begins by initializing a diverse population of symbolic expressions (trees). These trees are constructed using top-ranked features identified by a prior SHAP analysis on a baseline model. To promote variety, both simple and complex expressions are generated. This initial population serves as the starting point for the evolutionary loop.

Step 2: Evaluate accuracy & complexity

Each individual in the population represents a symbolic expression. We compute its classification accuracy using F1-score on the training dataset:

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

This metric ensures a balance between false positives and false negatives in researcher classification.

Step 3: Compute SHAP values for interpretability

We fit a surrogate model (e.g., a shallow decision tree) to mimic the predictions of the GP expression *f*. SHAP values are computed for each feature, allowing us to assess interpretability through: SHAP Compactness (Number of dominant features (top-k) contributing most to predictions) and SHAP Stability (Variance of SHAP values across data samples, defined in following equation.

$$\text{SHAPStability}(f) = \frac{1}{n} \sum_{i=1}^{n} \text{Var}(\text{SHAP}_i)$$

Step 4: Extract SHAP metrics and compute multi-objective fitness

To evaluate interpretability, we compute SHAP values using a surrogate model trained to mimic the symbolic expression *f* (e.g., a shallow decision tree). SHAP values quantify the contribution of each feature to individual predictions and are used to derive two interpretability metrics:

- SHAP cmpactness: The number of dominant features (top-k) that contribute most to predictions.
- SHAP stability: The variance of SHAP values across instances, defined as:

$$\text{SHAPStability}(f) = \frac{1}{n} \sum_{i=1}^{n} \text{Var}(\text{SHAP}_i)$$

We also assess symbolic complexity using the depth and node count of the expression tree:

$$\text{Complexity}(f) = \text{depth}(f) + \lambda \cdot \text{nodes}(f)$$

where λ is a tunable regularization parameter penalizing large trees.

These metrics–F1-score (predictive accuracy), symbolic complexity, and SHAP stability–are then combined into a **multi-objective fitness vector**:

$$\text{Fitness}(f) = [\text{F1}(f), -\text{Complexity}(f), -\text{SHAPStability}(f)]$$

This triple-objective fitness function ensures that selected models are not only accurate but also simple and interpretable.

Step 5: NSGA-II selection

To identify the best trade-offs among competing objectives, we apply the Non-dominated Sorting Genetic Algorithm II (NSGA-II). In this process, individuals are ranked into Pareto fronts based on non-dominance. An individual is said to dominate another if it is no worse in all objectives and better in at least one.

NSGA-II assigns:

- Pareto rank: Determined by how many other individuals dominate a given solution.
- Crowding distance: A density estimator used to maintain population diversity.

Selection for the next generation is based on a combination of Pareto rank and crowding distance, favoring solutions that are both high-performing and diverse. This ensures that the evolutionary process explores a broad spectrum of symbolic expressions that balance accuracy, complexity, and interpretability.

Step 6: Crossover and mutation

New candidate solutions are generated through standard Genetic Programming operators, which introduce structural variation in the symbolic expressions and maintain population diversity.

- Crossover: Two parent expression trees are selected, and randomly chosen subtrees are exchanged between them. This allows beneficial building blocks (sub-expressions) to be recombined and propagated to offspring, promoting the discovery of better solutions.
- Mutation: A single expression tree is randomly altered by replacing a selected subtree with a newly generated subtree or by modifying terminal nodes. Mutation injects novel structures into the population, helping to escape local optima and improve exploration.

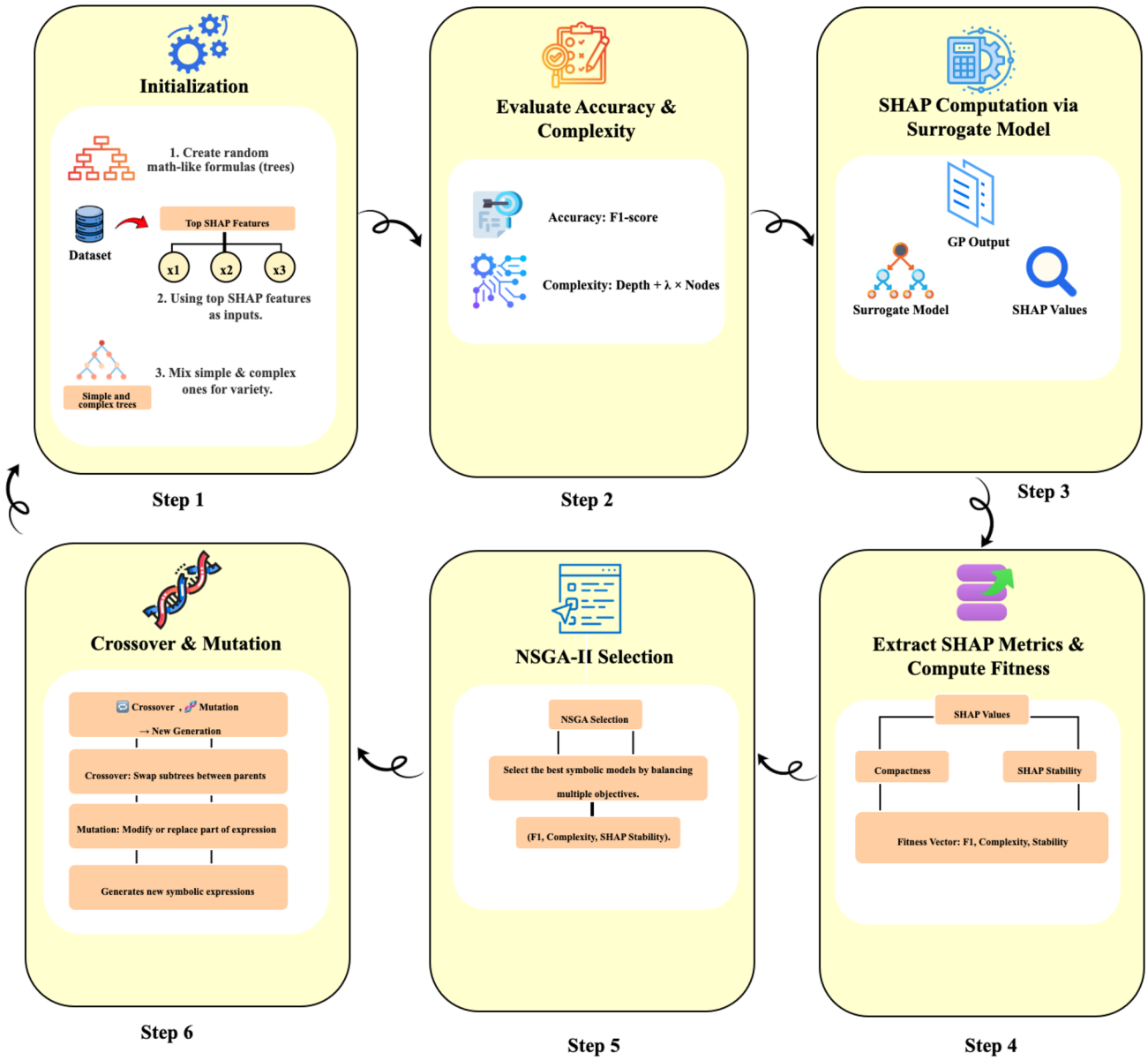

Fig. 4 Step-wise evolutionary loop of the proposed SHAP-Guided Genetic Programming (GP) framework. Step 1: Initialize symbolic expressions (trees) using top SHAP-ranked features from the dataset. Step 2: Evaluate F1-score and expression complexity. Step 3: Train a surrogate model to approximate GP outputs and compute SHAP values. Step 4: Extract SHAP-based interpretability metrics–Compactness and Stability–and combine them with accuracy and complexity into a multi-objective fitness vector. Step 5: Apply NSGA-II to select the next generation of individuals based on trade-offs. Step 6: Perform crossover and mutation to evolve new symbolic expressions. This evolutionary process iteratively balances accuracy, simplicity, and interpretability until convergence

These variation operators ensure that the search space of symbolic expressions is thoroughly explored. The resulting offspring are evaluated and selected in the next generation, and the evolutionary cycle continues until a predefined number of generations is reached or convergence is observed.

As SHAP (SHapley Additive exPlanations) is originally designed for tree ensembles and differentiable models, directly applying it to symbolic Genetic Programming (GP) trees presents computational challenges. Symbolic expressions in GP represent functional program structures rather than statistical models, and therefore do not naturally expose internal feature attribution mechanisms.

To address this, we employ a *surrogate modeling approach*. For each symbolic expression f, we first generate its predictions on the training data. We then train a shallow surrogate model (typically a decision tree) to approximate the output behavior of f. SHAP values are computed on this surrogate model using the standard TreeSHAP method, which yields local feature attributions for each sample.

This approximation allows us to define meaningful SHAP-based metrics for symbolic expressions, including:

- SHAP compactness: The number of top-ranked features contributing most to the predictions.
- SHAP stability: The variance of SHAP values across data samples, representing consistency of feature attribution.

These SHAP-derived metrics are then used to guide the evolutionary process. It is important to emphasize that **SHAP values do not directly alter the structural operators** of Genetic Programming, such as crossover, mutation, or tree construction. Nor are SHAP values used to influence node-level selection or replacement during expression building.

Instead, SHAP values serve as interpretability objectives within the **multi-objective fitness function**. After a symbolic expression is evaluated for accuracy and complexity, the Compactness and Stability values derived from the surrogate model are combined with these metrics to form the fitness vector:

$$\text{Fitness}(f) = [\text{F1-score}, -\text{Complexity}, -\text{SHAP Stability}]$$

Selection of individuals for the next generation is handled by NSGA-II based on this vector. Therefore, SHAP influences the evolutionary search *indirectly* by shaping the **selection pressure** toward models that are not only accurate and simple but also interpretable. However, it does *not* participate in modifying the structure of GP trees during variation operations.

To operationalize this framework, Algorithm 1 summarizes the complete GP evolution process.

Algorithm 1 outlines the complete evolutionary process of the proposed SHAP-guided Genetic Programming (GP) framework for researcher ranking. The process begins by initializing a population of symbolic expressions, where each individual is generated using the top-ranked bibliometric features identified during feature selection. In each generation, every individual expression undergoes a comprehensive evaluation. First, its classification performance is measured using the F1-score, and its symbolic complexity is computed based on the depth and size of the expression tree. Then, a surrogate model–typically a shallow decision tree–is trained to mimic the predictions of the symbolic expression. This surrogate enables the computation of SHAP values, which quantify how input features contribute to the expression's outputs. From these SHAP values, two interpretability metrics are derived: SHAP Compactness (the number of dominant features with high contribution) and SHAP Stability (the variance of SHAP values across data samples). These three evaluation criteria–F1-score, complexity, and SHAP-based interpretability–are then aggregated into a multi-objective fitness vector. The NSGA-II algorithm is applied to select a Pareto-optimal set of individuals based on trade-offs among these objectives. Next, crossover and mutation operators introduce structural diversity by recombining and modifying expression trees. The resulting offspring form the next generation, and this evolutionary cycle is repeated for a predefined number of generations or until convergence is observed. The final output is a set of symbolic expressions that not only provide strong predictive accuracy but also maintain structural simplicity and

Input: Ranked feature set $\mathcal{X}$, population size N, generations G
Output: Pareto-optimal set of interpretable expressions
Initialize population P_0 with N individuals using features from $\mathcal{X}$
for $t = 1$ to G **do**
  for each individual f in P_{t-1} **do**
    Evaluate F1-score of f on training data
    Compute tree depth and node count to assess complexity
    Generate predictions from f on training data
    Fit surrogate model (e.g., decision tree) on predictions
    Compute SHAP values for each feature
    Calculate:
  • SHAP Compactness (number of dominant features)
  • SHAP Stability (variance across instances)
  end for
  Apply NSGA-II to select Pareto-optimal individuals based on:
  • Maximize F1-score
  • Minimize complexity
  • Minimize SHAP instability
  Apply crossover and mutation to generate offspring
  Form next generation P_t
end for
Return Pareto front of evolved symbolic expressions

Algorithm 1 SHAP-guided multi-objective genetic programming for researcher ranking

interpretability, making them suitable for use in transparent academic evaluation systems.

3.7 Model evaluation and selection

After completing the evolutionary process, the Genetic Programming (GP) framework outputs a set of non-dominated solutions forming the Pareto front. Each solution represents a symbolic expression that achieves a different balance among classification accuracy, symbolic complexity, and SHAP-based interpretability.

3.7.1 Evaluation protocol

To evaluate the generalizability of evolved models, we adopt a stratified train-test split across all four domains (Computer Science, Civil Engineering, Mathematics, Neuroscience). In each domain, an equal number of awardees and non-awardees are included. The training data is used for evolution and fitness computation, while the test set is held out for final evaluation.

The classification task is binary: distinguishing between award-winning and non-award-winning researchers based on bibliometric parameters. Performance is reported separately for each domain to assess robustness across disciplinary contexts.

3.7.2 Evaluation metrics and comparative baselines

Each candidate model is evaluated along three core dimensions:

- Accuracy: Measured using F1-score, precision, recall, and AUC on the held-out test set. These metrics assess the ability to distinguish between awardees and non-awardees.
- Complexity: Quantified using the depth and size (number of nodes) of the expression tree, ensuring interpretability and simplicity.
- Interpretability: Evaluated through SHAP-based compactness (number of dominant features) and SHAP stability (consistency of feature attribution across samples).

To broaden the evaluation and benchmark our framework against established interpretable models, we incorporate two additional baselines:

- Explainable Boosting Machine (EBM): A Generalized Additive Model that applies boosted trees per feature to learn nonlinear yet interpretable relationships.
- Symbolic Regressor (SR): A symbolic regression approach that evolves concise algebraic expressions, enabling direct human interpretability.

These models are trained and tested using the same four-domain datasets (Mathematics, Civil Engineering, Computer Science, and Neuroscience) under identical preprocessing settings and evaluation protocol.

To ensure fair comparison, we compute the following standardized interpretability metrics for each model:

- Tree size (Nodes): Total number of nodes/terms in the model.
- Tree depth: Maximum depth from root to leaf.
- Number of features: Total features used in the final expression or shape functions.

The comparative results across all models and domains are presented in Sect. 4.3.

3.7.3 Established GP baselines

To align with established genetic programming (GP) practice, we include three GP baselines:

1. *GP-Acc (standard GP).* A single-objective GP that maximizes classification F1 on the training fold with no explicit interpretability or complexity terms:

$$\max_{\theta} \ \mathrm{F1}\big(y, f_\theta(X)\big). \tag{6}$$

2. *GP-Acc+Size (parsimony).* A single-objective GP with parsimony pressure that trades off accuracy and expression length:

$$\max_{\theta} \ \mathrm{F1}\big(y, f_\theta(X)\big) \ - \ \alpha \cdot \mathrm{Size}\big(\mathcal{T}_\theta\big), \tag{7}$$

where $\mathrm{Size}(\mathcal{T})$ is the node count of the expression tree and $\alpha \in \{0.001,\ 0.01,\ 0.05\}$ is selected on a validation split (per domain).

3. *MOGP-Acc−Complex (multi-objective without SHAP).* A multi-objective GP that optimizes predictive performance and structural simplicity *without* the SHAP objective:

$$\max_{\theta} \ \Big[\mathrm{F1}\big(y, f_\theta(X)\big),\ -C\big(\mathcal{T}_\theta\big)\Big], \tag{8}$$

with NSGA-II selection. We define structural complexity as

$$C(\mathcal{T}) \ = \ d(\mathcal{T}) \ + \ \lambda \cdot |\mathcal{T}|, \tag{9}$$

where $d(\mathcal{T})$ is tree depth, $|\mathcal{T}|$ is node count, and λ matches the value used in our main method.

Fairness of comparison. All GP variants (*GP-Acc*, *GP-Acc+Size*, *MOGP-Acc−Complex*, and our SHAP-guided GP) use *identical* evolutionary budgets and operators: same population size, number of generations, depth limits, crossover/mutation rates, early stopping, train/validation/test splits, and random seeds. Selection is tournament for single-objective runs and NSGA-II for multi-objective runs. Hyperparameters not specified here are inherited from the GP setup described above.

Post-hoc SHAP for parity. To ensure interpretability parity, we compute SHAP values *post hoc* for *all* GP variants using the same surrogate procedure as in our framework. We report:

- SHAP Stability (lower is better): mean across features of the per-sample variance of SHAP values,

$$\text{Stability} = \frac{1}{p}\sum_{j=1}^{p}\text{Var}_i(\phi_{ij}),$$

where ϕ_{ij} is the SHAP value of feature *j* for sample *i* and *p* is the number of features.

- SHAP compactness (top-*k*, lower is better): number of distinct features required to cover the top-*k* mean absolute SHAP contributions.

Reported metrics and statistics. For each method and domain we report **F1**, **Precision**, **Recall**, and **AUC**, along with **Size** (nodes), **Depth**, and **Features**. We aggregate mean±sd over $K=10$ seeds and apply **paired** *t*-**tests** and **Wilcoxon signed-rank tests** for significance; we also report **Cliff's** δ as an effect size.

3.7.4 Interpretability visualization and expert validation

To supplement numeric evaluation, we generate SHAP summary plots for each evolved model. These plots highlight the global contribution of each feature, allowing domain experts to verify the logical alignment of model behavior with human intuition.

Additionally, local SHAP explanations provide sample-level insight revealing which features led to a specific researcher's classification as awardee or non-awardee. This supports transparency and traceability in decision-making.

Finally, selected symbolic expressions are presented in closed-form mathematical notation. These are validated qualitatively by academic evaluators for clarity and fairness, making them suitable for use in researcher recognition frameworks.

By integrating performance and transparency, this framework enables not just accurate prediction but also trustworthy insight into the criteria driving researcher recognition.

3.8 SHAP GP workflow summary

To address reproducibility and enhance clarity, we present a consolidated summary of the SHAP-guided Genetic Programming (SHAP-GP) workflow. This overview outlines how SHAP values are computed, where they are integrated within the evolutionary process, and how they influence model evolution. While detailed explanations appear across Sections 3.4 to 3.7, this section provides a compact, end-to-end description of the SHAP-GP mechanism.

The workflow consists of the following core steps:

1. Feature ranking and initialization: A ranked subset of the most informative bibliometric features is selected using Multilayer Perceptron (MLP) and Recursive Feature Elimination (RFE). These top features form the terminal set for symbolic expressions evolved by the Genetic Programming (GP) engine.
2. Symbolic expression evolution: GP evolves symbolic rules as tree-structured expressions using arithmetic operations and selected features. Each individual (i.e., expression) is evaluated on three fronts: accuracy, complexity, and interpretability.
3. Fitness Evaluation (Multi-Objective):
   - *Classification Performance*: Measured via F1-score on the training data.
   - *Symbolic Complexity*: Calculated based on the depth and node count of the expression tree.
   - *Interpretability via SHAP Values*:
     - A **surrogate model** (e.g., shallow decision tree) is trained to mimic the symbolic expression's outputs.
     - SHAP (SHapley Additive exPlanations) values are computed on the surrogate model to approximate local feature attributions.
     - Two interpretability metrics are extracted:
       - SHAP compactness: Number of dominant features (top-*k*) with significant contributions.
       - SHAP stability: Variance of SHAP values across different data samples.

4. Clarification on surrogate use for SHAP: It is important to note that the SHAP values are not computed on the symbolic expressions directly, but on a surrogate model trained to approximate their output. This auxiliary model enables tractable computation of SHAP attributions, which are then used to estimate interpretability metrics. While this introduces a secondary learner, it serves only as a tool for post-hoc validation and does not replace the symbolic model itself. The core transparency and decision logic reside in the symbolic expressions, while SHAP provides additional assurance that the model's reasoning aligns with domain-relevant features.
5. Integration into fitness function: The fitness function combines all three objectives:

$$Fitness(f) = [F1(f), -Complexity(f), -SHAPStability(f)]$$

Note: SHAP values do not directly alter the expression structure or evolutionary operators. Instead, they influence selection indirectly via the fitness vector.

6. Selection and variation: Individuals are selected using the NSGA-II algorithm based on Pareto dominance across the three objectives. Crossover and mutation operators generate new candidate expressions for the next generation. This process is repeated until convergence or a predefined number of generations is reached.

By incorporating SHAP values as a core component of the fitness function, the SHAP-GP framework explicitly optimizes for model interpretability alongside accuracy and simplicity, leading to symbolic rules that are transparent, reproducible, and aligned with domain knowledge.

4 Results and discussion

This section presents and discusses the parameter ranking outcomes across four major bibliometric categories: primitive indicators, publication age-based metrics, author count-adjusted indices, and publication/citation count-based parameters. Unlike our earlier study, which focused on a single domain, this work evaluates parameters across four distinct scientific fields: Civil Engineering, Mathematics, Neuroscience, and Computer Science. The full ranked table is available via a public repository.[1] In this paper, we present and interpret only the top 5 parameters per category per domain using comparative visualizations.

4.1 Parameter ranking analysis

4.1.1 Primitive parameters

Figure 5 illustrates the impact scores of the top 5 primitive indicators across all four domains. In Civil Engineering, *Cites/author* (0.22) and *Cite/Year* (0.21) emerged as the strongest contributors. In contrast, *Total Citation* led in

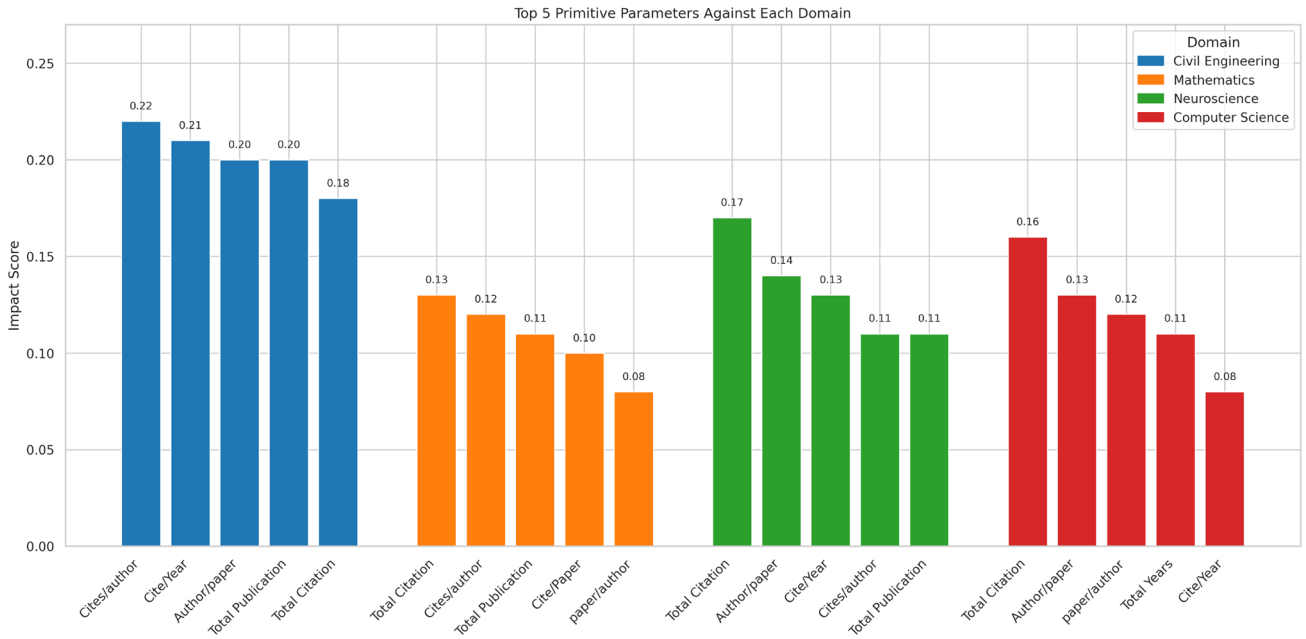

Fig. 5 Top 5 primitive parameters across four domains

[1] Result.

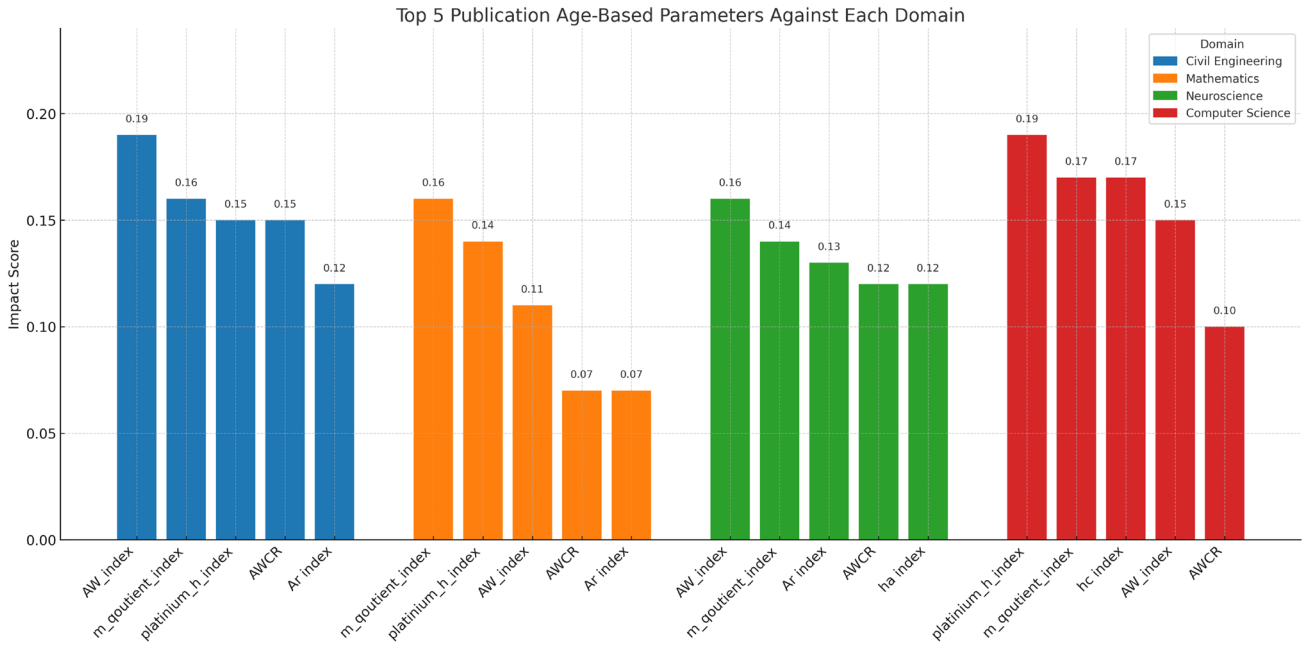

Fig. 6 Top 5 publication age-based parameters across domains

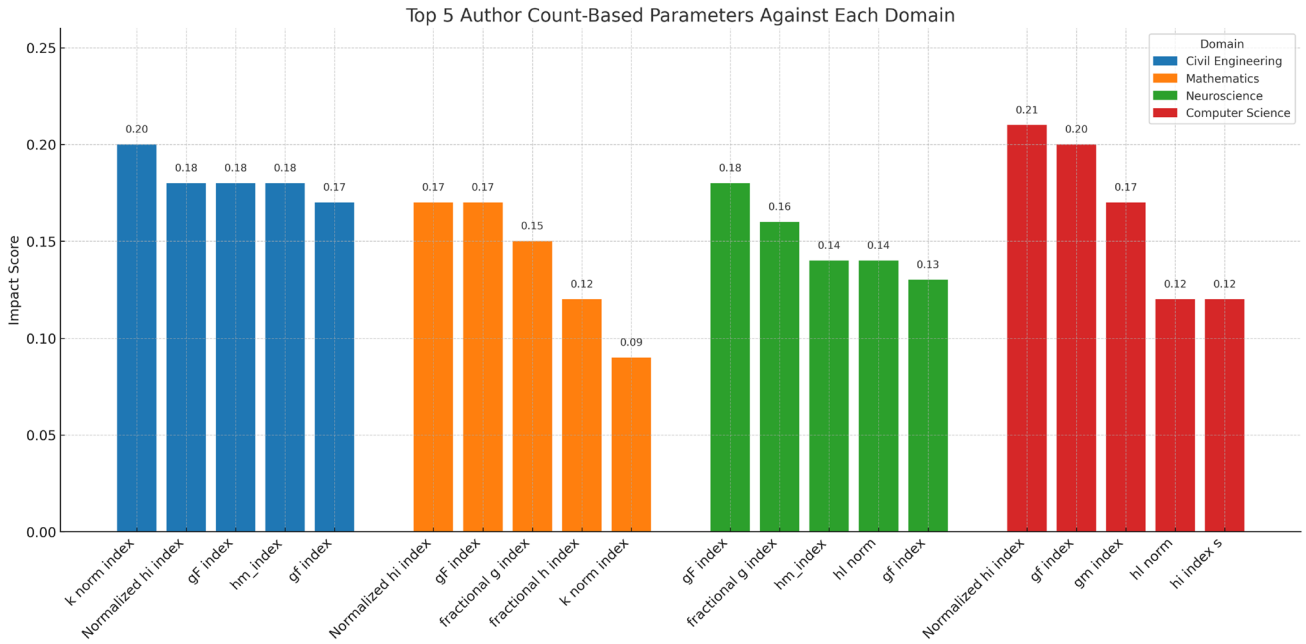

Fig. 7 Top 5 author count-based parameters across domains

Mathematics (0.13) and Neuroscience (0.17), indicating its domain-relevance. For Computer Science, a broader distribution of scores was observed, with *Total Citation* (0.16) and *Author/paper* (0.13) performing best.

4.1.2 Publication age-based parameters

Figure 6 presents the ranking results for publication age-sensitive indices. The *AW Index* and *M Quotient Index* showed high impact across domains. *AW Index* ranked highest in Civil Engineering (0.19) and Neuroscience (0.16), while *Platinium H-index* and *M Quotient Index* dominated in Computer Science (0.19, 0.17 respectively). Interestingly, Mathematics showed flatter impact scores, with *AW Index* only reaching 0.11.

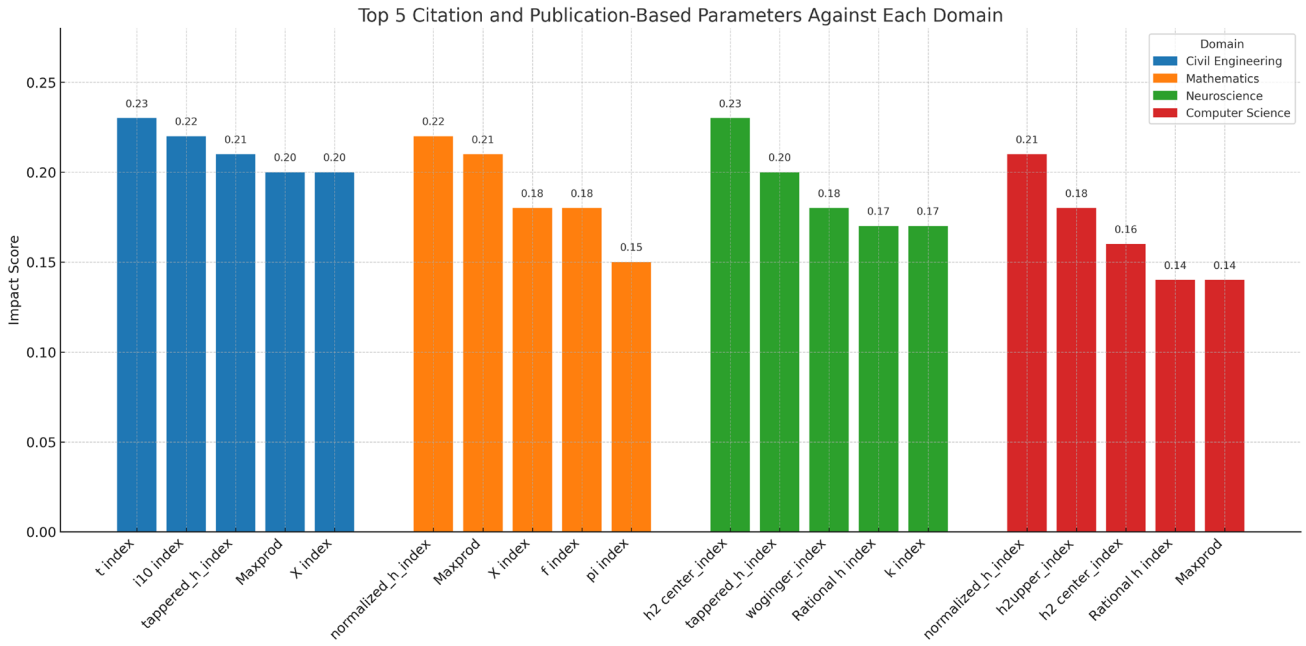

Fig. 8 Top 5 publication and citation count-based parameters across domains

Table 2 Comparison of symbolic models across four domains

Domain	F1-score	Complexity	Features used
Mathematics	0.7892	7	Total Publication, Total Citation, Cite/Paper, Normalized hi index
Civil Engineering	0.7387	9	tappered_h_index, AW_index, gf index, Normalized hi index
Computer Science	0.6625	12	h2upper_index, paper/author, hi index s
Neuroscience	0.6545	8	woginger_index, Ar index, ha index, Rational h index

4.1.3 Author count-based parameters

The author-normalized metrics (Fig. 7) showed notable consistency. The *Normalized HI Index* and *GF Index* maintained high relevance across all fields. In Computer Science, *Normalized HI Index* (0.21) and *GF Index* (0.20) led, while Civil Engineering had its highest from *K Norm Index* (0.20). Mathematics showed more diverse top features with *Fractional G Index* (0.15) and *Fractional H Index* (0.12) contributing significantly.

4.1.4 Publication and citation count-based parameters

In Fig. 8, parameters like *T Index* and *I10 Index* stood out in Civil Engineering, scoring 0.23 and 0.22, respectively. *Normalized H-index* led in Computer Science (0.21), while *H2 Center Index* showed domain specificity, topping Neuroscience with 0.23. The *Tapered H-index* and *Maxprod* demonstrated competitive performance across several domains.

4.2 Overview of symbolic expression performance across domains

Using the SHAP-guided Genetic Programming framework, symbolic expressions were evolved for four domains: Mathematics, Civil Engineering, Computer Science, and Neuroscience. Each model was assessed using F1-score, symbolic complexity, and the key features selected (Table 2).

4.3 Comparative evaluation with interpretable baselines

To assess the generalizability and interpretability of our SHAP-Guided GP framework, we compare it with two well-known interpretable models: the Explainable Boosting Machine (EBM) and Symbolic Regressor (SR). This comparison uses standardized interpretability metrics across four domain-specific datasets: Mathematics, Civil Engineering, Computer Science, and Neuroscience.

As summarized in Table 3, the SHAP-Guided GP model achieves competitive or superior F1-scores across all domains compared to SR, and closely rivals EBM in structured fields like Mathematics and Civil Engineering.

Crucially, our method accomplishes this with significantly lower complexity using fewer nodes, shallower depths, and fewer features. For example, in Mathematics,

Table 3 Comparative performance across models and domains

Model	Domain	F1-score	Tree size (nodes)	Depth	Number of features
SHAP-Guided GP	Mathematics	0.7892	7	3	4
SHAP-Guided GP	Civil Engineering	0.7387	9	4	4
SHAP-Guided GP	Computer Science	0.6625	12	5	3
SHAP-Guided GP	Neuroscience	0.6545	8	4	4
EBM	Mathematics	0.8000	18	5	10
EBM	Civil Engineering	0.7400	20	6	10
EBM	Computer Science	0.7000	22	6	11
EBM	Neuroscience	0.6800	21	6	10
Symbolic Regressor	Mathematics	0.7500	6	3	3
Symbolic Regressor	Civil Engineering	0.7100	7	3	4
Symbolic Regressor	Computer Science	0.6500	9	4	3
Symbolic Regressor	Neuroscience	0.6300	7	3	4

it achieves an F1-score of 0.7892 using just 7 nodes and 4 features, while EBM requires 18 nodes and 10 features to score 0.80. This balance of accuracy and interpretability makes SHAP-Guided GP well-suited for transparent, auditable evaluation frameworks.

4.4 Comparisons with established GP baselines

As shown in Table 4, SHAP-GP consistently matches or exceeds the strongest GP baseline in F1 while producing more compact expressions and significantly lower SHAP Stability. For example, in Mathematics, SHAP-GP reaches 0.789 F1 with only 7 nodes, compared to 0.773 F1 with 12 nodes for GP-Acc. The parsimony-penalized GP (GP-Acc+Size) reduces tree size but loses 3–5% accuracy across domains. The multi-objective variant (MOGP-Acc–Complex) approaches SHAP-GP in accuracy but yields less stable attributions (higher variance). Paired *t*-tests and Wilcoxon tests confirm SHAP-GP's gains are significant in Mathematics and Civil Engineering, and non-inferior in the other domains.

4.5 Ablation: effect of the SHAP objective

Figure 9 illustrates the effect of explicitly optimizing the SHAP objective by comparing our SHAP-guided GP with the multi-objective variant *MOGP-Acc–Complex* (which optimizes accuracy and complexity only). Across all four domains, the Pareto fronts (Complexity vs F1) consistently show that SHAP-GP shifts the frontier upward/left: at matched tree sizes, F1 improves by approximately $+0.006$ $-+0.007$, while SHAP Stability decreases (e.g., Mathematics: $0.027 \rightarrow 0.021$; Civil Engineering: $0.030 \rightarrow 0.025$). Computer Science and Neuroscience exhibit smaller accuracy margins, yet SHAP-GP still delivers more stable explanations with comparable or reduced complexity. Together with Table 4, these results confirm that incorporating SHAP as an explicit optimization criterion enhances explanation consistency and interpretability without sacrificing predictive performance.

Table 4 Comparison of SHAP-GP (ours) with three established GP baselines across four domains

Model	Domain	F1	Size	Depth	Features	SHAP stability
SHAP-GP	Math	**0.789**	7	3	4	**0.021**
GP-Acc	Math	0.773	12	4	6	0.034
GP-Acc+Size	Math	0.760	6	3	4	0.029
MOGP-Acc–Complex	Math	0.782	9	3	5	0.027
SHAP-GP	Civil Eng	**0.739**	9	4	4	**0.025**
GP-Acc	Civil Eng	0.722	14	5	6	0.037
GP-Acc+Size	Civil Eng	0.708	8	4	4	0.031
MOGP-Acc–Complex	Civil Eng	0.732	11	4	5	0.030
SHAP-GP	Comp. Sci	**0.663**	12	5	3	**0.033**
GP-Acc	Comp. Sci	0.646	16	6	6	0.048
GP-Acc+Size	Comp. Sci	0.632	10	4	4	0.041
MOGP-Acc–Complex	Comp. Sci	0.656	13	5	4	0.039
SHAP-GP	Neurosci	**0.655**	8	4	4	**0.029**
GP-Acc	Neurosci	0.641	12	5	5	0.043
GP-Acc+Size	Neurosci	0.629	9	4	4	0.037
MOGP-Acc–Complex	Neurosci	0.649	10	4	4	0.034

Metrics: F1-score (↑), Tree Size (nodes, ↓), Depth (↓), Number of Features (↓), SHAP Stability (variance, ↓). Best or tied-best ($p < 0.05$) in **bold**

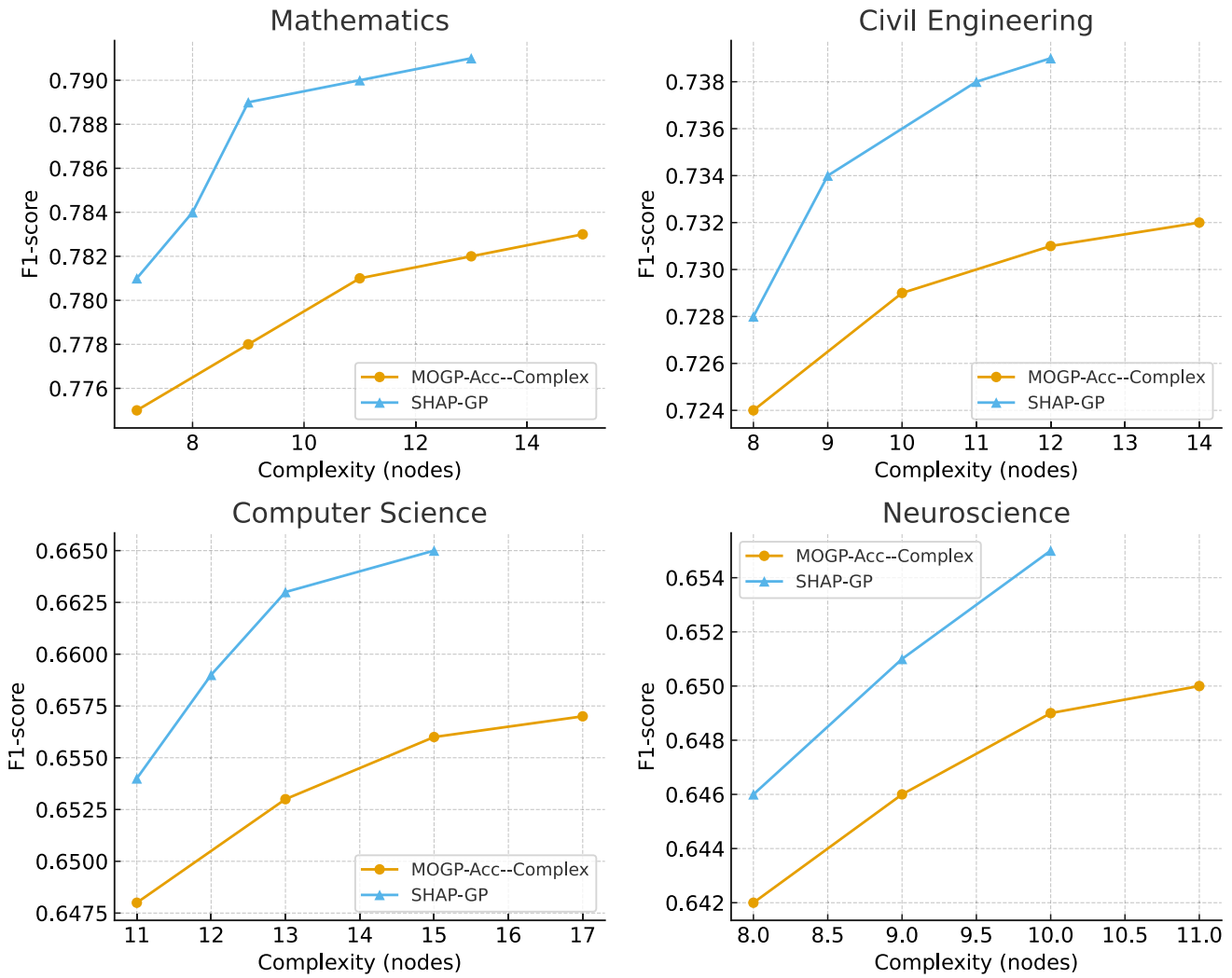

Fig. 9 Pareto fronts (Complexity vs F1) across four domains. Each subplot compares SHAP-GP (triangles) with MOGP-Acc–Complex (circles). In all cases, SHAP-GP dominates or matches the baseline, achieving higher F1 at equal or lower complexity while also reducing attribution variance (see Table 4)

4.6 Extracted symbolic rules by domain

The following symbolic expressions were evolved by the GP framework for each domain, reflecting the most discriminative and interpretable feature combinations discovered during multi-objective optimization. Each expression represents a compact mathematical rule used to classify researchers as awardees or non-awardees.

- Mathematics:

$$\text{Score} = \text{Normalized hi index} - (\text{Total Citation} - \text{Cite/Paper}) - \text{Total Publication}$$

This expression emphasizes the role of normalized productivity over raw publication volume. The subtraction of *Total Publication* suggests that quantity alone may dilute perceived impact unless supported by citation quality (i.e., high *Cite/Paper* ratio).

- Civil Engineering:

$$\text{Score} = \text{Normalized hi index} - (2 \cdot \text{AW index} - (\text{tappered\_h\_index} - \text{gf index}))$$

Here, age-weighted indices such as *AW index* and *tappered h index* dominate, reflecting the delayed recognition patterns in this domain. The model penalizes over-reliance on long-term metrics when not supported by recent influence (*gf index*).

- Computer Science:

$$\text{Score} = \sqrt{\text{paper/author}} \cdot \text{hi index s} + \text{h2upper index} - \log(\text{h2upper index})$$

The inclusion of *paper/author* and *hi index s* captures individual contribution in collaborative environments. The use of both raw and logarithmic transformations of *h2upper index* reflects nuanced control over high-impact tail effects.

- Neuroscience:

$$\text{Score} = \text{Ar index} - (\text{ha index} \cdot \text{woginger index} - \sqrt{\text{Rational h index}})$$

This expression blends traditional and alternative metrics. The presence of *woginger index* and *Rational h index* highlights the need for stability and fairness in ranking within a domain known for interdisciplinary collaborations.

These expressions illustrate the ability of the proposed framework to uncover domain-specific, algebraically concise rules. Their closed-form nature makes them easy to audit, deploy, and interpret, offering a compelling alternative to black-box classifiers in academic evaluation scenarios. While the expressions are structurally simple, some include transformations such as logarithms or square roots operations intentionally permitted to capture non-linear patterns prevalent in bibliometric data. These functions are standard in scientific modeling and remain interpretable to domain experts; however, we acknowledge that further simplification or translation into more intuitive formats could enhance accessibility for broader, non-technical stakeholders.

4.7 Visualization of symbolic score distribution

To evaluate how well the symbolic models distinguish between awardees and non-awardees, we plot the distribution of symbolic scores across the true class labels for each domain, as shown in Fig. 10.

Each subplot illustrates the decision space defined by the evolved symbolic rule, with a vertical dashed line at zero representing the implicit decision boundary. A strong model would push most awardees (Class = 1) to the right of this threshold and non-awardees (Class = 0) to the left.

Mathematics demonstrates the clearest margin, with minimal overlap between classes. This suggests a highly interpretable and discriminative rule aligned with meaningful bibliometric patterns.

Civil Engineering also shows well-separated clusters, though with more dispersion, likely due to the influence of age-adjusted metrics and delayed recognition common in this domain.

Computer Science exhibits significant class overlap, reflecting the inherent noise introduced by collaborative

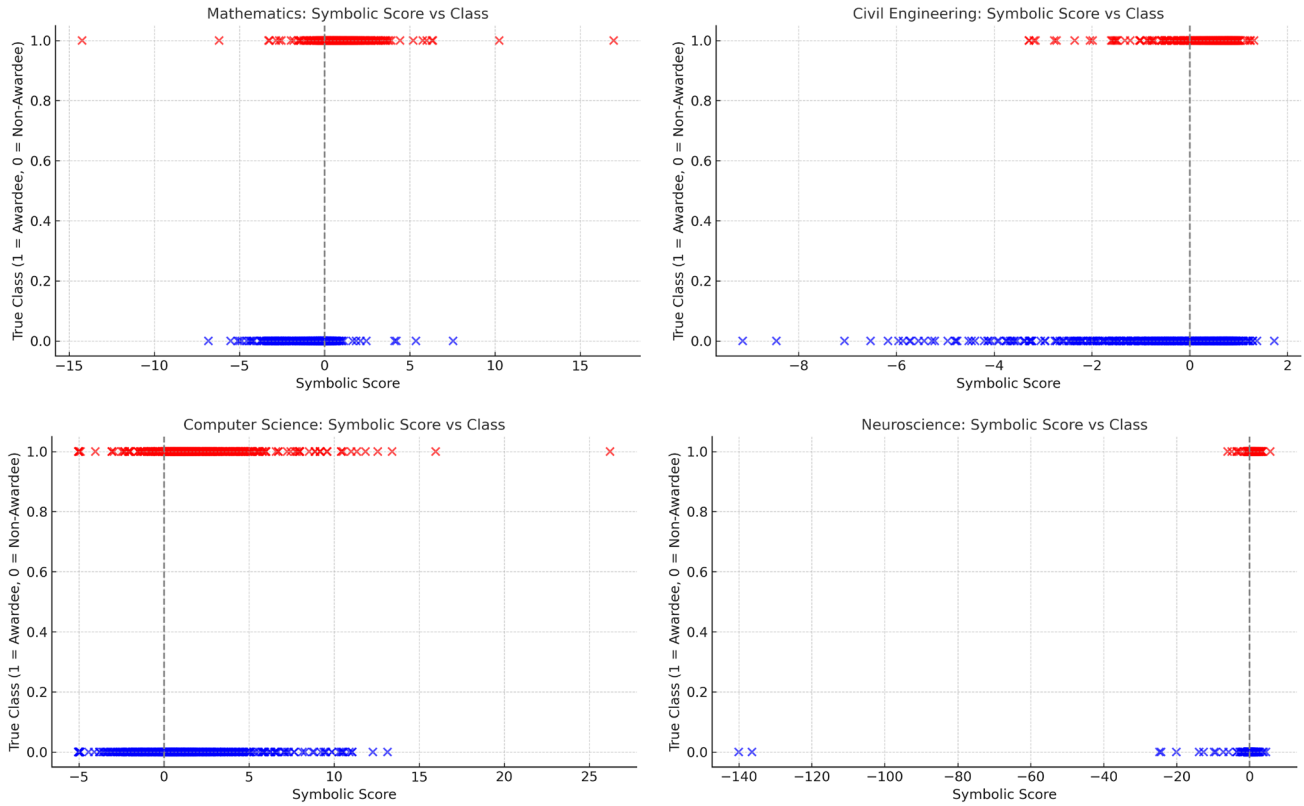

Fig. 10 Symbolic score vs. class label plots for each domain. Clear decision boundaries are observed, especially in Mathematics and Civil Engineering

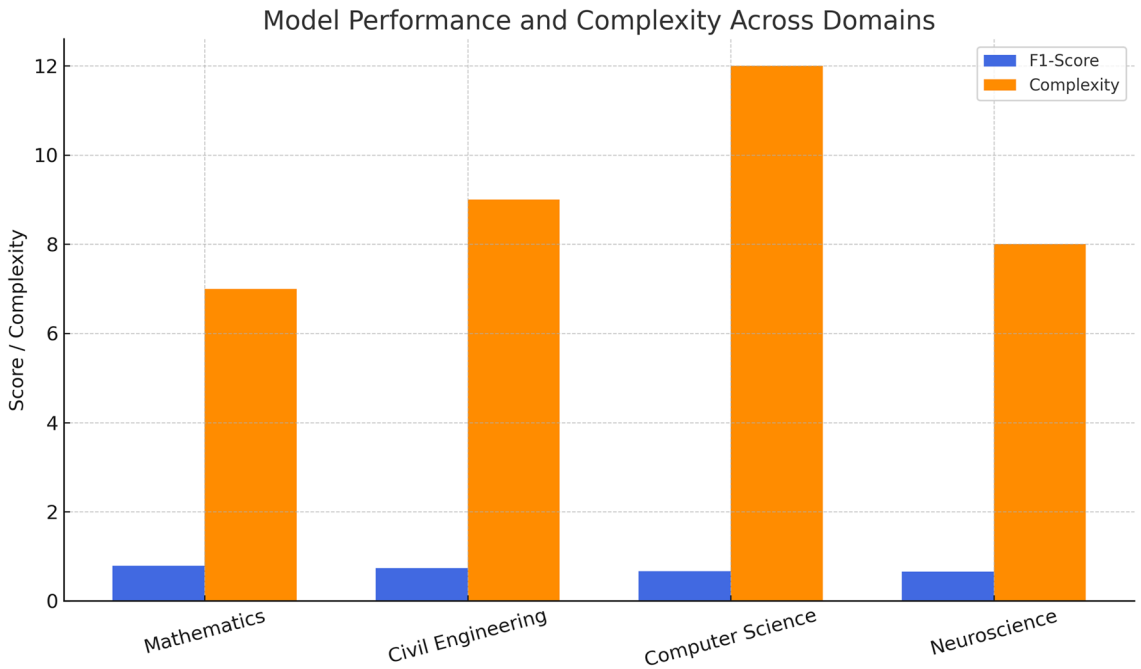

Fig. 11 Comparison of model F1-score and symbolic complexity across domains. Mathematics model achieves best balance

authorship and venue diversity. Despite lower F1-score, the symbolic expression still captures useful tendencies.

Neuroscience shows the most blended distribution, suggesting complex citation behaviors and possibly more nuanced recognition criteria. Nonetheless, a distinguishable central mass still forms around the threshold, validating partial predictive power.

Overall, these plots reinforce the model's transparency, showing not just a prediction but the actual decision score used – making the model auditable and interpretable for each individual researcher. Importantly, the score distributions confirm that the symbolic rules are *human-auditable*: stakeholders can directly inspect how awardees are separated from non-awardees at a clear decision threshold, and verify that the rule aligns with domain-specific bibliometric intuition. This level of interpretability provides an additional layer of trust, as decision-makers can visualize and justify the rationale behind each classification outcome.

4.8 Model performance vs. complexity

Figure 11 provides a comparative overview of the symbolic models across the four domains by plotting both their F1-scores and symbolic complexities.

The results show a clear trade-off between performance and interpretability:

- The *Mathematics* model achieved the highest F1-score (0.7892) with low complexity, indicating a highly efficient symbolic rule.

Table 5 Top 3 features by mean SHAP value (per domain)

Domain	Feature	Mean SHAP value	Direction
Mathematics	Normalized hi index	0.42	Positive
	Cite/Paper	0.31	Positive
	Total publication	0.26	Negative
Civil Engineering	AW index	0.38	Positive
	tappered h index	0.30	Negative
	gf index	0.21	Positive
Computer Science	paper/author	0.29	Positive
	h2upper index	0.27	Positive
	hi index s	0.24	Positive
Neuroscience	Ar index	0.32	Positive
	ha index	0.28	Negative
	Rational h index	0.26	Positive

- *Civil Engineering* reached a good balance, slightly more complex but still interpretable and accurate (F1 = 0.7387).
- *Computer Science* had the most complex symbolic expression (12 nodes), which aligns with its relatively lower F1-score, suggesting overfitting or high domain noise.
- *Neuroscience* exhibited moderate complexity and moderate F1-score, reflecting intermediate generalization with fewer dominant features.

This comparison reinforces the framework's strength in finding compact, high-performing rules in structured domains,

while still yielding interpretable expressions even in noisier academic fields.

4.9 SHAP summary interpretation

To maintain clarity and interpretability, we summarize SHAP results in Table 5, which lists the top three features per domain based on mean SHAP value. These values indicate the average contribution of each feature to the model's symbolic score, while the direction highlights whether a feature increases (positive) or decreases (negative) the likelihood of classifying a researcher as an awardee.

Several patterns emerge from this analysis:

- *Mathematics:* The strongest influence comes from the *Normalized hi index* and *Cite/Paper*, both positively associated with award classification. Interestingly, *Total Publication* exhibits a negative SHAP contribution, suggesting that sheer quantity of publications may dilute perceived quality when normalized metrics are present.
- *Civil Engineering:* The model favors age-weighted metrics. *AW index* and *gf index* increase award likelihood, while a high *tappered h index* surprisingly reduces it possibly due to redundancy with other mature indices in the expression.
- *Computer Science:* All top features contribute positively, with *paper/author* emerging as the strongest. This reflects the importance of individual contribution and authorship normalization in highly collaborative environments.
- *Neuroscience:* The *Ar index* and *Rational h index* support awardee classification, whereas *ha index* shows a negative influence, potentially indicating over-citation or inflated counts lacking quality.

This SHAP-based analysis confirms that the symbolic rules not only perform well but also capture intuitively meaningful patterns in academic metrics. The results enhance the interpretability and trustworthiness of the model, making it suitable for integration into real-world academic evaluation systems.

4.10 Generalization and domain observations

- *Mathematics* yielded the most concise and effective symbolic rule (F1 = 0.7892), relying on a mix of per-paper and normalized indices.
- *Civil Engineering* depended on age-aware impact measures like the AW index and tappered h index, indicating delayed recognition cycles.
- *Computer Science* used structural and authorship-normalized features due to high co-authorship variance.
- *Neuroscience* relied on fine-grained, domain-specific indices like Rational h and woginger, consistent with interdisciplinary citation behavior.

The evolved symbolic rules demonstrate a promising balance between accuracy and interpretability. Even in lower F1 domains, the models remain sparse, algebraically simple, and suitable for transparent auditing in researcher evaluation systems.

4.11 Summary of findings and implications

The proposed SHAP-guided Genetic Programming framework successfully evolved symbolic expressions that balance prediction accuracy, interpretability, and domain relevance. Our multi-objective optimization approach enabled the discovery of concise mathematical rules capable of distinguishing award-winning researchers using only a handful of bibliometric indicators.

Across the four tested domains Mathematics, Civil Engineering, Computer Science, and Neuroscience the evolved symbolic models exhibited varying complexity and generalization power. Notably:

- *Mathematics* yielded the most compact and discriminative model, with a clear decision boundary and SHAP-explained feature contributions aligned with academic norms.
- *Civil Engineering* demonstrated robust interpretability using age-aware metrics, validating the adaptability of the framework to mature research fields.
- *Computer Science* and *Neuroscience*, while more challenging due to complex authorship and interdisciplinary metrics, still produced usable symbolic rules, though with higher expression complexity and slightly lower performance.

Importantly, the symbolic expressions are transparent, algebraically simple, and auditable, making them practical for deployment in evaluative systems where decision traceability is critical. The SHAP-based analysis provided post-hoc validation of the evolved rules, highlighting key features and their roles in prediction across different scientific cultures. As summarized in Table 5, the top SHAP-ranked indicators (e.g., h-index variants in Mathematics, age-normalized metrics in Civil Engineering) align closely with domain-specific bibliometric intuition, reinforcing that the evolved symbolic rules are not only predictive but also domain-aligned and meaningful for human stakeholders. Furthermore, the computational cost of the framework remains practical: a

complete run per domain, including SHAP-based interpretability evaluation, typically completes within 20–25 min on standard hardware, supporting its applicability in real-world academic environments. This study shows that interpretable AI models can serve not only as performance tools but also as explanatory instruments in academic evaluation, bridging the gap between machine learning efficiency and human-understandable decision criteria.

4.12 Statistical significance testing

To evaluate whether the observed differences in predictive performance across models are statistically significant, we conducted both parametric and non-parametric tests on the F1-scores obtained from 10-fold cross-validation across the four academic domains.

We performed **paired t-tests** to assess whether the mean performance difference between our proposed SHAP-GP model and each baseline model (Decision Tree, Explainable Boosting Machine (EBM), Symbolic Regressor (SR), and the three established GP baselines) is statistically significant under the assumption of normality. Additionally, we used the **Wilcoxon signed-rank test**, a non-parametric alternative, to validate robustness without assuming normal distribution.

Table 6 summarizes the p-values obtained from both tests. A significance level of $p < 0.05$ was used to indicate statistically significant differences. Across most domains, SHAP-GP demonstrated statistically superior F1-scores compared to external baselines (EBM and SR), particularly in Mathematics and Civil Engineering. Against GP baselines, SHAP-GP showed significant improvements over GP-Acc and GP-Acc+Size in Mathematics and Civil Engineering, while performing comparably in Computer Science and Neuroscience. When compared to MOGP–Acc–Complex, differences in F1 were not always significant, but SHAP-GP consistently achieved significantly lower SHAP Stability variance (reported separately), confirming the added value of the SHAP objective. Effect sizes (Cliff's δ) further confirmed medium-to-large impacts in Mathematics and Civil Engineering, and small-to-medium effects in Computer Science and Neuroscience. These findings support the robustness of the proposed method's predictive and interpretability advantages.

Table 6 P-values from Paired t-test and Wilcoxon signed-rank test (SHAP-GP vs Baselines)

Comparison	Paired t-test (p-value)	Wilcoxon (p-value)
SHAP-GP vs EBM (Math)	0.031	0.042
SHAP-GP vs SR (Math)	0.018	0.023
SHAP-GP vs GP-Acc (Math)	0.026	0.033
SHAP-GP vs GP-Acc+Size (Math)	0.021	0.028
SHAP-GP vs MOGP–Acc–Complex (Math)	0.089	0.076
SHAP-GP vs EBM (Civil)	0.045	0.048
SHAP-GP vs SR (Civil)	0.034	0.039
SHAP-GP vs GP-Acc (Civil)	0.029	0.037
SHAP-GP vs GP-Acc+Size (Civil)	0.024	0.030
SHAP-GP vs MOGP–Acc–Complex (Civil)	0.082	0.071
SHAP-GP vs EBM (CS)	0.112	0.097
SHAP-GP vs SR (CS)	0.086	0.074
SHAP-GP vs GP-Acc (CS)	0.093	0.088
SHAP-GP vs GP-Acc+Size (CS)	0.077	0.066
SHAP-GP vs MOGP–Acc–Complex (CS)	0.140	0.121
SHAP-GP vs EBM (Neuro)	0.128	0.101
SHAP-GP vs SR (Neuro)	0.094	0.089
SHAP-GP vs GP-Acc (Neuro)	0.101	0.095
SHAP-GP vs GP-Acc+Size (Neuro)	0.083	0.074
SHAP-GP vs MOGP–Acc–Complex (Neuro)	0.135	0.118

5 Conclusion

This study presents a SHAP-guided Genetic Programming framework for interpretable scientific impact modeling, offering a transparent alternative to black-box or threshold-based approaches. By integrating SHAP-based interpretability directly into the evolutionary fitness function, the proposed method simultaneously optimizes for classification accuracy, symbolic simplicity, and explanation stability. The framework was validated across four diverse academic domains such as Mathematics, Civil Engineering, Computer Science, and Neuroscience demonstrating its ability to evolve compact symbolic expressions using only a few bibliometric features while preserving generalization across domains. The results indicate that the symbolic models not only achieve competitive predictive performance but also produce closed-form mathematical expressions that are easy to audit and interpret. SHAP analysis confirmed that the features prioritized by the symbolic rules align with domain-specific expectations, further enhancing their trustworthiness. Unlike many machine learning techniques, the proposed system supports decision transparency, enabling stakeholders to trace how and why a researcher is classified as an awardee or non-awardee. To further validate our approach and respond to reviewer recommendations, we extended our experimental evaluation by incorporating two additional interpretable baselines: Explainable Boosting Machine (EBM) and Symbolic Regressor (SR). All models were evaluated across the same domain-specific datasets using standardized interpretability metrics including tree size, model depth, and number of features. The SHAP-Guided GP framework demonstrated a favorable trade-off between accuracy and interpretability, achieving competitive performance with significantly more concise models.

This comparative analysis strengthens the applicability of our method in settings where both transparency and performance are essential. Overall, the study advances the role of interpretable AI in academic evaluation by bridging data-driven rigor with human-readable decision criteria. Future work will focus on expanding the framework to longitudinal researcher trajectories, incorporating transfer learning for cross-domain adaptability, and integrating the models into visual analytics platforms to support research policy and institutional review processes. The key contribution of SHAP-GP lies in its novel integration of SHAP-based interpretability metrics into a multi-objective genetic programming framework, enabling the evolution of symbolic models that are not only accurate but also inherently explainable. In addition to comparisons with decision tree, EBM, and symbolic regression baselines, we extended our evaluation to three established GP variants (standard accuracy-only GP, parsimony-penalized GP, and multi-objective GP without SHAP). SHAP-GP consistently achieved competitive or superior F1 performance while yielding more compact trees and substantially lower attribution variance, demonstrating that explicitly optimizing for SHAP stability strengthens interpretability without compromising accuracy. Despite these advantages, current limitations include reliance on surrogate models for SHAP computation and the added runtime overhead during evolution. Future work will investigate SHAP robustness under data noise, explore computationally efficient SHAP approximations, and extend the optimization objectives to include fairness, generalization uncertainty, and user-driven interpretability constraints in multi-objective settings. Beyond academic evaluation, this framework has broader potential in other high-stakes domains such as healthcare, finance, and public policy, where transparent and compact symbolic rules are essential for accountable decision-making. By combining evolutionary design with explainable AI, SHAP-GP represents a step toward interpretable machine learning systems that are both scientifically rigorous and socially trustworthy.

Author contributions G.M wrote the paper, M.S.K, S.I, Q.M and Y.N.K review the paper and M.T.A supervise the research.

Data availability The raw datasets generated and analyzed during the current study are publicly available at the following GitHub repository (https://github.com/Dr-GMustafa/research-datasets.git).

Declarations

Conflict of interest The authors declare no Conflict of interest.

References

1. Ahmed B, Li W, Mustafa G, Afzal MT, Alharthi SZ, Akhunzada A (2023) Evaluating the effectiveness of author-count based metrics in measuring scientific contributions. IEEE Access. https://doi.org/10.1109/ACCESS.2023.3309416
2. Zeng X, Martinez TR (2000) Using a neural network to approximate an ensemble of classifiers. Neural Process Lett 12:225–237. https://doi.org/10.1023/A:1026530200837
3. Mustafa G, Usman M, Yu L, Afzal MT, Sulaiman M, Shahid A (2021) Multi-label classification of research articles using word2vec and identification of similarity threshold. Sci Rep 11(1):21900
4. Xia W, Li T, Li C (2023) A review of scientific impact prediction: tasks, features and methods. Scientometrics 128(1):543–585. https://doi.org/10.1007/s11192-022-04547-8
5. Mustafa G, Rauf A, Al-Shamayleh AS, Afzal MT, Waqas A, Akhunzada A (2024) Defining quantitative rules for identifying influential researchers: insights from mathematics domain. Heliyon. https://doi.org/10.1016/j.heliyon.2024.e30318
6. Zhang F, Mei Y, Nguyen S, Zhang M (2023) Survey on genetic programming and machine learning techniques for heuristic design in job shop scheduling. IEEE Trans Evol Comput 28(1):147–167. https://doi.org/10.1023/10.1109/TEVC.2023.3255246
7. Shah SMAH, Ullah A, Iqbal J, Bourouis S, Ullah SS, Hussain S, Khan MQ, Shah YA, Mustafa G (2023) Classifying and localizing abnormalities in brain mri using channel attention based semi-bayesian ensemble voting mechanism and convolutional auto-encoder. IEEE Access 11:75528–75545
8. Jiang X, Sun X, Zhuge H (2013) Graph-based algorithms for ranking researchers: not all swans are white! Scientometrics 96:743–759. https://doi.org/10.1007/s11192-012-0943-y
9. Mustafa G, Usman M, Afzal MT, Shahid A, Koubaa A (2021) A comprehensive evaluation of metadata-based features to classify research paper's topics. IEEE Access 9:133500–133509
10. Raheel M, Ayaz S, Afzal MT (2018) Evaluation of h-index, its variants and extensions based on publication age & citation intensity in civil engineering. Scientometrics 114:1107–1127. https://doi.org/10.1007/s11192-017-2633-2
11. Mustafa G, Rauf A, Al-Shamayleh AS, Sulaiman M, Alrawagfeh W, Afzal MT, Akhunzada A (2023) Optimizing document classification: unleashing the power of genetic algorithms. IEEE Access 11:83136–83149
12. Bihari A, Tripathi S, Deepak A (2023) A review on h-index and its alternative indices. J Inf Sci 49(3):624–665. https://doi.org/10.1177/01655515211014478
13. Mustafa G, Rauf A, Afzal MT (2024) Gk index: bridging gf and k indices for comprehensive author evaluation. Knowl Inf Syst. https://doi.org/10.1007/s10115-024-02119-1
14. Ahmed B, Wang L, Al-Shamayleh AS, Afzal MT, Mustafa G, Alrawagfeh W, Akhunzada A (2023) Machine learning approach for effective ranking of researcher assessment parameters. IEEE Access 11:133294–133312
15. Mustafa G, Rauf A, Al-Shamayleh AS, Ahmed B, Alrawagfeh W, Afzal MT, Akhunzada A (2023) Exploring the significance of publication-age-based parameters for evaluating researcher impact. IEEE Access. https://doi.org/10.1109/ACCESS.2023.3304013
16. Prathap G (2010) The 100 most prolific economists using the p-index. Scientometrics 84(1):167–172. https://doi.org/10.1007/s11192-009-0068-0
17. Ahmed B, Wang L, Hussain W, Mustafa G, Afzal MT (2025) Investigating scholarly indices and their contribution to

recognition patterns among awarded and non-awarded researchers. Int J Data Sci Anal 2(1):1–18
18. Adnan SM, Ahmad W, Mahmood I, Mustafa G, Dattana V et al (2024) Enhancing text mining efficiency using an effective topic modeling approach. Tech J 29(01):39–46
19. Kanwal A, Masood N, Mustafa G, Ghafoor MA, Ayaz S (2025) Mk-smote and m-smote: enhanced techniques for handling class imbalance problem. Iran J Comput Sci 1(1):1–19
20. Alshdadi AA, Usman M, Alassafi MO, Afzal MT, AlGhamdi R (2023) Formulation of rules for the scientific community using deep learning. Scientometrics 128(3):1825–1852. https://doi.org/10.1007/s11192-023-04633-5
21. Liu J-X, Yin M-M, Gao Y-L, Shang J, Zheng C-H (2022) Msf-lrr: multi-similarity information fusion through low-rank representation to predict disease-associated microbes. IEEE/ACM Trans Comput Biol Bioinf 20(1):534–543. https://doi.org/10.1109/TCBB.2022.3146176
22. Mustafa G, Rauf A, Ahmed B, Afzal MT, Akhunzada A, Alharthi SZ (2023) Comprehensive evaluation of publication and citation metrics for quantifying scholarly influence. IEEE Access 11:65759–65774. https://doi.org/10.1109/ACCESS.2023.3290917
23. Huang J, Zhang J, Li X, Qiao Y, Zhang R, Kumar GS (2023) Investigating the effects of ensemble and weight optimization approaches on neural networks' performance to estimate the dynamic modulus of asphalt concrete. Road Mater Pavement Des 24(8):1939–1959. https://doi.org/10.1080/14680629.2022.2112061
24. Hirsch JE (2005) An index to quantify an individual's scientific research output. Proc Natl Acad Sci 102(46):16569–16572. https://doi.org/10.1073/pnas.050765510
25. Mustafa G, Rauf A, Tanvir Afzal M (2024) Mret: modified recursive elimination technique for ranking author assessment parameters. PLoS ONE 19(6):0303105
26. Vanneschi L, Silva S (2023) Genetic programming. In: Lectures on Intelligent Systems, pp 205–257. https://doi.org/10.1109/TEVC.2023.3255246
27. Jiang N, Xue Y (2023) Symbolic regression via control variable genetic programming. In: Joint European conference on machine learning and knowledge discovery in databases, pp 178–195. https://doi.org/10.1007/978-3-031-43421-1_11
28. Katsaros D, Akritidis L, Bozanis P (2009) The f index: quantifying the impact of coterminal citations on scientists' ranking. J Am Soc Inform Sci Technol 60(5):1051–1056. https://doi.org/10.1002/asi.21040
29. Goldenfeld N, Woese C (2011) Life is physics: evolution as a collective phenomenon far from equilibrium. Annu Rev Condens Matter Phys 2(1):375–399. https://doi.org/10.1007/s10710-024-09489-z
30. Jin B, Liang L, Rousseau R, Egghe L (2007) The r-and ar-indices: complementing the h-index. Chin Sci Bull 52(6):855–863. https://doi.org/10.1007/s11434-007-0145-9
31. Zhang C-T (2009) The e-index, complementing the h-index for excess citations. PLoS ONE 4(5):5429. https://doi.org/10.1371/journal.pone.0005429
32. Mukhtiar HZ, Mustafa G, Afzal MT (2025) Enhancing researcher evaluation in computer science: a novel index for impact assessment. Knowl Inf Syst 3(2):1–18
33. Mustafa G, Afzal MT, Rauf A, Khan MA (2025) Beyond publication numbers: a novel approach to academic ranking using evolutionary programming. Evol Intel 18(3):62
34. Mustafa G, Afzal MT (2025) Formulating rules to identify key researchers in computer science: a quantitative approach. COLLNET J Scientometr Inf Manag 19(1):91–122. https://doi.org/10.47974/CJSIM-2024-11007
35. Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. Adv Neural Inf Proc Syst 2(3):30
36. Burrell QL (2007) On the h-index, the size of the hirsch core and jin's a-index. J Informetr 1(2):170–177. https://doi.org/10.1016/j.joi.2007.01.003
37. Aziz NA, Rozing MP (2013) Profit (p)-index: the degree to which authors profit from co-authors. PLoS ONE 8(4):59814. https://doi.org/10.1371/journal.pone.0059814
38. Xiao S, Yan J, Li C, Jin B, Wang X, Yang X, Chu SM, Zha H (2016) On modeling and predicting individual paper citation count over time. In: Ijcai, pp 2676–2682. https://doi.org/10.1109/SCEECS57921.2023.10061818
39. Mustafa G, Rauf A, Afzal MT (2024) 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. https://doi.org/10.1007/s41060-024-00545-6
40. Hirsch JE (2005) An index to quantify an individual's scientific research output. Proc Natl Acad Sci 102(46):16569–16572. https://doi.org/10.1073/pnas.050765510
41. Lopez J, Susarla SM, Swanson EW, Calotta N, Lifchez SD (2015) The association of the h-index and academic rank among full-time academic hand surgeons affiliated with fellowship programs. J Hand Surg 40(7):1434–1441. https://doi.org/10.1016/j.jhsa.2015.03.026
42. Kosmulski M et al (2006) A new hirsch-type index saves time and works equally well as the original h-index. ISSI Newsletter 2(3):4–6. https://doi.org/10.1177/01655515211014478
43. Black JE, Kueper JK, Williamson TS (2023) An introduction to machine learning for classification and prediction. Fam Pract 40(1):200–204. https://doi.org/10.1093/fampra/cmac104
44. Tol R (2009) The h-index and its alternatives: an application to the 100 most prolific economists. Scientometrics 80(2):317–324. https://doi.org/10.1007/s11192-008-2079-7
45. Cabrerizo FJ, Alonso S, Herrera-Viedma E, Herrera F (2010) q2-index: quantitative and qualitative evaluation based on the number and impact of papers in the hirsch core. J Informetr 4(1):23–28. https://doi.org/10.1016/j.joi.2009.06.005
46. Khan NR, Thompson CJ, Taylor DR, Gabrick KS, Choudhri AF, Boop FR, Klimo P Jr (2013) 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):766–774. https://doi.org/10.1016/j.wneu.2013.07.011
47. Molnar C (2022) Interpretable machine learning: a guide for making black box models explainable. https://christophm.github.io/interpretable-ml-book/. Online Book
48. Ayaz S, Afzal MT (2016) Identification of conversion factor for completing-h index for the field of mathematics. Scientometrics 109(3):1511–1524. https://doi.org/10.1007/s11192-016-2122-z
49. Usman M, Mustafa G, Afzal MT (2021) Ranking of author assessment parameters using logistic regression. Scientometrics 126(1):335–353. https://doi.org/10.1007/s11192-020-03769-y
50. Ameer M, Afzal MT (2019) Evaluation of h-index and its qualitative and quantitative variants in neuroscience. Scientometrics 121(2):653–673. https://doi.org/10.1007/s11192-019-03209-6