computation

MDPI

Article

Interpretable Machine Learning for Research Evaluation: Transparent Bibliometric Rules and the Limits of Association with Institutional Recognition in Neuroscience

Ghulam Mustafa [1,*,†], Muhammad Tanvir Afzal [2,†], Ahmad Sami Al-Shamayleh [3], Hammad Abbasi [1,†], Isabel De la Torre Díez [4], Mirtha Silvana Garat de Marin [5,6,7,8] and Eduardo Silva Alvarado [5,9,10]

1 Faculty of Computing, Shifa Tameer-e-Millat University, Islamabad 44000, Pakistan; hammad.ssc@stmu.edu.pk
2 Department of Computing and Data Science, GISMA University of Applied Sciences, 14469 Potsdam, Germany; tanvir.afzal@gisma.com
3 Department of Data Science and Artificial Intelligence, Faculty of Information Technology, Al-Ahliyya Amman University, Amman 19111, Jordan; a.alshamayleh@ammanu.edu.jo
4 eHealth and Telemedicine Group, University of Valladolid, 47011 Valladolid, Spain; isator@uva.es
5 Department of Projects, Escuela Politécnica Superior, Universidad Europea del Atlántico, Isabel Torres 21, 39011 Santander, Spain; silvana.marin@unic.co.ao (M.S.G.d.M.) ; eduardo.silva@uneatlantico.es (E.S.A.)
6 Department of Projects, Universidad Internacional Iberoamericana, Arecibo, PR 00613, USA
7 Department of Projects, Universidade Internacional do Cuanza, Cuito EN250, Bié Province, Angola
8 Department of Projects, Fundación Universitaria Internacional de Colombia, Bogotá 111321, Colombia
9 Department of Projects, Universidad Internacional Iberoamericana, Campeche 24560, Mexico
10 Universidad de La Romana, La Romana 22000, Dominican Republic
* Correspondence: ghulam.mustafa.ssc@stmu.edu.pk
† These authors contributed equally to this work.

Abstract

The rapid growth of the scientific literature has intensified interest in how bibliometric information relates to scientific recognition, particularly in interdisciplinary fields such as neuroscience. Existing studies often rely on isolated indicators or black-box models. Here, the outcome is defined narrowly as institutionally recognized influence, operationalized as receipt of a prestigious neuroscience award from one of five societies (SFN, IBRO, FENS, CNS, ANS), rather than scientific influence in general. We compute 64 author assessment parameters in four categories for a balanced dataset of 1060 researchers (530 awardees, 530 non-awardees), using publications dated before the award year. A multilayer perceptron with leave-one-out feature ablation (MLP-FIS) ranks the parameters, and the top five per category feed Gini-based decision trees from which human-readable rules are extracted. On a random 20% test split the highlighted rules capture 33–49% of awardees at 83–95% specificity. Under repeated cross-validation with in-fold feature ranking, trees reach 66–70% accuracy against 68–73% for random forests, and the MLP-FIS top-five sets were unstable across folds. On a temporal hold-out of 106 researchers, rule accuracy was 52–67%. The benchmark results show limited incremental predictive advantage over strong conventional alternatives; accordingly, the contribution is interpretive rather than performance-based. The rules are descriptive benchmarks of bibliometric patterns associated with institutional recognition, not validated tools for judging scientific influence or individuals, and the analysis is subject to citation-timing leakage, award-selection bias and data-quality problems.

Check for updates

Academic Editor: Ali Cemal Benim

Received: 9 October 2026
Revised: 11 November 2026
Accepted: 12 November 2026
Published: 19 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: neuroscience; bibliometrics; researcher evaluation; explainable AI; decision trees; rule mining; h-index variants; scientometrics; multilayer perceptron; feature importance scoring; award-based benchmarking; author assessment parameters; citation analysis; interpretable machine learning; scholarly impact

1. Introduction

The exponential growth of scientific publications over the past few decades has profoundly transformed the global research landscape [1]. Advances in digital publishing, open-access initiatives, and interdisciplinary collaboration have resulted in an unprecedented volume of scholarly output across nearly all scientific domains [2]. While this expansion reflects the vitality of modern research ecosystems, it has simultaneously complicated the task of identifying researchers whose contributions have had sustained and meaningful scientific impact [3]. Formal recognition mechanisms, such as prestigious awards and fellowships, often lag behind actual scholarly influence, creating a temporal disconnect between measurable research performance and institutional acknowledgment [4].

This challenge is particularly pronounced in the field of neuroscience. As an inherently interdisciplinary domain, neuroscience integrates methodologies from biology, medicine, psychology, engineering, and computational sciences. Consequently, research trajectories, publication practices, and citation dynamics vary substantially across its subfields [3,5]. Influential contributions may accumulate recognition unevenly over time, and citation patterns differ between experimental, clinical, and theoretical research streams. These characteristics make the identification of impactful neuroscientists using simple quantitative indicators both non-trivial and susceptible to bias [6,7].

To facilitate large-scale evaluation, bibliometric platforms such as Google Scholar, Semantic Scholar, and Web of Science provide extensive author-level metadata, including publication counts, citation totals, and derivative indices such as the h-index [8]. These indicators have enabled quantitative assessments of scholarly impact and are widely used in academic evaluations [9]. However, reliance on individual metrics remains problematic. Publication counts may emphasize quantity over quality, while citation-based measures are influenced by self-citation practices, field-dependent citation norms, and collaboration patterns [6]. As a result, no single metric adequately captures the multidimensional nature of scientific influence [3,10].

The introduction of the h-index represented a significant advancement by attempting to balance productivity and citation impact within a single measure [8]. Despite its widespread adoption, numerous limitations have been documented. Additional citations to highly cited papers do not necessarily increase the index value, moderately cited yet influential work may be undervalued, and early-career researchers are systematically disadvantaged due to citation accrual delays [11]. These shortcomings have motivated the development of a wide range of alternative and complementary indices, including citation-intensity measures, age-adjusted metrics, and author-count–normalized indicators [7].

In response, several studies have conducted comparative evaluations of bibliometric indicators across disciplines and methodological settings [5,7,9,11]. While these efforts provide valuable insights, two critical gaps remain. First, many studies evaluate a limited subset of available parameters or rely on historical award data that predates the introduction of several modern indices. Second, most approaches emphasize predictive performance without yielding transparent and interpretable decision rules that can directly support researchers and evaluation committees.

The absence of interpretable frameworks is particularly important for neuroscience because opaque predictive models reveal little about which bibliometric patterns are associated with recognition or about the limits of those patterns. Although machine learning techniques have increasingly been applied to bibliometric analysis, many models operate as black boxes and provide limited insight into why particular profiles are classified as award-associated [9]. The present study therefore treats interpretability as an analytical objective rather than as a justification for automated screening.

In this study, we propose an explainable, rule-based framework for examining how bibliometric profiles relate to institutionally recognized influence in neuroscience. We compute a comprehensive set of 64 author assessment parameters organized into four categories: (i) primitive parameters, (ii) publication and citation count-based parameters, (iii) publication age-based parameters, and (iv) author count-based parameters. To rank the importance of these parameters, we employ a Multilayer Perceptron combined with a leave-one-out feature ablation strategy, referred to as *MLP-based Feature Importance Scoring (MLP-FIS)*. From each category, the top five parameters are selected and used to induce human-interpretable decision rules via a Gini impurity-based Decision Tree with post-pruning. The objective is not to claim superior prediction of scientific influence, but to determine what transparent multivariate patterns can be learned about award recognition and where the associations captured by bibliometric information reach their limits.

The proposed framework is evaluated on a balanced neuroscience dataset comprising 1060 researchers, including 530 award recipients and 530 matched non-awardees, drawn from five recognized neuroscience societies: the Society for Neuroscience (SFN), the International Brain Research Organization (IBRO), the Federation of European Neuroscience Societies (FENS), the Cognitive Neuroscience Society (CNS), and the Australasian Neuroscience Society (ANS). Rather than replacing expert judgment, this study investigates whether interpretable rules derived from bibliometric data can approximate recognition patterns observed in award decisions. The resulting rules are used to characterize combinations of bibliometric features associated with recognition and, equally importantly, to assess how much of that institutional outcome remains unexplained by citation-based profiles. They are therefore presented as descriptive evidence for Computational Social Science rather than as practical criteria for identifying influential researchers or screening individuals.

Scientific influence is a multidimensional construct. It can refer to recognition by peers and institutions, to intellectual contributions that change how a field thinks, to the diffusion of ideas through citation and reuse, to the building of research communities and trainee networks, and to the capacity to shape later research trajectories [2,3,12]. Bibliometric indicators capture only some of these dimensions, mainly diffusion (citations), productivity, longevity and collaboration, and they say little about mentoring, community building or the originality of ideas [10]. We therefore do not claim to measure scientific influence in general. Following the scientometric practice of using prizes as an external benchmark, we operationalize *institutionally recognized influence* as receipt of a prestigious neuroscience award, and we ask which bibliometric profiles are associated with that form of recognition.

The proposed pipeline is not a new bibliometric index. Its purpose is to map the large and redundant family of existing indices onto a recognition outcome, to find a compact subset that carries most of the discriminative signal, and to express that subset as thresholds that can be inspected, criticized and compared with committee practice. This matters for two reasons. Committees and early-career researchers already use h-index variants, and a transparent account of which variants and which combinations align with past recognition is more useful than another opaque score. In addition, such an account bears on a sociological question: whether recognition in a field follows citation volume, author-

normalized contribution or sustained activity, and how strongly any bibliometric profile separates recognized researchers from comparable non-recognized ones [13–15].

We address three research questions. **RQ1:** Which of the four parameter categories, and which individual indicators within them, are most discriminative for award recipients versus matched non-recipients? **RQ2:** What additional descriptive knowledge is provided by compact, human-readable multivariate rules beyond individual indicators and feature rankings, and how well do those rules hold for awardees from a later period? **RQ3:** To what extent are bibliometric profiles associated with institutional award recognition when interpretable trees are benchmarked against strong individual indicators and standard classifiers? Because the ground truth is award receipt, the answers concern the bibliometric correlates and limits of institutional recognition, not scientific merit or scientific influence in general.

The contributions are: (i) a category-structured set of 64 indicators computed for 1060 neuroscientists; (ii) an MLP-based ablation ranking within and across categories, with a Random Forest cross-check; (iii) transparent threshold rules that expose conditional combinations of bibliometric characteristics associated with award recognition rather than merely ranking variables one at a time; (iv) temporal hold-out and benchmark analyses that show these multivariate rules do not consistently outperform strong individual indicators or standard classifiers; and (v) an empirical account of the resulting substantive implication—bibliometric profiles capture an important dimension of citation-based visibility but are only imperfectly associated with institutional recognition. This positioning makes the limits of prediction part of the contribution rather than treating them only as a methodological shortcoming.

The methodological originality of this study therefore does not rest on predictive superiority or on a fundamentally new learning algorithm. It lies in using a common empirical design to connect three forms of evidence that are often reported separately: individual indicator performance, multivariate feature ranking, and explicit conditional rules. Feature rankings show which variables are individually influential to a model, whereas the decision rules reveal inspectable conjunctions and ranges of indicators under which recognition-associated profiles occur. The benchmark results are essential to this interpretation: because MLP-FIS and the rule-based models do not consistently outperform conventional methods or the strongest individual indicators, the value of the framework is primarily interpretive and comparative. It makes visible both the combinations associated with historical recognition and the limited incremental information gained from increasingly complex bibliometric profiles.

The results suggest that recognition-related signal is spread over several indicator families and not concentrated in one indicator. Citation-based, age-adjusted and author-normalized measures all appear among the highest-ranked parameters, but their discriminative power is modest: on a temporally held-out set, rule accuracy was between 52% and 67%, and only the publication and citation count-based rules were clearly above chance (Section 5.7). The extracted rules should be read as descriptive profiles of recognized researchers, shaped by the cumulative-advantage and selection effects discussed in Section 2, and not as criteria of scientific merit.

The remainder of this paper is organized as follows. Section 2 reviews related work and the conceptual foundations of scientific influence. Section 3 describes the methodology: dataset construction, parameter computation, feature ranking, rule mining and the evaluation metrics. Section 4 reports the parameter rankings. Section 5 presents the extracted rules, their coverage and specificity, the held-out validation and the baseline comparison. Section 6 analyzes how awardees are distributed in top-100 lists. Section 7 discusses the

findings, limitations and implications for responsible use. Section 8 concludes and Section 9 outlines future work.

2. Literature Review

The evaluation of scientific impact has long been a central topic within scientometrics and information science, evolving from simple quantitative indicators toward increasingly sophisticated multi-metric and machine learning-based frameworks. Despite this progress, several fundamental issues remain unresolved: the validity of bibliometric proxies as surrogates for genuine scholarly influence, the redundancy among the proliferating family of h-index variants, and the extent to which recognition mechanisms such as awards reflect merit rather than institutional or social factors. This review critically examines these issues and positions the proposed framework within the existing landscape.

2.1. Scientific Influence: Conceptual Foundations

Scientific influence has no single agreed operational definition. The literature distinguishes at least five aspects: scholarly recognition by peers and institutions, intellectual contribution to a field, diffusion of ideas through citation and reuse, the development of research communities, and the capacity to shape later research trajectories [2,3,12]. Quantitative indicators address these aspects unevenly. Citation counts and h-index variants mainly reflect diffusion, age-adjusted indices add a temporal dimension, author-normalized indices attempt to isolate individual contribution within teams, and primitive indicators describe raw productivity. Community building, mentoring and the originality of an idea are largely invisible to them [10].

Existing indicators are insufficient for two reasons that motivate the present study. First, they are numerous and strongly correlated, so that committees cannot easily tell which of them add information. Second, they are rarely related to an external recognition outcome in a way that yields explicit, inspectable decision thresholds. The aim here is therefore not another index but a transparent mapping from existing indices to institutionally recognized influence, the one aspect of influence for which a defensible external benchmark (prestigious awards) is available.

We evaluate this mapping against three benchmarks: (i) chance performance on a balanced sample (50% accuracy), (ii) single-index rankings, and (iii) standard classifiers trained on the same features. An improvement over these benchmarks is meaningful only with respect to recognition, and we make no claim about intrinsic scientific quality.

2.2. From Simple Metrics to H-Index Variants: Progress and Persistent Limitations

Early approaches to researcher assessment relied on total publication count, total citation count, and journal impact factors. Although widely available and easy to compute, these measures have been consistently criticized for rewarding quantity over quality, for susceptibility to strategic manipulation through self-citation, and for failing to capture the multidimensional nature of scholarly contribution [6,7]. A critical limitation rarely acknowledged in the literature is that citation counts are field-dependent: neuroscience, as a highly collaborative and interdisciplinary domain, exhibits systematically higher citation rates than mathematics or engineering, rendering cross-domain comparisons fundamentally problematic without normalization [5].

The introduction of the h-index represented a significant methodological advance by combining productivity and citation impact into a single robust measure [8]. Its rapid institutional adoption, however, preceded a thorough empirical validation of its properties. Subsequent studies exposed several structural limitations: the h-index fails to reward additional citations beyond the h-core, systematically disadvantages early-career researchers

due to citation accrual delays, and may over represent researchers who have become inactive but whose earlier publications continue accumulating citations [7,11]. Critically, the h-index is insensitive to highly cited outlier papers a property that may be particularly consequential in neuroscience, where landmark discoveries often generate citation distributions with heavy tails.

2.3. Proliferation of Variants: Redundancy and Fragmentation

In response to h-index limitations, a large number of alternative indices have been proposed, including citation-intensity measures (g-index, R-index), age-adjusted metrics (AR-index, m-quotient), and author-count normalized indicators (hI-index, fractional h-index) [7]. While each variant targets a specific weakness, their rapid proliferation has introduced a new problem: substantial empirical redundancy. Several studies have demonstrated high mutual correlations among h-index variants, raising the question of whether they capture meaningfully distinct dimensions of scholarly impact or merely reformulate the same underlying signal [5,7]. This redundancy has important practical implications: evaluation committees adopting multiple correlated indices may gain a false sense of comprehensiveness without genuine additional discriminative power.

A further underexplored issue concerns the conceptual validity of these indicators. Most bibliometric indices measure visibility and citedness rather than intrinsic scientific quality. High citation counts may reflect methodological accessibility, controversy, or replication rather than foundational contribution. This distinction between impact as influence and impact as quality is rarely made explicit in the bibliometric literature, and it applies directly to neuroscience, where highly cited papers span a wide spectrum from transformative discoveries to widely used but methodologically routine protocols.

2.4. Award-Based Benchmarking: Validity and Social Construction

A recurring methodological choice in bibliometric evaluation, including the present study, is the use of award receipt as a proxy for scientific influence. Recognition-oriented and domain-specific bibliometric evaluations, including work in neuroscience [5,16] and broader research-evaluation frameworks [3,10], show that bibliometric indicators can be informative while remaining sensitive to disciplinary and institutional context. However, this approach carries an important validity assumption that deserves critical scrutiny: awards reflect the judgments of specific committees within specific institutional and cultural contexts. Scientific recognition is not purely meritocratic it is shaped by visibility, network effects, geographic representation, and the prestige of nominating institutions [7].

A well-established theoretical account of this dynamic is Merton's concept of cumulative advantage, widely known as the Matthew effect [13]: researchers who accumulate early recognition attract disproportionately more citations, collaborators, and further awards, compounding their visibility independently of subsequent contribution quality. In bibliometric terms, this means that citation counts and derived indices may partly reflect accumulated prestige rather than ongoing scientific merit [17]. Complementary work on evaluation bias has shown that peer judgment in award and grant committees is shaped by institutional affiliation, gender, and network proximity, introducing systematic distortions into recognition outcomes [18,19]. Taken together, these frameworks imply that a predictive model trained on award receipt as its ground truth is, at least in part, learning to replicate the structural biases embedded in those awards rather than detecting scientific merit per se. Consequently, a framework that predicts award receipt is more precisely characterized as modeling the bibliometric correlates of institutionally recognized influence, rather than scientific impact per se. We explicitly acknowledge this distinction and adopt award status as

a pragmatic and externally validated benchmark consistent with established scientometric practice, while recognizing its inherent limitations as a ground truth.

2.5. Machine Learning Approaches: Performance Gains at the Cost of Interpretability

More recently, machine learning techniques have been applied to rank bibliometric parameters and predict scholarly recognition. Logistic regression, random forests, neural networks, and ensemble approaches have demonstrated improved predictive performance over single-metric evaluations [9]. However, a critical limitation of these studies is their emphasis on predictive accuracy at the expense of interpretability. Black-box models such as random forests and deep neural networks yield feature importance scores but do not necessarily expose explicit conditional relationships among indicators that can be inspected and critically evaluated. Moreover, existing machine learning studies typically evaluate parameters from a single category or a mixed but unstructured feature set, without systematically analyzing the relative contribution of distinct metric categories (primitive, citation-based, age-adjusted, author-normalized). This limits the theoretical insight these studies can provide about which dimensions of scholarly activity are most predictive of recognition.

2.6. Rule-Based and Interpretable Frameworks: Scope and Limitations

Decision tree-based frameworks have been proposed as interpretable alternatives to black-box models, producing human-readable rules that explicitly describe classification conditions [20]. However, existing rule-based studies suffer from two limitations that the present work directly addresses. First, they typically operate on a restricted and unstructured subset of bibliometric parameters, without category-aware feature selection. Second, they lack systematic validation of rule coverage and specificity, making it difficult to assess whether reported rules reflect genuine discriminative patterns or overfit to idiosyncratic dataset characteristics.

2.7. Research Gap and Theoretical Positioning

The literature comparison presented in Table 1 reveals four converging gaps that motivate the present work. First, to our knowledge, no existing study evaluates a comprehensive and category-structured set of bibliometric parameters within the neuroscience domain simultaneously. Second, the redundancy among bibliometric indicators has been identified but not systematically examined through category-aware feature selection and common benchmarks. Third, predictive studies often provide limited transparency about the multivariate conditions associated with recognition, while interpretable studies do not always establish whether their added complexity yields information beyond strong individual indicators. Fourth, existing rule-based frameworks often lack rigorous validation of coverage, specificity, temporal transfer, and comparative performance. These gaps motivate a study that evaluates both what transparent bibliometric rules reveal and how much descriptive or predictive value they add.

The present study addresses these gaps by proposing a framework that is simultaneously comprehensive (64 parameters across four categories), interpretable (human-readable decision rules with explicit thresholds), and validated on a temporal hold-out of recent awardees (with the limitations discussed in Section 7.1). Crucially, the framework does not claim to measure scientific quality directly; rather, it models the bibliometric correlates of institutionally recognized influence in neuroscience correlates that are themselves shaped by the cumulative advantage dynamics and evaluation biases described above [13,18]. The extracted rules should therefore be interpreted as benchmarks for the kind of bibliometric profile that award committees have historically recognized, rather than as objective criteria for scientific merit. Used with this awareness, the framework offers transparent and

actionable guidance for researchers and evaluation committees while acknowledging the structural limitations of award-based ground truth.

Table 1. Comparison of representative bibliometric studies for researcher evaluation.

Study	Domain	Parameter Scope	Methodology	Interpretable Rules
Hirsch [8]	General	Single metric	h-index formulation	No
Bornmann & Daniel [7]	General	Limited	Index comparison	No
Dienes [11]	General	Limited	Empirical evaluation	No
Schreiber [5]	Neuroscience	Moderate	Index-based analysis	No
Usman et al. [9]	Researcher evaluation	Moderate	Logistic regression/ML	No
Mustafa et al. [21]	Mathematics	Comprehensive (64)	MLP ranking + decision-tree rules	Yes
Alshdadi et al. [20]	General	Limited	Decision tree rules	Yes (restricted)
This study	Neuroscience	Comprehensive (64)	MLP-FIS + decision trees	Yes

3. Methodology

This section describes the proposed methodology for examining bibliometric patterns associated with institutional award recognition in neuroscience using an explainable, rule-based framework. The overall workflow, illustrated in Figure 1, consists of five major stages: (i) dataset collection and preprocessing, (ii) computation of author assessment parameters, (iii) ranking of parameters using a machine learning-based feature selection approach, (iv) prediction using top-ranked parameters, and (v) rule mining through interpretable decision tree models. Each stage is designed to ensure robustness, transparency, and domain relevance.

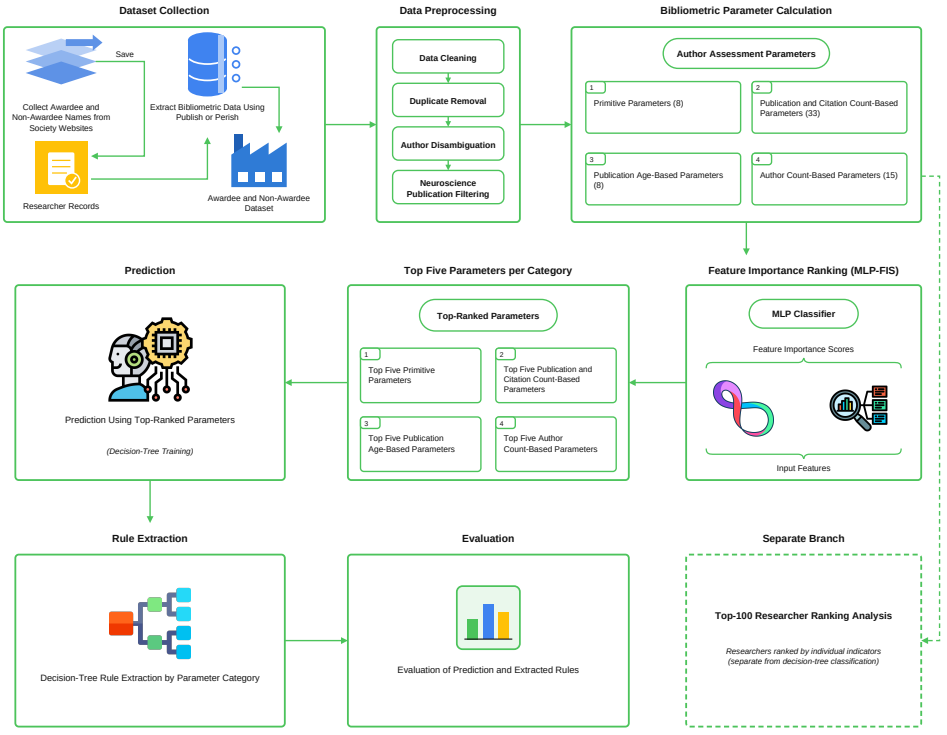

Figure 1. Overview of the proposed interpretable machine learning framework for identifying bibliometric patterns associated with neuroscience award recognition. The workflow illustrates dataset

collection, preprocessing, computation of 64 bibliometric indicators across four categories, MLP-based feature importance ranking (MLP-FIS), selection of the five highest-ranked indicators per category, decision-tree classification, rule extraction, and performance evaluation. A separate branch represents the top-100 researcher ranking analysis based on individual bibliometric indicators.

The subsequent sections describe each of these components in detail.

3.1. Domain Selection and Dataset Collection

For experimental evaluation, a domain-specific dataset was required. In this study, the *Neuroscience* domain was selected due to the substantial volume of research activity and its critical role in advancing the understanding of the brain, nervous system, and related disorders. Neuroscience investigates diseases, injuries, and functional abnormalities of the nervous system, with researchers continuously working to develop preventive and therapeutic strategies for conditions affecting cognitive, sensory, and motor functions. The scale and societal importance of research in this domain make the identification and ranking of influential neuroscientists a particularly important and challenging task.

Owing to its interdisciplinary nature and scientific significance, the neuroscience domain has been widely adopted in prior bibliometric and scientometric studies [5,16]. In this work, we constructed a benchmark neuroscience dataset comprising 1060 researcher profiles, including both awardees and non-awardees. The dataset is balanced, consisting of 530 awardees and 530 non-awardees.

The decision to construct a balanced dataset warrants explicit methodological justification. In the real-world research population, award recipients constitute a small minority of all active researchers, meaning that the natural class distribution is heavily skewed toward non-awardees. Training and evaluating classifiers on such an imbalanced distribution would allow a trivial majority class predictor one that labels every researcher as a non-awardee to achieve deceptively high accuracy while providing no discriminative value whatsoever. By enforcing a 1:1 ratio between awardees and non-awardees, any classifier that simply predicts the majority class would achieve no better than 50% accuracy on this dataset, ensuring that all reported accuracy figures reflect genuine discriminative power rather than class imbalance artifacts. This balanced design is consistent with established practice in bibliometric classification studies [9,16,21], where controlled matching is used to isolate the effect of bibliometric signals from distributional confounds. We acknowledge, however, that this design choice has implications for real-world deployment: because the operational class distribution is far more imbalanced than the experimental one, precision and recall figures reported here may not directly transfer to screening scenarios involving large researcher pools. Practitioners applying these rules in such contexts should recalibrate decision thresholds to account for the true prior probability of award receipt in the target population.

Awardee data were collected from well-established neuroscience related scientific societies that confer prestigious annual awards recognizing outstanding research contributions. These five societies include the Society for Neuroscience (SFN) (https://www.sfn.org/, accessed on 11 November 2026), the International Brain Research Organization (IBRO), the Federation of European Neuroscience Societies (FENS) (https://www.fens.org/, accessed on 11 November 2026), the Cognitive Neuroscience Society (CNS) (https://www.cogneurosociety.org/, accessed on 11 November 2026), and the Australasian Neuroscience Society (ANS) (https://www.austneuroscience.org.au/, accessed on 11 November 2026).

Researcher names and their corresponding award years were collected from the official websites of five neuroscience societies: SFN, IBRO, FENS, CNS, and ANS. The award collection period covered 1990 to 2023, inclusive (34 calendar years). The study distinguishes the historical model-development window (1990–2020) from the later temporal-evaluation

window (2021–2023); records associated with 2021 are designated for the latter window, not the development window. This describes the intended cohort assignment and must be checked against researcher-level records before a strictly non-overlapping temporal split can be claimed. The dataset includes recipients of neuroscience-related awards conferred by these societies during the specified period. For each awardee, only publications published before the corresponding award year were considered when computing bibliometric indicators. The number of awardees contributed by each society is given in Table 2.

Non-awardee data were compiled from the benchmark dataset introduced by Ameer and Afzal [16], which has been widely used for evaluating bibliometric indicators in the neuroscience domain. To ensure temporal consistency and prevent information leakage, a pre-award (hold-on) data collection strategy was adopted. Bibliometric data for both awardees and non-awardees were extracted using the *Publish or Perish* platform (https://harzing.com/resources/publish-or-perish, accessed on 11 November 2026), which employs advanced retrieval algorithms to collect publication records and author-level metadata from Google Scholar. Only publications published prior to the award year were considered for awardees. It is important to clarify how citation counts were handled within this pre-award framework. All bibliometric parameters, including citation-based indices, were computed exclusively from publications dated before each researcher's award year, and publications dated after the award year were excluded together with their citations. However, the citation counts of the retained pre-award publications were taken as indexed by Google Scholar at the time of data collection via the *Publish or Perish* platform. Citations that these publications accrued between the award year and the collection date are therefore included. This is a form of temporal information leakage: a prestigious award can itself increase the later visibility and citation of earlier work [13,17], so raw citation totals and the indices derived from them may be inflated for awardees relative to non-awardees. Google Scholar does not provide citation archives at past dates, so we could not remove this effect. We do not argue that the effect is small: author-normalized and age-weighted indices are computed from the same citation counts and would inherit any post-award boost, so their performance is not evidence against leakage. All reported discrimination should therefore be read as an upper bound on what strictly pre-award information would support. A sensitivity analysis that recomputes the indicators from year-resolved citation data truncated at the award year (for example from OpenAlex or Scopus) is the appropriate remedy and is a priority for future work (Section 9). As a partial check, Section 5.9 reports how well indicators that do not depend on citation counts discriminate.

To maintain equity within the dataset, non-awardee researchers were selected in the same proportion as awardees for each award year. For example, if 15 awardees were identified in a given year, data for 15 non-awardees with comparable publication timelines prior to that year were collected using the same criteria.

Matching was by award year and publication timeline only. It does not control for career age at award, subfield, institution, region or gender, so differences between the two groups may partly reflect these factors. We verified that none of the non-awardees later received any of the five awards. The matching strategy was designed primarily to maintain temporal comparability between awardees and non-awardees rather than to establish equivalence across all researcher characteristics. Award year and publication timeline were used as the principal matching criteria to reduce differences associated with publication periods and opportunities for research accumulation. Nevertheless, researchers with comparable publication timelines may differ in career stage, neuroscience specialization, geographic representation, institutional background, and publication practices. These characteristics were not explicitly controlled during matching and may influence the observed bibliometric differences independently of award status. Consequently, the balanced dataset should be

interpreted as a controlled comparison based on the available temporal criteria rather than a fully representative or demographically matched sample of neuroscience researchers. The potential influence of residual confounding and sampling bias should therefore be considered when interpreting the results.

Before conducting any analysis, a comprehensive data cleaning and preprocessing procedure was performed to enhance data quality and reliability. This process involved verifying bibliographic accuracy, removing duplicate records, and eliminating irrelevant or noisy entries that could adversely affect experimental outcomes. Additionally, two critical refinement steps were applied. First, domain-specific filtering was performed to ensure that all publications were relevant to neuroscience. Second, an author disambiguation process was conducted to identify and resolve duplicate researcher profiles arising from variations in author name spellings or formatting. These preprocessing steps collectively ensured the consistency, accuracy, and relevance of the final dataset.

The inclusion criteria required awardee records to correspond to neuroscience-related awards granted by the five selected societies between 1990 and 2023, with 2021–2023 designated as the temporal-validation years and 1990–2020 as the development years. Bibliometric information was restricted to publications predating the corresponding award year, while non-awardees were selected using comparable publication timelines and reference years. During dataset preprocessing, duplicate records were not identified in the final compiled dataset. However, awards received by the same individual in different years represent distinct award events rather than distinct researchers. Accordingly, the dataset statistics are interpreted at the researcher level, with the reported total of 530 awardees referring to individual researcher profiles.

Detailed statistics of the constructed neuroscience dataset are presented in Table 2.

Table 2. Neuroscience dataset statistics.

Neuroscience Dataset Statistics	**Count**
Total Researchers	1060
Total Awardees	530
Total Non-Awardees	530
Total Publications	166,871
Total Citations	25,855,493
Awardees Societies Statistics	
Society for Neuroscience (SFN)	291
International Brain Research Organization (IBRO)	32
Federation of European Neuroscience Societies (FENS)	33
Cognitive Neuroscience Society (CNS)	90
Australasian Neuroscience Society (ANS)	84

3.2. Calculation of Author Assessment Parameters

In this section, a total of sixty-four author assessment parameters are computed using the constructed neuroscience dataset. These parameters are designed to capture multiple dimensions of scholarly impact and have been extensively discussed in the literature, along with their mathematical formulations, strengths, and limitations. For systematic analysis, the parameters are organized into four distinct categories. All indices were implemented and computed using the Python (version 3.11) programming language. The categories and their corresponding parameters are described below.

Appendix A Table A1 provides computational descriptions and source references for the indices documented in the cited reference appendix. Some source notation and the 64-feature inventory require reconciliation with the actual implementation.

3.2.1. Selection Criteria for the 64 Parameters

The selection of the sixty-four author assessment parameters was guided by three explicit criteria rather than arbitrary enumeration. First, *literature coverage*: each parameter was required to have appeared in at least one peer-reviewed scientometric or bibliometric study, ensuring that only theoretically grounded and empirically evaluated indicators were included. The foundational set of h-index variants follows the taxonomy established by Bornmann and Daniel [7], which documents the most widely studied productivity and citation-based indices. Additional indices covering author normalization and age adjustment were drawn from subsequent comparative evaluations in neuroscience [5,16] and broader reviews of citation-impact indicators [3,7], ensuring domain relevance alongside breadth of coverage.

Second, *categorical completeness*: parameters were selected to ensure comprehensive representation across four functionally distinct dimensions of scholarly output raw productivity (primitive parameters), citation-distribution shape (publication and citation count-based parameters), individual contribution in collaborative settings (author count-based parameters), and temporal sustainability of impact (publication age-based parameters). This four-category structure reflects the consensus in the scientometric literature that no single dimension is sufficient to characterize scholarly influence [5,7], and it enables the present study to perform category-aware feature selection and comparative analysis across dimensions a capability absent from prior single-category or unstructured multi-metric evaluations.

Third, *computational feasibility*: all selected parameters must be computable from the author-level metadata available through Google Scholar via the *Publish or Perish* platform, specifically publication records, citation counts, author counts, and publication years. Indices requiring data not available through this pipeline such as journal impact factors, funding records, or institutional affiliations were excluded. This constraint ensures full reproducibility of the parameter computation pipeline from publicly accessible data sources.

Taken together, these three criteria justify both the scope (sixty-four parameters) and the structure (four categories) of the parameter set. The resulting collection is intentionally comprehensive rather than minimal: a broad initial set is necessary to allow the MLP-FIS feature ranking procedure (Section 3.3) to identify the most discriminative subset empirically, rather than relying on subjective pre-selection that could introduce investigator bias into the analysis.

3.2.2. Primitive Parameters

Primitive parameters represent basic bibliometric indicators that describe overall research productivity and citation behavior. The parameters included in this category are Total Publications, Total Citations, Total Years, Citations per Year, Citations per Paper, Authors per Paper, Citations per Author, and Papers per Author.

3.2.3. Publication and Citation Count-Based Parameters

This category comprises bibliometric indices that are primarily derived from publication counts and citation distributions. The parameters included are the h-index, g-index, e-index, h-core citations, A-index, R-index, P-index, M-index, F-index, T-index, Q2-index, tapered h-index, Maxprod, Wu-index, Pi-index, weighted h-index, h(2)-index, Woeginger index, gh-index, Rm-index, X-index, hg-index, h2 upper index, h2 center index, h2 lower index, k-dash index, rational h-index, real h-index, i10-index, normalized h-index, k-index, w-index, and h-dash index.

3.2.4. Author Count-Based Parameters

Author-based parameters normalize bibliometric impact by accounting for co-authorship patterns and individual contributions. The parameters in this category include the hI-index, normalized hI-index, hm-index, gm-index, hf-index, gf-index, GF-index, k-normalized index, w-normalized index, pure h-index, fractional g-index and fractional h-index.

3.2.5. Publication Age-Based Parameters

Publication Age-based parameters incorporate the temporal dimension of scholarly output by adjusting citation impact according to publication age. The parameters included in this category are the platinum h-index, m-quotient index, AW-index, AR-index, V-index, Ha-index, Hc-index (contemporary h-index), and the age-weighted citation rate (AWCR).

All author assessment parameters across the four categories were computed using the collected neuroscience dataset. To facilitate reproducibility and scalability, a dedicated application was developed to automate data extraction and parameter computation. The application retrieves researcher bibliometric data directly from Google Scholar using a provided profile link and computes all sixty-four author assessment parameters for the selected researcher.

The implementation was carried out in Python, and the complete source code for data extraction and parameter computation is publicly available online (https://github.com/ghulammustafacomsat/dataextraction, accessed on 11 November 2026).

3.3. Ranking of Author Assessment Parameters

In machine learning, feature ranking plays a critical role in identifying the most influential variables used for model construction and evaluation. Feature ranking serves multiple purposes, including dimensionality reduction, improved model interpretability, mitigation of overfitting, optimization of prediction performance, and refinement of feature engineering processes [22].

The overall feature ranking framework is illustrated in Figure 2. To assess parameter importance, we employed a multilayer perceptron classifier (MLP) combined with a leave-one-out feature ablation strategy to compute parameter importance scores. It is important to note that this procedure differs from iterative Recursive Feature Elimination (RFE) [23,24], which recursively retrains the model on progressively reduced feature sets across multiple elimination rounds. In our approach, the MLP is trained once on the full feature set to establish a baseline accuracy, and then each feature is removed independently and the model is retrained on the reduced set. The drop in validation accuracy when a given feature is excluded serves as its importance score. This one-shot ablation approach is computationally efficient and yields a ranking of all 64 parameters from a single set of retraining runs, without the iterative elimination of the standard RFE procedure. The stability of the ranking across training folds is examined in Section 5.8, where it turns out to be low. Hereafter we refer to this procedure as **MLP-based Feature Importance Scoring (MLP-FIS)**.

The selection of MLP-FIS over simpler feature ranking alternatives warrants explicit justification. Commonly used alternatives include mutual information, Pearson correlation-based ranking, and tree-based Gini importance derived from Random Forest classifiers. While these methods are computationally efficient and theoretically well-grounded in linear settings, they carry limitations that reduce their suitability for the present dataset. Mutual information and correlation-based methods rank features independently of one another and therefore fail to account for interaction effects among highly correlated bibliometric indices a well-documented characteristic of the h-index family, where variants such as the g-index, tapered h-index, and h2 center index are structurally related and exhibit high mutual correlation [5,7]. Ranking such features independently would yield unstable and

potentially misleading importance estimates, as the marginal contribution of any one index is confounded by its redundancy with others. MLP-FIS, by contrast, evaluates each feature's contribution within the context of the full feature set, capturing non-linear interactions and redundancy effects that linear ranking methods cannot model.

To verify the consistency of MLP-FIS rankings against a simpler alternative, we conducted a cross-check against Random Forest Gini importance scores computed on the same training partition. The top-5 parameters selected per category under MLP-FIS were compared with the top-5 parameters identified by Random Forest Gini importance under identical experimental conditions (100 estimators, default hyperparameters, fixed random seed 42). Agreement between the two methods was assessed at the category level. The top-5 selections were fully consistent across the Primitive- and Publication Age-Based categories, and four out of five parameters matched in the Publication and Citation Count-Based and Author Count-Based categories, with only minor rank-order differences among lower-ranked entries within each category. This concordance indicates that the MLP-FIS rankings are broadly consistent with an established alternative; it does not show that MLP-FIS is better than simpler rankings. A further caveat applies to any leave-one-out ablation: when indices are highly correlated, the remaining correlated indices can compensate for a removed one, so ablation scores can understate the importance of redundant but informative indices. The MLP-FIS procedure was therefore retained as the primary ranking method, with Random Forest Gini importance serving as a corroborating sanity check.

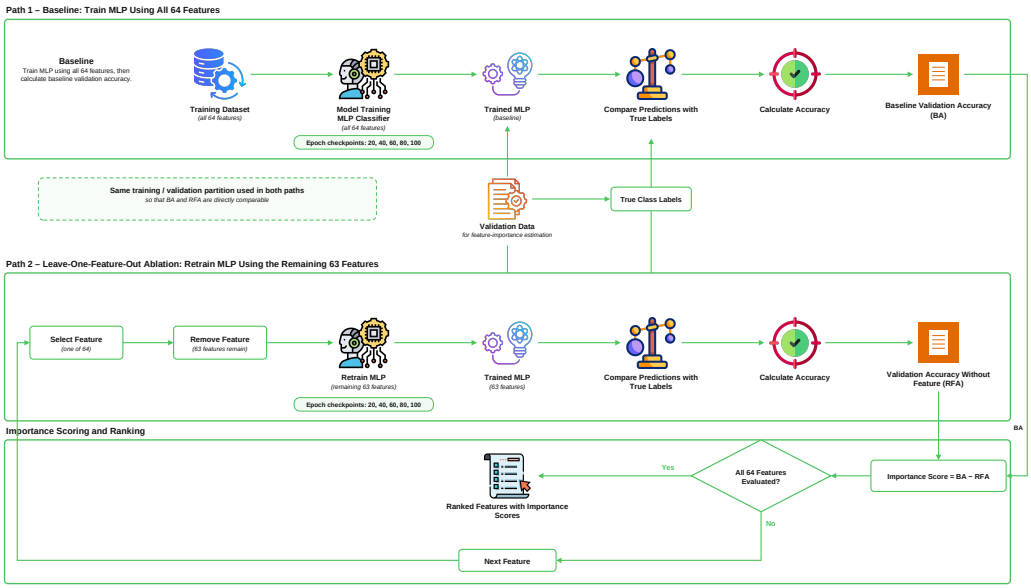

Figure 2. Detailed workflow of the multilayer perceptron-based feature importance scoring (MLP-FIS) procedure. The process comprises two parallel evaluation paths: baseline MLP training using all 64 bibliometric features and leave-one-feature-out ablation using the remaining 63 features. Both paths employ consistent training and validation partitions, with accuracy evaluated at predefined epoch checkpoints (20, 40, 60, 80, and 100). Feature importance is determined from the difference between baseline accuracy (BA) and reduced-feature accuracy (RFA). The five epochs indicate the prespecified comparison schedule. Since early stopping can shorten a run, feature-importance differences should be based on comparable baseline and omission-model checkpoints rather than treating a nonexistent epoch as a new measurement. This procedure is repeated for all 64 features to obtain their relative importance scores and final ranking.

However, agreement between feature-ranking methods on a single data partition should not be interpreted as evidence of ranking stability. As demonstrated by the repeated cross-

validation analysis in Section 5.8, the MLP-FIS top-five feature sets exhibit considerable variability across training folds, particularly within highly correlated bibliometric indicator categories. Accordingly, the resulting rankings are interpreted as exploratory indicators of relative feature contribution rather than stable or uniquely optimal feature selections.

MLP is widely recognized in the literature for its versatility and effectiveness across multiple problem domains, including classification, prediction, and regression tasks [25]. The MLP classifier is implemented as a feed-forward artificial neural network composed of multiple hidden layers [26]. In classification tasks, the number of neurons in the input layer corresponds to the number of input features, while the number of neurons in the output layer corresponds to the number of target classes. The hidden layers form a fully connected network trained using the backpropagation algorithm.

During forward propagation, the output of each layer is computed as a linear transformation of the previous layer's activations, followed by the addition of a bias term, as defined in Equation (1).

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

where X represents the output matrix, W denotes the weight matrix, A is the input activation matrix, and b represents the bias vector.

To ensure that the network output remains within a bounded and meaningful range, an activation function is applied to the linear output, as expressed in Equation (2).

$$C = g(X) \tag{2}$$

where C denotes the activated output matrix.

In this study, the Rectified Linear Unit (ReLU) activation function was employed in all hidden layers, while the Softmax function was used in the output layer. The ReLU activation function introduces nonlinearity by mapping negative values to zero while preserving positive values, as shown in Equation (3).

$$f(X) = \max(0, X) \tag{3}$$

The Softmax activation function, defined in Equation (4), is widely used in multi-class classification problems. It converts raw output scores into probability values that sum to one, allowing the most probable class to be identified.

$$\text{Softmax}(X_i) = \frac{e^{X_i}}{\sum_{j=1}^{J} e^{X_j}} \tag{4}$$

where J denotes the number of classes and X_i represents the output score for the i-th class.

Model performance was evaluated using the mean squared error (MSE) loss function, which quantifies the discrepancy between predicted and actual values, as defined in Equation (5).

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

where n represents the total number of samples, z_i denotes the predicted value, and $\hat{z}_i$ represents the actual value. Mean squared error (MSE) was adopted as the training loss in the implemented MLP configuration. Although MSE can be applied to classification problems by comparing predicted class probabilities with encoded target labels, categorical cross-entropy is generally more appropriate for Softmax-based classification because it directly optimizes the likelihood of the observed class labels. Therefore, the use of MSE is treated as an implementation choice rather than a methodological advantage. A systematic comparison of the two loss functions would be necessary to establish whether the choice influences predictive performance or the resulting feature importance rankings. To

mitigate overfitting and address issues such as gradient vanishing or explosion in deep neural networks, batch normalization was incorporated after each hidden layer [27]. Batch normalization normalizes the output activations using the mean and standard deviation, as shown in Equation (6).

$$X_i = \frac{X_i - \mu_i}{\sigma_i} \tag{6}$$

where X_i denotes the output of the i-th hidden layer, μ_i is the mean, and σ_i is the standard deviation.

The MLP classifier was configured with ten hidden layers, each consisting of ten neurons. The selection of the number of layers and neurons was determined empirically through multiple experiments. Although this architecture may appear overparameterized relative to the roughly 850 training samples, two design choices were made to address this concern. First, early stopping with a patience of 40 epochs was employed to terminate training before overfitting could occur, retaining only the best-performing model parameters. Second, batch normalization was applied after each hidden layer to stabilize activations and improve generalization. Together, these mechanisms constrain the effective model complexity during training.

The adopted MLP architecture should be understood as an empirically selected configuration rather than a theoretically optimal network structure. Given the relatively limited number of training observations, the use of multiple hidden layers introduces potential overfitting and model-complexity concerns. Batch normalization and early stopping were incorporated as regularization and training-stabilization mechanisms, although these techniques do not guarantee the absence of overfitting. Furthermore, the purpose of the MLP in the proposed framework is feature importance estimation rather than final classification or interpretable rule generation. We acknowledge that a systematic comparison with shallower architectures would be required to determine whether comparable feature rankings could be obtained with fewer network layers.

Crucially, the MLP is employed exclusively for feature *ranking* it is not the final interpretable output of the proposed framework. The interpretable component consists entirely of the decision tree rules extracted in Section 5, which are derived from a transparent, human-readable model. The role of the MLP FIS procedure is therefore analogous to a preprocessing filter that identifies the most informative subset of the 64 parameters before the interpretable decision tree is trained. We acknowledge that alternative feature importance methods such as permutation importance and SHAP values would provide more theoretically principled estimates and would map more directly onto the explainability framing of this work; the cross-check against Random Forest Gini importance reported above provides an initial empirical check, and a more systematic comparison with permutation importance and SHAP values remains a promising direction for future investigation.

ReLU activation and batch normalization were applied after each hidden layer to enhance stability and generalization. The output layer employed the Softmax activation function to generate class probabilities. During the prediction phase, a probability vector is produced for each sample. The final class label is determined using the argmax operation, as shown in Equation (7).

$$\text{Output} = \arg\max(\text{Predicted Vector}) \tag{7}$$

The model was trained using the Adam optimization algorithm with a learning rate of 0.0003, a batch size of 64, and a maximum of 100 epochs. Early stopping was employed to prevent overfitting, with a patience value of 40 epochs. Training was terminated if the validation loss failed to improve over 40 consecutive epochs, and the best-performing model parameters were retained.

To compute feature importance, the leave-one-out ablation procedure described above was applied systematically. The dataset was initially divided into training and validation sets using an 80:20 ratio. The MLP classifier was trained using all 64 features to obtain the baseline accuracy. Subsequently, one feature was removed at a time, and the model was retrained using the reduced feature set. The resulting validation accuracy was recorded and compared with the baseline accuracy. The difference between the baseline accuracy and the accuracy obtained without the parameter served as the importance score for the removed feature. This process was repeated at five epoch checkpoints at intervals of 20, specifically at epochs 20, 40, 60, 80, and 100, and the importance score was averaged across these checkpoints as defined in Equation (8). The formula represents the five scheduled comparison epochs; it presumes that paired baseline and feature-omission accuracies exist at those epochs. If early stopping ends a run before a scheduled epoch, an accuracy at that epoch should not be treated as independently observed. The appropriate general form averages over the set of checkpoints actually available for both runs, as clarified below. This distinction concerns the interpretation of the checkpoint average and does not change the reported feature-ranking results; confirmation of the precise historical checkpoint-selection behavior requires the original training logs or implementation.

$$F_{IS}(j) = \frac{1}{|K_j|} \sum_{t \in K_j} \left(BLA_t - WOPA_{j,t} \right), \quad K_j \subseteq \{20, 40, 60, 80, 100\} \tag{8}$$

Here $F_{IS}(j)$ is the importance score for omitted feature j, BLA_t is baseline validation accuracy at epoch t, $WOPA_{j,t}$ is validation accuracy for the model omitting feature j at the same epoch, and K_j contains only paired checkpoints at which both accuracies are defined. When all five checkpoints exist, $|K_j| = 5$ and Equation (8) reduces to the originally stated five-term average. This is a general mathematical clarification, not a retrospective assertion that a specific unavailable-checkpoint policy was implemented in the historical runs. Averaging across five checkpoints is intended to reduce sensitivity to epoch-specific fluctuations in validation accuracy. With early stopping, the corresponding average is defined over paired checkpoints only; this avoids imputing unobserved future-epoch accuracies. The historical ranking values are retained, and their exact checkpoint provenance remains an implementation-level reproducibility qualification.

The MLP-FIS importance score represents the change in validation accuracy when an individual feature is removed. A positive score indicates a decrease in predictive accuracy following feature removal, whereas a negative score indicates that the reduced feature set achieved higher validation accuracy. Scores close to zero indicate little measurable marginal contribution under the evaluated model and data partition. However, near-zero scores do not necessarily imply that a feature is uninformative, because highly correlated indicators may provide overlapping information. Consequently, the magnitude and sign of these scores should be interpreted cautiously, particularly when differences are comparable to normal validation variability.

This procedure was repeated for all sixty-four parameters. Upon completion, two lists were generated: one containing the parameter names and the other containing their corresponding importance scores. The parameters were then sorted in descending order of importance scores, yielding a ranked list of author assessment parameters for each of the four categories.

3.4. Prediction Using Top-Ranked Parameters

After completing the parameter ranking process, the top five parameters from each category namely primitive parameters, publication and citation count-based parameters, author count-based parameters, and publication age-based parameters were selected for

further analysis. These selected parameters represent the most influential indicators within their respective categories.

Using each top-ranked parameter independently, all 1060 researchers in the dataset were ranked to evaluate the presence and distribution of awardees within the top-100 ranked researcher list. For each parameter, Recall@100 was computed as the proportion of the 530 awardees in the full dataset who appear within the top-100 researcher list when researchers are sorted in descending order of that parameter's value (Equation (9)). Precision@100 was computed as the proportion of the top-100 list that consists of awardees (Equation (10)).

$$\text{Recall@100} = \frac{\text{Number of awardees in top-100 list}}{530} \tag{9}$$

$$\text{Precision@100} = \frac{\text{Number of awardees in top-100 list}}{100} \tag{10}$$

Because the dataset is balanced, the expected Precision@100 under random ranking is 50%. A list of 100 can contain at most 100 awardees, so Recall@100 cannot exceed $100/530 = 18.9\%$, and its chance level is $50/530 = 9.4\%$. Precision@100 must therefore be compared with 50% and Recall@100 with 9.4%.

When multiple researchers share identical values on a given parameter which occurs frequently for integer-valued indices such as the h-index and k-index ties are broken by secondary ranking on Total Citations in descending order. If ties persist after this criterion, a tertiary sort on Total Publications in descending order is applied. This deterministic tie-breaking rule is applied consistently across all parameter rankings to ensure full reproducibility of Recall@100 and Precision@100 values reported in Result Section.

The procedural steps involved in selecting the top-ranked parameters, ranking researchers, and computing category-wise Recall@100 and Precision@100 values are detailed in Algorithm 1.

Algorithm 1 Prediction Using Top-Ranked Parameters

Require: Dataset $\mathcal{D}$ of 1060 researchers with pre-computed parameter values; ranked parameter list $\mathcal{P}_c$ for each category c; total awardee count $A = 530$; top-k threshold $k = 100$

Ensure: Recall@100 and Precision@100 for each parameter across all categories

Divide parameters into four categories $c \in$ {Primitive, Publication & Citation Count–Based, Author Count–Based, Publication Age–Based}
for each category c **do**
  Select the top-5 ranked parameters $\{p_1, p_2, p_3, p_4, p_5\} \leftarrow \mathcal{P}_c[1{:}5]$
  for each parameter $p_i \in \{p_1, p_2, p_3, p_4, p_5\}$ **do**
    Sort all 1060 researchers in $\mathcal{D}$ in descending order of p_i value
    Tie-breaking: For researchers with equal p_i values, apply secondary sort by Total Citations (descending); if ties persist, apply tertiary sort by Total Publications (descending)
    Extract the top-k ($k = 100$) researchers from the sorted list $\rightarrow \mathcal{T}_{p_i}$
    Count awardees in top-k list: $n \leftarrow |\mathcal{T}_{p_i} \cap \text{Awardees}|$
    Compute Recall@100(p_i) $\leftarrow n$ / $A = n$ / 530
    Compute Precision@100(p_i) $\leftarrow n$ / $k = n$ / 100
    Record Recall@100(p_i) and Precision@100(p_i) for category c
  end for
  Report category-wise Recall@100 and Precision@100 for all five parameters in $\mathcal{P}_c[1{:}5]$
end for
Compare Recall@100 and Precision@100 distributions across all four parameter categories
return Category-wise Recall@100 and Precision@100 values for the analysis in Section 6

The resulting rankings provide insight into the predictive strength and discriminative capability of each parameter category. The explicit formulation of Recall@100 and

Precision@100, combined with the deterministic tie-breaking rule, ensures that all results reported in Result Section are fully reproducible from the publicly available dataset (The complete ranking and Recall@100 computation scripts are available at https://github.com/Dr-GMustafa/research-datasets, accessed on 11 November 2026).

3.5. Rule Mining Using Top Five Parameters per Category

In machine learning, decision trees are widely recognized as effective models for a variety of tasks, including prediction, classification, and rule mining [28,29]. Owing to their inherent interpretability and ease of implementation, decision trees were adopted in this study to extract explainable decision rules from the ranked bibliometric parameters.

The construction of a decision tree begins with partitioning the source dataset into subsets based on a selected attribute. This process follows a recursive learning strategy in which each node is further divided into child nodes until one of two stopping criteria is met: either all samples within a node belong to the same class, or further splitting no longer contributes to improved classification performance. Decision trees are well suited for identifying patterns and categorizing data through statistical measures, without requiring prior domain-specific assumptions. Their transparency and competitive accuracy make them particularly valuable for exploratory analysis and rule extraction.

The splitting of decision tree nodes is typically guided by an impurity measure, most commonly information gain or the Gini index. In this study, the Gini index was employed due to the numerical nature of the bibliometric features and its suitability for handling continuous valued data [30]. Information gain, while effective in identifying informative attributes, can be less appropriate when datasets contain attributes with many distinct values [31]. In contrast, the Gini index provides a robust measure of node impurity by quantifying the probability of incorrect classification when samples are randomly assigned to classes.

The Gini index is defined in Equation (11).

$$GI = 1 - \sum_{j=1}^{n} (P_j)^2 \tag{11}$$

where P_j denotes the probability of an instance belonging to class j. The Gini index takes values between 0 and 1, and its interpretation is as follows:

- A Gini index value of 0 indicates perfect purity, where all instances within a node belong to a single class.
- A Gini index value closer to 1 indicates higher impurity, reflecting a more uniform distribution of instances across classes.

Given the large number of author assessment parameters in the dataset, the resulting decision trees tend to be highly complex and dense, which can lead to overfitting and reduced interpretability. To address this issue and eliminate redundant or insignificant rules, a post-pruning strategy was applied. Post-pruning involves initially growing the decision tree to its full depth, followed by the systematic removal of subtrees that contribute marginally to predictive performance.

This pruning process evaluates each subtree based on a cost parameter that accounts for both classification accuracy and model complexity, measured by the number of nodes. Subtrees associated with higher costs were pruned, yielding a more compact and generalizable tree structure.

It should be noted that the fully grown and pruned decision trees reach depths of five to six levels with approximately 40 nodes, which exceeds the depth threshold (typically ≤ 3) that the rule-extraction literature generally associates with direct human interpretability. In this work, interpretability is therefore a property of the *extracted rules*, the single prominent root-to-node paths presented as decision equations in Section 5, rather than of the full

tree structure. Each extracted rule consists of at most five conditions (three to four after overlapping thresholds are consolidated) drawn from the top-ranked parameters of its category. The decision-tree figures in Section 5 display the part of each tree that contains the reported rule, with sibling branches omitted where needed for legibility, so that each rule can be checked against an actual split path; the node identifiers quoted in Section 5 refer to these figures. The highlighted rules are paths to internal or terminal nodes of trees fitted on the training partition, so the coverage and specificity values in Section 5.6 are training-partition values. Constraining trees to a maximum depth of 3 would give more compact, directly interpretable tree diagrams and is identified as a direction for future work. Clarification of Feature Selection and Evaluation Protocol: The experimental framework distinguishes between exploratory feature ranking and rule extraction, and the common-protocol predictive evaluation. The initial MLP-FIS feature rankings and decision rules were derived using the training and validation procedures described in Sections 3.3–3.5. Because the resulting training-partition rule statistics may be optimistic, these findings are interpreted as exploratory rather than as independent estimates of generalization performance. For the comparative evaluation reported in Section 5.8, repeated stratified five-fold cross-validation with three repetitions was employed. Feature importance scoring and top-five parameter selection were performed independently within each training fold, with the corresponding held-out fold excluded from feature selection and model fitting. All compared classifiers were evaluated using identical folds. Although the manuscript describes the adopted model configurations and pruning approach, a separate nested hyperparameter optimization procedure is not reported. Consequently, the cross-validation results provide a more reliable assessment than the exploratory training-partition results, while some uncertainty associated with model selection remains.

3.6. Evaluation Metrics and Denominators

For a rule or classifier that labels a researcher as "awardee" or "non-awardee", let TP, FP, FN and TN be the numbers of awardees correctly labeled, non-awardees wrongly labeled as awardees, awardees missed, and non-awardees correctly labeled. We report accuracy, precision, recall (sensitivity), specificity and F_1:

$$\mathrm{Acc} = \frac{TP + TN}{TP + FP + FN + TN}, \qquad \mathrm{Prec} = \frac{TP}{TP + FP}, \qquad \mathrm{Rec} = \frac{TP}{TP + FN},$$
$$\mathrm{Spec} = \frac{TN}{TN + FP}, \qquad F_1 = \frac{2\,\mathrm{Prec} \cdot \mathrm{Rec}}{\mathrm{Prec} + \mathrm{Rec}}.$$

The denominators depend on the data on which a metric is computed and are stated for every table. For rules evaluated on the training partition (Section 5.6) the partition contains 415 awardees and 437 non-awardees, so *coverage* is $TP/415$ and specificity is $1 - FP/437$. For the temporal hold-out (Section 5.7) there are 53 awardees and 53 non-awardees. For the top-100 analysis (Section 6) the denominators are 100 for Precision@100 and 530 for Recall@100. Metrics computed on the data used to grow a tree are resubstitution estimates and are optimistic. Confidence intervals for proportions are Wilson 95% intervals, and chance levels are tested with one-sided binomial tests.

4. Results and Discussion

In this section, we present and discuss the experimental results obtained using the proposed methodology on the neuroscience dataset. The results cover parameter ranking, prediction analysis and rule mining, together with the limits of the evidence. A detailed discussion is provided to highlight the contribution and interpretability of the selected bibliometric indicators.

4.1. Parameter Ranking

This subsection presents the ranking of author assessment parameters obtained using the MLP-based Feature Importance Scoring (MLP-FIS) technique, which employs a leave-one-out feature ablation strategy in conjunction with a Multilayer Perceptron (MLP) classifier. The ranking reflects the relative importance of each parameter in distinguishing awardees from non-awardees within the neuroscience domain.

The top-ranked parameters, identified based on their computed importance scores, were subsequently utilized for downstream analysis, including prediction experiments and rule extraction using decision tree models. These ranked parameters form the foundation for deriving explainable and domain-specific insights into researcher influence.

Two cautions apply to the rankings. The importance scores are differences in validation accuracy (Equation (8)) obtained from a single 80:20 partition with about 212 validation samples, so one sample changes accuracy by about 0.5 percentage points. Differences of a few hundredths between neighbouring ranks are therefore within what a single partition can produce, and rank orders should be treated as approximate. In addition, the single-parameter accuracies reported below are the accuracies of a one-split threshold classifier (a decision stump) fitted on the training folds and tested on the held-out folds of repeated stratified 5-fold cross-validation (3 repeats); the chance level is 50%.

4.1.1. Primitive Parameter Ranking

Within the primitive parameter category, the Total Citation count emerged as the most influential parameter, achieving the highest importance score of 0.17, as illustrated in Figure 3. This indicates that total citations contribute more significantly to the model's predictive performance than other primitive indicators. The Author per Paper parameter ranked second with an importance score of 0.14, followed by Citations per Year, which secured the third position with an importance score of 0.13. The remaining parameters—Citations per Author, Citations per Paper, Total Publications, Papers per Author, and Total Years—achieved importance scores of 0.11, 0.11, 0.11, 0.11, and 0.09, respectively.

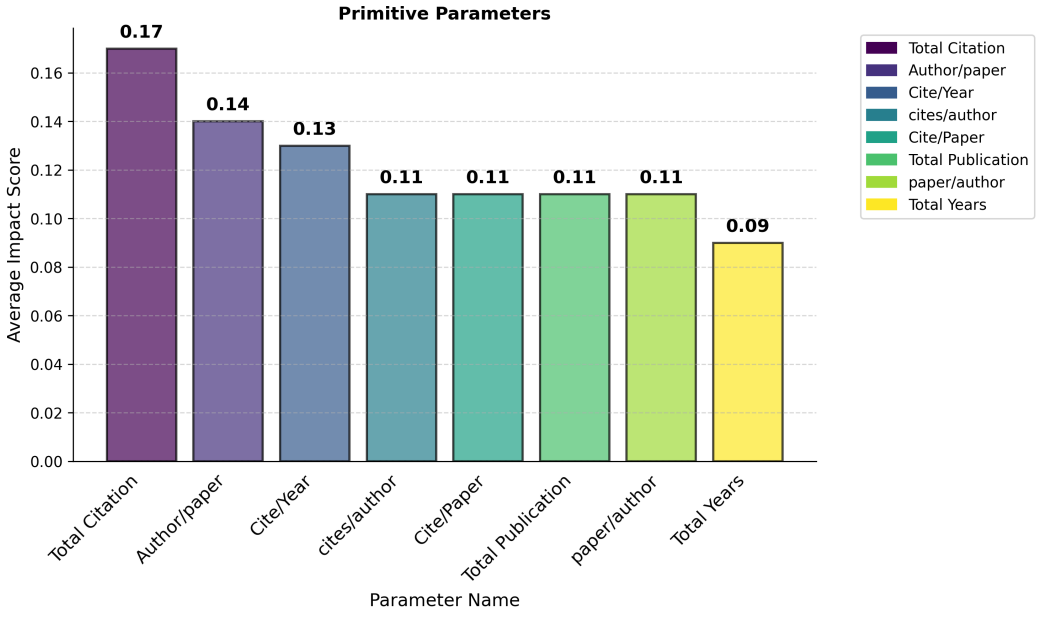

Figure 3. Ranking of primitive parameters.

To further evaluate their discriminative capability we computed the single-parameter accuracy defined above. Cite/Year (64.8%) and Total Citation (62.5%) were the most

discriminative primitive parameters, followed by Cite/Paper (61.5%) and Cites per Author (61.3%). Author per Paper (56.0%), Papers per Author (53.5%), Total Publications (51.9%) and Total Years (50.8%) were close to chance.

These findings indicate that citation-related measures contain a substantial share of the bibliometric signal associated with award recognition. In particular, a higher total citation count is associated with a higher likelihood of being an award recipient, although the single-parameter accuracy of 62.5% is only moderately above the 50% chance level and should not be interpreted as a general measure of scientific influence.

4.1.2. Author Count-Based Parameter Ranking

Within the author count-based parameter category, the gf-index emerged as the most influential indicator, achieving the highest importance score of 0.18, as shown in Figure 4. This result indicates that the fractional g-index normalized by author contribution plays a crucial role in enhancing model performance. The fractional g-index ranked second with an importance score of 0.16, followed by the hl norm index in third position with an importance score of 0.14. The hm-index also achieved a score of 0.14, tying for third place. The gF-index, k norm index, w norm index, h_m_index, hf index, and fractional h index followed with scores of 0.13, 0.12, 0.12, 0.12, 0.12, and 0.12, respectively. The normalized hI-index ranked last in this category with an importance score of 0.10.

To further validate the top-ranked parameters we computed their single-parameter accuracy. The fractional g-index (60.7%) and the hl norm index (60.4%) were the most discriminative, followed by the gf-index (59.1%) and the hm-index (57.3%); the gF-index was close to chance (51.9%). The gf-index, which had the highest importance score, was therefore not the most discriminative single parameter.

These values are modest. Author-normalized indices carry some information about recognition, but none is clearly better than the citation-based indicators of the other categories.

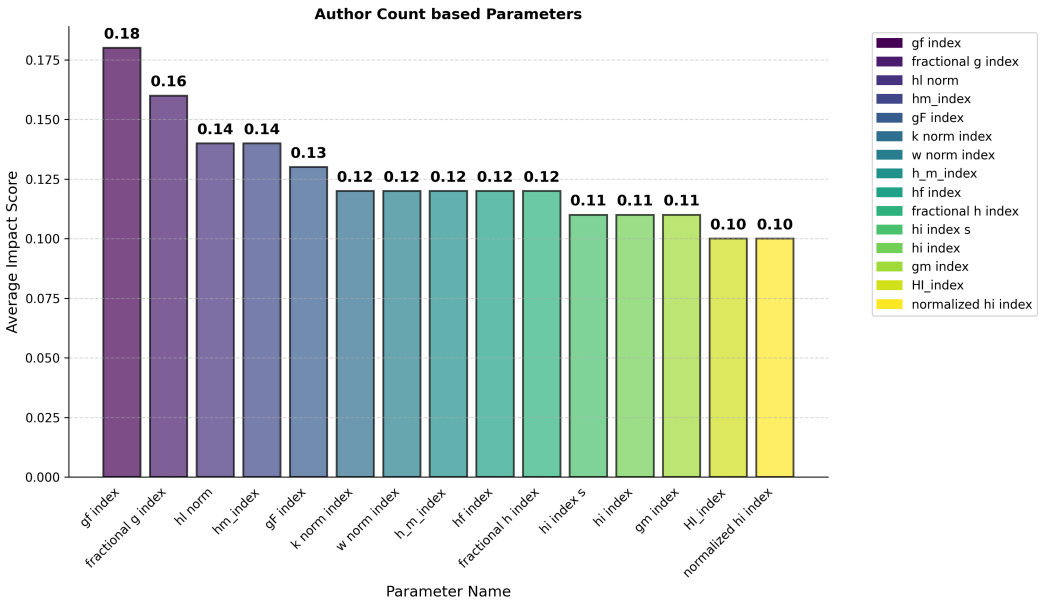

Figure 4. Ranking of author count-based parameters.

4.1.3. Publication Age-Based Parameter Ranking

Within the publication age-based parameter category, the AW-index emerged as the most influential indicator, achieving the highest importance score of 0.16, as shown in Figure 5. This result highlights the importance of age-weighted citation rates in capturing

sustained scholarly impact over time. The m-quotient index ranked second with an importance score of 0.14, followed by the AR-index in third position with an importance score of 0.13. The AWCR and ha index both achieved scores of 0.12, while the V index, platinum h-index, and hc index scored 0.11, 0.09, and 0.06, respectively.

To further assess these parameters we computed their single-parameter accuracy. The AR-index (65.0%) was the most discriminative, followed by the ha index (63.5%) and the AW-index and AWCR (62.5% each); the m-quotient index reached 56.4%.

These findings suggest that age-adjusted indicators carry some information about sustained research contributions in neuroscience, comparable to that of simple citation-based indicators.

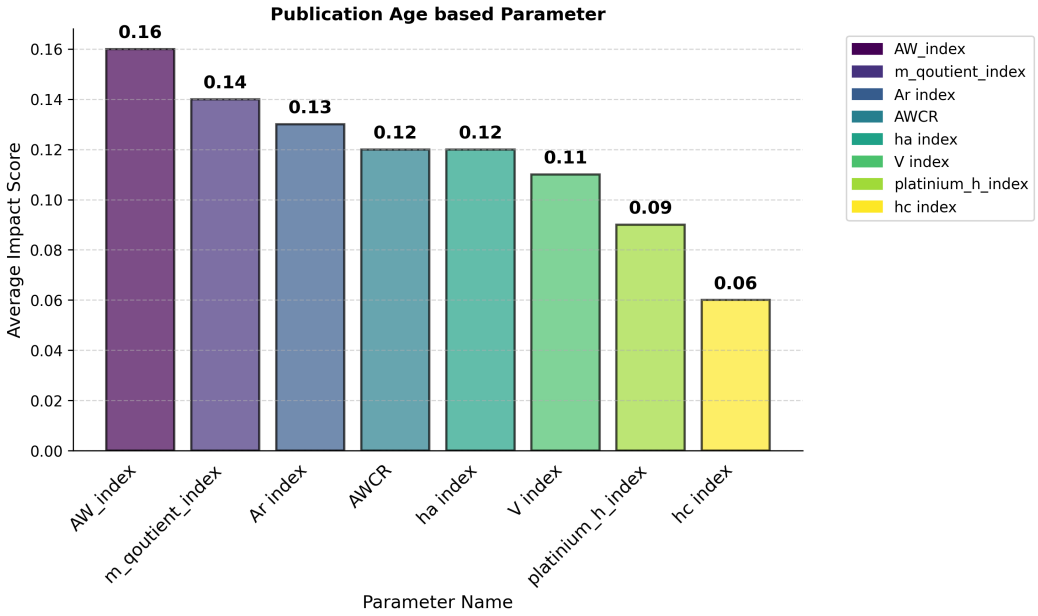

Figure 5. Ranking of publication age-based parameters.

4.1.4. Publication and Citation Count-Based Parameter Ranking

Within the publication and citation count-based parameter category, the h2 center index emerged as the most influential parameter, achieving the highest importance score of 0.23, as shown in Figure 6. This indicates that the h2 center index, which captures the core citation impact around the h-index, contributes substantially to the model's discriminative performance. The tapered h-index ranked second with an importance score of 0.20, followed by the woginger index and k index in third and fourth positions, each attaining an importance score of 0.18 and 0.17, respectively. The Rational h index and X index achieved scores of 0.17 each, while the remaining parameters -h dash index, h2upper index, pure h index, and w index -scored 0.16, 0.15, 0.13, and 0.13, respectively. Maxprod, pi index, and the normalized h index ranked among the lower performers in this category, with scores of 0.11, 0.11, and 0.10, respectively.

To validate these parameters we computed their single-parameter accuracy. The k index (66.4%) was the most discriminative, followed by the Woginger index (63.3%) and the Rational h index (59.4%); the h2 center index, which had the highest importance score, reached 57.0% and the tapered h-index 56.0%.

Overall, the best single-parameter accuracies of the top-five parameters are similar across categories (primitive 64.8%, author count-based 60.7%, age-based 65.0%, publication and citation count-based 66.4%). Across all 64 indicators the single-parameter accuracy ranged from 50.8% to 70.5% (best: P index 70.5%, platinum h-index 67.1%, k index 66.4%).

The MLP-FIS ranking is therefore only loosely related to single-parameter discrimination: the first-ranked h2 center index reached 57.0% and the first-ranked gf-index 59.1%.

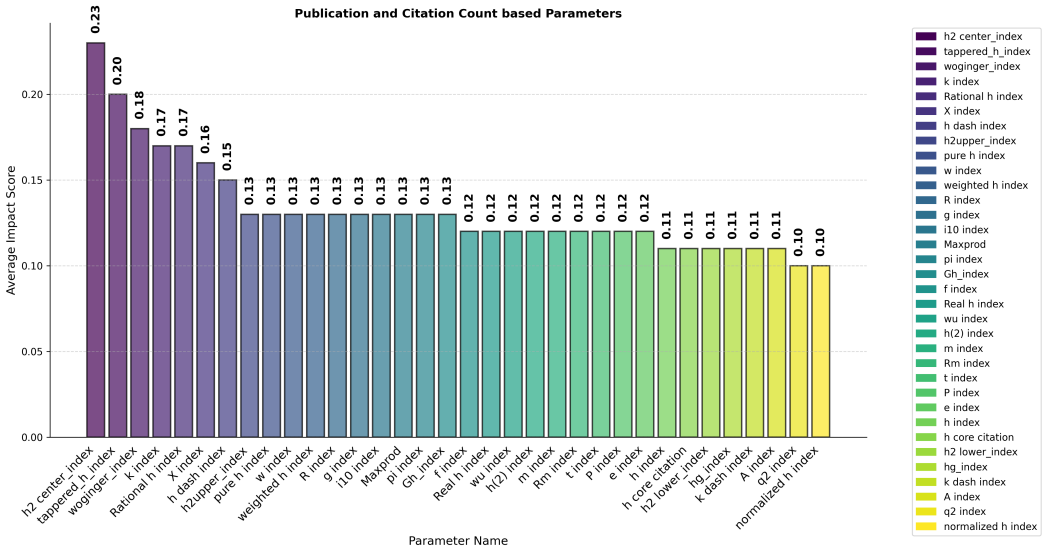

Figure 6. Ranking of publication and citation count-based parameters.

4.1.5. Combination of All Parameters

In this subsection, parameters from all four categories are combined to analyze their collective influence on model performance. Combining all categories results in a larger set of parameters, which makes comprehensive visualization in a single figure less effective. Therefore, only the top ten highest-ranked parameters are presented in Figure 7 for clarity.

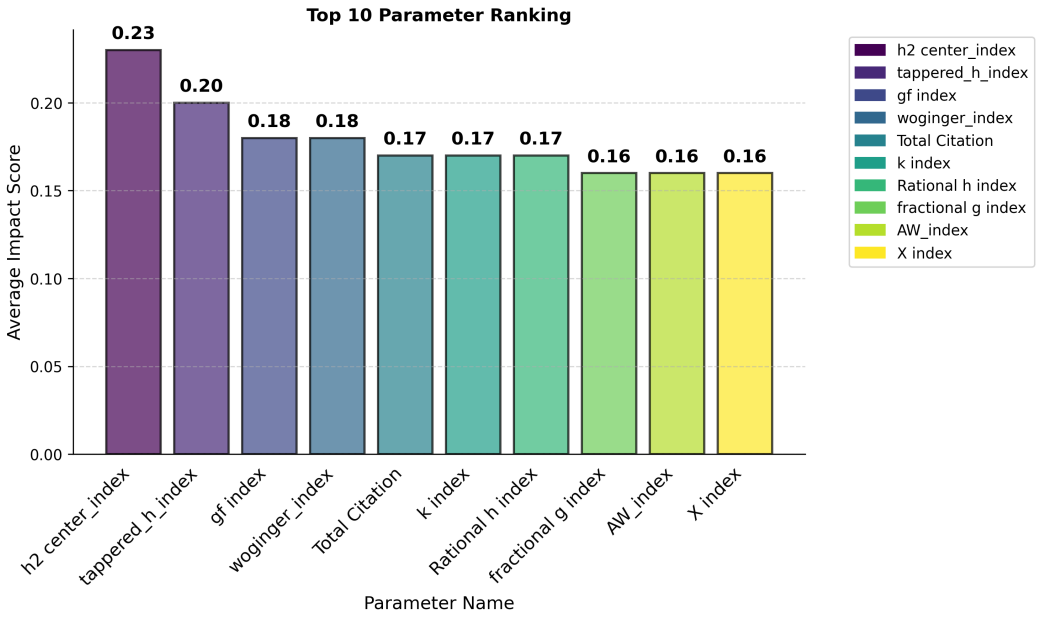

Figure 7. Top-ranked parameters across all categories.

From the combined ranking, it is observed that the h2 center index outperforms all other parameters across categories with an importance score of 0.23, indicating its strong and consistent contribution to identifying influential researchers. The tapered h-index ranked second with a score of 0.20, followed by the gf index and woginger index in third

and fourth positions, each scoring 0.18. The Total Citation, k index, and Rational h index followed with scores of 0.17 each, while the fractional g index, AW_index, and X index completed the top ten with scores of 0.16 each. These results suggest that citation-core and author-normalized parameters occur most often among the highest-ranked parameters when all categories are considered together.

4.2. Redundancy Among the 64 Parameters

The literature reports strong correlations among h-index variants (Section 2). We quantified the redundancy in our data with Spearman rank correlations and a principal component analysis of the rank-transformed, standardized indicators (Table 3). Of the 2016 pairs of parameters, 874 (43%) have $|\rho| \geq 0.9$ and the median absolute correlation is 0.85; hierarchical clustering groups the 64 parameters into 17 clusters at $|\rho| \geq 0.9$, 11 at $|\rho| \geq 0.8$ and 8 at $|\rho| \geq 0.7$. Three principal components explain 90% of the variance and five explain 95%.

The author count-based indicators are almost interchangeable (median $|\rho| = 0.97$; one component explains 96% of their variance), the publication and citation count-based and age-based indices each share one dominant component, and the primitive indicators are the most heterogeneous. Between groups the median correlation ranges from 0.64 (primitive versus publication and citation count-based) to 0.93 (publication and citation count-based versus author count-based). The 64 parameters therefore contain only a few dominant dimensions of variation. This has two consequences for the analysis: it supports the earlier finding that the choice of any single index within a group matters little for discrimination, and it explains why the selection of a fixed set of five indices by leave-one-out ablation is fragile (Section 5.8), because a removed index is largely replaced by its correlated neighbours. Thresholds for individual indices should accordingly not be over-interpreted.

Table 3. Redundancy among the 64 parameters. ρ: Spearman rank correlation over all researchers. PCs for 90%: number of principal components (of rank-transformed, standardized indicators) needed to explain 90% of the variance; PC1: variance share of the first component.

Parameter Group	Indicators	Median $\|\rho\|$	PCs for 90%	PC1 Share
Primitive	8	0.53	3	60%
Publication and citation count-based	33	0.81	2	75%
Publication age-based	8	0.83	2	86%
Author count-based	15	0.97	1	96%
All 64 parameters	64	0.85	3	77%

5. Rule Mining

In this section, we present the extracted decision rules for each parameter category based on the top five ranked parameters within that category. Figures 8–12 illustrate the most representative and high-coverage rules derived for each category. These rules were extracted directly from the split paths of fully grown decision trees trained on the 852-sample training partition (437 non-awardees and 415 awardees, as shown at the root node of each tree), which inherently contain multiple decision paths. All coverage and specificity values in this section are therefore training-partition (resubstitution) values and are optimistic; held-out results are reported in Section 5.7. The upper bounds in the highlighted rules are produced by the way the tree partitions the data: observations above a bound are assigned to other branches, and those branches can contain many awardees (for example, node #31 in Figure 10 holds 78 awardees among 96 samples with AR index above 101.402). The bounds must not be read as evidence that very high values are detrimental.

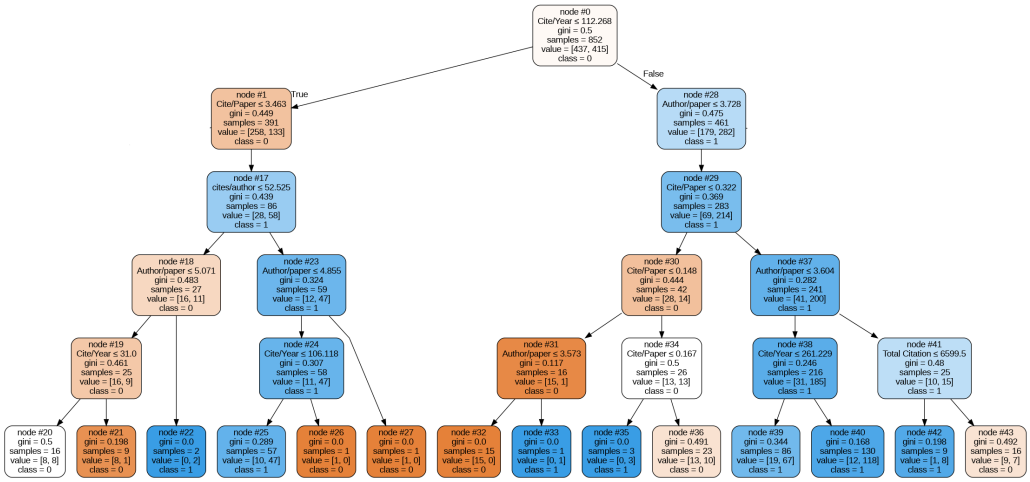

Figure 8. Decision rule derived from primitive parameters. Numbers preceded by # are node identifiers of the decision tree, as referred to in the text.

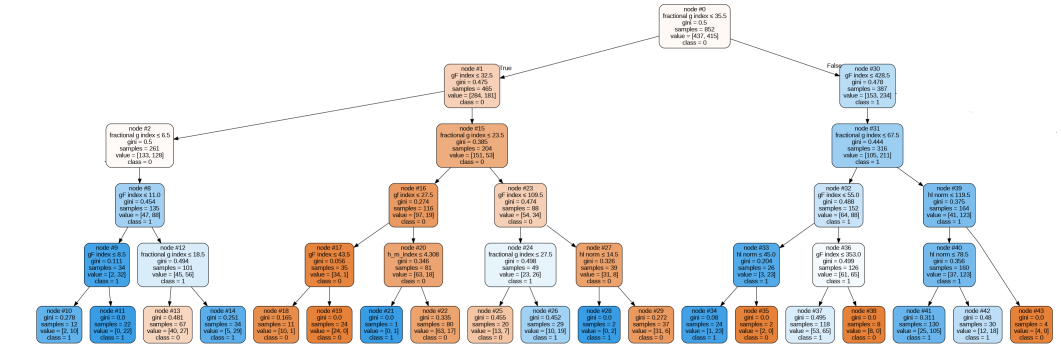

Figure 9. Decision rule derived from author count-based parameters. Numbers preceded by # are node identifiers of the decision tree, as referred to in the text.

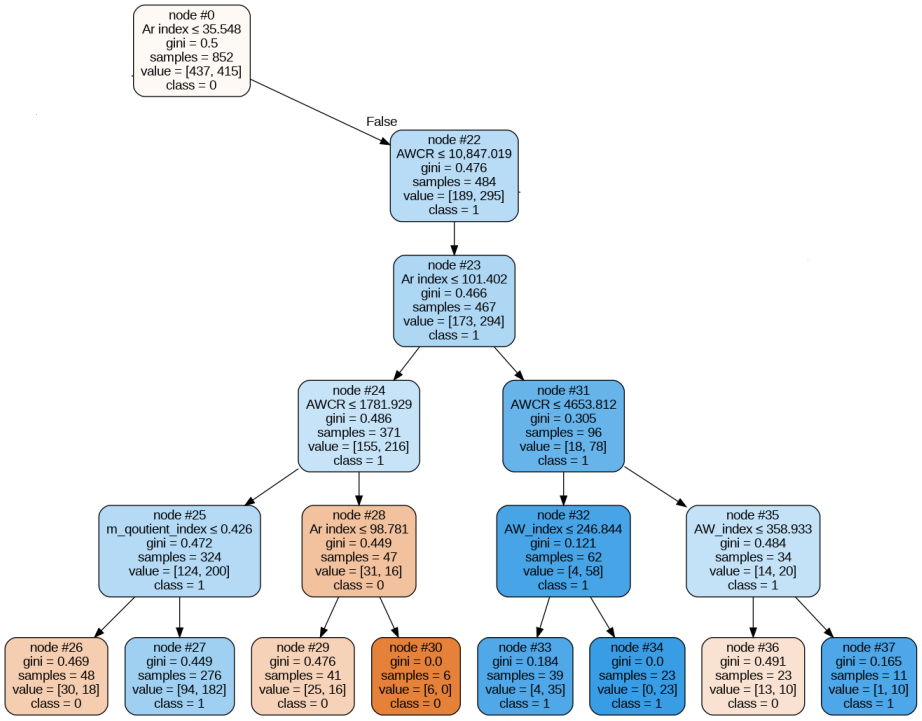

Figure 10. Decision rule derived from age-based parameters. Numbers preceded by # are node identifiers of the decision tree, as referred to in the text.

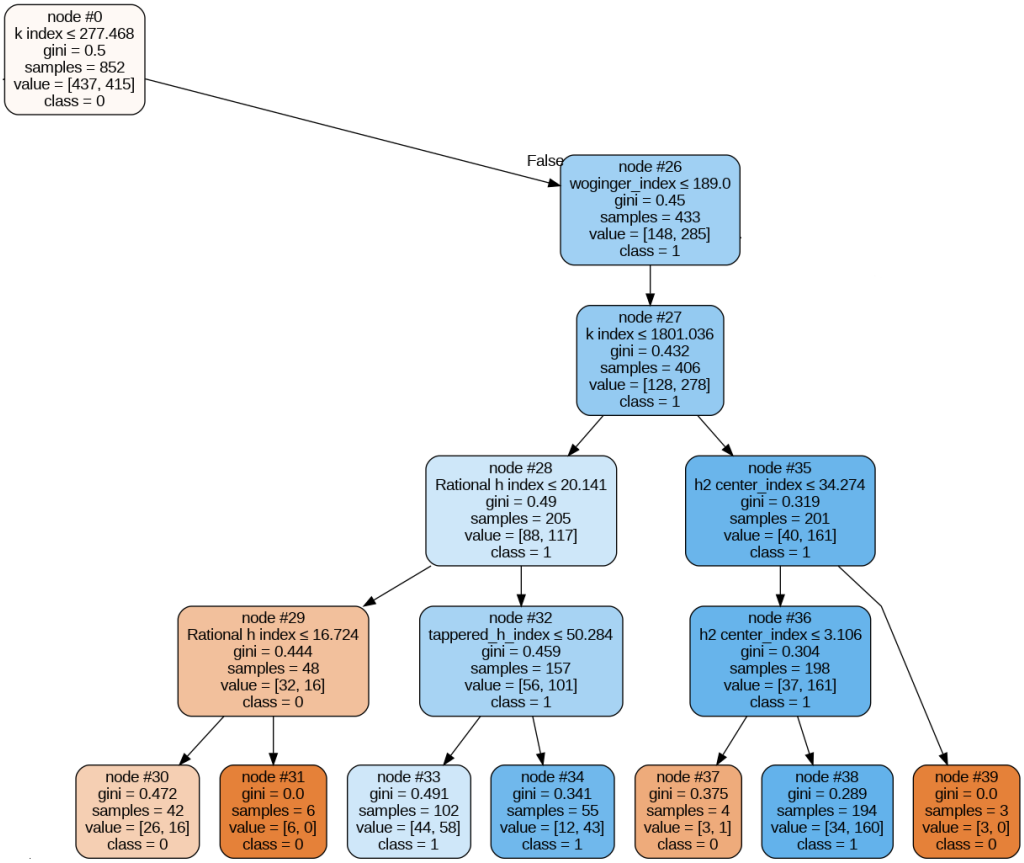

Figure 11. Decision rule derived from publication and citation count-based parameters.

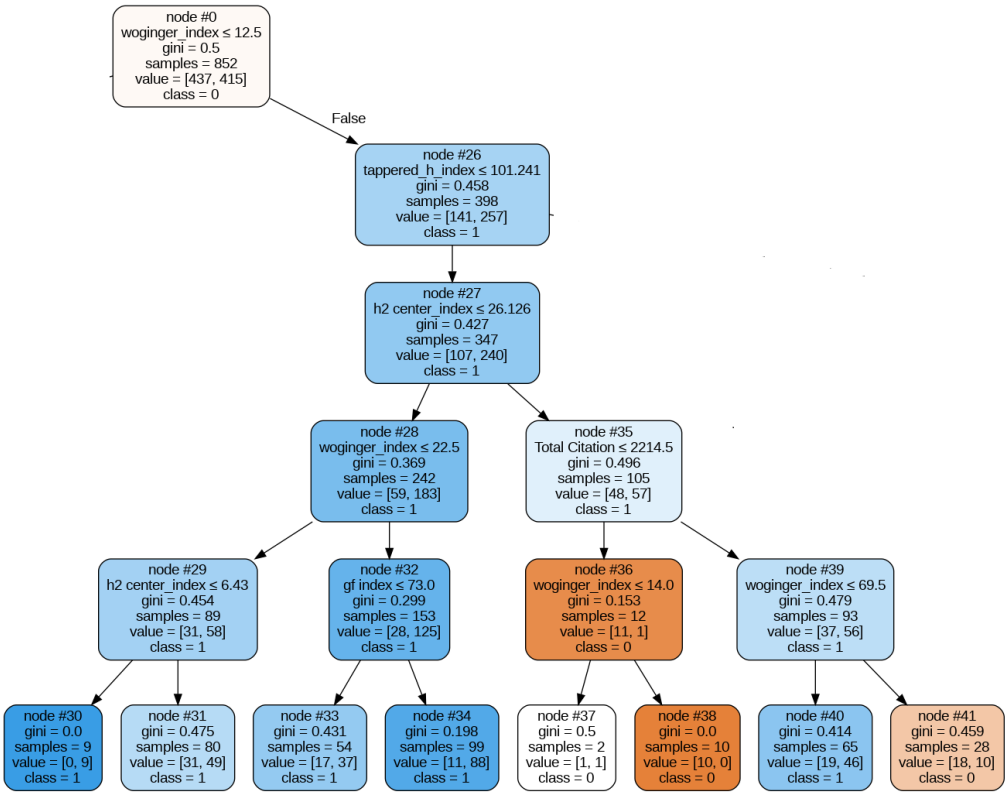

Figure 12. Decision rule derived from combined parameters. Numbers preceded by # are node identifiers of the decision tree, as referred to in the text.

Rather than reporting all generated rules, we selectively highlight those that capture dominant patterns associated with award recipients and offer meaningful insights for the scientific community. The presented rules play a significant role in understanding how different bibliometric characteristics contribute to the ranking and recognition of influential researchers.

The reported rules represent selected awardee-dominated paths extracted from the decision trees rather than an exhaustive evaluation of all generated rules. Their selection emphasizes substantial coverage of award recipients and human-readable threshold conditions. This selective presentation may introduce reporting bias; therefore, the highlighted rules should not be interpreted as uniquely optimal or as a complete classification rule set. Training-partition support, coverage, precision, and specificity are reported separately from performance on the independent random test partition, allowing the exploratory rule characteristics to be distinguished from held-out results.

5.1. Rule for Primitive Parameters

Among the primitive parameters, the top five ranked features identified through feature importance analysis are *Total Citations*, *Author per Paper*, *Citations per Year*, *Citations per Author*, and *Citations per Paper*. To uncover dominant patterns associated with award recipients, a decision tree classifier was trained using the Gini impurity criterion for node splitting.

The resulting decision tree achieved an overall classification accuracy of 78% in distinguishing awardees from non-awardees based solely on primitive parameters; this figure is a whole-dataset (resubstitution) accuracy that includes the training data, and held-out results are given in Section 5.7. Figure 8 illustrates the structure of the induced decision tree. The most prominent rule follows a four-condition path (Citations per Year $>$ 112.268, Author per Paper $\leq$ 3.728, Citations per Paper $>$ 0.322, Author per Paper $\leq$ 3.604). Consolidating the two Author per Paper thresholds, where the stricter bound subsumes the looser one, leaves three non-redundant conditions and leads to node #38, which holds 185 awardees among 216 training samples (31 non-awardees). Node #38 is an internal node that the tree splits further (Citations per Year $\leq$ 261.229); the rule is stated at this node because it is the largest awardee-dominated node on the path. Its coverage is 185/415 = 44.6% of training awardees. The consolidated rule is:

$$\textbf{IF} \begin{cases} \text{Citations per Year} > 112.268, \\ \text{Author per Paper} \leq 3.604, \\ \text{Citations per Paper} > 0.322 \end{cases} \quad \textbf{THEN} \quad \text{Award Recipient} \tag{12}$$

The rule describes researchers who accumulate more than about 112 citations per year and whose papers have comparatively few authors on average (at most about 3.6). The condition on Cite/Paper is an additional, weak condition: about 20% of the researchers in the dataset have a value of at most 0.322, so the main content of the rule is Citations per Year $>$ 112.268 together with Author per Paper $\leq$ 3.604.

5.2. Rule for Author Count-Based Parameters

Within the author count-based parameter category, the top five ranked indicators identified through feature importance analysis are *gf index*, *fractional g index*, *hl norm*, *hm index*, and *gF index*. To extract dominant patterns associated with award recipients, a decision tree classifier was trained using the Gini impurity criterion for node splitting.

The induced decision tree achieved an overall classification accuracy of 73% in identifying award recipients based on author-normalized metrics (whole-dataset accuracy, optimistic). Figure 9 illustrates the structure of the decision tree. The most prominent awardee path follows a four-condition path (fractional g index $>$ 35.5, gF index $\leq$ 428.5,

fractional g index $>$ 67.5, hl norm $\leq$ 119.5). Consolidating the two fractional g index conditions, where the stricter bound subsumes the looser one, leaves three non-redundant conditions and leads to node #40, which holds 123 awardees among 160 training samples (37 non-awardees; coverage 123/415 = 29.6%). Node #40 is an internal node that the tree splits further on hl norm ($\leq$78.5). The rule is:

$$\textbf{IF} \begin{cases} \text{fractional g index} > 67.5, \\ \text{gF index} \leq 428.5, \\ \text{hl norm} \leq 119.5 \end{cases} \quad \textbf{THEN} \quad \text{Award Recipient} \tag{13}$$

The rule selects researchers with a high author-normalized citation contribution (fractional g index above 67.5) whose gF index and hl norm values lie below the caps of 428.5 and 119.5. The caps remove the upper tail of these two indices from this branch of the tree; we have no evidence on why recognized researchers should fall below them, and they may partly reflect correlation among the author-normalized indices. The rule is interpreted only as a descriptive profile.

5.3. Rule for Age-Based Parameters

Within the age-based parameter category, the top five ranked indicators identified through feature importance analysis are *AW-index*, *m-quotient index*, *AR-index*, *AWCR*, and *ha index*. To extract dominant patterns associated with award recipients, a decision tree classifier was trained using the Gini impurity criterion for node splitting.

The induced decision tree achieved an overall classification accuracy of 68% in distinguishing awardees from non-awardees. Figure 10 illustrates the structure of the resulting decision tree and the extracted decision rules. The most prominent awardee path follows a five-condition path in the tree. After consolidating overlapping conditions specifically, the two AR index thresholds ($>$35.548 and $\leq$101.402), which together form a bounded range, and the two AWCR thresholds ($\leq$10,847.019 and $\leq$1781.929), where the stricter bound subsumes the looser one the rule simplifies to three non-redundant conditions, yielding a leaf node with 182 awardees out of 276 samples (node #27, class = 1). This rule captures 182/415 = 43.9% of training awardees, and 94 non-awardees also satisfy it. It is expressed as follows:

$$\textbf{IF} \begin{cases} 35.548 < \text{AR index} \leq 101.402, \\ \text{AWCR} \leq 1781.929, \\ \text{m-quotient index} > 0.426 \end{cases} \quad \textbf{THEN} \quad \text{Award Recipient} \tag{14}$$

The rule combines three age-related conditions. The bounded AR-index range (35.548 $<$ AR index $\leq$ 101.402) selects researchers with a moderate-to-high age-adjusted citation impact. The lower bound on the m-quotient ($>$0.426) selects a comparatively high rate of h-index growth per year of activity, and the ceiling on AWCR ($\leq$1781.929) bounds the age-weighted citation rate. The two upper bounds arise from the tree partition: researchers above them are assigned to other branches, which can contain many awardees (see the note at the beginning of this section). The rule is a descriptive profile and not a causal explanation. The available methodological record does not verify that the 149 researchers with anomalous Total Years values were removed before feature selection or tree induction; consequently, their potential influence on the learned feature order and split thresholds cannot be ruled out. In addition, the observed anomalies in Total Years for 149 researchers may distort duration-dependent indicators, especially the m-quotient. Although omitting these records reportedly alters single-index AUCs by less than 0.02, this check does not establish the stability of the exact AR-index, AWCR, or m-quotient thresholds in Equation (14). These

cutoffs should not be treated as generalizable career-age standards without a date-quality audit and refitting sensitivity analysis (Section 7.1).

5.4. Rule for Publication and Citation Count-Based Parameters

Within the publication and citation count-based parameter category, the top five ranked indicators identified through feature importance analysis are *h2 center index*, *tapered h-index*, *woginger index*, *k index*, and *Rational h index*. To extract dominant patterns associated with award recipients, a decision tree classifier was trained using the Gini impurity criterion for node splitting.

The induced decision tree achieved an overall classification accuracy of 71% (whole-dataset accuracy, optimistic). Figure 11 illustrates the structure of the decision tree. The most prominent awardee path follows a five-condition path (k index $>$ 277.468, Woginger index $\leq$ 189.0, k index $>$ 1801.036, h2 center index $\leq$ 34.274, h2 center index $>$ 3.106). Consolidating the two k index thresholds (the stricter bound subsumes the looser one) and the two h2 center index thresholds (which together form a bounded range) leaves three non-redundant conditions and leads to node #38, which holds 160 awardees among 194 training samples (34 non-awardees). This rule captures 160/415 = 38.6% of training awardees and is expressed as follows:

$$\textbf{IF} \begin{cases} \text{k index} > 1801.036, \\ \text{woginger index} \leq 189.0, \\ 3.106 < \text{h2 center index} \leq 34.274 \end{cases} \quad \textbf{THEN} \quad \text{Award Recipient} \tag{15}$$

The rule selects researchers with a very large k index (above about 1800) and a Woginger index of at most 189. The bounds on the h2 center index (3.106 and 34.274) come from the tree partition and cut off very few researchers: only 4 and 3 samples fall into the sibling nodes #37 and #39. The substantive content of the rule is therefore the combination of a high k index with a bounded Woginger index. It is a descriptive profile, and no claim is made here about the shape of the underlying citation distributions.

5.5. Rule for Combined Parameters

In this subsection, parameters from all categories were combined to analyze their collective influence on identifying award recipients. Based on the combined feature importance ranking, the top five parameters were identified as *h2 center index*, *tapered h-index*, *gf index*, *woginger index*, and *Total Citation*. To extract dominant patterns associated with award recipients, a decision tree classifier was trained using the Gini impurity criterion for node splitting.

The induced decision tree achieved an overall classification accuracy of 78% in distinguishing awardees from non-awardees (whole-dataset accuracy, optimistic). Figure 12 illustrates the structure of the decision tree. The most prominent awardee path follows a five-condition path (Woginger index $>$ 12.5, tapered h-index $\leq$ 101.241, h2 center index $\leq$ 26.126, Woginger index $>$ 22.5, gf index $>$ 73.0). Consolidating the two Woginger index thresholds (the stricter bound subsumes the looser one) leaves four non-redundant conditions and leads to node #34, which holds 88 awardees among 99 training samples (11 non-awardees). This rule captures 88/415 = 21.2% of training awardees and is expressed as follows:

$$\textbf{IF} \begin{cases} \text{woginger index} > 22.5, \\ \text{tapered h-index} \leq 101.241, \\ \text{h2 center index} \leq 26.126, \\ \text{gf index} > 73.0 \end{cases} \quad \textbf{THEN} \quad \text{Award Recipient} \tag{16}$$

The rule draws on a citation-spread index (Woginger), a citation-core index (h2 center), a tapered h-index and an author-normalized index (gf). Among the highlighted rules it has the highest precision (88.9%) and specificity (97.5%) but the lowest coverage (21.2%). The ceilings on the tapered h-index and h2 center index are partition boundaries: researchers above them are assigned to other branches of the tree, which are not shown, so they should not be read as penalizing high values. As with the other rules, the rule is descriptive.

5.6. Rules Coverage and Specificity

After generating decision rules for all parameter categories, a comparative analysis of rule coverage (sensitivity on training data) and specificity was performed. Coverage alone is an insufficient indicator of discriminative power a trivial rule such as "Total Publications ≥ 1" would achieve near-100% coverage of awardees while providing no discrimination. Therefore, specificity (the proportion of non-awardees correctly excluded by the rule) is reported alongside coverage (and precision) for each highlighted rule in Table 4.

Table 4. Training-partition performance of the highlighted decision rules (852 training samples: 415 awardees and 437 non-awardees). TP and FP are the numbers of awardees and non-awardees that reach the node at which each rule is stated (node counts read from Figures 8–12). Coverage = TP/415; specificity = 1 − FP/437; precision = TP/(TP + FP). These are resubstitution values and are optimistic.

Rule (Node)	TP	FP	Coverage	Specificity	Precision
Primitive (node #38)	185	31	44.6%	92.9%	85.6%
Author count-based (node #40)	123	37	29.6%	91.5%	76.9%
Age-based (node #27)	182	94	43.9%	78.5%	65.9%
Pub. & citation count-based (node #38)	160	34	38.6%	92.2%	82.5%
Combined (node #34)	88	11	21.2%	97.5%	88.9%

Each highlighted rule describes a single path of a tree, so its coverage is limited by construction: the combined-parameter rule has the highest specificity (97.5%) and precision (88.9%) but the lowest coverage (21.2%), the primitive and age-based rules cover about 44–45% of training awardees, and the age-based rule has the lowest specificity (78.5%) and precision (65.9%). Because these single-node rules are evaluated on the same partition that was used to grow the trees, they overstate performance. Several complementary paths would be needed to cover most awardees, and we have not evaluated such rule sets. Held-out performance is reported in the next subsection.

Table 5 applies the same rules to the 214 researchers that were set aside before the trees were grown. On this split the rules cover 33–49% of awardees at 83–95% specificity, with precision between 75.7% and 88.4% and accuracy between 62.1% and 69.2% when used as classifiers. Specificity and precision stay close to their training-partition values, which suggests that the individual rule paths are not heavily overfitted to the training partition, but each rule misses more than half of the awardees.

Table 5. Performance of the highlighted rules on the 20% test split of the same data (114 awardees, 100 non-awardees; `random_state=42`), which was not used to grow the trees. This is a random split, not the temporal hold-out of Table 6. Coverage = recall; the rule is treated as a classifier that predicts "awardee" only for researchers who satisfy it.

Rule (Node)	TP	FP	FN	TN	Cov.	Spec.	Prec.	Acc.
Primitive (#38)	56	8	58	92	49.1%	92.0%	87.5%	69.2%
Author count-based (#40)	38	5	76	95	33.3%	95.0%	88.4%	62.1%
Age-based (#27)	53	17	61	83	46.5%	83.0%	75.7%	63.6%
Pub. & citation count-based (#38)	42	7	72	93	36.8%	93.0%	85.7%	63.1%
Combined (#34)	38	5	76	95	33.3%	95.0%	88.4%	62.1%

Table 6. Reported temporal hold-out performance of extracted rules on a balanced set of recent awardees (53 awardees + 53 non-awardees, years 2021–2023, designated for validation; researcher-level non-overlap is not independently verified here). The 95% CI is the Wilson interval for accuracy with $n = 106$.

Parameter Category	Acc.	95% CI (Acc.)	Prec.	Recall	F_1	Spec.
Primitive parameters	55%	45.5–64.1%	0.57	0.48	0.52	0.62
Author count-based	60%	50.5–68.8%	0.63	0.52	0.57	0.68
Age-based parameters	52%	42.6–61.3%	0.53	0.43	0.48	0.60
Pub. & citation count-based [a]	67%	57.6–75.2%	0.64	0.78	0.70	0.57

[a] Accuracy, precision and F_1 were recomputed from the reported recall (0.78) and specificity (0.57), taking the nearest integer counts for 53 + 53 samples (TP = 41, TN = 30). The first version reported 68%, 0.71 and 0.74 for this row, which are not consistent with these two rates. The other three rows agree with their recall and specificity to within about one percentage point. The displayed *TP* and *TN* are an aggregate consistency reconstruction from rounded rates; they are not presented as a newly recovered case-level prediction file.

5.7. External Validation on Held-Out Recent Awardees

To validate the robustness of the extracted decision rules against data not seen during any stage of model training or rule extraction, a held-out external validation set was constructed comprising award recipients and non-awardees from the most recent three years (2021, 2022, and 2023). This external validation set comprised 53 awardees and 53 matched non-awardees (106 total), maintaining the same balanced class distribution as the training data. The reported evaluation protocol designates awards dated 1990–2020 for model development and awards dated 2021–2023 for temporal validation; hence 2021 belongs to the validation period rather than the development period. This separation is a description of the intended protocol, not a verified researcher-level overlap audit. In particular, the supplied manuscript does not provide the cohort identifiers needed to independently establish whether all 2021 observations and repeat award recipients were assigned exclusively to the validation cohort. Accordingly, the reported results are described as temporal hold-out results, with strictly independent cohort membership requiring confirmation from the original data.

Under the stated non-overlapping year-window protocol, this temporal hold-out assesses performance on award years later than those designated for development. Researcher-level separation must nevertheless be confirmed before describing the set as fully independent external validation. The use of recent awardees (2021–2023) is particularly stringent, as citation profiles of early-career or recently recognized researchers may differ systematically from those of the longer-established awardees who dominate the training set, providing a conservative and meaningful test of rule generalizability.

The reported evaluation applied the category-specific rules to the temporal validation sample. The available summary provides accuracy, precision, recall (sensitivity), F1-score, and specificity, as presented in Table 6. Table 6 summarizes evaluation at the category level on the stated balanced hold-out (53 awardees and 53 non-awardees). For a binary classifier, recall equals $TP/53$ and specificity equals $TN/53$; thus these two rates determine the corresponding aggregate accuracy, precision and F1 once integer counts are specified. This arithmetic consistency check makes the reported summary transparent without implying that the original per-researcher predictions have been regenerated. Because rounded rates can admit approximate rather than uniquely verified counts, the reconstructed row is identified in the table note and should be interpreted at the aggregate level. A case-level prediction archive would permit more complete independent verification, but the absence of that additional artifact does not by itself establish an error in the reported aggregate evaluation.

Rules based on publication and citation count-based parameters achieved the strongest performance, with an accuracy of 67% (95% CI 57.6–75.2%), recall of 0.78, specificity of

0.57, precision of 0.64 and F1-score of 0.70. Primitive parameter rules achieved an accuracy of 55% (95% CI 45.5–64.1%) and age-based rules 52% (95% CI 42.6–61.3%); both intervals include the 50% chance level. Author count-based rules achieved 60% (95% CI 50.5–68.8%), only marginally above chance. With 106 researchers the intervals are wide, so differences between categories such as 55% and 60% are not statistically distinguishable, and only the publication and citation count-based rules are clearly above chance. A trivial classifier that labels every researcher as an awardee would obtain 100% recall, 50% accuracy, 50% precision and 0% specificity on this balanced set.

Training-partition values (Table 4) and held-out values cannot be compared directly, but they differ in a way that matters. For the publication and citation count-based category the highlighted rule had a specificity of 92.2% on the training partition, whereas the held-out specificity was 0.57. This drop indicates that the training-partition values are optimistic and that the rules do not transfer to later awardees without loss. Held-out sets from a single time window are also sensitive to changes in citation practices and to the matching of non-awardees.

It is important to acknowledge the scope of this external validation. The held-out set, while temporally separated from the training data, remains within the same disciplinary domain neuroscience and draws from the same five award societies used to construct the training set. Cross-domain validation, for instance evaluating whether the extracted rules and parameter rankings generalise to researchers in Computer Science, Civil Engineering, or Biomedical Engineering, was not conducted in the present study. Such generalisation cannot be assumed: citation norms, co-authorship patterns, and the bibliometric correlates of institutional recognition vary substantially across disciplines [3,10], and rules calibrated on neuroscience awardees may require recalibration before deployment in other domains. A further practical constraint is that constructing equivalent cross-domain validation sets requires access to clean, society-validated ground-truth award lists comparable to those available from SFN, IBRO, FENS, CNS, and ANS a data collection effort that is non-trivial and domain-specific. Cross-domain validation is therefore identified as the primary direction for future work (Section 9), and the present results should be interpreted as neuroscience-specific findings pending replication in other disciplinary contexts.

5.8. Baseline Comparison

To contextualize the discriminative capability of the proposed decision tree framework, the classification accuracy achieved by the decision trees in each parameter category (reported in Section 5) was benchmarked against two standard baseline classifiers: Logistic Regression and Random Forest. This comparison addresses whether the interpretability constraint imposed by the decision tree structure incurs a cost in predictive performance relative to simpler or more powerful alternatives.

Both baseline models were trained and evaluated on the same neuroscience dataset using an identical stratified 80:20 train test split and the same top-five parameters per category as used for the decision trees. Logistic Regression features were standardized using z-score normalization prior to training (L2 regularization, $C = 1.0$). Random Forest comprised 100 estimators with default hyperparameters. A fixed random seed (42) was applied throughout to ensure reproducibility. Performance was evaluated using accuracy, precision, recall, F_1-score, and AUC-ROC. The complete comparison is presented in Table 7.

Table 7 must be read with an important caveat. The Decision Tree accuracies were obtained in the rule-mining runs, whereas Logistic Regression and Random Forest were evaluated on a separate stratified 20% test split. The two sets of numbers come from different data partitions and evaluation procedures, and the accuracy of a tree on the data used

to grow it is optimistic. We therefore do not claim that the Decision Tree outperforms the baselines.

Table 7. Classification performance of Decision Tree, Logistic Regression and Random Forest per parameter category (top-5 features). Logistic Regression and Random Forest were evaluated on a stratified 20% test split. † Decision Tree accuracies are those of the rule-mining runs in Section 5 (training partition); they come from a different partition and procedure and are *not* comparable with the other rows.

Category	Model	Acc.	Prec.	Recall	F_1	AUC
Primitive	Decision Tree †	78.0%	—	—	—	—
	Logistic Regression	59.8%	0.632	0.453	0.527	0.655
	Random Forest	74.8%	0.783	0.679	0.727	0.788
Author count-based	Decision Tree †	73.0%	—	—	—	—
	Logistic Regression	58.9%	0.641	0.387	0.482	0.641
	Random Forest	68.7%	0.714	0.613	0.660	0.769
Publication age-based	Decision Tree †	68.0%	—	—	—	—
	Logistic Regression	65.0%	0.678	0.557	0.611	0.712
	Random Forest	65.0%	0.678	0.557	0.611	0.709
Pub. & citation count-based	Decision Tree †	71.0%	—	—	—	—
	Logistic Regression	59.3%	0.594	0.566	0.580	0.652
	Random Forest	69.6%	0.699	0.679	0.689	0.779
Combined	Decision Tree †	78.0%	—	—	—	—
	Logistic Regression	62.6%	0.620	0.632	0.626	0.691
	Random Forest	71.5%	0.737	0.660	0.697	0.786

What the table does show is that, on a held-out split, Random Forest reached 65.0–74.8% accuracy (AUC 0.709–0.788) and Logistic Regression 58.9–65.0% (AUC 0.641–0.712). For reference, the rule-based classifiers reached 52–67% on the temporal hold-out (Table 6), which is not higher than these test-split values, although the two evaluations use different data. Whether the interpretability constraint carries a cost in predictive performance therefore remains an open question. In the age-based category the two baselines obtained identical test metrics (65.0%).

To compare the models under equivalent conditions, we re-evaluated everything with repeated stratified 5-fold cross-validation [32] (Table 8). In each training fold all 64 parameters were ranked and the top five per category were selected, so the test folds play no role in feature selection; all models were trained and tested on identical folds. Under this protocol the pruned decision trees reached 66–70% accuracy (AUC 69–72%), Random Forests on the same features 68–73% (AUC 73–79%), and logistic regression 58–68%. Trees were therefore within about 1–4 points of Random Forests and ahead of logistic regression in four of five parameter sets (all except the age-based set), but they were not better than Random Forests in any set, so the earlier statement that the tree outperforms the baselines is not supported. Using all 64 parameters gave higher accuracy for every model (Random Forest 75.4%, AUC 82.4%), and single-index stumps reached only 58.9% (h index) and 62.5% (Total Citation). The restriction to five parameters per category therefore costs predictive accuracy and is justified by interpretability, not by performance.

We also compared feature rankers by the accuracy of a decision tree on the combined top-five set: MLP-FIS 68.0%, Random Forest Gini importance 71.5%, permutation importance 70.6%, mutual information 69.9% and L1-logistic coefficients 67.0%. Pooled over the five parameter sets, the decision-tree accuracy with MLP-FIS selection was lower than with Gini importance (Wilcoxon over folds, $p = 0.005$, mean difference 1.4 points; folds overlap across repeats, so p-values are approximate) and not distinguishable from the other three rankers; in the combined set alone the difference to Gini importance was

3.4 points ($p = 0.026$). The selected sets were also unstable: across the 15 training folds the mean pairwise Jaccard overlap of the MLP-FIS top-five sets was 0.47 (primitive), 0.47 (age), 0.21 (author count) and 0.09 (publication and citation count), compared with 0.96, 0.74, 0.62 and 0.66 for Gini importance, and the MLP-FIS sets overlapped the top-five sets reported in Section 4 by only 0.9–3.1 of 5 parameters. The MLP-FIS rankings in Section 4, obtained from a single partition, should therefore not be regarded as stable, and we do not claim an advantage of MLP-FIS over conventional rankers.

Table 8. Common-protocol evaluation: repeated stratified 5-fold cross-validation (3 repeats, 15 folds). In each training fold the parameters were ranked with MLP-FIS and the top five per category were used; all models see identical folds. Values are mean ± SD of the fold accuracy (%), with AUC (%) in parentheses. The last rows use all 64 parameters, and a one-split stump on a single index.

Parameters	DT (Pruned)	DT (Depth ≤ 3)	Logistic Reg.	Random Forest
Primitive (top 5)	70.0 ± 3.1 (71.8)	70.4 ± 3.2 (72.0)	57.5 ± 3.1 (63.6)	73.3 ± 3.4 (79.4)
Pub. & citation count-based (top 5)	67.7 ± 3.2 (70.5)	67.6 ± 3.3 (70.3)	65.3 ± 4.7 (70.5)	69.0 ± 2.7 (75.7)
Age-based (top 5)	67.2 ± 5.2 (69.0)	66.5 ± 4.2 (69.0)	67.9 ± 4.3 (72.7)	70.7 ± 4.6 (76.9)
Author count-based (top 5)	66.0 ± 3.9 (69.3)	64.4 ± 3.9 (67.4)	64.1 ± 4.0 (69.5)	67.8 ± 3.1 (73.0)
Combined (top 5)	68.0 ± 3.4 (69.8)	67.7 ± 3.1 (69.8)	63.9 ± 4.7 (68.0)	69.4 ± 3.0 (76.4)
All 64 parameters	70.7 ± 3.4 (72.3)	72.8 ± 2.5 (76.8)	72.0 ± 2.4 (79.1)	75.4 ± 1.9 (82.4)
Single index: h index (stump)		58.9 ± 2.8		
Single index: Total Citation (stump)		62.5 ± 3.9		

We note that precision, recall, F_1-score, and AUC-ROC for the Decision Tree rules are reported separately in the validation experiment (Table 6, Section 5.7), where the rules are evaluated on a balanced held-out set containing both awardees and non-awardees.

5.9. Sensitivity to Citation Timing

Citations that accrue after the award year can inflate only citation-based indicators; counts of publications, career span and the number of authors per paper are not affected. As a partial check on the leakage described in Section 3, we therefore compared models that use only these four citation-independent indicators with models that use the 60 citation-based indicators (Table 9).

With the four citation-independent indicators a Random Forest reached 61.0% accuracy (AUC 66.4%), a pruned tree 58.8% and logistic regression 53.9%, whereas the citation-based indicators reached 74.4% (Random Forest), almost the full 75.4% obtained with all 64 indicators. About 43% of the accuracy above chance of the full Random Forest model ($(61.0 - 50)/(75.4 - 50)$) is thus available without any citation count, so the discrimination is not entirely produced by a possible post-award citation boost. However, most of the signal comes from citation-based indicators, which cannot be separated from such a boost with the present data. This check bounds the problem but does not replace the analysis with year-resolved citation data listed in Section 9.

Table 9. Sensitivity to citation timing: cross-validated accuracy (%, mean ± SD over 15 folds) with AUC (%) in parentheses, using indicators that do not depend on citation counts (Total Publication, Total Years, Author per Paper, Papers per Author), the remaining 60 citation-based indicators, and all 64 indicators. Same folds and models as in Table 8.

Indicators Used	DT (Pruned)	Logistic Reg.	Random Forest
Citation-independent (4)	58.8 ± 2.5 (61.3)	53.9 ± 2.1 (56.1)	61.0 ± 3.1 (66.4)
Citation-based (60)	69.2 ± 2.7 (69.7)	72.1 ± 2.7 (79.4)	74.4 ± 2.2 (82.0)
All 64 indicators	70.7 ± 3.3 (72.3)	72.0 ± 2.3 (79.1)	75.4 ± 1.8 (82.4)

6. Awardee Trends in the Top 100 Researchers

In this section we examine how awardees are distributed among the top-100 researchers when all 1060 researchers are ranked by a single parameter. The two quantities, defined in Section 3.4, are Precision@100 (share of the top-100 list that are awardees) and Recall@100 (share of the 530 awardees that appear in the list). A list of 100 can contain at most 100 awardees, so Recall@100 cannot exceed 100/530 = 18.9%. Under random ranking of this balanced dataset the expected values are a Precision@100 of 50% and a Recall@100 of 50/530 = 9.4%. Figures 13–16 plot Precision@100 for the top-five parameters of each category; Table 10 gives the values together with the implied Recall@100, and the dashed lines mark the 50% chance level.

Table 10. Precision@100 and implied Recall@100 recomputed from the dataset file for the top-five parameters of each category (Section 4), ranking in descending order with the tie-breaking rule of Section 3.4. Recall@100 = n/530 with n the number of awardees in the list. p: nominal (unadjusted) one-sided binomial p-value for $P@100 > 50\%$; significance across the 20 comparisons is assessed using the Bonferroni-adjusted threshold 0.05/20 = 0.0025; values far below 50% mean that awardees tend to have *low* values of the parameter.

Category	Parameter	P@100	R@100	p
Primitive	Total Citation	55%	10.4%	0.18
	Author per Paper	26%	4.9%	>0.99
	Cite/Year	53%	10.0%	0.31
	Cites per Author	60%	11.3%	0.028
	Cite/Paper	71%	13.4%	<0.001
Author count-based	gf index	64%	12.1%	0.003
	Fractional g index	67%	12.6%	<0.001
	hl norm	64%	12.1%	0.003
	hm index	54%	10.2%	0.24
	gF index	37%	7.0%	0.997
Publication age-based	AW index	55%	10.4%	0.18
	m-quotient	50%	9.4%	0.54
	AR index	70%	13.2%	<0.001
	AWCR	51%	9.6%	0.46
	ha index	58%	10.9%	0.067
Pub. & citation count-based	h2 center index	30%	5.7%	>0.99
	Tapered h-index	52%	9.8%	0.38
	Woginger index	56%	10.6%	0.14
	k index	71%	13.4%	<0.001
	Rational h index	51%	9.6%	0.46

We recomputed the top-100 analysis from the dataset file for the top-five parameters of each category (Table 10 and Figures 13–16). Seven of the 20 parameters place significantly more awardees than chance in the top 100 (nominal $p < 0.05$): Cite/Paper and the k index (71%), the AR index (70%), the fractional g index (67%), the gf index and hl norm (64%) and cites per author (60%). To control the family-wise error rate across the 20 comparisons, we applied the Bonferroni correction ($\alpha = 0.05/20 = 0.0025$). The p-values displayed in Table 10 are nominal (unadjusted); seven comparisons meet the nominal $p < 0.05$ criterion, whereas only four remain significant at the adjusted threshold: Cite/Paper, k index, AR index, and fractional g index. Statistical significance is interpreted using the adjusted threshold. Total Citation (55%) is not distinguishable from chance. Three parameters fall far *below* chance (Author per Paper 26%, h2 center index 30%, gF index 37%): in descending order their top-100 lists contain mostly non-awardees, so for these parameters awardees tend to have lower values. This includes the h2 center index, which had the highest MLP-FIS importance score (0.23) but is a poor ranker in this analysis. Single-index rankings are in general weak and parameter-specific discriminators of award recipients.

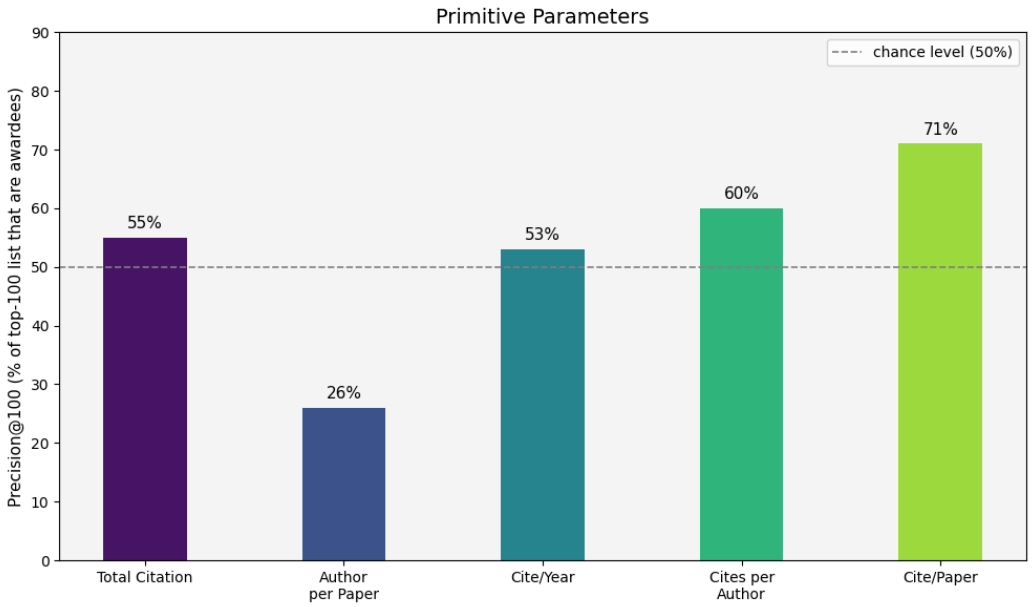

Figure 13. Precision@100 for the top-five primitive parameters (share of the top-100 list that are awardees; dashed line: 50% chance level; values in Table 10).

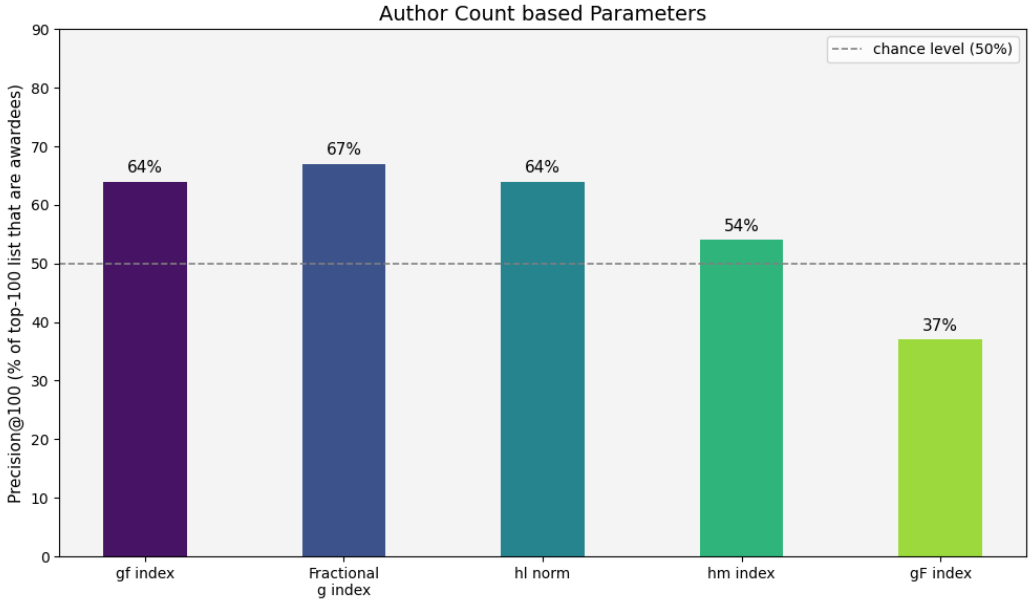

Figure 14. Precision@100 for the top-five author count-based parameters (dashed line: 50% chance level; values in Table 10).

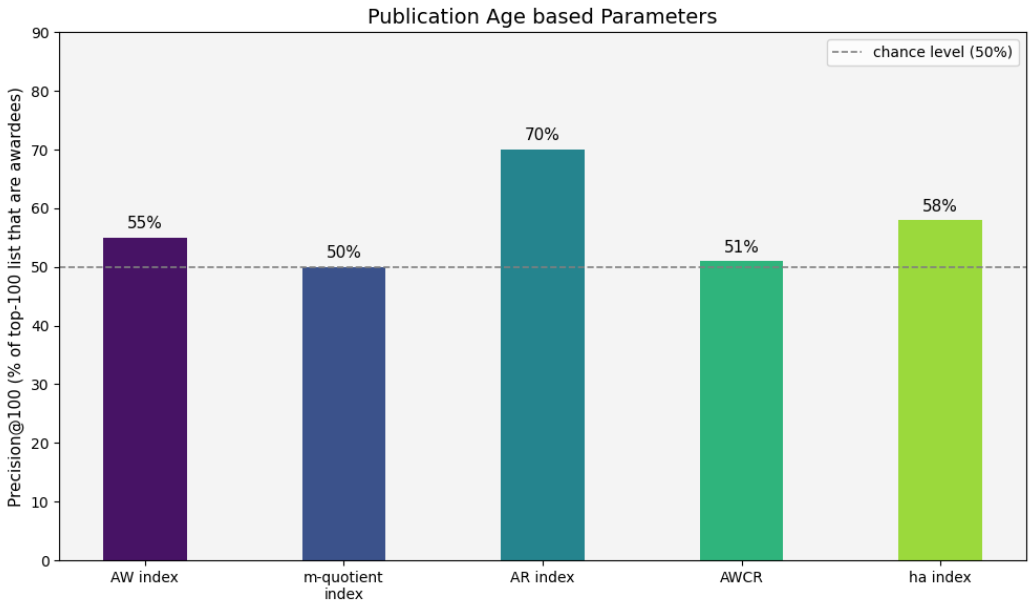

Figure 15. Precision@100 for the top-five publication age-based parameters (dashed line: 50% chance level; values in Table 10).

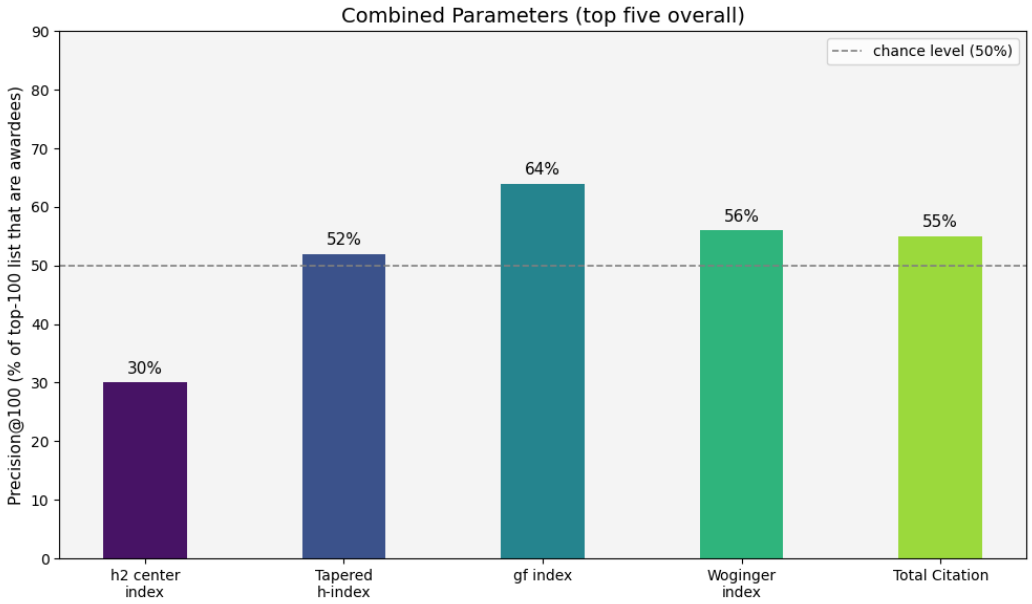

Figure 16. Precision@100 for the top-five parameters of the combined ranking (dashed line: 50% chance level; values in Table 10).

7. Discussion

This study presents a comprehensive, explainable framework for examining the relationship between bibliometric profiles and institutionally conferred recognition in neuroscience by integrating parameter ranking, predictive modeling, benchmarking, and rule mining. The central question is not whether the framework can identify scientific influence in a general sense, but how far citation-based and related bibliometric signals account for an award outcome and what transparent patterns can be extracted from that relationship.

The parameter ranking results show that citation intensity, author normalization, publication productivity, and career-age information all contain recognition-related signal, but no category provides a complete account of award outcomes. Citation-related metrics rank highly across several analyses, indicating that award recognition is associated with citation-based visibility. However, the moderate held-out performance and the competitive performance of strong single indicators show that these associations should not be interpreted as a comprehensive measure of scientific influence. Author-normalized indicators such as the normalized hI-index scored well in the top-100 analysis, although this index ranked last among the author count-based parameters in the MLP-FIS ranking, so the evidence for its superiority is mixed.

While the association between citation impact and award recognition is broadly established in the bibliometric literature, the present study contributes three exploratory observations that sharpen the interpretation of this relationship in neuroscience. First, the category-aware feature ranking identifies which specific variants within the h-index family carry recognition-related signal beyond raw totals, although instability across folds cautions against treating any particular ranking as definitive. Second, the rule-mining phase exposes explicit conditional combinations—for example, conjunctions involving Citations per Year and the AR index—that cannot be obtained from a univariate ranking alone. These thresholds are descriptive partitions learned from the present sample, not actionable criteria or causal mechanisms. Third, the benchmark analyses show that such multivariate structure yields limited incremental predictive benefit over strong individual indicators and standard classifiers. This negative or modest incremental result is substantively informative: it suggests that much of the measurable association with awards is already contained in citation-based visibility, while a substantial part of institutional recognition remains outside the bibliometric feature space. In the recomputed top-100 analysis (Table 10), the fractional g index (67%), the gf index and hl norm (64%) and the AR index (70%) placed more awardees in the top 100 than raw Total Citation (55%, not distinguishable from chance), but the differences between neighbouring indices are small relative to the sampling variation of a list of 100.

Beyond confirming known patterns, the results reveal several non-trivial findings that merit explicit discussion. First, the h2 center index which captures citation density around the h-threshold ranks above the standard h-index among the 64 parameters in the single-partition MLP-FIS ranking (importance score 0.23), but it is a poor ranker in the top-100 analysis (30%) and the top-five sets were unstable across folds (Section 5.8), despite the h-index being the dominant metric in practice. This suggests that recognized neuroscientists are distinguished not by the breadth of their h-core but by the citation intensity within it: a few landmark contributions cited far above the h-threshold, rather than a large number of moderately cited papers. Institutions relying solely on the h-index may therefore systematically undervalue researchers whose influence is concentrated in a smaller number of high-impact works. Second, the rule in Equation (15) combines a very high k-index (above about 1800) with a Woginger index of at most 189. A speculative reading is that this describes a citation distribution concentrated in relatively few highly cited papers. This has not been tested with distribution-shape measures, and the reading remains a hypothesis. Third, the highlighted rules trade coverage against precision. On the training partition the combined-parameter rule has the highest specificity (97.5%) and precision (88.9%) but the lowest coverage (21.2%), whereas the primitive and age-based rules cover 44–45% of training awardees at lower specificity (92.9% and 78.5%). Because each rule describes a single path, several complementary rules would be needed to cover most awardees; such rule sets were not evaluated.

Single-parameter accuracies were modest (up to 66.4% among the top-five parameters and 70.5% among all 64 indicators) and similar across categories, and rule-level accuracy on the held-out set was highest for the publication and citation count-based rules (67%) and not distinguishable from chance for the primitive and age-based rules (Table 6). We therefore do not conclude that any one category plays a decisive role in distinguishing awardees from non-awardees.

The rule mining phase translates model behavior into human-readable decision rules. The highlighted rules cover 21–45% of training awardees at 78–98% specificity (Table 4) and 33–49% of awardees at 83–95% specificity on a random 20% test split (Table 5), and their thresholds are explicit and can be inspected. The upper bounds in the rules arise from the tree partition and should not be read as penalizing high values. Validation on recent awardees (Table 6) shows a clear drop in specificity relative to the training partition, so the rules do not transfer without loss, and with only 106 held-out researchers the evidence on generalization is limited. It is important to acknowledge two structural bias considerations that qualify the interpretation of these results. First, the five award societies used to define the awardee class (SFN, IBRO, FENS, CNS, and ANS) may systematically favor researchers from well-resourced institutions with high publishing output. Consequently, the extracted rules may partly reflect institutional privilege alongside individual scholarly contribution, rather than separating the two. Second, bibliometric data were retrieved from Google Scholar without filtering for self-citations. A portion of the citation signal may therefore reflect field-specific citation norms or collaborative citation practices rather than broad community-wide impact. Both constraints are inherent to award-based bibliometric designs and apply equally to prior work in this area. They do not invalidate the comparative framing of the framework, but they appropriately qualify the extracted rules as benchmarks for institutionally recognized influence rather than ground-truth measures of scientific merit. Future work incorporating self-citation filtering and institutional stratification would help disentangle these confounds.

Two further analyses qualify the interpretation. The 64 parameters are highly redundant (Section 4.2): three principal components explain 90% of the variance, which helps to explain why the selection of five indices by MLP-FIS was unstable and why single indices of the same group perform similarly. Indicators that do not depend on citation counts discriminate only modestly (about 61% accuracy for a Random Forest, Section 5.9), so most of the signal lies in citation-based indicators that are exposed to post-award citation accrual.

Overall, the results support a narrower and more consequential interpretation than "identifying influential researchers." Bibliometric indicators capture citation-based visibility and related career characteristics that are associated with institutional recognition, but these associations are only partial. The residual gap is not simply model error: it is consistent with award decisions being socially and institutionally constructed outcomes that may also reflect disciplinary hierarchies, professional networks, institutional prestige, cumulative advantage, committee priorities, and other factors not represented in the present data. The framework is therefore most useful for making the bibliometric component of recognition transparent and for demonstrating its limits, not for reducing recognition or scientific influence to a classification score.

7.1. Limitations and Threats to Validity

The findings should be read with the following limitations in mind.

- *Ground truth.* Award receipt reflects committee judgment and is shaped by visibility, networks, geography and cumulative advantage [13–15]. A model trained on it partly learns those selection effects.

- *Temporal leakage.* Citations accrued after the award year to pre-award publications are included (Section 3). Leakage of this kind can inflate apparent performance [33]; the reported discrimination is therefore an upper bound for strictly pre-award information, although indicators that do not depend on citation counts alone reach about 61% accuracy (Section 5.9). Consequently, the present results should not be interpreted as fully prospective predictions of award recognition. A strictly prospective assessment would require year-resolved citation histories truncated at each corresponding evaluation date.
- *Sampling and prevalence.* The 1:1 sample is matched by award year only and is not representative of an operational population. As an illustration, a rule with recall 0.78 and specificity 0.57 (Table 6) would have a precision of only about 1.8% at an award prevalence of 1%. Precision-recall evaluation under realistic prevalence is needed [34].
- *Ranking and selection procedure.* Feature ranking used a single 80:20 partition and a single initialization, with 64 correlated features and roughly 850 training samples. Feature selection on data that are later used for evaluation inflates estimates [35], and the MLP-FIS top-five sets proved unstable across folds (Section 5.8). Leave-one-out ablation can understate the importance of redundant indices.
- *Evaluation.* The temporal-validation results are reported as aggregate metrics for a balanced cohort; for the indicated row, a consistency check reconstructs approximate integer counts from rounded recall and specificity. This makes the arithmetic auditable but does not substitute for releasing case-level predictions or original confusion matrices, which would further strengthen independent replication. Coverage and specificity of the highlighted rules are resubstitution estimates. The temporal hold-out has only 106 researchers, so confidence intervals are wide, and the baseline comparison in Table 7 is not like-for-like. The top-100 ranking analysis (Section 6) reports nominal (unadjusted) one-sided binomial p-values for 20 comparisons, but evaluates significance using the Bonferroni-adjusted family-wise threshold ($0.05/20 = 0.0025$). Seven indicators satisfy the nominal $p < 0.05$ criterion, whereas four remain significant after correction. Consequently, claims of statistical significance are based on the adjusted threshold, although the exploratory analysis and dependence among correlated indicators still warrant cautious interpretation.
- *Data quality and scope.* Google Scholar data were not filtered for self-citation. The published analysis does not document a separate pre-selection exclusion of the 149 records with anomalous Total Years values; therefore, the reported feature rankings and tree paths should be interpreted as results of the stated analysis dataset rather than as results from a verified anomaly-free subset. A limited AUC sensitivity check is available, but no claim is made that the exact age-based tree thresholds were re-estimated on a cleaned subset. The Total Years field contains implausible values (zero or above 100) for 149 researchers. These values directly undermine career-duration-dependent quantities such as Citations per Year and the m-quotient; the extent to which other publication age indicators are affected depends on their calculation from publication-level dates and cannot be established from the Total Years field alone. The reported exclusion sensitivity check changes single-index AUCs by less than 0.02, indicating limited changes in aggregate single-index discrimination, but does not test the stability of the age-based tree splits or their numerical thresholds. In particular, the age-based rule in Equation (14) should be interpreted as an exploratory, data-dependent description, not a validated career-age criterion; recomputation after auditing the underlying dates would be needed to establish threshold stability. All results concern neuroscience and five award societies. The numerical decision-tree thresholds are therefore data- and domain-dependent descriptors rather than universal bibliometric cutoffs; their stability

and transportability require independent validation across disciplines, institutions, award systems and time periods.

7.2. Implications and Responsible Use

The results speak mainly to how recognition relates to bibliometric profile. The modest separation between awardees and matched non-awardees on held-out data indicates that these indicators capture only a limited share of the observed differences associated with recognition, which is consistent with the literature on selection and cumulative advantage. The pipeline does not address how long before an award a researcher crosses the rule thresholds (recognition lag) or whether different societies reward different profiles; both questions require year-resolved indicators and society-level analysis and are left for future work.

Interpretability does not by itself make a model reliable or fair. Before any use in evaluation, a rule set would need calibrated uncertainty estimates, an audit for subgroup differences (for example by region, gender and institution type), and a governance process in which human judgment retains the final say [10,36]. Work on decision systems that integrate predictive performance, uncertainty and governance offers a useful perspective for such a process [37]. In particular, the rules are not suitable as a standalone basis for hiring, promotion or funding decisions; their appropriate uses are exploratory benchmarking, self-assessment context for researchers and the study of recognition patterns.

8. Conclusions

The assessment of scientists' achievements has gained increasing importance over the past two decades, driven by the need for objective, transparent, and data-driven mechanisms for evaluating scholarly impact. To support decision-making processes such as researcher ranking, award nomination, and academic recognition, numerous bibliometric parameters have been proposed in the literature. However, the relative effectiveness and interpretability of these parameters remain an open challenge, particularly within collaborative and data-intensive research domains such as neuroscience.

In this study, we proposed an explainable machine learning framework to assess how bibliometric profiles relate to award recognition across four distinct categories: primitive parameters, publication and citation count-based parameters, author count-based parameters, and publication age-based parameters. The primary objective was to identify which bibliometric signals are associated with institutional recognition, determine whether their combinations provide information beyond strong individual indicators, and express those associations in a transparent form. A multilayer perceptron classifier combined with a leave-one-out feature ablation strategy was employed to compute importance scores and rank parameters within each category, followed by interpretable tree-based rule extraction and benchmark comparison.

The experimental results indicate that citation-related indicators rank highly in the MLP-FIS ranking: total citation count in the primitive category, the h2 center index and tapered h-index in the publication and citation count-based category, the gf-index in the author count-based category, and the AW-index in the age-based category. Discriminative power was moderate: single-parameter accuracies reached at most 70.5% across all 64 indicators, and rule-level accuracy on a temporally held-out set was 52–67%, with only the publication and citation count-based rules clearly above chance; under repeated cross-validation, interpretable trees reached 66–70% accuracy against 68–73% for Random Forests on the same features and 75% for a Random Forest on all 64 parameters. The evidence is exploratory and is limited by residual citation-timing leakage, the dependence on award selection and the small held-out sample.

Beyond parameter ranking, the rule-mining phase produced transparent, threshold-based patterns that reveal how combinations and ranges of bibliometric indicators co-occur with award recognition. This information differs from a single indicator or feature ranking because it makes conditional multivariate profiles inspectable; however, it does not establish causal mechanisms and did not consistently improve prediction over established alternatives. The rules are therefore best understood as descriptive characterizations of the bibliometric component of institutional recognition. Award receipt is not treated as an unproblematic proxy for scientific influence: it is a socially and institutionally conferred outcome that may reflect citation visibility, disciplinary hierarchies, professional networks, cumulative advantage, institutional prestige, and committee-specific selection processes. Consequently, the framework identifies correlates of recognized visibility, not determinants of scientific merit.

The principal contribution is therefore not an alternative classification procedure that outperforms established methods. Rather, it is a transparent empirical account of what bibliometric information is and is not associated with in relation to institutional recognition. The framework shows that interpretable multivariate rules can expose recognizable combinations of indicators, while the benchmark results simultaneously show that their incremental predictive value is limited. This combination of transparency and negative comparative evidence is substantively important for Computational Social Science because it separates citation-based visibility from the broader social process through which recognition is conferred. The current rules should not be used for preliminary screening or other consequential assessment of individuals; their most defensible uses are methodological analysis, exploratory benchmarking, and the study of recognition processes in scientometrics.

Future work should extend this framework to additional scientific domains and award systems to test whether the observed distinction between citation-based visibility and institutional recognition is stable across contexts. It should also incorporate self-citation filtering, year-resolved citation histories, institutional stratification, and—where ethically and practically appropriate—non-bibliometric variables related to disciplinary position, professional networks, mentoring, interdisciplinary reach, and other contextual dimensions of recognition. Such extensions would help determine which components of award outcomes are associated with bibliometric visibility and which are related to broader social and institutional processes.

9. Future Work

The most important next steps address the limitations identified above. (i) Recompute all indicators from year-resolved citation data truncated at the award year, to remove citation-timing leakage and permit a genuinely prospective evaluation. (ii) Evaluate the rules under a unified protocol with repeated stratified cross-validation, feature ranking inside each training fold, identical folds for all models, confidence intervals and paired tests, and compare MLP-FIS with permutation importance, SHAP values [38], mutual information and Random Forest importance [39]. (iii) Report precision and recall under realistic prevalence. (iv) Add a second operationalization of influence (for example highly cited researcher lists or other major prizes) to test whether the findings are specific to these societies. (v) Analyze recognition lag and society-level differences using year-resolved indicators, audit subgroup differences, and test whether the extracted parameter rankings and numerical thresholds remain stable across disciplines, institutions, award systems and time periods.

Several promising directions can be explored to extend the scope and applicability of this research. First, future studies will expand the proposed framework to multiple scientific domains, such as Computer Science, Civil Engineering, and Biomedical Engineer-

ing, to evaluate the generalizability and domain sensitivity of the identified parameters. Cross-domain analysis will help determine whether parameter importance patterns remain stable or vary across disciplinary cultures.

Second, there is a need to design novel bibliometric indices that more comprehensively capture an author's lifetime scholarly contributions. Future metrics may incorporate factors such as long-term productivity trends, collaboration influence, mentorship activities, and interdisciplinary impact, which are not fully addressed by existing indicators.

Third, advanced deep learning architectures will be explored to further enhance predictive accuracy while maintaining interpretability. Hybrid models that integrate neural networks with explainable components may provide improved performance without sacrificing transparency.

Additionally, future work will consider incorporating subjective and contextual factors into scientometric evaluation. For instance, recently proposed indicators such as the Q-factor account for elements including time dynamics, project contribution, dissertation supervision, and funding impact. Integrating such factors will require the construction of richer and more diverse datasets.

Finally, future research will address broader questions in scientometrics, including the feasibility of deploying these metrics in digital libraries, their role in supporting fair and transparent evaluation systems, and whether hybrid quantitative–qualitative indicators can better assist subjective decision-making processes in academic and research institutions.

Author Contributions: Conceptualization, G.M. and M.T.A.; methodology, G.M.; software, G.M.; validation, G.M., M.T.A., H.A., A.S.A.-S., I.D.l.T.D., M.S.G.d.M. and E.S.A.; formal analysis, G.M. and H.A.; investigation, G.M.; resources, G.M.; data curation, G.M. and H.A.; writing—original draft preparation, G.M.; writing—review and editing, M.T.A., H.A., A.S.A.-S., I.D.l.T.D., M.S.G.d.M. and E.S.A.; visualization, G.M.; supervision, M.T.A.; project administration, G.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset associated with this study is publicly available in an online repository. The data have been uploaded to the Dr-GMustafa repository and can be accessed at https://github.com/Dr-GMustafa/research-datasets.git (accessed on 11 November 2026). The complete source code for data extraction and parameter computation is available at https://github.com/ghulammustafacomsat/dataextraction (accessed on 11 November 2026), and the ranking and Recall@100 computation scripts are available at https://github.com/Dr-GMustafa/research-datasets (accessed on 11 November 2026).

Acknowledgments: During the preparation of this manuscript, the authors used Grammarly, ChatGPT (GPT-5, OpenAI) and Claude (Sonnet 4.5, Anthropic) (Anthropic) solely for the purpose of improving the English language, including grammar, spelling, clarity and readability of the text. These tools were not used to generate scientific content, study design, data, analysis, results or references. The authors reviewed and edited all AI-assisted output and take full responsibility for the content of the publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SFN	Society for Neuroscience
FENS	Federation of European Neuroscience Societies
CNS	Cognitive Neuroscience Society

ANS	Australasian Neuroscience Society
IBRO	International Brain Research Organization
MLP	Multilayer Perceptron
MLP-FIS	MLP-based Feature Importance Scoring
RFE	Recursive Feature Elimination
ReLU	Rectified Linear Unit
MSE	Mean Squared Error
AR	Age-adjusted h-index
AW	Age-weighted h-index
AWCR	Age-Weighted Citation Rate
AUC	Area Under the Curve
BA	Baseline Accuracy
RFA	Remove Feature Accuracy
IS	Importance Score

Appendix A. Bibliometric Indicator Definitions

Table A1. Concise computational definitions and source reference identifiers. Concise computational definitions and source reference identifiers. Category abbreviations: PP, Primitive Parameters; PCP, Publication and Citation Count-Based Parameters; ACP, Author Count-Based Parameters; AP, Age-Based (Publication Age-Based) Parameters.

Category	Indicator	Definition/Computational Interpretation	Ref.
PP	Total publications	$P = \lvert\mathcal{P}\rvert$ (number of included papers)	[40]
PP	Total citations	$C = \sum_{i=1}^{P} c_i$	[40]
PP	Total years	Assessment year minus first publication year (per source convention)	[20]
PP	Citations/year	C/L, where L is career length	[20]
PP	Citations/paper	C/P for $P > 0$	[20]
PP	Authors/paper	$P^{-1} \sum_i a_i$ for $P > 0$	[20]
PP	Citations/author	$\sum_i c_i/a_i$	[20]
PP	Papers/author	$\sum_i 1/a_i$	[20]
PCP	h-index	$h = \max\{k : c_k \geq k\}$	[8]
PCP	g-index	$g = \max\{k : \sum_{i=1}^{k} c_i \geq k^2\}$	[41]
PCP	e-index	$e = \sqrt{\sum_{i=1}^{h} c_i - h^2}$	[42]
PCP	h-core citations	$C_h = \sum_{i=1}^{h} c_i$	[20]
PCP	A-index	$A = C_h/h$ for $h > 0$	[43]
PCP	R-index	$R = \sqrt{C_h}$	[44]
PCP	P-index	Combine total citations and mean citations per paper via the p-index power rule	[40]
PCP	m-index (h-core median)	$m = \text{median}(c_1, \ldots, c_h)$	[45]
PCP	f-index	Rank threshold based on harmonic mean of citations among leading papers	[46]
PCP	T-index	Rank threshold using geometric mean citations among leading papers	[46]
PCP	Q2-index	$Q_2 = \sqrt{h\,m}$	[47]

Table A1. *Cont.*

Category	Indicator	Definition/Computational Interpretation	Ref.
PCP	Tapered h-index	Sum tapered, rank-dependent credit for cited papers	[48]
PCP	Maxprod	$\max_{1\le i\le P}(ic_i)$	[49]
PCP	Wu-index	Largest w supported by at least w papers each with 10w citations	[50]
PCP	Pi-index	One hundredth of citations accumulated by the leading square-root-of-P papers	[51]
PCP	Weighted h-index	Citation-weighted ranking variant of the h-core	[52]
PCP	h(2)-index	$h^{(2)} = \max\{k : c_k \ge k^2\}$	[53]
PCP	Woeginger index	Largest w meeting ordered rank-dependent citation thresholds	[54]
PCP	gh-index	Generalized h-type threshold depending on the h-core profile	[55]
PCP	Rm-index	Aggregate square-root-transformed citations in the h-core	[56]
PCP	X-index	Maximum of rank-weighted citations under the x-index definition	[57]
PCP	hg-index	$h_g = \sqrt{hg}$	[58]
PCP	h2 upper index	Share of total citations above the h-core square contribution	[59]
PCP	h2 center index	Share of total citations corresponding to h squared	[59]
PCP	h2 lower index	Share of total citations in the h-tail	[59]
PCP	k-dash index	Ratio combining publication and h-core/h-tail citation totals	[60]
PCP	Rational h-index	Interpolate between adjacent integer h values based on missing citations	[61]
PCP	Real h-index	Real-valued h interpolation using citations of ranks h and h + 1	[62]
PCP	i10-index	$i10 = \lvert\{i : c_i \ge 10\}\rvert$	[63]
PCP	Normalized h-index	h/P for $P > 0$	[64]
PCP	k-index	Ratio constructed from h-tail and h-core citation shares	[65]
PCP	w-index	Count of highly cited papers meeting its rank-specific threshold	[66]
PCP	h-dash index	h-type adjustment based on head-to-tail excess-citation ratio	[67]
ACP	hI-index	h-index corrected by mean authorship in the h-core	[68]
ACP	Normalized hI-index	Author-normalized h measure; source notation must be checked against code	[69]
ACP	hm-index	Fractionally weighted paper-rank h-type threshold	[70]
ACP	gm-index	g-type index computed using effective fractional publication ranks	[71]
ACP	hf-index	Citation-adjusted h threshold accounting for each paper's authors	[72]
ACP	gf-index	g-type cumulative citation threshold adjusted for coauthor count	[72]

Table A1. *Cont.*

Category	Indicator	Definition/Computational Interpretation	Ref.
ACP	gF-index	Generalized fractional authorship/citation g-type threshold	[72]
ACP	k-normalized index	k-type adjustment to author-normalized h-core measures	[73]
ACP	w-normalized index	w-type adjustment using author-normalized total citation counts	[73]
ACP	Pure h-index	h-index discounted by effective author count	[74]
ACP	Fractional g-index	g-threshold computed from citations divided by number of coauthors	[72]
ACP	Fractional h-index	h-threshold computed from citations divided by number of coauthors	[72]
ACP	hI-norm index	h-index variant using paper-level authorship normalization	[75]
AP	Platinum h-index	Combine h-index and citation intensity with career-length adjustment	[76]
AP	m-quotient	h/L for career length $L > 0$	[77]
AP	AW-index	Square root of age-weighted citation accumulation	[78]
AP	AR-index	Square root of age-weighted citations restricted to h-core	[44]
AP	V-index	h-index adjusted by career-year publication output	[79]
AP	Ha-index	Largest threshold supported by papers' mean yearly citations	[80]
AP	Hc-index	h-type threshold using citations discounted by publication age	[64]
AP	AWCR	$\sum_i c_i/\text{age}_i$; age convention as in code	[81]

Reconciliation note. The source appendix contains 62 distinctly enumerated indicator entries in the grouping reproduced here (8 primitive, 33 publication/citation, 13 author-count, 8 age-based). The current manuscript enumerates only 61 indicator names (8, 33, 12, 8 respectively), despite referring to 64 computed features. Therefore, the missing or differently named features cannot be verified from these documents alone. Please reconcile the manuscript's 64-column code/data inventory before asserting that the appendix table fully documents 64 distinct implemented metrics.

References

1. Bornmann, L.; Mutz, R. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *J. Assoc. Inf. Sci. Technol.* **2015**, *66*, 2215–2222. [CrossRef]
2. Fortunato, S.; Bergstrom, C.T.; Börner, K.; Evans, J.A.; Helbing, D.; Milojević, S.; Petersen, A.M.; Radicchi, F.; Sinatra, R.; Uzzi, B.; et al. Science of science. *Science* **2018**, *359*, eaao0185. [CrossRef] [PubMed]
3. Waltman, L. A review of the literature on citation impact indicators. *J. Inf.* **2016**, *10*, 365–391. [CrossRef]
4. Zaheer, A.; Mustafa, G.; Afzal, M.T. A hybrid framework for predicting scholarly recognition of authors using network centrality and generative AI embeddings. *Scientometrics* **2026**, *131*, 2885–2925. [CrossRef]
5. Schreiber, W.E.; Giustini, D.M. Measuring scientific impact with the h-index: A primer for pathologists. *Am. J. Clin. Pathol.* **2019**, *151*, 286–291. [PubMed]
6. Moed, H.F. *Citation Analysis in Research Evaluation*; Springer: Berlin/Heidelberg, Germany, 2005.
7. Bornmann, L.; Mutz, R.; Daniel, H.D. Do we need the h index and its variants in addition to standard bibliometric measures? *J. Am. Soc. Inf. Sci. Technol.* **2009**, *60*, 1286–1289. [CrossRef]
8. Hirsch, J.E. An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 16569–16572. [CrossRef] [PubMed]
9. Usman, M.; Mustafa, G.; Afzal, M.T. Ranking of author assessment parameters using logistic regression. *Scientometrics* **2021**, *126*, 335–353. [CrossRef]

10. Hicks, D.; Wouters, P.; Waltman, L.; de Rijcke, S.; Rafols, I. Bibliometrics: The Leiden Manifesto for research metrics. *Nature* **2015**, *520*, 429–431. [CrossRef] [PubMed]
11. Dienes, K.R. Completing h. *J. Inf.* **2015**, *9*, 385–397. [CrossRef]
12. Sinatra, R.; Wang, D.; Deville, P.; Song, C.; Barabási, A.L. Quantifying the evolution of individual scientific impact. *Science* **2016**, *354*, aaf5239. [CrossRef] [PubMed]
13. Merton, R.K. The Matthew Effect in Science. *Science* **1968**, *159*, 56–63. [CrossRef]
14. Zuckerman, H. *Scientific Elite: Nobel Laureates in the United States*; Free Press: New York, NY, USA, 1977.
15. Allison, P.D.; Stewart, J.A. Productivity differences among scientists: Evidence for accumulative advantage. *Am. Sociol. Rev.* **1974**, *39*, 596–606. [CrossRef]
16. Ameer, M.; Afzal, M.T. Evaluation of h-index and its qualitative and quantitative variants in neuroscience. *Scientometrics* **2019**, *121*, 653–673. [CrossRef]
17. Perc, M. The Matthew Effect in Empirical Data. *J. R. Soc. Interface* **2014**, *11*, 20140378. [CrossRef] [PubMed]
18. Lamont, M. *How Professors Think: Inside the Curious World of Academic Judgment*; Harvard University Press: Cambridge, MA, USA, 2009.
19. Münch, R. *Academic Capitalism: Universities in the Global Struggle for Excellence*; Routledge: London, UK, 2014.
20. Alshdadi, A.A.; Usman, M.; Alassafi, M.O.; Afzal, M.T.; AlGhamdi, R. Formulation of rules for the scientific community using deep learning. *Scientometrics* **2023**, *128*, 1825–1852. [CrossRef]
21. Mustafa, G.; Rauf, A.; Al-Shamayleh, A.S.; Afzal, M.T.; Waqas, A.; Akhunzada, A. Defining quantitative rules for identifying influential researchers: Insights from mathematics domain. *Heliyon* **2024**, *10*, e30318. [CrossRef] [PubMed]
22. Rahangdale, A.; Raut, S. Machine learning methods for ranking. *Int. J. Softw. Eng. Knowl. Eng.* **2019**, *29*, 729–761. [CrossRef]
23. Kannari, P.R.; Chowdary, N.S.; Biradar, R.L. An anomaly-based intrusion detection system using recursive feature elimination technique for improved attack detection. *Theor. Comput. Sci.* **2022**, *931*, 56–64. [CrossRef]
24. Theerthagiri, P. Predictive analysis of cardiovascular disease using gradient boosting based learning and recursive feature elimination technique. *Intell. Syst. Appl.* **2022**, *16*, 200121. [CrossRef]
25. Zhang, J.; Li, C.; Yin, Y.; Zhang, J.; Grzegorzek, M. 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.* **2023**, *56*, 1013–1070. [CrossRef] [PubMed]
26. Shirazi, D.H.; Toosi, H. Deep Multilayer Perceptron Neural Network for the Prediction of Iranian Dam Project Delay Risks. *J. Constr. Eng. Manag.* **2023**, *149*, 04023011. [CrossRef]
27. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the International Conference on Machine Learning*; PMLR: New York, NY, USA, 2015; pp. 448–456. [CrossRef]
28. Bilal, M.; Israr, H.; Shahid, M.; Khan, A. Sentiment classification of Roman-Urdu opinions using Naive Bayesian, Decision Tree and KNN classification techniques. *J. King Saud Univ.–Comput. Inf. Sci.* **2016**, *28*, 330–344. [CrossRef]
29. Panhalkar, A.R.; Doye, D.D. Optimization of decision trees using modified African buffalo algorithm. *J. King Saud Univ.–Comput. Inf. Sci.* **2022**, *34*, 4763–4772. [CrossRef]
30. Tangirala, S. Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, 612–619. [CrossRef]
31. Fayaz, S.A.; Zaman, M.; Butt, M.A. Performance evaluation of GINI index and information gain criteria on geographical data: An empirical study based on JAVA and Python. In *Proceedings of the International Conference on Innovative Computing and Communications: Proceedings of ICICC 2021*; Springer: Singapore, 2021; Volume 3, pp. 249–265.
32. Arlot, S.; Celisse, A. A survey of cross-validation procedures for model selection. *Stat. Surv.* **2010**, *4*, 40–79. [CrossRef]
33. Kapoor, S.; Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* **2023**, *4*, 100804. [CrossRef] [PubMed]
34. Saito, T.; Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **2015**, *10*, e0118432. [CrossRef] [PubMed]
35. Ambroise, C.; McLachlan, G.J. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 6562–6566. [CrossRef] [PubMed]
36. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef] [PubMed]
37. Khalili, N.; Sedghpour, A.S.; Jahanbakht, M.; Hosseini, Z. From Credit Scoring to Governed Decision Systems: Integrating Predictive Performance, Uncertainty, and Governance. *Mach. Learn. Knowl. Extr.* **2026**, *8*, 308. [CrossRef]
38. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In *Proceedings of the Advances in Neural Information Processing Systems 30*; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 4765–4774.
39. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
40. Prathap, G. The 100 most prolific economists using the p-index. *Scientometrics* **2010**, *84*, 167–172. [CrossRef]

41. Egghe, L. Theory and practise of the g-index. *Scientometrics* **2006**, *69*, 131–152. [CrossRef]
42. Zhang, C.T. The e-index, complementing the h-index for excess citations. *PLoS ONE* **2009**, *4*, e5429. [CrossRef] [PubMed]
43. Jin, B. H-index: An evaluation indicator proposed by scientist. *Sci. Focus* **2006**, *1*, 8–9.
44. Jin, B.; Liang, L.; Rousseau, R.; Egghe, L. The R- and AR-indices: Complementing the h-index. *Chin. Sci. Bull.* **2007**, *52*, 855–863. [CrossRef]
45. Bornmann, L.; Mutz, R.; Daniel, H.D. Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine. *J. Am. Soc. Inf. Sci. Technol.* **2008**, *59*, 830–837. [CrossRef]
46. Tol, R. The h-index and its alternatives: An application to the 100 most prolific economists. *Scientometrics* **2009**, *80*, 317–324. [CrossRef]
47. Cabrerizo, F.J.; Alonso, S.; Herrera-Viedma, E.; Herrera, F. q2-index: Quantitative and qualitative evaluation based on the number and impact of papers in the Hirsch core. *J. Inf.* **2010**, *4*, 23–28. [CrossRef]
48. Anderson, T.R.; Hankin, R.K.S.; Killworth, P.D. Beyond the Durfee square: Enhancing the h-index to score total publication output. *Scientometrics* **2008**, *76*, 577–588. [CrossRef]
49. Kosmulski, M. MAXPROD—A new index for assessment of the scientific output of an individual, and a comparison with the h-index. *Cybermetrics Int. J. Sci. Inf. Bibliometr.* **2007**, *11*, 5.
50. Wu, Q. The w-index: A measure to assess scientific impact by focusing on widely cited papers. *J. Am. Soc. Inf. Sci. Technol.* **2010**, *61*, 609–614. [CrossRef]
51. Vinkler, P. The π-index: A new indicator for assessing scientific impact. *J. Inf. Sci.* **2009**, *35*, 602–612. [CrossRef]
52. Egghe, L.; Rousseau, R. An h-index weighted by citation impact. *Inf. Process. Manag.* **2008**, *44*, 770–780. [CrossRef]
53. Kosmulski, M. A new Hirsch-type index saves time and works equally well as the original h-index. *ISSI Newsl.* **2006**, *2*, 4–6.
54. Woeginger, G.J. An axiomatic characterization of the Hirsch-index. *Math. Soc. Sci.* **2008**, *56*, 224–232. [CrossRef]
55. Xu, F.; Liu, W.; Mingers, J. New journal classification methods based on the global h-index. *Inf. Process. Manag.* **2015**, *51*, 50–61. [CrossRef]
56. Panaretos, J.; Malesios, C. Assessing scientific research performance and impact with single indices. *Scientometrics* **2009**, *81*, 635–670. [CrossRef]
57. Fenner, T.; Harris, M.; Levene, M.; Bar-Ilan, J. A novel bibliometric index with a simple geometric interpretation. *PLoS ONE* **2018**, *13*, e0200098. [CrossRef] [PubMed]
58. Alonso, S.; Cabrerizo, F.J.; Herrera-Viedma, E.; Herrera, F. hg-index: A new index to characterize the scientific output of researchers based on the h- and g-indices. *Scientometrics* **2010**, *82*, 391–400. [CrossRef]
59. Bornmann, L.; Mutz, R.; Daniel, H.D. The h index research output measurement: Two approaches to enhance its accuracy. *J. Inf.* **2010**, *4*, 407–414. [CrossRef]
60. Chen, D.Z.; Huang, M.H.; Ye, F.Y. A probe into dynamic measures for h-core and h-tail. *J. Inf.* **2013**, *7*, 129–137. [CrossRef]
61. Ruane, F.; Tol, R. Rational (successive) h-indices: An application to economics in the Republic of Ireland. *Scientometrics* **2008**, *75*, 395–405. [CrossRef]
62. Guns, R.; Rousseau, R. Real and rational variants of the h-index and the g-index. *J. Inf.* **2009**, *3*, 64–71. [CrossRef]
63. Cornell University Library. Measuring Your Research Impact: I10-Index. Research Guide Describing the Google Scholar i10-index. Available online: https://guides.library.cornell.edu/impact/author-impact-10 (accessed on 11 November 2026).
64. Sidiropoulos, A.; Katsaros, D.; Manolopoulos, Y. Generalized Hirsch h-index for disclosing latent facts in citation networks. *Scientometrics* **2007**, *72*, 253–280. [CrossRef]
65. Ye, F.; Rousseau, R. Probing the h-core: An investigation of the tail-core ratio for rank distributions. *Scientometrics* **2010**, *84*, 431–439. [CrossRef]
66. Wu, Q. The w-index: A significant improvement of the h-index. *arXiv* **2008**, arXiv:0805.4650.
67. Zhang, C.T. The h'-index, effectively improving the h-index based on the citation distribution. *PLoS ONE* **2013**, *8*, e59912. [CrossRef] [PubMed]
68. Batista, P.D.; Campiteli, M.G.; Kinouchi, O. Is it possible to compare researchers with different scientific interests? *Scientometrics* **2006**, *68*, 179–189. [CrossRef]
69. Wohlin, C. A new index for the citation curve of researchers. *Scientometrics* **2009**, *81*, 521–533. [CrossRef]
70. Schreiber, M. A modification of the h-index: The hm-index accounts for multi-authored manuscripts. *J. Inf.* **2008**, *2*, 211–216. [CrossRef]
71. Schreiber, M. Fractionalized counting of publications for the g-index. *J. Am. Soc. Inf. Sci. Technol.* **2009**, *60*, 2145–2150. [CrossRef]
72. Egghe, L. Mathematical theory of the h- and g-index in case of fractional counting of authorship. *J. Am. Soc. Inf. Sci. Technol.* **2008**, *59*, 1608–1616. [CrossRef]
73. Anania, G.; Caruso, A. Two simple new bibliometric indexes to better evaluate research in disciplines where publications typically receive less citations. *Scientometrics* **2013**, *96*, 617–631. [CrossRef]

74. Wan, J.K.; Hua, P.H.; Rousseau, R. The pure h-index: Calculating an author's h-index by taking co-authors into account. *COLLNET J. Sci. Inf. Manag.* **2007**, *1*, 1–5. [CrossRef]
75. Alonso, S.; Cabrerizo, F.J.; Herrera-Viedma, E.; Herrera, F. h-index: A review focused on its variants, computation and standardization for different scientific fields. *J. Inf.* **2009**, *3*, 273–289. [CrossRef]
76. Smith, D.R. "Platinum H": Refining the H-index to more realistically assess career trajectory and scientific publications. *Arch. Environ. Occup. Health* **2015**, *70*, 67–69. [CrossRef] [PubMed]
77. Burrell, Q.L. Hirsch's h-index: A stochastic model. *J. Inf.* **2007**, *1*, 16–25. [CrossRef]
78. Agarwal, A.; Durairajanayagam, D.; Tatagari, S.; Esteves, S.C.; Harlev, A.; Henkel, R.; Roychoudhury, S.; Homa, S.; Puchalt, N.G.; Ramasamy, R.; et al. Bibliometrics: Tracking research impact by selecting the appropriate metrics. *Asian J. Androl.* **2016**, *18*, 296. [CrossRef] [PubMed]
79. Vaidya, J.S. V-index: A fairer index to quantify an individual's research output capacity. *BMJ* **2005**, *331*, 1339.
80. Hagen, N.T. Harmonic publication and citation counting: Sharing authorship credit equitably—Not equally, geometrically or arithmetically. *Scientometrics* **2010**, *84*, 785–793. [CrossRef] [PubMed]
81. Cucchetti, A.; Mazzotti, F.; Pellegrini, S.; Cescon, M.; Maroni, L.; Ercolani, G.; Pinna, A.D. The use of the Hirsch index in benchmarking hepatic surgery research. *Am. J. Surg.* **2013**, *206*, 560–566. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.