

BenchCouncil Transactions

TBench

Volume 4, Issue 3

2024

on Benchmarks, Standards and Evaluations

Editorial

☉ Five Axioms of Things

Jianfeng Zhan

☉ Fundamental concepts and methodologies in evaluatology

Jianfeng Zhan

Original Articles

☉ Evaluation of mechanical properties of natural fiber based polymer composite

Tarikur Jaman Pramanik, Md. Rafiquzzaman, Anup Karmakar,
Marzan Hasan Nayeem, ... Md. Ragib Abid

☉ Could bibliometrics reveal top science and technology achievements and researchers? The case for evaluatolo- gy-based science and technology evaluation

Guoxin Kang, Wanling Gao, Lei Wang, Chunjie Luo, ... Jianfeng Zhan

☉ Analyzing the obstacles to the establishment of sustain- able supply chain in the textile industry of Bangladesh

Md. Hasibul Hasan Hemal, Farjana Parvin, Alberuni Aziz

☉ Exploring the Orca Predation Algorithm for Economic Dispatch Optimization in Power Systems

Vivi Aida Fitria, Arif Nur Afandi, Aripriharta

ISSN: 2772-4859

Copyright © 2024 International Open Benchmark Council (BenchCouncil); spon-
sored by ICT, Chinese Academy of Sciences. Publishing services by Elsevier B.V.
on behalf of KeAi Communications Co. Ltd.

BenchCouncil Transactions on Benchmarks, Standards and Evaluations (TBench) is an open-access multi-disciplinary journal dedicated to benchmarks, standards, evaluations, optimizations, and data sets. This journal is a peer-reviewed, subsidized open access journal where The International Open Benchmark Council pays the OA fee. Authors do not have to pay any open access publication fee. However, at least one of the authors must register BenchCouncil International Symposium on Benchmarking, Measuring and Optimizing (Bench) (<https://www.benchcouncil.org/bench/>) and present their work. It seeks a fast-track publication with an average turnaround time of one month.

Contents

Five Axioms of Things

Jianfeng Zhan..... 1

Fundamental concepts and methodologies in evaluatology

Jianfeng Zhan..... 4

Evaluation of mechanical properties of natural fiber based polymer composite

Tarikur Jaman Pramanik, Md. Rafiquzzaman, Anup Karmakar, Marzan Hasan Nayeem, ... Md. Ragib Abid 9

Could bibliometrics reveal top science and technology achievements and researchers? The case for evaluatology-based science and technology evaluation

Guoxin Kang, Wanling Gao, Lei Wang, Chunjie Luo, ... Jianfeng Zhan..... 21

Analyzing the obstacles to the establishment of sustainable supply chain in the textile industry of Bangladesh

Md. Hasibul Hasan Hemal, Farjana Parvin, Alberuni Aziz..... 36

Exploring the Orca Predation Algorithm for Economic Dispatch Optimization in Power Systems

Vivi Aida Fitria, Arif Nur Afandi, Aripriharta..... 44

MultiPoint: Enabling scalable pre-silicon performance evaluation for multi-task workloads

Chenji Han, Xinyu Li, Feng Xue, Weitong Wang, ... Fuxin Zhang..... 56



Editorial

Five Axioms of Things

Jianfeng Zhan*

The International Open Benchmark Council, DE, USA
ICT, Chinese Academy of Sciences, Beijing, China
University of Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Keywords:

Thing
Model
Truth
Observation
Experiment
Evaluation
Measurement
Testing

ABSTRACT

This article explicitly defines several concepts, such as variables, models, and truth of a thing, that are fundamental to natural and social sciences. I present a generalized methodology for understanding a thing, categorically defining six foundational understanding approaches based on the nature of the thing and diverse perspectives: conjecture, observation, experiment, evaluation, measurement, and testing. I extend my previous work on the five axioms of evaluation to understanding a thing, which I call the five axioms of things. Also, I comment on five paradigms of science.

1. The formal definitions of fundamental concepts in understanding a thing

In this section, I rigorously delineate several fundamental concepts in understanding a thing, drawing upon definitions from key Refs. [1–5].

An *individual* can be defined as the object described by a given set of properties [5]. A *system* is a coherent entity comprising interacting or interdependent individuals, regardless of their likeness or diversity, culminating in a unified whole [1,2,5]. A system could be recursive. A *thing* could be an individual or a system. It could be a life, a natural phenomenon, an artifact, an abstract, or even a policy in natural or social sciences.

A *quantity* or *variable* is "a property of a thing whose instances can be compared by ratio or only by order [5,6]". One or more variables provide a partial understanding of a thing. The *truth* is a thing's facts or inherent properties that can be proven true or verified objectively.

A *model* serves as a streamlined representation of a thing that would otherwise be too intricate to analyze in exhaustive detail [5,7]. A model provides a full understanding of a thing, though it is simplified. A model can manifest as a physical, mathematical, or other construct. A mathematical model embodies a mathematical representation, frequently expressed through functions or equations, that captures the essence of a thing. In general, a model appears in a mathematical form. Regrettably, many things cannot be well-defined. When asserting that something is not well-defined, it implies that its structure and functions

remain incompletely comprehended. For example, a human body is not well-defined.

Take the function as an instance. According to [3,5], "a function, denoted as f , is a rule that assigns a unique element, referred to as $f(x)$, from a set R to each element in a set D ". In this context, "the domain, denoted as D , refers to the set of all possible values for which the function is defined [3]". On the other hand, "the range of the function, denoted as $f(x)$, consists of all the possible values that $f(x)$ can take as x varies within the domain [3]". The *independent variable* is represented by "a symbol that encompasses any arbitrary number within the domain of the function [3]". A *dependent variable*, represented by a symbol, "is used to denote a number within the range of the function [3]".

As a special instance of a model, a causal model is a *causal* explanation grounded in a model to understand a thing and infer its behavior [3,8].

A model is used to predict the outcome of understanding a thing. As these predicted outcomes increasingly align with the truth, the model is regarded as more precise, approaching a state of perfection.

2. Fundamental methodologies for understanding a thing

To obtain a model of a thing, it is essential to identify and isolate a system conducive to understanding the thing. This system must meet two criteria: first, it can operate autonomously. Second, it incorporates

* Correspondence to: The International Open Benchmark Council, DE, USA.
E-mail address: jianfengzhan.benchcouncil@gmail.com.
URL: <http://www.zhanjianfeng.org>.

	Thing	Self-contained Research System (SRS)	Methodology
	1 Parallel universes Soul	• Unknown	• Conjecture
	2 Cosmology Astronomy	• Partially unknown	• Observation
	3 Lots of life and nature phenomena	• Known but not well-defined	• Experiment
	4 Lots of natural phenomena	• Known, well-defined, but not subject to arbitrary manipulation	• Experiment
	5 Computer	• Known, well-defined, subject to arbitrary manipulation	• Experiment

Fig. 1. Five Categories of Self-Contained Research Systems (SRS) and Their Corresponding Methodologies.

the primary factors that determine the outcomes of understanding a thing, which I refer to as essential factors. I denote this system as a Self-contained Research System (abbreviated as SRS). If it is impossible to isolate an SRS, the impacts of other external factors will affect the outcomes of understanding a thing. After identifying and isolating an SRS, it is plausible to investigate the effect of the essential factors on the thing. I refer to the proposed methodology as SRS.

In comparison with the causal model methodology proposed in [8], the SRS methodology offers several advantages. Firstly, even if a thing's SRS is known, it may not be well defined, making it challenging to derive its causal models. Secondly, comprehending an SRS serves as a fundamental and robust basis for establishing its causal model.

Dealing with the diverse nature of an SRS presents various challenges as shown in Fig. 1. The first kind is when an SRS is unknown, e.g., in the case of investigating parallel universes, or soul. The second kind is when an SRS is only partially known, e.g., in the case of investigating a thing in cosmology and astronomy. The third kind is when an SRS is known but cannot be well defined, e.g., a human body. The fourth kind is when an SRS is known and well-defined but not subject to arbitrary manipulation. If a system can be modeled in a function, arbitrary manipulation entails setting its independent variables to any arbitrary number within the function's domain. The fifth kind is when an SRS is known, well-defined, and subject to arbitrary manipulation. For example, a computer nearly falls into this category. Moving from the fifth kind to the first kind of SRS, the challenge level increases.

I formally define six methodologies for understanding a thing: conjecture, observation, experiment, evaluation, measurement, and testing.

The conjecture is obtaining the model of a thing when it is impossible to identify and isolate an SRS. The observation aims to derive a model of a thing in cases where there is only a partial understanding of an SRS. The experiment aims to derive a model of a thing in cases where it is feasible to identify and isolate an SRS. In the context of an experiment, there are subtly different scenarios: an SRS may not be well-defined; an SRS is well-defined but not subject to arbitrary manipulation; and an SRS is well-defined and subject to arbitrary manipulation.

Experiments and observations fall into two primary categories: those conducted independently and those involving stakeholders. Involving stakeholders classifies an experiment or observation as an evaluation, while those without stakeholders are categorized as natural experiments or observations.

My previous work [9] distinguished between the concepts of evaluation, measurement, and testing. Measurement is experimentally obtaining one or more values attributed to a quantity of a thing [6]. A test oracle is a fact or inherent property of a thing and its SRS. Testing

is a verification process to determine whether (1) a thing conforms to the test oracles (the first category) and/or (2) When a thing operates within an SRS, both the thing and its SRS conform to the test oracles (the second category) [5,9]. As per the findings of [9], evaluation entails causal inferring the impact and value of a thing within an SRS tailored to meet the evaluation requirements of stakeholders, relying on measurements and/or testing of the SRS.

Measurements and testing offer a foundational methodology for gaining partial insights into a thing by focusing on specific properties or facts. In contrast, observation, experiments, and evaluation endeavor to achieve a full understanding of the thing.

Viewed from another perspective, understanding a thing can be seen as intentionally imposing a research condition (RC) upon it in order to establish an SRS. Building on the previous discussion, an RC can be envisioned as the SRS from which the thing under investigation is removed. I formally define an RC as the context that is applied to the thing, playing a crucial role in guaranteeing independent operation and incorporating the essential factors. Within a particular methodology for understanding a thing, I designate a specific RC, such as an evaluation condition for evaluation, an observation condition for observation, or an experimental condition for experiment.

3. Five axioms of things

Derived from the essence of understanding a thing, I propose five axioms focusing on key aspects of the outcomes of understanding a thing, including observation, experiment, evaluation, measurement, and testing, as the foundational theory. These axioms serve as the bedrock upon which universal theories and methodologies are built for understanding a thing.

The Axiom of Metric Essence asserts that in the absence of stakeholder involvement, the essence of a metric holds intrinsic physical significance. Alternatively, when stakeholders are engaged, the essence of the metric may possess intrinsic physical significance or be solely determined by the value function. In the latter scenario, a value function establishes a composite metric that normalizes metrics of different dimensions based on stakeholder perspectives.

The Axiom of True Outcomes declares that when an SRS is known, the outcome of understanding a thing has true value.

The Axiom of Traceable Outcomes declares that when its SRSes are known, the divergence in the outcomes of understanding a thing can be attributed to disparities in RCs, thereby establishing traceability.

The Axiom of Consistent Outcomes posits that when a thing is examined under various samples from a known RC population, the outcomes tend to converge towards the true value under the RC population.

The Axiom of Comparable Outcomes declares when equipped with equivalent well-defined experimental conditions, the outcomes of understanding different things are comparable.

4. Comments to five paradigms of science

There have been several insightful discussions on the diverse paradigms of science in previous works [10,11].

Ioannidis [10] elegantly encapsulated the evolving paradigm shifts in science, delineating the transition from the traditional paradigms of empirical/experimental science (practiced for millennia) to theoretical model science (spanning centuries), followed by computational science (over decades), and the emergence of data-driven science as envisioned by Jim Gray (over the past 15 years), ultimately culminating in the advent of the 5th paradigm: AI-driven science.

Building upon the precise definitions of fundamental concepts and methodologies outlined in Sections 1 and 2, I offer concise reflections on Ioannidis's narrative regarding the paradigm shifts within the domain of science.

Following the formal definition of an “experiment”, I express reservations about the conflation of “empirical and experimental”. I propose the usage of “empirical practice” instead of “empirical science”. Empirical practice, akin to observation and conjecture, operates within a context where the SRS is either unknown or only partially known. In such scenarios, essential factors may be absent, impeding the attainment of truth.

It is crucial to underscore that in an experiment, an SRS is known; irrespective of its manifestation, understanding its behavior within a controlled environment enables the pursuit of truth through experimentation.

As elucidated in Section 1, a model serves as the culmination of observations or experiments. On the other side, computational, data-driven, and AI-driven sciences predominantly function as novel tools or methodologies that complement observations or experiments. From this perspective, it is arguable that they do not constitute an independent methodology for understanding a thing.

References

- [1] System, 2024, <https://www.merriam-webster.com/dictionary/system>. (Accessed: 6 February 2024).
- [2] A. Backlund, The definition of system, *Kybernetes* 29 (4) (2000) 444–451.
- [3] J. Stewart, *Single Variable Calculus: Concepts and Contexts*, Cengage Learning, 2018.
- [4] D.S. Starnes, D. Yates, D.S. Moore, *The Practice of Statistics*, Macmillan, 2010.
- [5] J. Zhan, L. Wang, W. Gao, H. Li, C. Wang, Y. Huang, Y. Li, Z. Yang, G. Kang, C. Luo, H. Ye, S. Dai, Z. Zhang, Evaluatolgy: The science and engineering of evaluation, *BenchCouncil Trans. Benchmarks Stand. Eval.* 4 (1) (2024) 100162.
- [6] I. BiPM, I. IFCC, I. IUPAC, O. ISO, The international vocabulary of metrology—basic and general concepts and associated terms (VIM), *JCGM 200* (2012) 2012.
- [7] H.D. Young, R.A. Freedman, L.A. Ford, *University Physics with Modern Physics*, 2020.
- [8] J. Pearl, D. Mackenzie, *The book of why: the new science of cause and effect*, Basic books, 2018.
- [9] J. Zhan, A short summary of evaluatolgy: The science and engineering of evaluation, *BenchCouncil Trans. Benchmarks Stand. Eval.* (2024) 100175.
- [10] Y. Ioannidis, The 5th paradigm: AI-driven scientific discovery, *Commun. ACM* (2024).
- [11] T. Hey, S. Tansley, K.M. Tolle, et al., *The Fourth Paradigm: Data-Intensive Scientific Discovery*, vol. 1, Microsoft research Redmond, WA, 2009.



Dr. Jianfeng Zhan is a Full Professor at the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), and University of Chinese Academy of Sciences (UCAS), the director of the Research Center for Distributed Systems, ICT, CAS. He received his B.E. in Civil Engineering and MSc in Solid Mechanics from Southwest Jiaotong University in 1996 and 1999 and his Ph.D. in Computer Science from the Institute of Software, CAS, and UCAS in 2002. His research areas focus on evaluatolgy, evaluatolgy-based design automation, and optimization automation. His exceptional expertise is exemplified by his introduction to the discipline of evaluatolgy, an endeavor that encompasses the science and engineering of evaluation; within this discipline, his proposition of a universal framework for evaluation encompasses essential concepts, terminologies, theories, and methodologies for application across various disciplines. He has made substantial and effective efforts to transfer his academic research into advanced technology to impact general-purpose production systems. Several technical innovations and research results, including 35 patents from his team, have been adopted in benchmarks, operating systems, and cluster and cloud system software with direct contributions to advancing parallel and distributed systems in China or worldwide. Over the past two decades, he has supervised over ninety graduate students, post-doctors, and engineers. Dr. Jianfeng Zhan is the founder and chairman of BenchCouncil. He also holds the role of Co-EIC of BenchCouncil Transactions on Benchmark, Standards and Evaluations, alongside Prof. Tony Hey. Dr. Zhan has served as an Associate Editor for IEEE TPDS (IEEE Transactions on Parallel and Distributed Systems) from 2018 to 2022. In recognition of his exceptional contributions, he has been honored with several prestigious awards. These include the second-class Chinese National Technology Promotion Prize in 2006, the Distinguished Achievement Award of the Chinese Academy of Sciences in 2005, the IISWC Best Paper Award in 2013, and the Test of Time Paper Award from the Journal of Frontier of Computer Science.



Editorial

Fundamental concepts and methodologies in evaluatology

Jianfeng Zhan*

The International Open Benchmark Council, DE, USA
 Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
 University of Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Keywords:

Evaluatology
 Fundamental concepts
 Fundamental methodology
 Testbed
 Experimental platform
 Simulation environment

ABSTRACT

While I have authored three articles introducing Evaluatology, a novel discipline that encompasses the science and engineering of evaluation across various domains, I have struggled to fully depict this challenging yet promising field.

This article delves into the fundamental concepts and methodologies within Evaluatology. I aim to provide a complete picture of evaluation problems in Evaluatology based on my proposed fundamental methodology of understanding a thing. In diverse engineering fields, testbeds, experimental platforms, or simulation environments are commonly utilized to evaluate design or implementation decisions. However, a rigorous methodology is often lacking. I propose a rigorous methodology rooted in Evaluatology for testbeds, experimental platforms, or simulation environments.

1. Why am I drafting this article?

I have drafted three articles to present a new discipline named Evaluatology [1–3], among which I coauthored with my colleagues or students on the first article [1]. However, there are three flaws in the previous work [1–3].

First, I failed to draw a complete picture of evaluation problems in Evaluatology. For example, in the first Evaluatology article [1], I focus on the scenario where we can well define an evaluation condition and emphasize how to construct equivalent evaluation conditions where it is almost impossible to achieve in other scenarios. In Section 3.2, I will formally define what is an evaluation condition. In [3], I discussed other essential scenarios. However, I fail to provide a unified methodology framework for different scenarios.

Second, I fail to propose a generalized methodology for evaluating or understanding (in a much broader sense) a thing in the previous work [1,2]. Later, in [3], I proposed a generalized methodology for understanding a thing, which can provide a solid basis for presenting a complete picture of evaluation problems in Evaluatology.

Third, I fail to propose a generalized methodology to define evaluation conditions. The methodology proposed in [1] is limited and only applied to some scenarios. The above reasons justified my motivation for drafting this article.

2. Fundamental concepts in evaluatology

I reuse the concepts in [3] most of the time. Fig. 1 summarizes the primary entities within Evaluatology.

An *individual* can be defined as “the object described by a given set of properties” [1,3]. A *system* is “a coherent entity comprising interacting or interdependent individuals and/ or systems, regardless of their likeness or diversity, culminating in a unified whole” [1,3–5].

The evaluation subject [1] (in short, subject) is a *thing* that could be an individual or a system. Typical subjects could be “a life, an artifact, an abstract, or even a policy in natural and social sciences”.

A *quantity* “embodies any property of a thing whose instances can be compared by ratio or only by order [1,3,6]”. The *truth* is “a thing’s facts or inherent properties that can be proven true or verified objectively [3]”. A *model* is “a streamlined representation of a thing that would otherwise be too intricate to analyze in exhaustive detail” [1,3,7]. A model can manifest as “a physical, mathematical, or other construct” [1,3,7]. As a special model instance, a causal model is “a causal explanation grounded in a model to understand a thing and infer its behavior” [3,8]. Quantity and truth provide partial insights into a thing, while a model gains a full understanding of a thing in a simplified manner.

* Correspondence to: The International Open Benchmark Council, DE, USA.
 E-mail address: jianfengzhan.benchcouncil@gmail.com.
 URL: <https://www.zhanjianfeng.org>.

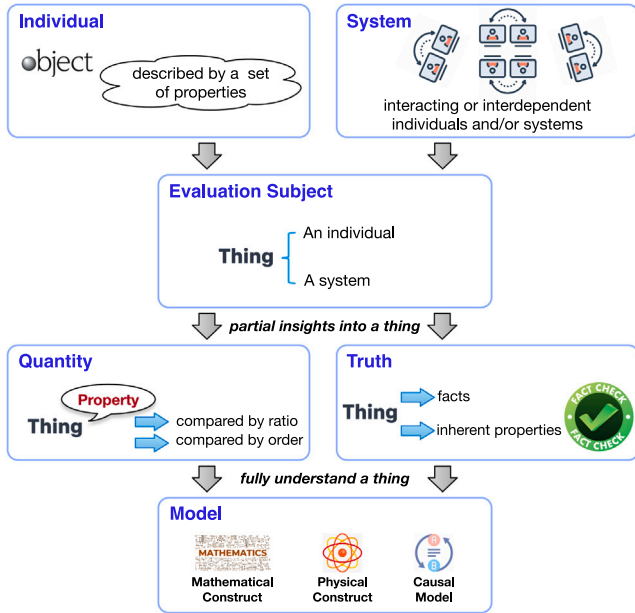


Fig. 1. The primary entities within Evaluatology.

3. Fundamental methodologies in evaluatology

Evaluation is one of the methodologies used to understand a thing. Other methodologies listed in [3] include conjecture, observation, experiment, measurement, and testing.

3.1. The generalized methodology understanding a thing

In my publication [3], I have introduced a generalized methodology for understanding a thing, with the focal point being a concept termed the Self-contained Research System (abbreviated as SRS). As detailed in [3], an SRS must adhere to two key criteria: firstly, it can operate autonomously, and secondly, it should encompass the primary factors that influence the understanding of the thing, known as essential factors. Fig. 2 summarizes the generalized methodology of understanding a thing.

In various contexts, an SRS can be designated differently. For instance, within the realm of evaluation, an SRS may be referred to as a self-contained evaluation system (abbreviated as SES).

To obtain a model of a thing, it is essential to identify and isolate an SRS that is conducive to understanding it. I explained the reason in [3]. If isolating an SRS is unfeasible, external factors may significantly influence understanding a thing. However, once an SRS is identified and isolated, examining the impact of essential factors on the thing becomes viable. The methodology I introduced in [3] is referred to as SRS.

3.2. The relationships among observation, experiment, measurement, testing, and evaluation

If only some of the essential factors are identified within the SRS, I classify it as an observation. Conversely, if all essential factors are identified within the SRS, I classify it as an experiment. The distinction between observation and experiment lies in the presence of hidden or unknown factors that can influence understanding the thing in the former scenario.

Unlike observation and experiments that fully understand the thing, measurements, and testing gain partial insights into a thing by focusing on specific properties, facts, or inherent properties [3]. Measurement is

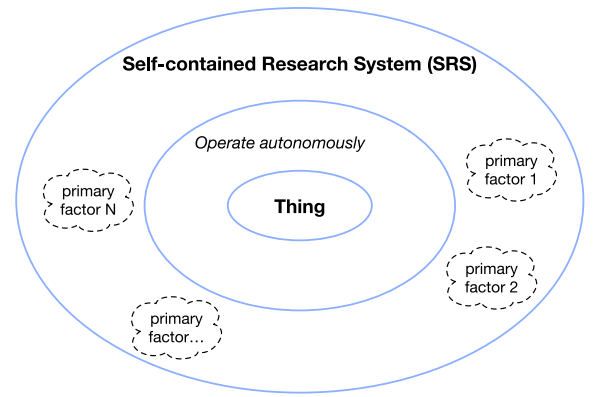


Fig. 2. The Generalized Methodology Understanding a Thing.

“experimentally obtaining one or more values attributed to a quantity of a thing” [3,6]. A test oracle is “a fact or inherent property of a thing and its SRS [3]”. Testing is a “verification process to determine whether (1) a thing conforms to the test oracles and/or (2) When a thing operates within an SRS, both the thing and its SRS conform to their test oracles [1,2]”.

Building upon the notions of experiment and observation, I elucidate the evaluation concept. If an experiment or observation engages stakeholders, it falls under the evaluation category, whereas those conducted without stakeholder involvement are classified as natural experiments or observations [3]. As elucidated in the discoveries of [2], evaluation involves “causal inference regarding the impact and value of a subject within an SES customized to fulfill stakeholders’ evaluation needs, relying on measurements and/or testing of the SES”. Fig. 3 depicts the differences between observation, experiment, and evaluation. Fig. 4 depicts the differences between measurement and testing.

Considered from an alternative angle, the process of evaluating a subject can be depicted as “deliberately imposing an evaluation condition (EC) upon it to establish an SES” [3]. Building on the previous discussion, an EC can be envisioned as the SES from which the subject is removed. We formally delineate an EC as the context that is crucial in guaranteeing independent operation and incorporating the essential factors when applied to the subject.

3.3. Standardized evaluation methodology

As shown in Fig. 5, I summarize a standardized evaluation methodology as follows.

The initial step involves defining and characterizing the thing slated for evaluation, which constitutes the subject. Often, evaluations aim to compare different subjects. Without a clear definition and characterization, it is challenging to ensure that the subjects under investigation can be classified into one category and compared. An integral aspect of this phase is modeling the thing, which includes delineating its structure. In many instances, stakeholders harbor specific requirements for evaluating a component of the thing, such as a branch predictor module within the CPU. Failure to formally define the thing’s structure and establish a consensus renders evaluating a designated component futile. Divergent stakeholder perspectives on structures or differing functionalities assigned to the same structure can obfuscate the evaluation process.

The subsequent step involves defining the SES for the specified thing. As per the SES definition, it must fulfill two criteria. Firstly, an SES should have autonomous functionality. Secondly, an SES should encompass the essential factors. Constructing an SES is a complex endeavor that necessitates a trial-and-error approach to attain an optimal or viable SES.

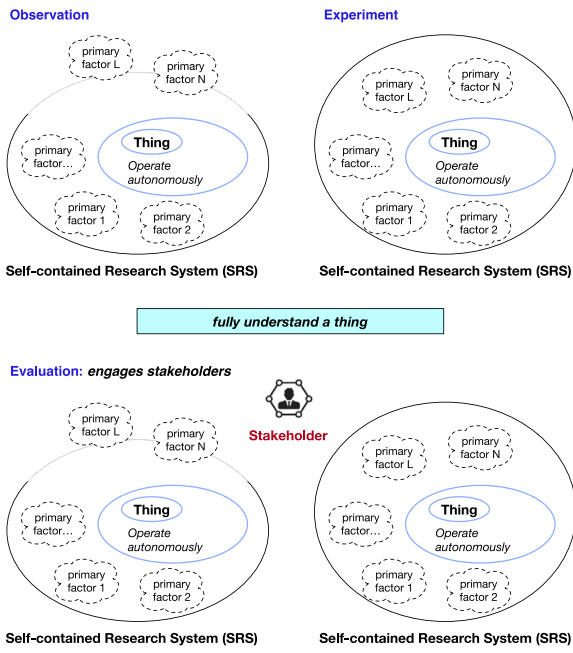


Fig. 3. The Relationships Among Observation, Experiment, and Evaluation.

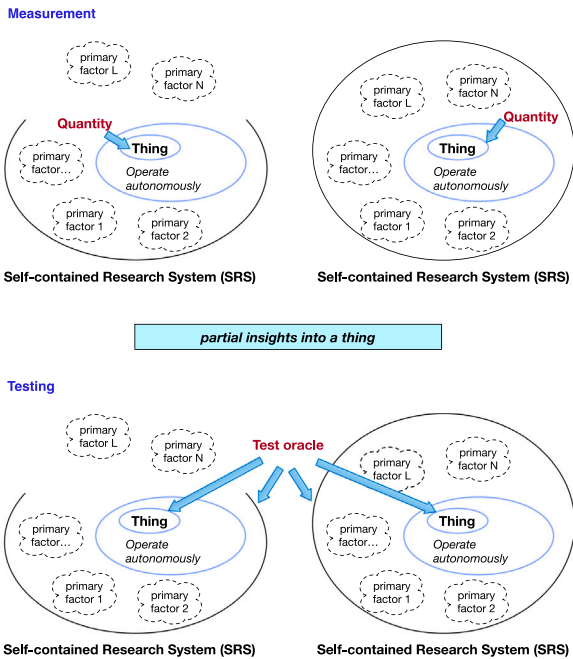


Fig. 4. The Relationship Between Measurement and Testing.

The third step involves acquiring the EC. This can be accomplished by removing the subject under examination from the SES.

Following the establishment of an SES, the fourth step entails analyzing its nature and determining the appropriate evaluation methodologies, which are open issues worth investigating.

Promising evaluation approaches include establishing equivalent EC for the known, well-defined SES [1,9], the causal model approach [8] and the statistical approaches [10].

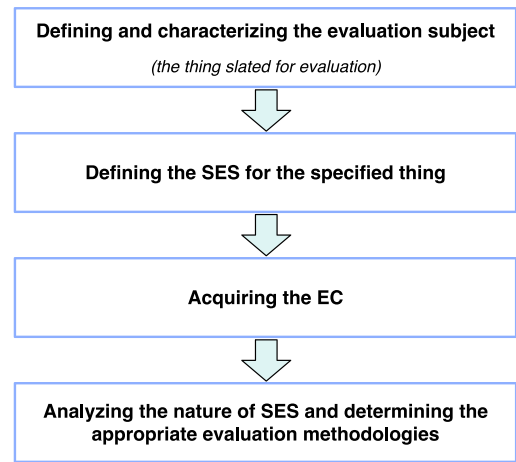


Fig. 5. Standardized Evaluation Methodology.

3.4. The full picture of evaluation problems in evaluatology

My previous research has delved into the diverse natures of SESes [3]. I have slightly adjusted my presentations to demonstrate the full picture of evaluation problems in Evaluatology.

As shown in Fig. 6, the diverse nature of an SES presents various challenges to evaluation [3]. The first kind is when an SES is unknown, e.g., in the case of investigating parallel universal or soul. The second kind is when an SES is only partially known, e.g., when investigating a thing in cosmology and astronomy. The third kind is when an SES is known.

In the third case, there are three different sub-categories. The first subcategory is when an SES is very complex and cannot be well-defined. “When asserting that something is not well-defined, it implies that its structure and functions remain incompletely comprehended. For example, the human body is not well-defined”. [3]. The second subcategory is when an SES is known and well-defined but not subject to arbitrary manipulation for different reasons, such as realization limitations, unaffordable costs, unaffordable consequences, or ethical reasons. “If a system can be modeled in a function, arbitrary manipulation entails setting its independent variables to any arbitrary number within its domain. [3]”. The third subcategory is when an SES is known, well-defined, and subject to arbitrary manipulation. For example, a computer nearly falls into this category [3].

4. Fundamental methodologies in testbed

Testbeds, experimental platforms, or simulation environments are widely used in different engineering fields to evaluate or test design or implementation decisions. However, they lack a rigorous methodology. In the rest of this article, I use the testbed concept to refer to those systems. When I use the concept of a testbed, it depicts a system that can vary diverse ECs to evaluate design or implementation decisions.

In [1], I introduced a universal evaluation methodology for complex scenarios in collaboration with my colleagues and students. Initially, I outlined the methodology proposed in [1]. Subsequently, I will highlight its limitations. Finally, I will introduce a rigorous methodology built upon my proposed approach in [1].

Fig. 7 illustrates the original universal evaluation methodology in complex scenarios. I previously referred to the complete set of real-world systems utilized for assessing specific subjects as the real-world evaluation system (ES). Nevertheless, the notion of a real-world ES is rather vague. In its place, we have formally articulated the self-contained evaluation system (SES). I have adopted the concept of SES to supplant the real-world ES. Fig. 8 presents the upgraded universal

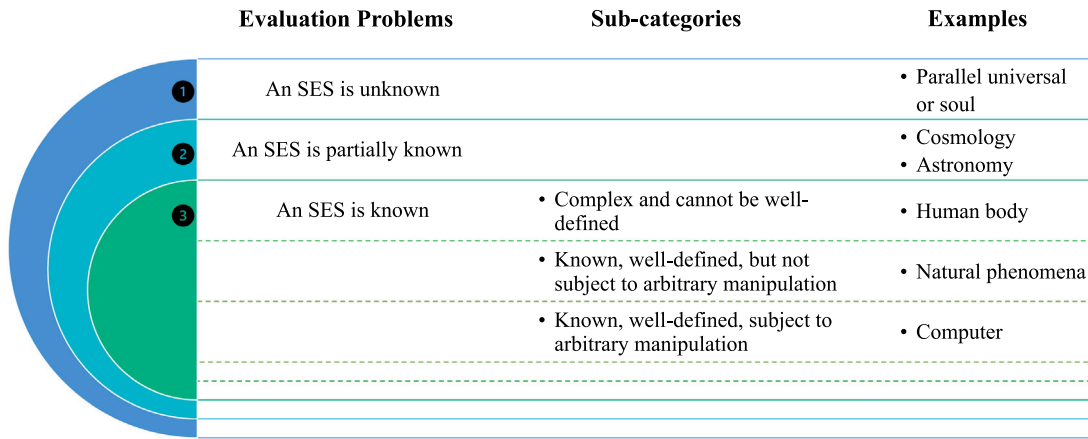


Fig. 6. The Full Picture of Evaluation Problems in Evaluatology.

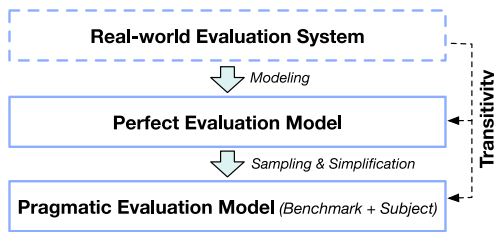


Fig. 7. The original universal evaluation methodology in complex scenarios [1], with the permission of the authors.

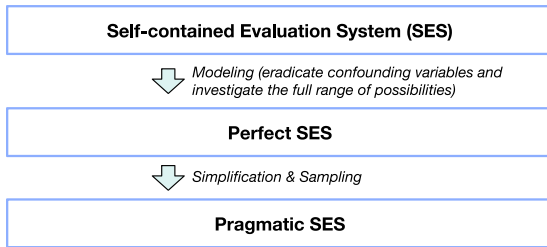


Fig. 8. The upgraded universal evaluation methodology in complex scenarios, based on the concept of the Self-contained Evaluation System (SES).

evaluation methodology in complex scenario, based on the concept of SES.

An SES encounters various hurdles like that of a real-world ES outlined in [1]. Initially, dealing with confounding variables within an SES presents a significant obstacle. Completely eradicating these confounding variables is frequently challenging, if not unattainable.

Furthermore, manipulating the SES proves to be daunting, rendering the establishment of controlled environments for subject evaluation nearly impossible.

Moreover, it is crucial to recognize that SES, irrespective of its characteristics, tends to exhibit a predisposition towards specific groupings. Instead, it should be subject to arbitrary manipulation.

I propose the concept of a Perfect SES that could replicate the SES with the highest level of fidelity. In theory, a Perfect SES would possess three characteristics that enhance the evaluation of subjects.

First, a Perfect SES would streamline manipulation, enabling a free setting of diverse EC. This adaptability would empower researchers to delve into multiple scenarios and appraise subjects under varying conditions, enriching the evaluation process in both depth and breadth.

Secondly, a Perfect SES could effectively eradicate confounding variables. By isolating and controlling variables of interest, researchers

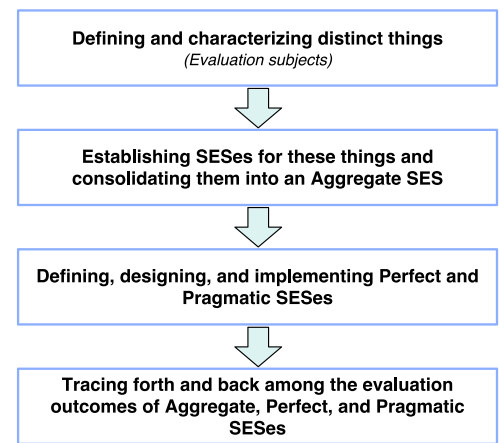


Fig. 9. A standardized evaluation methodology for a testbed.

could attain more precise insights into how specific variables influence the subjects under scrutiny.

Moreover, a Perfect SES would possess the capacity to comprehensively investigate and grasp the full range of possibilities within an SES.

The nature of the Perfect SES, which involves accommodating extensive populations of ECs along with numerous independent variables, may result in substantial evaluation costs. Yet, to tackle this issue, it is crucial to introduce a Pragmatic SES that streamlines the Perfect SES through two essential modifications.

To reduce evaluation costs associated with multiple independent variables, it is crucial to identify and focus on those variables that substantially impact evaluation outcomes. By identifying and ranking these crucial variables, researchers can streamline evaluations, utilize resources more effectively, and exclude or regulate insignificant variables, reducing complexity and costs. It is important to note that simplification in creating a Pragmatic SES will likely decrease its accuracy.

Moreover, employing sampling techniques can efficiently handle extensive populations of ECs. Instead of assessing every possible scenario, researchers can choose representative samples that encompass the population's diversity and breadth. This method ensures a more manageable evaluation process while maintaining adequate coverage and representation.

In its essence, a testbed functions as a pragmatic SES, offering support for evaluating various categories of things. As shown in Fig. 9, I propose a standardized evaluation methodology for a testbed as

outlined below:

The initial phase involves defining and characterizing distinct things (subjects). Next, it entails establishing SESes for these things and consolidating them into an Aggregate SES. Subsequently, the process includes defining, designing, and implementing Perfect and Pragmatic SESes. Finally, it necessitates tracing forth and back among the evaluation outcomes of Aggregate, Perfect, and Pragmatic SESes.

Acknowledgments

I am very grateful to Dr. Wanling Gao for preparing all the figures and to Dr. Lei Wang and Dr. Wanling Gao for the discussions.

References

- [1] J. Zhan, L. Wang, W. Gao, H. Li, C. Wang, Y. Huang, Y. Li, Z. Yang, G. Kang, C. Luo, H. Ye, S. Dai, Z. Zhang, Evaluatolgy: The science and engineering of evaluation, *BenchCouncil Trans. Benchmarks, Stand. Eval.* 4 (1) (2024) 100162.
- [2] J. Zhan, A short summary of evaluatolgy: The science and engineering of evaluation, *BenchCouncil Trans. Benchmarks, Stand. Eval.* (2024) 100175.
- [3] J. Zhan, Five axioms of things, *BenchCouncil Trans. Benchmarks, Stand. Eval.* (2024) 100184.
- [4] System. <https://www.merriam-webster.com/dictionary/system>, Accessed: February 6, 2024.
- [5] A. Backlund, The definition of system, *Kybernetes* 29 (4) (2000) 444–451.
- [6] I. BiPM, I. IFCC, I. IUPAC, O. ISO, The international vocabulary of metrology—basic and general concepts and associated terms (VIM), *JCGM* 200 (2012) 2012.
- [7] H.D. Young, R.A. Freedman, L.A. Ford, *University Physics with Modern Physics*, 2020.
- [8] J. Pearl, D. Mackenzie, *The Book of Why: the New Science of Cause and Effect*, Basic books, 2018.
- [9] C. Wang, L. Wang, W. Gao, Y. Yang, Y. Zhou, J. Zhan, Achieving consistent and comparable CPU evaluation outcomes, 2024, arXiv preprint [arXiv:2411.08494](https://arxiv.org/abs/2411.08494).
- [10] G.W. Imbens, D.B. Rubin, *Causal Inference in Statistics, Social, and Biomedical Sciences*, Cambridge University Press, 2015.



Dr. Jianfeng Zhan is a Full Professor at the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), and University of Chinese Academy of Sciences (UCAS), the director of the Research Center for Distributed Systems, ICT, CAS. He received his B.E. in Civil Engineering and M.Sc. in Solid Mechanics from Southwest Jiaotong University in 1996 and 1999 and his Ph.D. in Computer Software and Theory from the Institute of Software, CAS, and UCAS in 2002. His research areas focus on evaluatolgy, evaluatolgy-based design automation, and optimization automation. His exceptional expertise is exemplified by his introduction to the discipline of evaluatolgy, an endeavor that encompasses the science and engineering of evaluation; within this discipline, his proposition of a universal framework for evaluation encompasses essential concepts, terminologies, theories, and methodologies for application across various disciplines. He has made substantial and effective efforts to transfer his academic research into advanced technology to impact general-purpose production systems. Several technical innovations and research results, including 35 patents from his team, have been adopted in benchmarks, operating systems, and cluster and cloud system software with direct contributions to advancing parallel and distributed systems in China or worldwide. Over the past two decades, he has supervised over ninety graduate students, post-doctors, and engineers. Dr. Jianfeng Zhan is the founder and chairman of BenchCouncil. He also holds the role of Co-EIC of BenchCouncil Transactions on Benchmark, Standards, and Evaluations alongside Prof. Tony Hey. He has been honored with several prestigious awards for his exceptional contributions. These include the second-class Chinese National Technology Promotion Prize in 2006, the Distinguished Achievement Award of the Chinese Academy of Sciences in 2005, the IISWC Best Paper Award in 2013, and the Test of Time Paper Award from the Journal of Frontier of Computer Science.



Full Length Article

Evaluation of mechanical properties of natural fiber based polymer composite

Tarikur Jaman Pramanik^a, Md. Rafiquzzaman^a, Anup Karmakar^a, Marzan Hasan Nayeem^{b,*}, S M Kalbin Salim Turjo^c, Md. Ragib Abid^d

^a Department of Industrial Engineering and Management, Khulna University of Engineering & Technology, Khulna, Bangladesh

^b Department of Industrial & Production Engineering, National Institute of Textile Engineering and Research (NITER), Dhaka, Bangladesh

^c Department of Materials Science and Engineering, Khulna University of Engineering & Technology, Khulna, Bangladesh

^d Department of Industrial and Production Engineering, Bangladesh Army University of Science and Technology (BAUST), Saidpur, Bangladesh



ARTICLE INFO

Keywords:

Numerical simulation

Jute fiber

Composites

Mechanical evaluation

Optimized fabrication

ABSTRACT

Natural fiber based polymer composites are eco-friendly alternatives to synthetic materials, with greater mechanical properties, biodegradability, availability, ease of access, and affordability. Jute fiber is widely recognized as one of the most important and beneficial natural fibers due to its strength, durability, and biodegradability. In this study, the jute composite is designed and fabricated using a 5-layer jute and epoxy resin, utilizing the manual hand lay-up technique. The combination of 52.5 % jute and 47.5 % of epoxy resin and harder is found optimized to achieve the goals of improving the tensile strength and flexural strength, reducing the cost of epoxy resin, and promoting eco-friendliness and sustainability. Tensile testing was performed on a universal testing machine, while flexural testing was done with a three-point bending test. Experimentally, the composites reinforced with jute and epoxy resin were capable of achieving the required levels of tensile strength (42.91 MPa) and bending strength (69.30 MPa). To validate and visualize specimens, numerical analysis was performed on the ABAQUS simulation software. The numerical simulation utilized ASTM D3039 and ASTM D7264 as the specified requirements for tensile and flexural behavior. For validation, these tensile and flexural test results were then numerically analyzed and compared to the experimental data. Finally, composite design, fabrication, and optimization can improve mechanical properties, reduce composite weight, lower resin cost, and increase sustainability. The proposed design and composition can be implemented to achieve lightweight properties in various applications, such as car components, door handle sheets, bicycle seat backs, and luggage covers.

1. Introduction

A composite is created by combining two or more components with various qualities. Composite materials are created by encasing high load-bearing augmentation in softer materials (matrix). The two critical categories of differentiation of materials are matrix and other is reinforcement. One of the matrix's main roles is transferring stresses between the reinforcing fibers or particles. A composite's mechanical qualities, such as its impact strength, flexural strength, tensile strength, elasticity, etc., are increased when fibers or particles are present. Mechanical and natural damage can also be prevented. The matrix material may be reinforced before or after being inserted into the mold cavity.

Undoubtedly, one of the most important advancements in material evolution is the creation of composed fibers with associated models and production methods. Composites are a type of material with unique mechanical and physical qualities that are employed in many different industries. The advantages of composite materials over traditional materials include their tensile stress, impact resistance, bending strength, stiffness, and fatigue appearances. Because of their various benefits, applied in the aerospace sector, advertisement mechanical design systems such as equipment apparatuses, vehicles, diesel engines, and moving parts such as crankshafts, reservoirs, brake systems, compressors, and drivetrains, thermal protection and electronics industries, railway coaches, and aerostructures, etc. [1]. Polymers like thermosets

* Corresponding author.

E-mail addresses: tarikurtonmoy@gmail.com (T.J. Pramanik), rafiq123@iem.kuet.ac.bd (Md. Rafiquzzaman), anupkarmakar1711031@gmail.com (A. Karmakar), mhnyayem@niter.edu.bd (M.H. Nayeem), kalbinsalimturjo@gmail.com (S.M.K.S. Turjo), md.ragibabid@gmail.com (Md.R. Abid).

<https://doi.org/10.1016/j.tbench.2024.100183>

Received 22 September 2024; Accepted 1 October 2024

Available online 3 October 2024

2772-4859/© 2024 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

and thermoplastics combine with continuous and noncontinuous reinforcements or fillers to create polymer composites. Composites frequently incorporate polymeric materials to improve the material's effectiveness. Polymer composites are being used in an increasing number of technical fields. Jute fiber is entirely renewable and environmentally friendly because it is biodegradable. Due to its golden and silky shine, it is a natural fiber known as Golden Fiber. Usually, worldwide creation, manufacture, and accessibility, along with vegetable fibers, come in second behind cotton. It promotes increased fabric permeability while having a high tensile property and limited adaptability. Jute is, therefore, perfect for packing agricultural items in bulk [2].

Composites with natural fiber compounds are becoming increasingly popular because they can replace traditional synthetic material composites while being more environmentally friendly. The stiffness-to-weight ratio of the resultant composites is improved because natural fibers are lighter than glass fibers, which have a lower density ($\rho = 1.3 \text{ g/cm}^3$). The principal stem-type natural fibers that are native to India, Bangladesh, and Nepal are jute fibers, which are also utilized quite extensively. Cotton is the other top producer, followed by jute and related fibers, according to the 2019 World Natural Fiber Production Report [3]. Normally, composite materials can be classified into different types. These types are shown in Fig. 1.

Composite materials, polymers, and ceramics have recently been the most popular developing engineering materials. Organic fiber is popular and environmentally friendly. Natural fiber's obtainable characteristics and simplicity of production have motivated researchers all around the world. They were able to test out locally accessible, less expensive fiber options to see how much they met the criteria for a well-reinforced polymer composite for structural use. Natural fibers were generally employed in composite materials to increase bulk and lower costs as opposed to increasing mechanical qualities. However, the manufacturing and usage of synthetic fibers, combined with environmental issues, have altered the scenario. Historically, both organic and compostable matrices have regularly used natural fibers as reinforcement components. Despite having superior flexural and impact qualities, minor improvements in tensile strength of natural fiber reinforcements have been a focus of research. Numerous initiatives have been made to enhance mechanical characteristics, including the addition of filler and chemical treatment [4]. Like any other natural fiber, jute fiber exhibits natural variability in its exterior and inner mechanical properties; it is influenced by various variables, such as increasing circumstances (such as air temp, moisture, and surface status), 'retting' (fluid, fog, and enzymatic activity) and fiber separation procedures, fiber shape and size, natural substances, and the proportionate amounts of each. The fiber's structural, physiological, and environmental properties are also influenced by the fiber's microstructural features. The overall architecture of the jute fiber is covered in the first part of this section, which is followed by examples of how it performs as fiber, yarn, and woven or nonwoven fabric. To increase its effectiveness for a particular application, jute fiber is functionally treated in some cases. Natural fiber-based goods are drawing a lot of interest from academic and industrial researchers looking to produce sustainable products because of their low carbon footprint. The development of vegetable crops, seed and plant entomology deviation evaluation at various situations, retting process,

plant mineral treatment, biological DNA series, and multifaceted of extensively used natural fibers, uses of jute fiber in research and innovation, which would include material for apparel, have recently sparked renewed study interest. Natural, social, and environmental progress are all interconnected and can be directly linked to the rising popularity of jute fibers [5]. In this study, we designed and fabricated the jute composite using a 5-layer jute composite with epoxy resin, utilizing the manual hand lay-up technique. We evaluate the combination of 52.5 % jute and 47.5 % epoxy resin to improve the tensile strength and flexural strength, reduce the weight of the composite, reduce the cost of epoxy resin, and promote eco-friendliness and sustainability. The specimens are tested experimentally utilizing various tests. The evaluation is also carried out using numerical simulation in ABAQUS software.

2. Literature review

2.1. Natural fiber-based polymer composites

Natural fibers composites have been increasingly popular in recent years due to their numerous appealing qualities, including biocompatibility, lack of abrasion resistance, adaptability, accessibility, affordability, and ease of production. Researchers have conducted many studies to enhance the mechanical properties of organic nutrient composite materials. Cazaurang et al. investigated henequen fiber's characteristics thoroughly, and it was noted that these fibers had mechanical qualities that make them acceptable for reinforcing in thermoplastic resins [6]. Sweeti Shahinur et al. explored that organic, recyclable, and biopolymers are critically needed to replace environmentally hazardous synthetic fabrics from a sustainability perspective. One of the natural fibers, jute, is essential in creating composite materials that have the potential to be used in a range of applications, including home, industrial, and medical devices [5]. Schneider and Karmaker inquired about the mechanical behavior of polypropylene matrix based on jute and kenaf fiber, stating that jute fiber offers superior mechanical qualities to kenaf fiber [7]. Joseph et al. observed fibers, such as silk, pineapple fiber, an empty bunch of fruit fiber from the oil palm, etc., exhibit physical and mechanical activity [8]. George et al. examined how well cellulose fiber performed in polypropylene cellulose composites to increase stiffness and decrease damping [9]. Gowda et al. looked into the physical behavior of jute fiber composites and found that jute fibers composites exhibit larger strengths than those composed of wood [10]. Pavithran et al. reported the fracture energies for polyester composites reinforced with sisal, pineapple, banana, and coconut fibers, and it was observed that except for coconut fiber, an increase in fiber toughness was accompanied by a rise in fracture energy. They also demonstrated the mechanical characteristics of flax/polypropylene composites [11]. Rafiquzzaman et al. employed notched and unnotched specimens to experimentally and quantitatively analyze how composite layering systems behave mechanically. Then, a mathematical procedure incorporating the finite element technique was used to evaluate the overall corrosion behavior of the uniaxial and open-hole polymer thermoplastic composite composites under experimentally applied stress [12]. Aditya et al. found that hybrid FRC composed of Sisal and Pineapple exhibits a higher elastic modulus, whereas FRC with date palm demonstrates enhanced impact strength [13]. Gassan et al. found that the improved quality of the fiber-matrix adhesion reduced the loss of energy on non-penetration impact-tested jute fiber composites [14]. Rajesh et al. found that natural fiber composites using synthetic fiber hybridization can be included in automobile sectors and bullet proof vest [15]. Harish et al. used coir fiber reinforced composite in the mechanical test evaluation and found that coir/epoxy composites exhibit average values for the tensile strength, flexural strength and impact strength of 17.86 MPa, 31.08 MPa and 11.49 kJ/m², respectively [16].

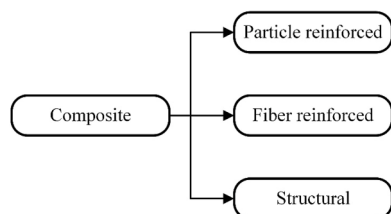


Fig. 1. Identification of composite.

2.2. Jute fiber-based polymer composites

The study [3] showed jute fiber characteristics and surface alterations to improve the presentation of their compatibility with the polymer matrices. A survey of jute-based polymer composites is the focus of this paper. The mechanical properties of the various thermosets, thermoplastic, biobased resin, and hybrid jute composites, as well as their composition, are explained. Muhammad Yasir Khalid et al. experimented with the tensile characteristics of hybrid composites reinforced with natural and synthetic fibers. Different glass and jute fiber stacking sequences were used using hand-prepared glass-jute hybrid composites. The experimental findings show that lower jute fiber concentrations were the only ones that had an impact on the tensile characteristics of glass fiber-reinforced polymer [17]. Shahinur et al. observed that chemically treated fibers were found to absorb less heat than untreated fibers. In every instance involving the treated fibers, the heat flow went negative, as did the jute fiber. For the production of composite materials based on polymers, this study offers crucial information regarding the thermal properties of the treated jute fibers [18]. Prasath et al. utilized computer-assisted universal testing machines and charpy impact testing machines, and mechanical properties of manufactured composite plates subjected to tests such as flexural strength and impact strength of the various specimens are estimated. Based on the findings, it can be concluded that a combination of pure basalt fibers retains better values in both flexural and tensile tests [2]. Gupta et al. focused on jute fiber-reinforced polymer composite's mechanical characterization. In this work, the mechanical characteristics of JFRPC—including its tensile, flexural, and impact characteristics—are examined. Additionally, it describes how several factors, including fiber content, fiber size, stacking sequence, and chemical modification, influence the mechanical characteristics of JFRPC [19]. Ovali et al. established the lack and existence of acrylic acid additions, and the effects of jute fabric surface modifications on the strength properties, flexural modulus, and higher strength properties of the LDPE/jute composites were examined [20]. Ramakrishnan et al. reviewed that for structural applications requiring low to medium strength, jute composite can be a good alternative material. Based on the encouraging findings of the current investigation, it was intended to create green composites and examine their static and dynamic mechanical characteristics [21]. Sakthi et al. studied various mechanical features of fly ash-weaved natural fibers. The samples were constructed manually using Taguchi's orthogonal arrays, jute fibers, fly ash, and various chemical fiber exposure period amounts. Different machine-learning regression models are used to identify the relationship between input and output properties [22]. After being alkali-treated, Sajin et al. characterized the jute fiber optic composites' thermal, mechanical, and morphology properties. The analysis above urges a large impact on the polymer industries by utilizing the developed ecological composites in diverse lightest and greater hardness applications [23]. Jute composites showed lower impact results due to the higher interface adhesion. The higher interface adhesion between the matrix and the fibers produces a lack of energy absorption mechanism in the impact test. Jute composites showed good mechanical properties compared to other natural fibers because of the higher wettability of the fibers by the low initial viscosity thermoset resin [24]. Balcioglu et al. investigated the mechanical properties of SiC filler jute fiber composites and found that tensile behavior is superior to the impact test. Additionally, filler can be used to increase the lifetime of compression in jute fiber composites [25]. The 5 % NaOH-treated fiber-reinforced polyester composites have a 15.6 % increase in flexural strength compared to the 10 % NaOH-treated jute fiber-reinforced polyester composites. In contrast, it was 20 percentage for jute-epoxy composites. The jute-polyester composite seemed to have better impact energy than jute-epoxy composites [26]. It is found that 0° composite orientations are capable of absorbing sufficient impact energy for 5 ms⁻¹ but not for velocity greater than 10 ms⁻¹. When fiber orientations were used between 15° – 45°, the composite impact

resistance increased, indicating two significant peak forces [27].

2.3. Numerical simulation on composites

Alemi-Ardakani et al. used Abaqus/Explicit to simulate the 200 J collision of composites made of fiberglass and polypropylene. The fabric was progressively harmed when using the constructed failure criterion damage criterion. The preliminary simulation, built on the material characteristics from nonlinear static test cases, differed significantly from the outcomes of the destructive testing [28]. Jensen et al. evaluated a full-scale composite wind energy blade for fracture against tendon pressure. The development of local displacement measurement technology allowed for the recording of displacements throughout the loading history. Local displacement measurements were used to locate the point at which the catastrophic failure was initiated [29]. Torre et al. investigated the sandwich construction's ability to absorb energy when hit by a single impact and the creation of criteria that can be used to choose materials. Compared to conventional sandwich structures, corrugated sandwich panels have demonstrated superior strength and energy absorption capabilities [30]. Fish et al. described several techniques regarding matrix nutrients and foundation. The topic of selecting the right scale is covered and discussed with matrix nutrients among the various temporal applications [31].

2.4. Morphology of jute fiber

Jute grows significantly and has very important features, such as the external plants being "individually tailored" to produce fabric, and the interior stalk and external plants being divided. The parts are divided and washed to get rid of dust from the plant. The fiber is sent to jute mills for conversion into hessian and jute yarn after cleaning. Due to government organizations' assistance for R&D and also because of the jute, a variety of lifestyle items are manufactured from it and expanded into several forms [32]. Since jute fiber is entirely biodegradable, reusable, and green, it is a good choice for the environment. The term "Golden Fiber" refers to the natural fiber's golden and silky sheen. It guarantees that fabrics are better breathable and have a high yield strength and minimal flexibility. Because of this, jute is ideal for bulk packaging of agricultural products. Making the highest quality commercial yarn, fabric, net, and bags is made easier by this. It is among the most adaptable natural fibers utilized as raw materials for the packing, textile, nontextile, building, and agricultural industries. When yarn is bulked, the resulting ternary blend has a lower breaking tensile strength and a higher breaking elasticity [2].

Composites that were treated with NaOH, and supplemented with nano-clay had their dynamic mechanical and physical properties and vibration properties examined by Ramakrishnan et al. It is assumed that sodium hydroxide treatment (NaOH) enhances the mechanical characteristics by partially expelling hemicellulose and lignin and roughening the fiber surface, which produces an adhesion between the polymer and fiber that functions as an anchor. Another expectation is that the fiber and polymer will form a strong bond as a result of the hydroxyl group reaction with sodium hydroxide [21]. The mechanical characteristics of polymer hardness are the complete list of thermoplastics demonstrated by LDPE, followed by polylactide and PVC, while the higher impact strength is demonstrated by polypropylene. PVC, followed by poly (lactic acid) and Polyethylene, has the highest density of any thermoplastic material. Of the thermoset polymers addressed, resin does have the best strength properties, followed by thermoplastics and thermosetting polymers, while polyphenol has the ultimate tensile flexibility. Table 1 shows the mechanical properties of jute fiber [19].

2.5. Epoxy and binding element

Epoxy resins are the thermoset material most frequently utilized in polymer matrix composites. They are a class of thermoplastic plastic

Table 1
Mechanical properties of jute fiber.

Properties	Amount
Moisture content (%)	1.1
Tensile strength (MPa)	393–773
Pectin (%)	0.2
Diameter of fiber(μm)	5–25
Density (g/cm ³)	1.46
Hemi-cellulose(%)	12
Elongation (%)	1.16–1.5
Micro-fibrillar angle (°)	8.0
Cellulose (%)	64.4
Young modulus (GPa)	13–26.5
Fiber length (mm)	0.8–6
Lignin (%)	11.8
Price (EUR/kg)	0.3
Waxes (%)	0.5

materials that do not emit reaction products during curing. As a result, they have low cure deformation, good adhesion, chemical and insulating capabilities, and enhanced biological and chemical resistance. Since the polymerization reaction is needed to create their results in varying chain lengths, epoxy resins are polymorphic or semi-polymeric compounds rarely found as pure substances. For some uses, highly pure grades can be produced. Purified liquid grades can form crystal solids due to their extremely regular structure, which necessitates melting to process them. A type of thermosetting resin known as epoxy resins is created through the hexagonal polymerization of substances with, on average, over one epoxy component per molecule.

2.6. Comparison of other composite material property

Jute, sisal, banana, and coir are the most common natural fibers produced around the world. These fibers are commonly utilized for various applications, such as cordage, sacks, fishnets, matting, and rope, as well as stuffing for mattresses and cushions. Cellulosic fibers can be obtained from several sections of plants. The economical, biodegradable jute goods combine with the soil after use, adding to the soil's accretion. Jute burns with no harmful fumes because it is formed of cellulose. Due to its low density and relatively stiff and robust behavior, jute fiber's unique characteristics can be compared to those of glass and some other fibers. As compared to other natural fibers, jute has a high tenacity and aspect ratio. Jute is a type of cellulosic fiber, and its composites have intermediate tensile and flexural strength with good impact strength. The world focuses on the inherent qualities of jute fiber, such as its low density, low elongation at break, and unique stiffness and strength

comparable to those of glass fiber. Composites with the same system of reinforcing materials may not perform better since they are subjected to a variety of loading circumstances over the course of their service life. Hybrid composites are the greatest answer for these applications in order to address this issue. In a hybrid composite, one type of fiber balances the lack of another fiber by combining two or more different types of fiber. Hybridization aims to produce a new substance that will retain the positive attributes of its parts while excluding their negative ones. Based on the types of reinforcement, polymer composites can be divided into particle-reinforced polymer composites and fiber-reinforced polymer composites. Particulate composites, also known as particle-reinforced composites, include reinforcing material in the form of particles. There may be different reinforcing particles, such as spherical, platelet-shaped, cubic, tetragonal, or of other regular or irregular geometry. Table 2 shows the comparison of properties of different composite materials.

In most of the cases, tensile strength and flexural strength were measured to determine the mechanical properties of the fabricated composites. Some studies focused on impact strength also. So, tensile and flexural tests are the most critical tests to focus on for composites, as they offer comprehensive data on the composites' elasticity, durability, and resistance to deformation under various loading conditions.

3. Methodology

3.1. Study design

The step-by-step procedure for this study is shown in Fig. 2. It starts

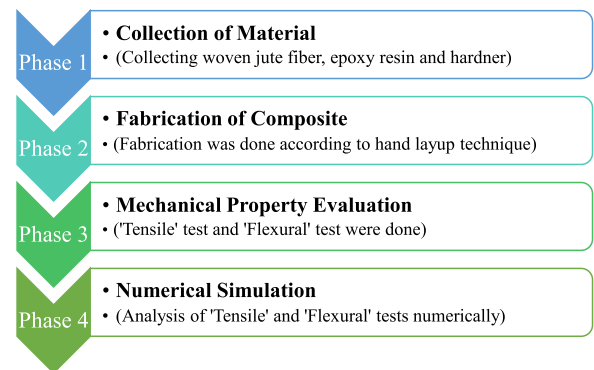


Fig. 2. Flow diagram of the working procedure.

Table 2
Comparison of properties of other composite materials.

Reinforced materials	Tensile strength (MPa)	Flexural strength (MPa)	Impact strength (J/m ²)	Applications	References
Basalt + jute	130	410	22	Roofing tiles, insulation panels	[33]
Glass + carbon	286.4	314.4	–	Marine industry, sports equipment	[34]
Kenaf + jute (K/J/K)	43.21	75.7	–	Packing material and material that absorbs oil and liquids	[35]
Carbon + basalt	354	400	–	Pipelines, beams, various car parts	[36]
Oil palm + kenaf	62	110	1.3	Building materials and animal feed	[37]
40 % Jute + resin	39.67	65.87	178.56	Textile, automobile	[1]
Coconut leaf sheath (CLS) + jute + glass	12.25	59.7	22.8	Roofing tiles, wall panels, furniture industry	[38]
Aramid + kenaf	202	15	34.8	Textile industries, insulation	[39]
Carbon + kevlar	388	2029.2	–	Defense industry, automotive industry	[40]
Carbon + flax	–	–	37	Printed banknotes and rolling paper for cigarettes	[41]
20 % Magnesium hydroxide + 40 % of kenaf	39	88	29	Building materials, packaging materials	[42]
Carbon fiber + 10 % carbon black	108.2	103.3	18.7	Automotive and aerospace industry	[43]
24 % kenaf + 16 % banana (plain woven)	140	170	–	Furniture, boxes	[44]
29 % Carbon + 14 % flax	222.63	–	–	Renewable energy	[45]

with the collection of materials and ends with the analysis of the result. Mechanical and numerical analyses were done for both tensile and flexural/bending tests.

3.2. Collection of material

3.2.1. Woven jute fiber

Woven Jute Fiber was obtained from the nearby jute mills. The woven jute fiber was acquired from nearby jute mills in Khulna, Bangladesh, with an overall body mass of 1.3 g/cm^2 and a depth of 3 mm. Jute fibers have significant benefits, including minimal cost, environmental friendliness, and reasonable physical qualities. Fig. 3 depicts a sample of woven jute fiber.

3.2.2. Epoxy resin and hardener

Epoxy resin and hardener were used in this fabrication. The resin-hardener mixture had a 3:1 ratio. Epoxy resin is a type of resin with excellent mechanical qualities, excellent resistance to chemicals, and high adhesion strength, making it extremely useful. It has numerous applications in technological and industrial areas. Curing occurs at huminitic conditions with the use of hardener. In the current study, epoxy obtained from a local chemical plant is used.

3.3. Fabrication of composite

Significantly, many different techniques and methods are used to create composites. The manual mixture of fabrication is one of the most straightforward ways to make composites. The manual mixture process was used to create all the composite layers. As reinforcement, five layers of jute fiber and epoxy are utilized as the matrix substance. Jute of 52.5 % and 47.5 % of resin and harder combination are used for making composite materials. Three respects to one epoxy are used to create the matrix. To strengthen matrix adhesive properties and provide strength to the composites, Epoxy resin was utilized as the matrix, along with hardener. The fabrication procedures performed are shown in Fig. 4.

3.3.1. Step-by-step procedure of fabrication

The fabrication procedures were performed sequentially as below:

- Step 1: The experimental bench was covered with wrapping paper to create a smooth surface for construction. All fibers cut according to the design were laid on the bench.
- Step 2: The fibers were cleaned and sun-dried. A 3:1 mixture of resin and hardener was combined in a ceramic dish. A spinner was used to dilute the epoxy and hardener until the hardener was entirely combined with the epoxy. An open mold was designed for the fabrication procedure.
- Step 3: Next, a matrix layer is applied to the fabrication area inside the mold cavity. A roller, such as a pen, was employed to ensure layer consistency.
- Step 4: Jute fiber is laid on the matrix layer, and a die is used to provide pressure to fix the matrix layer correctly.
- Step 5: Again, the matrix layer is deposited on the jute mat using a roller, and the jute mat is placed on top of the matrix layer. This method was repeated, and five layers of jute fiber were used. A



Fig. 3. Woven jute fiber.

combination of 52.5 % jute and 47.5 % epoxy resin and harder was used.

- Step 6: The sample was then wrapped in plastic wrappers and crushed with a couple of blocks.
- Step 7: To obtain excellent composite strength, a cure period of at minimum 70–72 hours was specified. The molds were shattered, and the components were withdrawn after they had thoroughly cured. The fabrication apparatus was completely dismantled.
- Step 8: All specimens were scaled to the desired dimension for various mechanical tests.
- Step 9: Finally, specimens were cut using grinding equipment from a nearby machine shop and were ready for tensile and flexural/bending tests.

3.4. Mechanical property evaluation

Mechanical characterization, tensile and flexural test were performed following fabrication. As per the literature review the impacat test shows less significance than tensile and flexural, the study focused on identifying optimum fiber composition in evaluating best mechanical property. This was accomplished through the use of a tensile and flexural test. Many researchers based their study on the results of these experiments.

3.4.1. Tensile test

Composite materials are tested in several ways, such as the tensile test. This test is used to evaluate elastic and plastic deformation. It determines the needed force as well as the elongation point at break. Nowadays, the uniaxial test is widely performed. Several characteristics were required to examine the specimen. Again, tensile testing is a basic form of resting process used by scientists and researchers. This test is necessary for examiners for new product development, design, manufacture, and prototype testing. This approach is used to measure stress and strain parameters. This is also used to generate a stress-strain curve. It is essential throughout study and innovation to determine acceptable materials. It may also be performed to ensure that substances meet the required hardness and elasticity specifications. Table 3 shows the tensile test result.

ASTM D3039 was used as the standard for specimen size. To begin, a grinding machine was used to cut the specimen into standard size. Then, it was set into the universal testing machine's jaws. The lower half of the specimen was then fixed, and the upper part of the specimen was loaded. Determine the force required and the elongation point of break.

3.4.2. Flexural test

Flexure experiments are commonly used to assess a product's bending stress or fracture toughness. Deflection examinations are less cost-prohibitive than other testing; however, the results can differ slightly. A sample is placed parallel to the ground above two different interaction locations (minimum operating frame). Then, stress is transmitted to the material's top via one or more locations of interaction (axial load frame) till the sample fails. Also, the sample's strength and stiffness are represented by the measured maximum force.

In this experiment, the three-point bend test is used. The specimen was placed horizontally at the top two points, and the force was delivered to the sample's upper surface through a single point such that the sample was curved in the shape of a "V." The three-point flexure test is excellent for assessing a single sample location. ASTM D7264 was used for sample dimensions $120 \text{ mm} \times 20 \text{ mm} \times 5 \text{ mm}$ and cut by using a grinding machine. Fig. 5 represents the specimen for the bending test.

The wheel was then used to progressively apply load to the center of the specimen. The sample is broken, i.e., fractured, at a particular load. The displacement was measured using a scale placed in the center of the specimen. The corresponding load is noted for the gradually rising distortion of the specimen, and then calculation is required to determine bend stress.

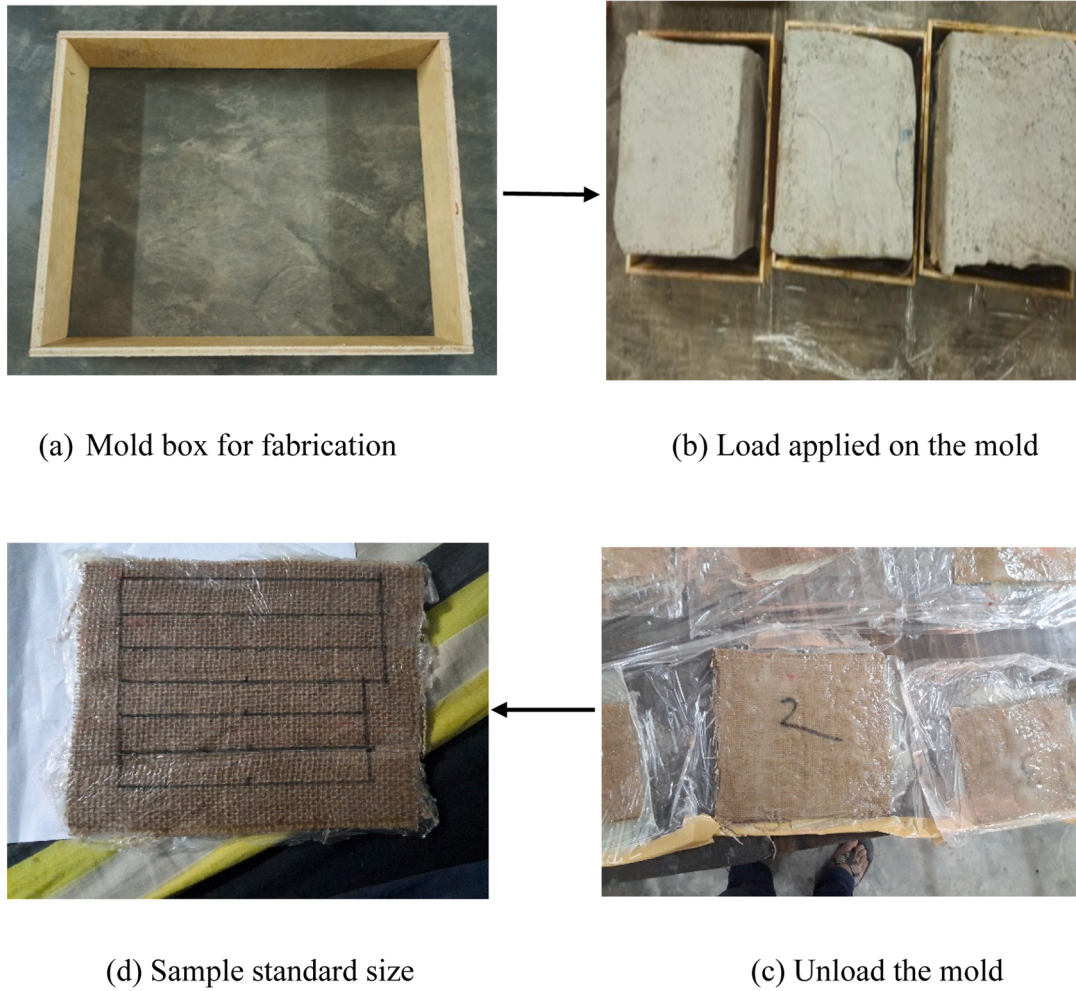


Fig. 4. Fabrication of jute fiber composite.

Table 3
Tensile test result.

Load (kN)	Displacement (mm)	Stress (MPa)
0.310	0.026	7.48
0.635	0.052	15.21
0.900	0.074	21.64
1.095	0.080	26.32
1.310	0.101	31.47
1.490	0.123	35.77
1.650	0.136	39.67
1.785	0.147	42.91

$$\begin{aligned}
 A &= \text{Cross Sectional Area} \\
 &= \text{Width} \times \text{Thickness} \\
 &= b \times h
 \end{aligned}$$

Tensile test results for different specimens are shown in Table 3. The displacement versus stress curve is given in Fig. 6.

As shown in Fig. 6, the stress rises as the displacement value rises. As a result, a high displacement rate implies higher tensile strength. The standard value is compared to the research work of a journal paper that is cited. Table 4 shows the tensile strength for different jute weights [1].



Fig. 5. Specimen for bending test.

4. Results and discussions

4.1. Experimental investigation of tensile test

Tensile Strength, $S_t = F / A$
Here, F = Force

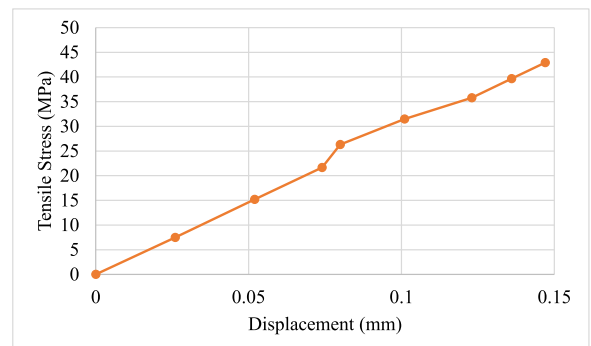


Fig. 6. Displacement versus tensile strength curve.

Table 4
Reference tensile strength of jute fiber.

Jute weight (%)	Tensile strength (MPa)
40	39.67
30	35.77

4.2. Experimental investigation of flexural test

Bending Strength, $\sigma = MC / I$

Here, M = Internal bending moment about the sections of the neutral axis

= Force \times Distance

= $P \times a$

C = Perpendicular distance from the neutral axis to the furthest point on the section

= thickness / 2

= $h / 2$

I = Moment of Inertia

= $1/12 \times \text{width} \times (\text{thickness})^3$

= $1/12 bh^3$

Bending test results for different specimens are shown in Table 5. The displacement versus flexural strength curve is given in Fig. 7.

As shown in Fig. 7, the flexural stress increases as the displacement value increases. As a result, a high displacement rate indicates greater flexural strength. The standard value is compared to the research work of a cited journal paper. Table 6 shows the flexural strength of different jute weights [1].

5. Numerical analysis

5.1. Tensile test

5.1.1. Numerical model

Fig. 8 shows the dimensions of the numerical model. The geometry and material information listed below are needed to model this scenario. Used standard is ASTM D3039 [46–49].

Layer of the Specimen:

Thickness: total 5 mm, each layer 0.5 mm

Layer 1,3,5,7,9 = Jute fiber

Layer 2,4,6,8,10 = Epoxy Resin

5.1.2. Defining the geometry

The main geometric model is created by using ABAQUS Workbench. The geometry created by using Abaqus is displayed in Fig. 9.

5.1.3. Material properties

The mechanical behavior of the finite element model's parts is defined by material models. Young modulus of 26,500 MPa and Poisson's ratio of 0.4 were selected as the specimen mechanical properties of jute fiber. Abaqus was used to modify the properties. Table 7 shows the properties of jute fiber and epoxy resin [1].

Table 5
Bending test result.

Load (kN)	Displacement (mm)	Stress (MPa)
0.015	0.112	8.13
0.025	0.197	14.27
0.045	0.394	28.54
0.06	0.549	41.27
0.08	0.709	51.32
0.085	0.776	56.18
0.95	0.864	62.53
0.105	0.957	69.3

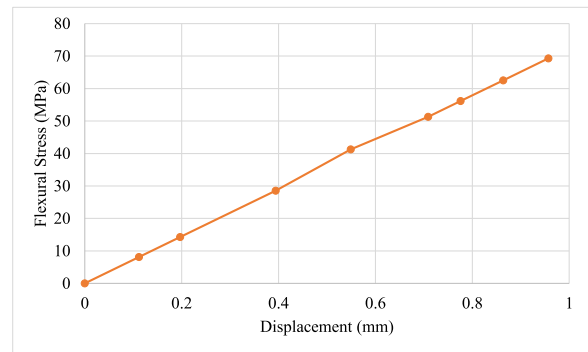


Fig. 7. Displacement versus flexural strength curve.

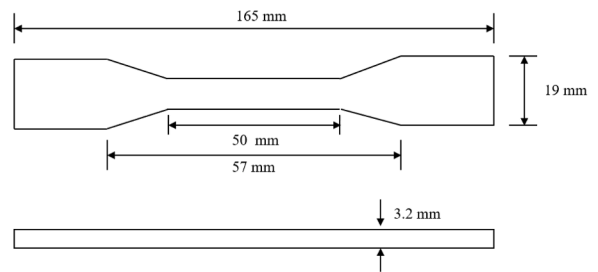


Fig. 8. Dimension for the numerical model.

Table 6
Reference flexural strength of jute fiber.

Jute weight (%)	Flexural strength (MPa)
40	65.87
30	62.87

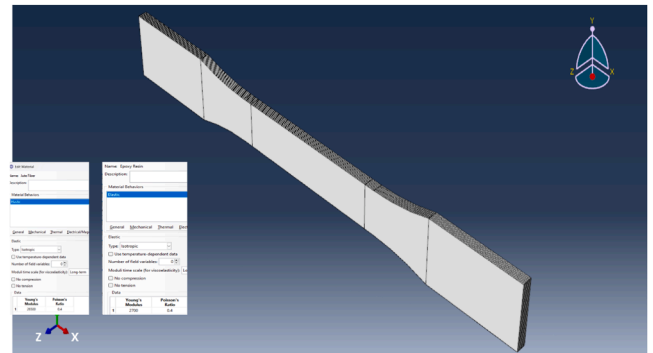


Fig. 9. Specimen geometry for tensile test.

Table 7
Properties of jute fiber and epoxy resin.

Properties	Young modulus (GPa)	Poisson's ratio	Density (g/cm ³)	Specific Gravity (gm/cc)
Jute Fiber	26.5	0.4	1.3	1.3
Epoxy Resin	2.7	0.4	1.2	1.8

5.1.4. Meshing

The follow-up interviews stage helps to divide the uninterrupted rigid face geometry, also known as meshing. Here, the general mesh is employed. In total, there are 8040 elements and 9996 nodes. The model

for numerical analysis meshing is displayed in Fig. 10.

5.1.5. Boundary conditions

First, one fixed support is placed on one end of the specimen in the other half to provide force to the geometry. This aids in limiting the degrees of freedom between any ends.

The upper part of the board is then subjected to a force in the Positive x direction. Here, the force was applied to one side of the body. The right face and force are applied on the geometry, with the right magnitude and direction [46] [49]. The boundary conditions are shown in Figs. 11 and 12.

5.1.6. Total deformation

The contour graphic represents the overall deformation in Fig. 13. Upon first inspection, the anticipated displacements appeared to be perfect. The experimental findings indicate a maximum deformation of around 0.147 mm.

Also, the simulated specimen was thought to be a homogenous material; the greatest deformation in this FEM result is 0.147 mm, which stress differs slightly from the experimental value.

The numerical results of the tensile test are shown in Table 8. Fig. 14 shows the displacement versus tensile stress curve.

Fig. 15 shows the numerical simulation data of fabricated composite laminates. For this the material property had not enough plasticity, therefore the curve was straight. The substance was broken down at its peak.

5.1.7. Comparison between numerical and experimental results

A comparison of experimental and numerical results is shown in Fig. 16. The deviation of the numerical and experimental analysis is acceptable. In both analyses, the displacement value for stress is relatively similar. As a result, we can conclude that the experiment and numerical results are identical.

5.2. Bending test

5.2.1. Numerical model

The geometry shown in Fig. 17 is needed to model this scenario. Used standard is ASTM D7264 [47,50–52].

5.2.2. Defining the geometry

The main geometric model is created by using ABAQUS Workbench. The geometry created by using Abaqus is displayed in Fig. 18.

5.2.3. Material properties

The numerical simulation model's parts' mechanical behavior is defined by material models. Jute fiber's Young Modulus of 26,500 MPa and Poisson ratio of 0.4 was chosen as the specimen's mechanical characteristics. Utilizing Abaqus, the attributes were altered.

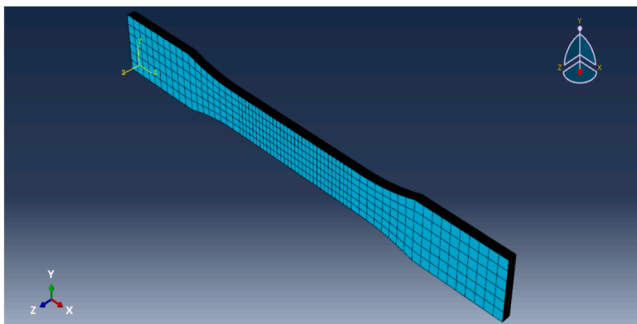


Fig. 10. Meshed specimen for tensile test.

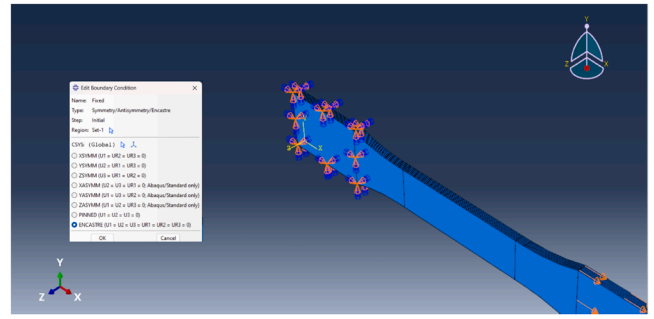


Fig. 11. Fixed support of the specimen for tensile test.

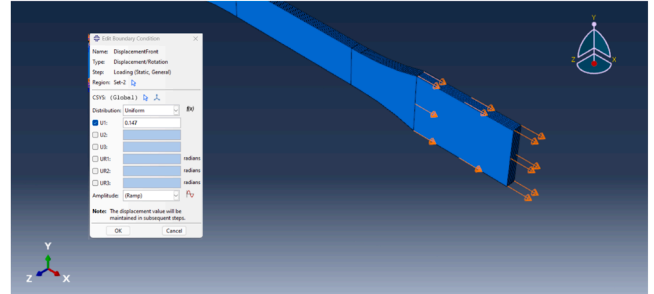


Fig. 12. Forced applied one side of the body for tensile test.

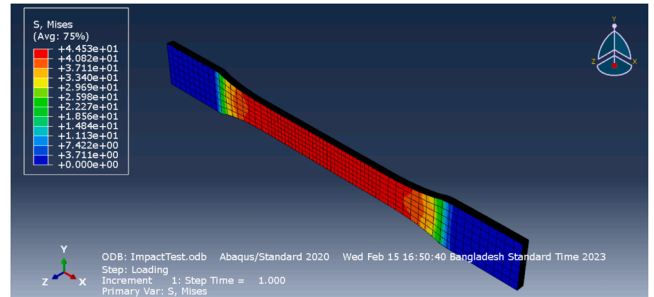


Fig. 13. Contour plot of total displacements for tensile test.

Table 8

Numerical results of tensile test.

Displacement (mm)	Stress (MPa)
0.026	9.16
0.052	16.83
0.074	23.57
0.080	27.94
0.101	33.42
0.123	38.02
0.136	41.67
0.147	44.53

5.2.4. Meshing

The follow-up interviews stage helps to divide the uninterrupted rigid face geometry, also known as meshing. Here, the general mesh is employed. In total, there are 9398 elements and 10,818 nodes. The model for numerical analysis meshing is displayed in Fig. 19.

5.2.5. Boundary conditions

To give the geometric force, two connecting pillars are first put on the two corners of the platform in the lower half. As a result, the boundary conditions between any two corners are restricted. The result

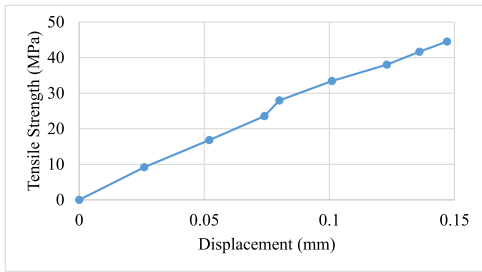


Fig. 14. Displacement versus tensile stress curve.

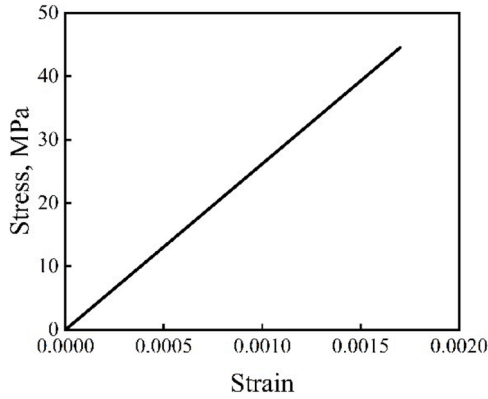


Fig. 15. Numerical stress-strain curve of tensile test.

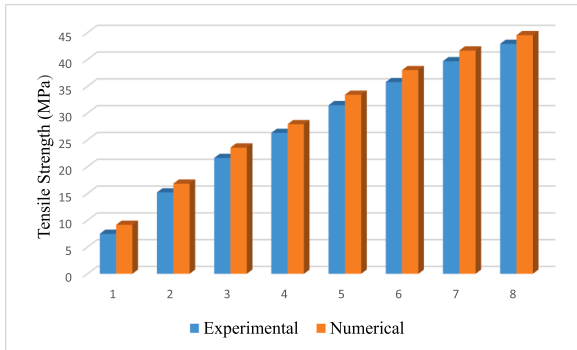


Fig. 16. Comparison of numerical and experimental results.

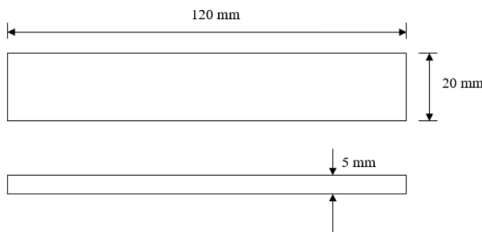


Fig. 17. Numerical model for bending test.

is comparable to Fig. 20. The loading nose was applied in the negative y direction shown in Fig. 21.

The upper part of the board is then subjected to a force in the negative z direction. Here, the force was applied using the 1 mm of the central part of the body. On the geometry, the right face and force are applied, with the right magnitude and direction [50,53,54].

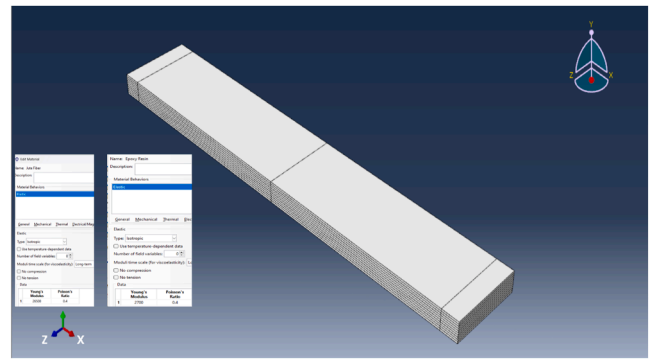


Fig. 18. Specimen geometry for bending test.

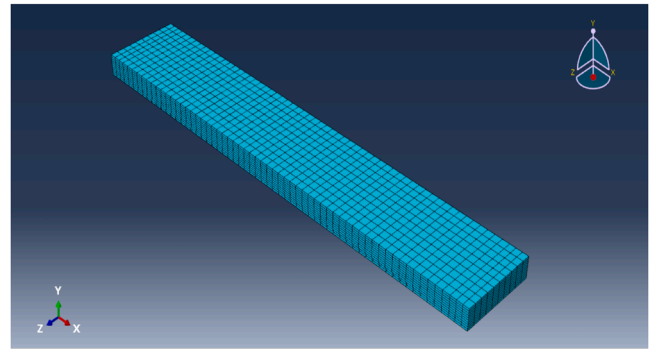


Fig. 19. Meshed specimen for bending test.

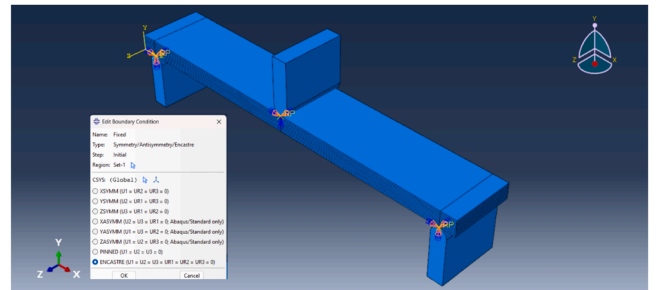


Fig. 20. Fixed support of the specimen for bending test.

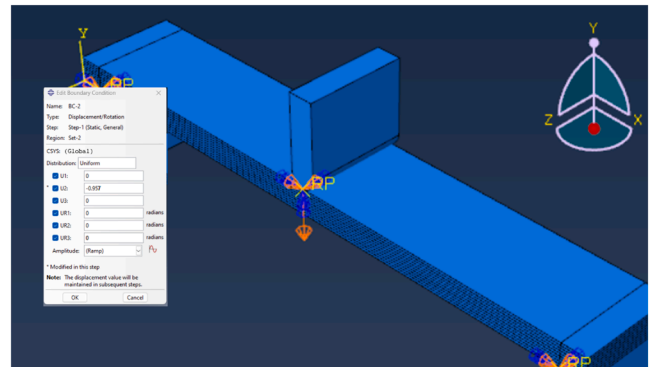


Fig. 21. Forced applied on the middle portion of the body for bending test.

5.2.6. Total deformation

A contour graphic representing overall deformation is shown in Fig. 22.

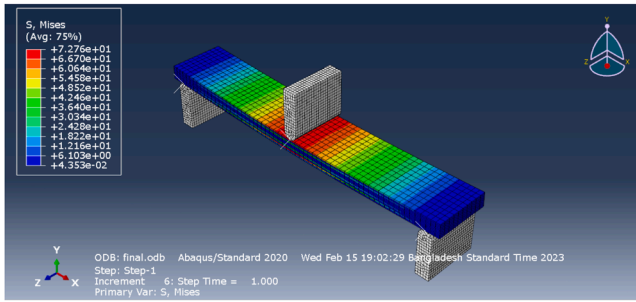


Fig. 22. Contour plot of total displacements for bending test.

On first inspection, the anticipated displacements appear to be perfect. The experimental findings indicate a maximum deformation of around 0.957 mm. Also, the simulated specimen was thought to be a homogenous material; the greatest deformation in this FEM result is 0.957 mm, which stress differs slightly from the experimental value.

Here, the numerical result of bending test is shown in Table 9. Fig. 23 shows the displacement versus bending stress curve.

Fig. 24 is a graph of numerical simulation data, and the stress-strain curve is generated using this data. For this the material property had not enough plasticity, therefore the curve was straight. This material behavior is similar to jute-glass reinforced epoxy composites [55]. The substance was broken down at its peak.

5.2.7. Comparison between the numerical and experimental results

Here is a comparison of experimental and numerical results. There is a very slight variation here. In both analyses, the displacement value with respect to stress is relatively similar. As a result, we can conclude that the experiment and numerical results are identical. Fig. 25 shows the comparison of numerical and experimental results.

Measurement entails acquiring quantitative data regarding a single property of a subject. In contrast, evaluation incorporates the larger context of using a well-defined evaluation condition (EC) to assess the subject's overall performance. For relevant subject comparisons and analyses, a well-defined EC is necessary [56]. We establish equivalent evaluation conditions (EECs) by ensuring that the same mechanical tests (e.g., tensile strength test, flexural strength test) are applied uniformly across all composite samples. This approach guarantees that the evaluation outcomes are comparable. Authentic and consistent evaluation results are achieved through the rigorous application of EECs.

In previous research work, Rafiquzzaman et al. [1] used 40 % jute and 60 % epoxy resin and harder. Their flexural strength and tensile stress were measured at 65.87 MPa and 39.47 MPa, respectively. The high level of epoxy resin content in their fabricated composite resulted in several limitations, leading to increased costs. When we found significant limitations to using a combination of jute fiber and other materials, we designed and fabricated 52.5 % jute and 47.5 % epoxy material. For the optimization of the layer design, we combined five layers of jute fiber and epoxy resin, the design of which also increases the value of mechanical properties. In our optimistic design and fabrication, we finally found that the mechanical properties of flexural stress

Table 9

Numerical results of bending test.

Displacement (mm)	Stress (MPa)
0.112	12.02
0.197	17.78
0.394	32.16
0.549	43.32
0.709	54.71
0.776	59.64
0.864	65.85
0.957	72.76

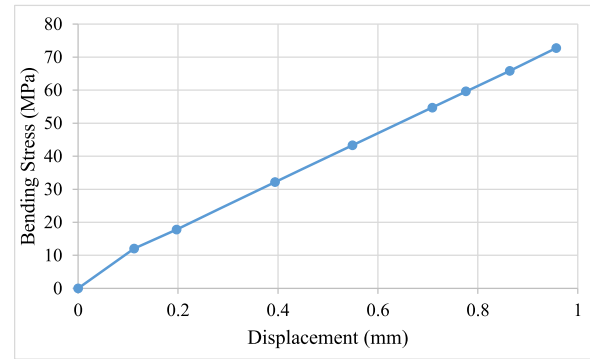


Fig. 23. Displacement versus bending stress curve.

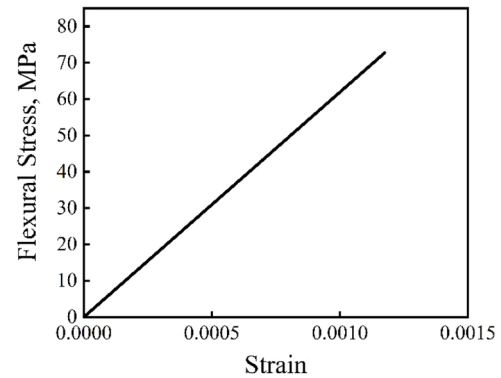


Fig. 24. Numerical stress-strain curve for flexural test.

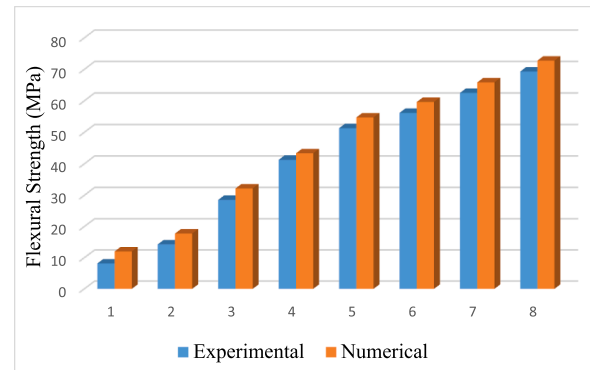


Fig. 25. Comparison of numerical and experimental results.

and tensile stress are 69.30 MPa and 42.91 MPa. So, our design and manufacturing boosted flexural strength by 5.2 % and tensile strength by 8.7 %. We observed that there were significant limitations in using the percentage of jute fiber and resin. Epoxy resin is more expensive than natural fibers like jute. Our design uses a higher percentage of jute, which will impact the cost. This jute is available both in Bangladesh and worldwide. Our composite materials increase the percentage of jute materials, which will reduce the cost of various applications. Jute fiber absorbs moisture levels that are high in the environment. We design and fabricate our composite in an optimized way. That is why resin can protect from swelling, degradation, or microbial growth. Jute fiber is bio-friendly and biodegradable, which contributes to environmental sustainability. We optimize the proportion of jute fibers in the composite. We prefer sustainable materials over petroleum-based products like epoxy resin. This optimized design and fabrication material aligns with global trends toward greater environmental friendliness. Jute

reduces the use of alternative synthetic materials like nylon, glass, and polyester. This will reduce plastic waste and carbon emissions linked to the SDG goal. A lighter, customized, and optimized design also contributes to energy efficiency, specifically in transportation applications, by reducing fuel consumption and emissions. Our composite design and fabrication effectively balance mechanical performance, cost-efficiency, and environmental sustainability.

6. Conclusions and recommendations

Throughout the work, composites made from jute fiber were constructed, and their mechanical capabilities were assessed. The study's findings are as follows: Epoxy effectively fabricates new bio-composites reinforced with jute fiber. The current experiment's findings demonstrated that composites reinforced with jute and epoxy resin can achieve the required levels of tensile strength (42.91 MPa) and bending strength (69.30 MPa). The numerical results differ somewhat from the experimental results. It is a result of the specimen being treated as a homogeneous material throughout numerical analysis. However, numerical analysis of various natural fibers with different compositions can also be used without creating a physical shape. This would undoubtedly aid in lowering the significant quantity of manufacturing costs. Finally, our composite design, fabrication, and optimization have the potential to improve mechanical properties, decrease composite weight, reduce resin cost, and increase material sustainability. The proposed design and composition will be adapted to obtain lightweight features in various applications, including components for automobiles, door handle sheets, bicycle seat backs, and baggage covers. For some light load-bearing tasks, the bending strength of jute fiber-based biodegradable polymers can be beneficial. Based on the precise hardness that this compound will deliver, designers can use the results of this research to create products using jute fiber-based polymer composites. The most important finding of this study is that jute, which is regarded as an environmental contaminant, may be used to create goods that could replace expensive glass fiber-based composites and contribute to the development of healthier ecosystems for humans and the environment.

Different organic materials may be applied to increase mechanical features. Future scholars will have much freedom to conduct additional research in this field. Optimization and cost function analysis may also be added in the future. Impact tests can also be done for further study using filler materials, as tensile and flexural tests play a more vital role, as observed in the discussion and literature review. The orientation angle will also be an important element for this type of investigation, and it can be modified. The composites can be used in interior furniture, automobile parts, building materials, and marine transportation sectors. Hence with this conclusion, it is sure that the technology shows composite is the most wanted material in the recent trend.

CRedit authorship contribution statement

Tarikur Jaman Pramanik: Data curation, Formal analysis, Investigation, Methodology, Software, Writing – original draft. **Md. Rafiquzzaman:** Conceptualization, Formal analysis, Investigation, Methodology, Software, Supervision, Writing – review & editing. **Anup Karmakar:** Writing – review & editing. **Marzan Hasan Nayeem:** Writing – review & editing. **S M Kalbin Salim Turjo:** Writing – review & editing. **Md. Ragib Abid:** Writing – review & editing.

Declaration of competing interest

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

References


- [1] M. Rafiquzzaman, M. Islam, H. Rahman, S. Talukdar, N. Hasan, Mechanical property evaluation of glass–jute fiber reinforced polymer composites, *Polym. Adv. Technol.* 27 (10) (2016) 1308–1316.
- [2] K.A. Prasath, B.R. Krishnan, C.K. Arun, Mechanical properties of woven fabric Basalt/jute fibre reinforced polymer hybrid composites, *Int. J. Mech. Eng* 2 (4) (2013) 279–290.
- [3] H. Chandekar, V. Chaudhari, S. Waigankar, A review of jute fiber reinforced polymer composites, *Mater. Today Proc.* 26 (2020) 2079–2082.
- [4] S.M.K.S. Turjo, M.F. Hossain, M.S. Rana, M. Al-Mamun, M.S. Ferdous, Durability and mechanical characteristics of unidirectional jute/banana and synthetic fiber reinforcement epoxy composite, *Hybrid Adv.* (2024) 100232.
- [5] S. Shahinur, M.M.A. Sayeed, M. Hasan, A.S.M. Sayem, J. Haider, S. Ura, Current development and future perspective on natural jute fibers and their biocomposites, *Polymers (Basel)* 14 (7) (2022) 1445.
- [6] M.N. Cazaaurang-Martinez, P.J. Herrera-Franco, P.I. Gonzalez-Chi, M. Aguilar-Vega, Physical and mechanical properties of henequen fibers, *J. Appl. Polym. Sci.* 43 (4) (1991) 749–756.
- [7] A.C. Karmaker, J.P. Shneider, Mechanical performance of short jute fibre reinforced polypropylene, *J. Mater. Sci. Lett.* 15 (3) (1996) 201–202.
- [8] K. Joseph, S. Thomas, C. Pavithran, Viscoelastic properties of short-sisal-fiber-filled low-density polyethylene composites: effect of fiber length and orientation, *Mater. Lett.* 15 (3) (1992) 224–228.
- [9] J. George, S.S. Bhagawan, S. Thomas, Thermogravimetric and dynamic mechanical thermal analysis of pineapple fibre reinforced polyethylene composites, *J. Therm. Anal. Calorim.* 47 (4) (1996) 1121–1140.
- [10] T. Munikenche Gowda, A.C.B. Naidu, R. Chhaya, Some mechanical properties of untreated jute fabric-reinforced polyester composites, *Compos. Part A Appl. Sci. Manuf.* 30 (3) (1999) 277–284, [https://doi.org/10.1016/S1359-835X\(98\)00157-2](https://doi.org/10.1016/S1359-835X(98)00157-2).
- [11] C. Pavithran, P.S. Mukherjee, M. Brahmakumar, A.D. Damodaran, Impact properties of natural fibre composites, *J. Mater. Sci. Lett.* 6 (1987) 882–884.
- [12] M. Rafiquzzaman, S. Abdullah, A.M.T. Arifin, Behavioural observation of laminated polymer composite under uniaxial quasi-static and cyclic loads, *Fibers Polym.* 16 (2015) 640–649.
- [13] P.H. Aditya, K.S. Kishore, D. Prasad, Characterization of natural fiber reinforced composites, *Int. J. Eng. Appl. Sci.* 4 (6) (2017) 257446.
- [14] J. Gassan, A.K. Bledzki, Possibilities to improve the properties of natural fiber reinforced plastics by fiber modification–Jute polypropylene composites–, *Appl. Compos. Mater.* 7 (2000) 373–385.
- [15] S. Rajesh, B. Vijayarajam, C. Elanchezian, S. Vivek, M.H. Prasad, M. Kesavan, Experimental investigation of tensile and impact behavior of aramid-natural fiber composite, *Mater. Today Proc.* 16 (2019) 699–705.
- [16] S. Harish, D.P. Michael, A. Bensely, D.M. Lal, A. Rajadurai, Mechanical property evaluation of natural fiber coir composite, *Mater. Charact.* 60 (1) (2009) 44–49.
- [17] M.Y. Khalid, A. Al Rashid, Z.U. Arif, M.F. Sheikh, H. Arshad, M.A. Nasir, Tensile strength evaluation of glass/jute fibers reinforced composites: an experimental and numerical approach, *Results Eng.* 10 (2021) 100232.
- [18] S. Shahinur, M. Hasan, Q. Ahsan, J. Haider, Effect of chemical treatment on thermal properties of jute fiber used in polymer composites, *J. Compos. Sci.* 4 (3) (2020) 132.
- [19] M.K. Gupta, R.K. Srivastava, H. Bisaria, Potential of jute fibre reinforced polymer composites: a review, *Int. J. Fiber Text. Res* 5 (3) (2015) 30–38.
- [20] S. Ovali, E. Sancak, Investigation of mechanical properties of jute fiber reinforced low density polyethylene composites, *J. Nat. Fibers* 19 (8) (2022) 3109–3126.
- [21] S. Ramakrishnan, K. Krishnamurthy, G. Rajeshkumar, M. Asim, Dynamic mechanical properties and free vibration characteristics of surface modified jute fiber/nano-clay reinforced epoxy composites, *J. Polym. Environ.* 29 (2021) 1076–1088.
- [22] G.S. Balan, et al., Flame resistance characteristics of woven jute fiber reinforced fly ash filled polymer composite, *J. Nanomater.* 2022 (1) (2022) 9704980.
- [23] J.B. Sajin, et al., Impact of fiber length on mechanical, morphological and thermal analysis of chemical treated jute fiber polymer composites for sustainable applications, in: *Curr. Res. Green Sustain. Chem.*, 5, 2022 100241.
- [24] E. Rodríguez, R. Petrucci, D. Puglia, J.M. Kenny, A. Vazquez, Characterization of composites based on natural and glass fibers obtained by vacuum infusion, *J. Compos. Mater.* 39 (3) (2005) 265–282.
- [25] H.E. Balcioglu, An investigation on the mechanical strength, impact resistance and hardness of SiC filled natural jute fiber reinforced composites, *Res. Eng. Struct. Mater.* 5 (3) (2019) 213–231.
- [26] A. Gopinath, M.S. Kumar, A. Elayaperumal, Experimental investigations on mechanical properties of jute fiber reinforced composites with polyester and epoxy resin matrices, *Procedia Eng.* 97 (2014) 2052–2063.
- [27] A.E. Ismail, M.H. Bin Zainulabidin, M.N. Roslan, A.L. Mohd Tobi, N.H. Muhd Nor, Effect of velocity on the impact resistance of woven jute fiber reinforced composites, *Appl. Mech. Mater.* 465 (2014) 1277–1281.
- [28] M. Alemi-Ardakani, A.S. Milani, S. Yannacopoulos, On complexities of impact simulation of fiber reinforced polymer composites: a simplified modeling framework, *Sci. World J.* 2014 (1) (2014) 382525.
- [29] F.M. Jensen, B.G. Falzon, J. Ankersen, H. Stang, Structural testing and numerical simulation of a 34 m composite wind turbine blade, *Compos. Struct.* 76 (1–2) (2006) 52–61.
- [30] L. Torre, J.M. Kenny, Impact testing and simulation of composite sandwich structures for civil transportation, *Compos. Struct.* 50 (3) (2000) 257–267.
- [31] J. Fish, Multiscale modeling and simulation of composite materials and structures, *Multiscale Methods Comput. Mech. Prog. Accomplishments* (2011) 215–231.

- [32] M. Ramesh, K. Palanikumar, K.H. Reddy, Mechanical property evaluation of sisal-jute-glass fiber reinforced polyester composites, *Compos. B Eng.* 48 (2013) 1–9.
- [33] P. Amuthakkannan, V. Manikandan, J.T.W. Jappes, M. Uthayakumar, Influence of stacking sequence on mechanical properties of basalt-jute fiber-reinforced polymer hybrid composites, *J. Polym. Eng.* 32 (8–9) (2012) 547–554.
- [34] R. Murugan, R. Ramesh, K. Padmanabhan, Influence of stacking sequence on mechanical properties and vibration characteristics of glass/carbon hybrid plates with different fabric areal densities, in: *Structural Integrity Assessment: Proceedings of ICONS 2018*, Springer, 2020, pp. 87–97.
- [35] T. Khan, M.T.H. Sultan, A.U.M. Shah, A.H. Ariffin, M. Jawaid, The effects of stacking sequence on the tensile and flexural properties of kenaf/jute fibre hybrid composites, *J. Nat. Fibers* 18 (3) (2021) 452–463.
- [36] G. Sun, S. Tong, D. Chen, Z. Gong, Q. Li, Mechanical properties of hybrid composites reinforced by carbon and basalt fibers, *Int. J. Mech. Sci.* 148 (2018) 636–651.
- [37] F. Hanan, M. Jawaid, P.M. Tahir, Mechanical performance of oil palm/kenaf fiber-reinforced epoxy-based bilayer hybrid composites, *J. Nat. Fibers* (2020).
- [38] K.N. Bharath, M.R. Sanjay, M. Jawaid, Harisha, S. Basavarajappa, S. Siengchin, Effect of stacking sequence on properties of coconut leaf sheath/jute/E-glass reinforced phenol formaldehyde hybrid composites, *J. Ind. Text.* 49 (1) (2019) 3–32.
- [39] X. Hou, Z. Cao, L. Zhao, L. Wang, Y. Wu, L. Wang, Microstructure, texture and mechanical properties of a hot rolled Mg–6.5 Gd–1.3 Nd–0.7 Y–0.3 Zn alloy, *Mater. Des.* 34 (2012) 776–781.
- [40] S. Vongpaisal, G. Li, R. Pakalnis, T. Brady, New development of expert system module for a decision-making on mine stope stability in underground blasthole mining operations, *Int. J. Min. Reclam. Environ.* 25 (1) (2011) 41–51.
- [41] Z. Al-Hajaj, B.L. Sy, H. Bougherara, R. Zdero, Impact properties of a new hybrid composite material made from woven carbon fibres plus flax fibres in an epoxy matrix, *Compos. Struct.* 208 (2019) 346–356.
- [42] N. Saba, O.Y. Allothman, Z. Almutairi, M. Jawaid, Magnesium hydroxide reinforced kenaf fibers/epoxy hybrid composites: mechanical and thermomechanical properties, *Constr. Build. Mater.* 201 (2019) 138–148.
- [43] S. Chauhan, R.K. Bhushan, Improvement in mechanical performance due to hybridization of carbon fiber/epoxy composite with carbon black, *Adv. Compos. Hybrid Mater.* 1 (2018) 602–611.
- [44] A. Alavudeen, N. Rajini, S. Karthikeyan, M. Thiruchitrambalam, N. Venkateshwareen, Mechanical properties of banana/kenaf fiber-reinforced hybrid polyester composites: effect of woven fabric and random orientation, *Mater. Des.* (1980–2015) 66 (2015) 246–257.
- [45] U. Kureemun, M. Ravandi, L.Q.N. Tran, W.S. Teo, T.E. Tay, H.P. Lee, Effects of hybridization and hybrid fibre dispersion on the mechanical properties of woven flax-carbon epoxy at low carbon fibre volume fractions, *Compos. B Eng.* 134 (2018) 28–38.
- [46] S. Anand Kumar, Y. Shivraj Narayan, Tensile testing and evaluation of 3D-printed PLA specimens as per ASTM D638 type IV standard, in: *Innovative Design, Analysis and Development Practices in Aerospace and Automotive Engineering*, 2, Springer, 2018, pp. 79–95. I-DAD2019.
- [47] C. Elanchezhian, B.V. Ramnath, G. Ramakrishnan, M. Rajendrakumar, V. Naveenkumar, M.K. Saravanakumar, Review on mechanical properties of natural fiber composites, *Mater. Today Proc.* 5 (1) (2018) 1785–1790.
- [48] E.A. Franco-Urquiza, Y.R. Escamilla, P.I. Alcántara Llanas, Characterization of 3D printing on jute fabrics, *Polymers (Basel)* 13 (19) (2021) 3202.
- [49] A. Gupta, H. Singh, R.S. Walia, Hybrid filler composition optimization for tensile strength of jute fibre-reinforced polymer composite, *Bull. Mater. Sci.* 39 (2016) 1223–1231.
- [50] S. Deshpande, T. Rangaswamy, Effect of fillers on E-glass/jute fiber reinforced epoxy composites, *Int. J. Eng. Res. Appl.* 4 (8) (2014) 118–123.
- [51] B.Y. Mekonnen, Y.J. Mamo, Tensile and flexural analysis of a hybrid bamboo/jute fiber-reinforced composite with polyester matrix as a sustainable green material for wind turbine blades, *Int. J. Eng.* 33 (2) (2020) 314–319.
- [52] C. Sivakandhan, G. Murali, N. Tamiloli, L. Ravikumar, Studies on mechanical properties of sisal and jute fiber hybrid sandwich composite, *Mater. Today Proc.* 21 (2020) 404–407.
- [53] A. Ali, et al., Experimental and numerical characterization of mechanical properties of carbon/jute fabric reinforced epoxy hybrid composites, *J. Mech. Sci. Technol.* 33 (2019) 4217–4226.
- [54] M. Venkatasudhahar, P. Kishorekumar, N. Dilip Raja, Influence of stacking sequence and fiber treatment on mechanical properties of carbon-jute-banana reinforced epoxy hybrid composites, *Int. J. Polym. Anal. Charact.* 25 (4) (2020) 238–251.
- [55] S.M.K.S. Turjo, M.F. Hossain, M.S. Rana, M.S. Ferdous, Mechanical characterization and corrosive impacts of natural fiber reinforced composites: an experimental and numerical approach, *Polym. Test.* 125 (2023) 108108.
- [56] J. Zhan, et al., Evaluatolgy: the science and engineering of evaluation, *BenchCouncil Trans. Benchmarks Stand. Eval.* 4 (1) (2024) 100162.



Full length article

Could bibliometrics reveal top science and technology achievements and researchers? The case for evaluatology-based science and technology evaluation

Guoxin Kang^{b,a,c}, Wanling Gao^{b,a,c}, Lei Wang^{b,a,c}, Chunjie Luo^{b,a,c}, Hainan Ye^{a,b,c}, Qian He^a, Shaopeng Dai^c, Jianfeng Zhan^{a,b,c} ^{*}

^a The International Open Benchmark Council, China

^b ICT, Chinese Academy of Sciences, Beijing, China

^c University of Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Keywords:

Evaluatology
Science and technology evaluation
Top achievements and researchers
Bibliometrics
Extended evaluation condition
Evaluation systems or models
Top N @X @Y

ABSTRACT

By utilizing statistical methods to analyze bibliographic data, bibliometrics faces inherent limitations in identifying the most significant science and technology achievements and researchers. To overcome this challenge, we present an evaluatology-based science and technology evaluation methodology. At the heart of this approach lies the concept of an extended evaluation condition (EC), encompassing nine crucial components derived from a field. We define four relationships that illustrate the connections among various achievements based on their mapped extended EC components, as well as their temporal and citation links. Within a relationship under an extended EC, evaluators can effectively compare these achievements by carefully addressing the influence of confounding variables. We establish a real-world evaluation system encompassing an entire collection of achievements, each of which is mapped to several components of an extended EC. Within a specific field like chip technology or open source, we construct a perfect evaluation model that can accurately trace the evolution and development of all achievements in terms of four relationships based on the real-world evaluation system. Building upon the foundation of the perfect evaluation model, we put forth four-round rules to eliminate non-significant achievements by utilizing four relationships. This process allows us to establish a pragmatic evaluation model that effectively captures the essential achievements, serving as a curated collection of the top N achievements within a specific field during a specific timeframe. We present a case study on the top 100 Chip achievements to demonstrate the effectiveness of our science and technology evaluatology. The case study highlights its practical application and efficacy in identifying significant achievements and researchers that otherwise cannot be identified by using bibliometrics.

1. Introduction

Science and technology (S&T) evaluation is a meticulous and comprehensive process. One of its paramount goals is to identify the most remarkable accomplishments in each field, duly recognize the individuals, institutions, or nations that have made significant contributions to these achievements, and delve deeper into the effective and efficient mechanisms and policies within the S&T ecosystems that profoundly shape the evolution of these achievements [1]. This article focuses on the first half of the task.

While bibliometrics methodologies have long relied on observable metrics such as publication numbers, citation counts, and the H-index

to assess correlations and impact [2–5], as illustrated in Fig. 1, it is essential to recognize their inherent three limitations and the need for alternative approaches.

First, bibliometrics commonly employs publication numbers, citation counts, and related metrics to gauge scholarly works' quality, influence, and significance. However, various confounding variables can significantly impact citation counts. Moreover, citation counts are vulnerable to manipulation by malicious networks.

Second, bibliometrics often fails to consider critical non-bibliometric metrics, making them insufficient for evaluating significant technological achievements that may have limited publication outputs. For instance, the Linux operating system in computer science has made

* Corresponding author at: ICT, Chinese Academy of Sciences, Beijing, China.

E-mail address: jianfengzhan.benchcouncil@gmail.com (J. Zhan).

URL: <http://www.zhanjianfeng.org> (J. Zhan).

<https://doi.org/10.1016/j.tbench.2024.100182>

Received 22 August 2024; Accepted 20 September 2024

Available online 18 October 2024

2772-4859/© 2025 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

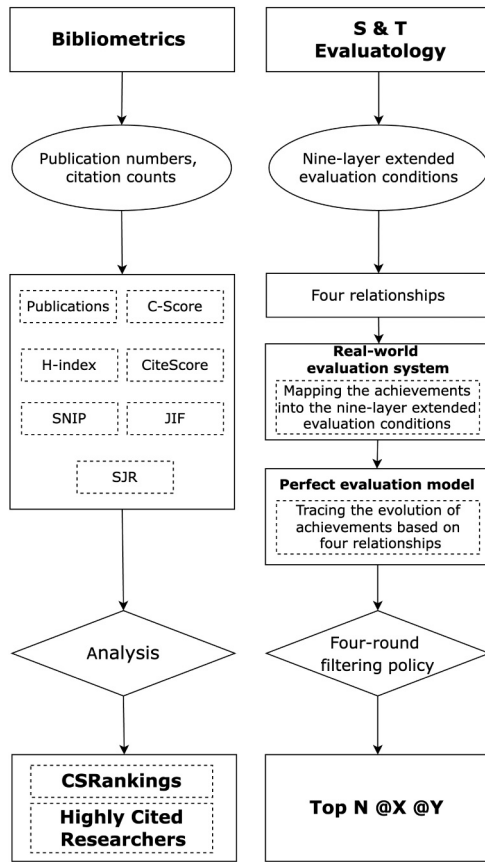


Fig. 1. Fundamental differences between bibliometrics and S&T evaluatology.

a substantial impact despite having a modest publication record.

Third, many bibliometrics methodologies prioritize the quantity over the quality of publications, which can result in an incomplete assessment of the true value and impact of scholarly work.

To address these shortcomings, we introduce the S&T evaluatology, which exemplifies the application of evaluatology in evaluating S&T achievements. The S&T evaluatology is illustrated in Fig. 1 and presented in detail in [6,7]. The fundamental principle of evaluatology is to implement a well-defined evaluation condition (EC) on particular subjects to establish evaluation models or systems.

At the core of the S&T evaluatology is the notion of an extended EC, which comprises nine key components: (1) the field that can be broken down into several problem domains; (2) The set of problem domains, each of which can be broken down into various sub-problem domains; (3) the sub-problem domains, each of which can be decomposed into several problems; (4) the set of a collective of equivalent problems, each of which can be broken down into multiple sub-problems; (5) the set of a collective of equivalent sub-problems; (6) the set of a collective of problems or sub-problem instances; (7) the algorithms or the algorithm-like mechanisms that tackles a problem or a sub-problem; (8) the implementations of algorithms or the algorithm-like mechanisms; (9) the support systems that provide necessary resources and environments [6,7].

We define four relationships that illustrate the connections among various achievements based on their mapped extended EC components, as well as their temporal and citation links. We define two primary relationships: *pioneering and progressive* and two auxiliary relationships: *parallel and related but not connected*. Within a pioneering or progressive relationship under an extended evaluation condition, evaluators can effectively compare these achievements by carefully addressing the influence of confounding variables.

Disadvantages of C-Score

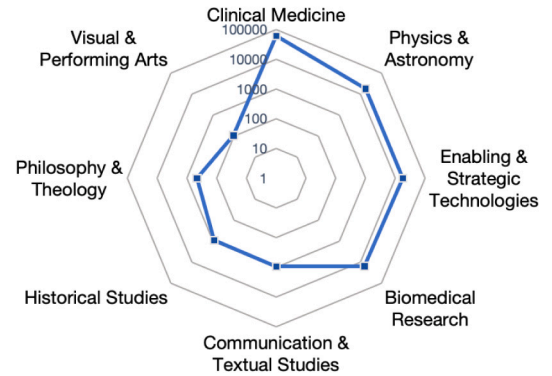


Fig. 2. Comparison of the number of scientists selected for the global top 2% in different disciplines.

We establish a real-world ES encompassing the complete collection of S&T achievements, each of which is mapped to several components of an extended EC. In line with the aim of identifying the top N S&T achievements, the proposed real-world S&T ES ignores the other components of the real-world S&T ecosystems, e.g., the mechanisms and policies that profoundly shape the evolution of these achievements [1].

Under the premise that all evaluated achievements belong to the same field, e.g., chip technology or open source, we construct “a perfect S&T EM” that can accurately trace the evolution and development of all achievements in terms of four relationships based on the real-world ES. We compare achievements that have a specific relationship under the extended EC they involve. Utilizing four relationships, we employ four rounds of rules to prune non-significant achievements to establish a pragmatic S&T EM that captures the fundamental S&T achievements. Essentially, the pragmatic S&T EM is a collection of top N achievements within a field during a timeframe.

The International Open Benchmark Council (BenchCouncil) utilized the S&T evaluatology principles and the instantiated Top N @X @Y methodology to systematically recognize the most 100 groundbreaking and influential achievements in chip technology (Chip100) [8]. The case study demonstrates the effectiveness of our proposed methodology compared to bibliometrics.

In the following sections, we will provide an in-depth examination of the S&T evaluatology. Section 2 enumerates the existing bibliometrics methodologies and analyzes their weakness. Section 3 presents the S&T evaluatology in detail. Section 4 provides an instantiated Top N @X @Y methodology. Section 5 introduces a case study on the Top 100 Chip Achievements. Section 6 concludes.

2. Motivation and related work

Bibliometrics is a field that applies statistical methods to analyze bibliographic data.

In this subsection, we first present the overall weakness of bibliometrics in Section 2.1. Then, we introduce the representative bibliometrics methodologies. Finally, we introduce the fundamental concept, theory, and methodology in evaluatology [6], based on which we will present the S&T evaluatology.

2.1. Motivation: The limitations of bibliometrics

Due to the nature of bibliometrics, there are several inherent drawbacks associated with its application.

First, bibliometrics commonly employs publication numbers, citation counts, and related metrics to gauge the quality, influence, and

significance of scholarly works. However, it is crucial to acknowledge that publication numbers and citation counts can be significantly impacted by various confounding variables. These may include the diverse disciplines involved, the reputation and networks linked to researchers and their institutional affiliations, as well as notable differences in researcher numbers and publication volumes across various fields. Moreover, citation counts may even be vulnerable to manipulation by malicious networks (Limitation One).

- (Limitation One-One). The same or similar works published in the same or different periods can be impacted by various confounding variables, such as the reputation and network of the researchers and their institutions, leading to significant variations in citation counts. Moreover, citation counts may be subject to manipulation by malicious networks.
- (Limitation One-Two). In fundamental disciplines like mathematics, once a problem has been effectively solved, there may be limited follow-up research on that specific topic. Consequently, the citation count for the original work in fundamental disciplines may not increase significantly. Hence, it cannot accurately reflect the impact or influence of the research in the field.
- (Limitation One-Three). Citation counts fail to account for the significant disparities in researcher numbers and publication volumes across different fields. In fields with fewer researchers and publications, citation counts are naturally lower, regardless of the quality of the research being conducted.
- (Limitation One-Four). Bibliometrics prioritize well-established disciplines, potentially overlooking emerging fields or unconventional research outputs that may have a significant impact but lower citation counts. The effectiveness of citation metrics is limited in representing contributions within emerging or specialized fields. Groundbreaking research in these domains might initially receive few citations due to the novelty of the subject matter or the field's limited scope. Consequently, as shown in Fig. 2, pivotal advancements in such areas risk being undervalued, as seen in the “top scientists” list created by Stanford University and the Elsevier data repository [3], which predominantly features scientists from well-established fields like Clinical Medicine and Physics & Astronomy. This bias is particularly harmful to innovators who spearhead new research directions, as their contributions may not be accurately captured by citation-based metrics.
- (Limitation One-Five). Self-citations occur when authors cite their previous work, potentially inflating the impact of their research. This practice skews the representation of a paper's or a researcher's genuine influence within the academic community. For instance, metrics like the H-index [4] are unable to circumvent the issue of self-citations, resulting in a biased assessment that may unfairly favor those who self-cite frequently.

Second, bibliometrics significantly ignores other fundamental non-bibliometric metrics and hence cannot be applied to significant technological achievements that have few or no publication outputs (Limitation Two). Bibliometrics primarily relies on analyzing published works. Limitation Two arises when considering groundbreaking technological advancements that may not be adequately represented in traditional scholarly publications.

In practical fields like computer science, substantial contributions frequently occur outside the conventional academic publishing framework. A prime example of this is the Linux operating system within the realm of computer science. As an open-source software, the Linux operating system boasts numerous contributors who may not publish extensively. Similarly, the computer mouse, one of the most universally adopted human-computer interaction technologies, demonstrates that significant impact does not necessarily stem from published research. Table 1 presents several significant technological achievements that are overlooked by bibliometrics. Therefore, bibliometrics alone may not

Table 1

Summary of significant achievements overlooked by bibliometrics.

Field	Achievements	Published paper	Citations
Chip	X86 ISA	No	N/A
	PCB	No	N/A
Open-sources systems	Linux Kernel	No	N/A
	Git	No	N/A
	MySQL	No	N/A
Benchmarks	Whetstone	No	N/A
	TPC-C	No	N/A
	TPC-H	No	N/A
	FIO	No	N/A

fully capture the impact and significance of these achievements. Consequently, non-academic metrics should be considered in the grading process to select the top-impact achievements.

Third, bibliometrics prioritizes the quantity over the quality of publications (Limitation Three). High citation counts of a researcher can result from either a large volume of modestly impactful publications or from many surveys on trending topics, such as timely topics on large language models. Although these works might garner significant attention, they do not necessarily represent substantial advancements within their disciplines. The emphasis on publication counts can lead to a skewed representation of research impact, as it fails to consider the significance, rigor, and originality of individual publications.

In summary, while bibliometrics provides a quantitative metric, like citation counts, for academic evaluation, they are beset with limitations that result in biased and incomplete assessments. Thus, the S&T evaluation urgently needs more nuanced and comprehensive evaluation metrics and methodologies that go beyond bibliometrics. Such metrics would ensure a fairer and more accurate depiction of scholarly impact, truly reflecting the multifaceted nature of academic contributions.

2.2. The representative bibliometrics methodologies

2.2.1. Csrankings in the computer science field

CSRankings is a specialized method for evaluating computer science achievements, favoring the conference publication over the journal. CSRankings adopts the metric of the number of publications at so-called top-tier conferences for gauging the academic influence of researchers or their affiliated institutions in computer science. Utilizing this metric, Emery Berger pioneered CSRankings [2], a tailored academic leaderboard specifically designed for the realm of computer science. CSRankings selects the Digital Bibliography & Library Project (DBLP) [9] as its data source, ensuring up-to-date and relevant rankings with quarterly updates.

However, this methodology has several serious flaws. First and foremost, it places a higher emphasis on publication quantity than quality, as outlined in Section 2.1 (Limitation Three), thereby having flaws in recognizing top researchers or groundbreaking achievements.

For example, David Patterson's influential works in chip technology, particularly with RISC, RISC-V, and RAID, have substantially shaped the field. Notwithstanding their extensive influence, Patterson is conspicuously absent from CSRankings, a glaring omission highlighting a significant shortcoming in the ranking system's ability to acknowledge key contributors even in leading institutions.

In addition, CSRankings cannot discern the varying impacts of different achievements. CSRankings quantifies the number of papers presented at top-tier conferences, but this approach fails to identify who pioneered a field. For instance, although the groundbreaking “Transformer” model [10] was presented at the 31st Conference on Neural Information Processing Systems (NeurIPS), it is erroneous to assume that all papers at this conference exert an influence comparable to that of the Transformer.

This situation underscores a fundamental flaw in the CSRankings system: overemphasizing top-tier conference publications can lead to

misleading representations, bypassing the real depth and enduring impact of substantial contributions.

Second, many influential works like the Linux operating system have never even sought publication in a so-called top-tier conference (Limitation Two). Table 1 provides other examples. The current metric focusing on publications fails to recognize significant achievements that are not encapsulated in conference papers. The Linux operating system's development and its widespread adoption stand as a prime example, achieving monumental impact without the endorsement of traditional academic publications.

Third, CSRankings overlooks the significant disparities in researcher numbers, publication frequency, and volumes across different fields within computer science (Limitation One-Three). This oversight has resulted in a skewed ranking from 1970–2022, where four out of the top seven institutions are led by faculty specializing in vision, a field known for its high paper acceptance volumes. For example, the field of computer vision, known for higher publication volumes, is overrepresented. As shown in Fig. 3, in 2022, The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR), a leading conference in computer vision, accepted 2065 papers [11], whereas The IEEE/ACM International Symposium on Microarchitecture Conference (MICRO), a top conference in computer architecture, accepted just 83 [12].

The discrepancies in publication volume across different fields may lead to potential misleading outcomes when using CSRankings. These disparities raise concerns about the accuracy of CSRankings in providing equitable representation for all fields within computer science.

Fourth, as a consequence of being accepted by a so-called top-tier conference, this metric is impacted by various confounding variables, such as the reputation and network of the researchers and their institutions, and is subject to manipulation by malicious networks. Collusion among reviewers is not an isolated incident in numerous computer science conferences.

2.2.2. The standardized citation metrics (c-score)

The c-score, developed by John Ioannidis [3], assesses the influence of scientists. This standardized indicator amalgamates various elements, including citations, h-index, co-authorship-adjusted hm-index, and authorship-position-specific citations. Leveraging this metric, Ioannidis's team curated a global database for ranking scientists, categorized into career-long and single-year impacts based on the Scopus data. The former category spans citations from 1996 to now, while the latter focuses on the current calendar year alone. This innovative metric transcends traditional citation metrics, avoiding the evaluation biases introduced by self-citations. However, its primary focus on publications cannot completely encompass the wider spectrum of a scientist's influence, particularly in areas such as practical applications or cross-disciplinary collaborations. These critical dimensions, essential to the fabric of scientific progress, are often understated in conventional bibliometric measures (Limitation Two).

Despite its popularity in the scientific field, the standardized citation metric has limitations in acknowledging the impact of researchers in emerging disciplines (Limitation One-Four), leading to an underrepresentation of their contributions. The metric's proclivity to privilege well-established, voluminous fields is evidenced by the fact that over half of the top-ranked influential scientists in 2021 originated from fields like Clinical Medicine, Physics & Astronomy, Biomedical Research, and Enabling & Strategic Technologies. This trend reveals an inherent bias, favoring areas with more substantial publication frequencies and higher citation volumes (Limitation One-Three).

In addition, Limitation One remains a challenge that cannot be mitigated by the standardized citation metrics (c-score). Factors such as the reputation and network of researchers and their affiliated institutions can confound the evaluation process. For instance, even when two researchers from different institutions achieve similar achievements, the level of attention and recognition their work receives can vary significantly. In some cases, the earlier work may receive limited

Disadvantages of CSRankings

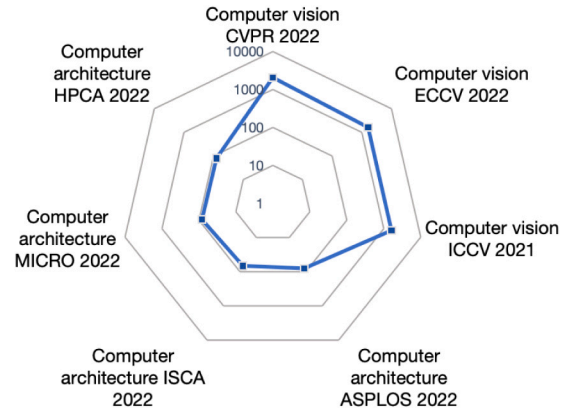


Fig. 3. Comparison of accepted papers by top conferences in the fields of computer vision and computer architecture.

attention, while subsequent work gains widespread acclaim. These disparities can be attributed to various factors, including the visibility and influence that researchers and their institutions hold within the academic community.

2.2.3. H-index

H-index [4] is a useful metric proposed by Jorge E. Hirsch to characterize a researcher's scientific output. The objective is to determine the highest value of h , where there are at least h papers with a citation number equal to or greater than h . The mathematical representation of H-index for a scientist is $h_index(f) = \max\{i \in \mathbb{N} : f(i) \geq i\}$ [5]. Here, f is an array that contains the number of citations for the scientist's publications in decreasing order [5]. Instead of relying solely on single-number criteria like the total number of papers, H-index takes a more holistic approach by considering both productivity and academic impact. In addition to the limitations that we have discussed extensively in Section 2.1, in practice, vast self-citations can raise the H-index value easily.

2.2.4. CiteScore metrics

CiteScore Metrics [13], developed by Elsevier, extensively evaluates academic journals' citation impact and influence. These metrics are calculated yearly, considering a three-year citation window and considering the volume, quality, and field-normalized citation rates of articles published in a specific journal. Featuring indicators such as average citations per document, quartile ranking, and overall standing, CiteScore Metrics provides a transparent and comprehensive tool for researchers and institutions. While CiteScore metrics are designed to assess the quality and impact of scholarly journals rather than evaluate the quality of research within specific fields. In addition to the limitations that we have discussed in Section 2.1, it has another serious limitation. It is based on a three-year citation window. Consequently, achievements with a substantial long-term impact but relatively few citations in the short term may be undervalued.

2.2.5. Source Normalized Impact per Paper (SNIP)

The Source Normalized Impact per Paper [14] is a metric employed in assessing the influence of scholarly journals. It is determined by dividing an article's citation count within a journal by the anticipated citation rate within its particular field. SNIP considers the citation potential within the journal's discipline, enabling equitable comparisons across diverse areas of study. In essence, SNIP serves as a valuable gauge for evaluating the impact of a journal relative to its field. It provides researchers and institutions with a standardized measure to evaluate the influence of scholarly journals rather than the impact

Extended Evaluation Condition (extended EC)	
Component 1	the field
Component 2	the set of problem domains
Component 3	the set of sub-problem domains
Component 4	the set of a collective of equivalent problems
Component 5	the set of a collective of equivalent sub-problems
Component 6	the set of a collective of problems or sub-problem instances
Component 7	the algorithms or the algorithm-like mechanisms that tackles a problem or a sub-problem
Component 8	the implementations of algorithms or the algorithm-like mechanisms
Component 9	the support systems that provide necessary resources and environments

Fig. 4. The overview of an extended EC.

of the specific research achievement. Furthermore, SNIP compares a journal's citation count with the citation frequency in its field. However, it fails to consider the variations in citation practices across different subject areas. In addition, it has many inherent bibliometrics limitations we discussed in Section 2.1.

2.2.6. Journal impact factor (JIF)

The journal impact factor, devised by Eugene Garfield, is used by Clarivate's Web of Science to evaluate a journal's impact. The impact factor is calculated as $C / \sum_{i=0}^n P_i$, where C is the number of citations received in a given year for publications in a journal that were published in the n preceding years, and $\sum_{i=0}^n P_i$ is the total number of citable items published in that journal during the n preceding years.

2.2.7. SCImago Journal Rank (SJR)

SCImago Journal Rank (SJR) indicator, developed by the Scimago Lab, is a measure of the prestige of journals. SJR is calculated by using an algorithm similar to Google's PageRank, which assumes that important websites are linked to other important websites. Citations are used to link the journals. The algorithm begins by setting an identical amount of prestige to each journal, then using an iterative procedure to transfer each journal's achieved prestige to each other through citations until each journal's update reaches a minimum threshold. The limitations of SJR include the algorithm's complexity, the degree of transparency, and the reproducibility of the results.

Besides, Kevin W. Boyack [15] utilizes data mining and analysis techniques to map knowledge domains, specifically applying them to 20 years of PNAS publications. It combines various data sources to analyze the input-output ratio and diffusion between disciplines. However, its reliance on raw citation counts as the primary measure of impact, without adjusting for self-citations, potentially leads to a skewed and less meaningful assessment of true scholarly influence.

2.3. The basic concepts, theories, and methodologies in evaluatology

According to [6], an individual or system being evaluated is a subject. A stakeholder is defined as an entity that holds a stake of responsibility or interest in the subject matter. Evaluation is "the process of inferring the impact of subjects indirectly within evaluation conditions (EC) that cater to the requirements of stakeholders, relying on objective measurements and/or testing" [7].

The fundamental methodology for evaluating a single subject is outlined as follows. Zhan et al. [6] propose a universal methodology to define an EC, which consists of five basic components [6]: "(1) a set of equivalent definitions of problems; (2) the set of a collective of equivalent problem instances; (3) the algorithms or algorithm-like mechanisms; (4) the implementations of algorithms or algorithm-like mechanisms; (5) support systems that provide necessary resources and environments [7]".

Subsequently, it becomes crucial to implement a well-defined EC for a precisely defined subject, forming a well-defined evaluation model (EM) or system (ES).

In terms of complex scenarios, the evaluation methodology is to establish a series of EMs that ensure transitivity from a real-world ES to a perfect EM and a pragmatic EM [6].

Zhan et al. [6,7] characterize the real-world ES, perfect or pragmatic EMs. Because our S&T evaluation methodology is based on those concepts, we give a concise summary based on [6,7].

The real-world ES refers to "the entire population of real-world systems that are used to evaluate specific subjects". The real-world ES has several significant obstacles: "the presence of numerous confounding, prohibitive evaluation costs resulting from the huge state spaces".

A perfect EM replicates the real-world ES with utmost fidelity: "It eliminates irrelevant problems and has the capability to thoroughly explore and comprehend the entire spectrum of possibilities of an EC". However, it also has serious limitations: "possesses huge state space, entails a vast number of independent variables, and hence results in prohibitive evaluation costs".

Providing a means to estimate the parameters of the real-world ES or a perfect EM, a pragmatic EM simplifies the perfect EM in two ways: "reduce the number of independent variables that have negligible effect and sample from the extensive state space".

3. The science and technology evaluatology

This section presents the essence of S&T evaluatology.

3.1. The overview

Understanding the development of S&T is highly challenging. Sometimes, practice leads the way; at other times, theory does. Some individuals pose a significant problem and offer a preliminary solution, while others provide state-of-the-practice solutions without explicitly stating the problems. The relentless efforts of scientists and engineers make the landscape of S&T achievements intricate and dense, much like an interwoven forest, thereby making the objective evaluation of S&T contributions extremely challenging.

To tackle this challenge, we have adopted the evaluatology framework developed by Zhan et al. [6] as the theoretical foundation for our research. This framework serves as the basis for developing S&T evaluatology. The core principles and methodologies of S&T evaluatology are outlined as follows:

First, building upon the definition of an EC proposed in the referenced paper [1], we introduce the concept of an extended EC, as shown in Fig. 4.

With respect to the EC definition [7], an extended EC introduces several extra components to accommodate the new requirements of S&T evaluation, including the field that can be broken down into several problem domains, the set of problem domains, the set of sub-problem domains, and the set of a collective of equivalent sub-problems. The definition of the extended EC serves as the foundation for the proposed S&T evaluatology. It provides the framework upon which the evaluation of S&T achievements is based.

Second, in the realm of S&T evaluation, a subject refers to an accomplishment that can mapped onto the nine components of an extended EC.

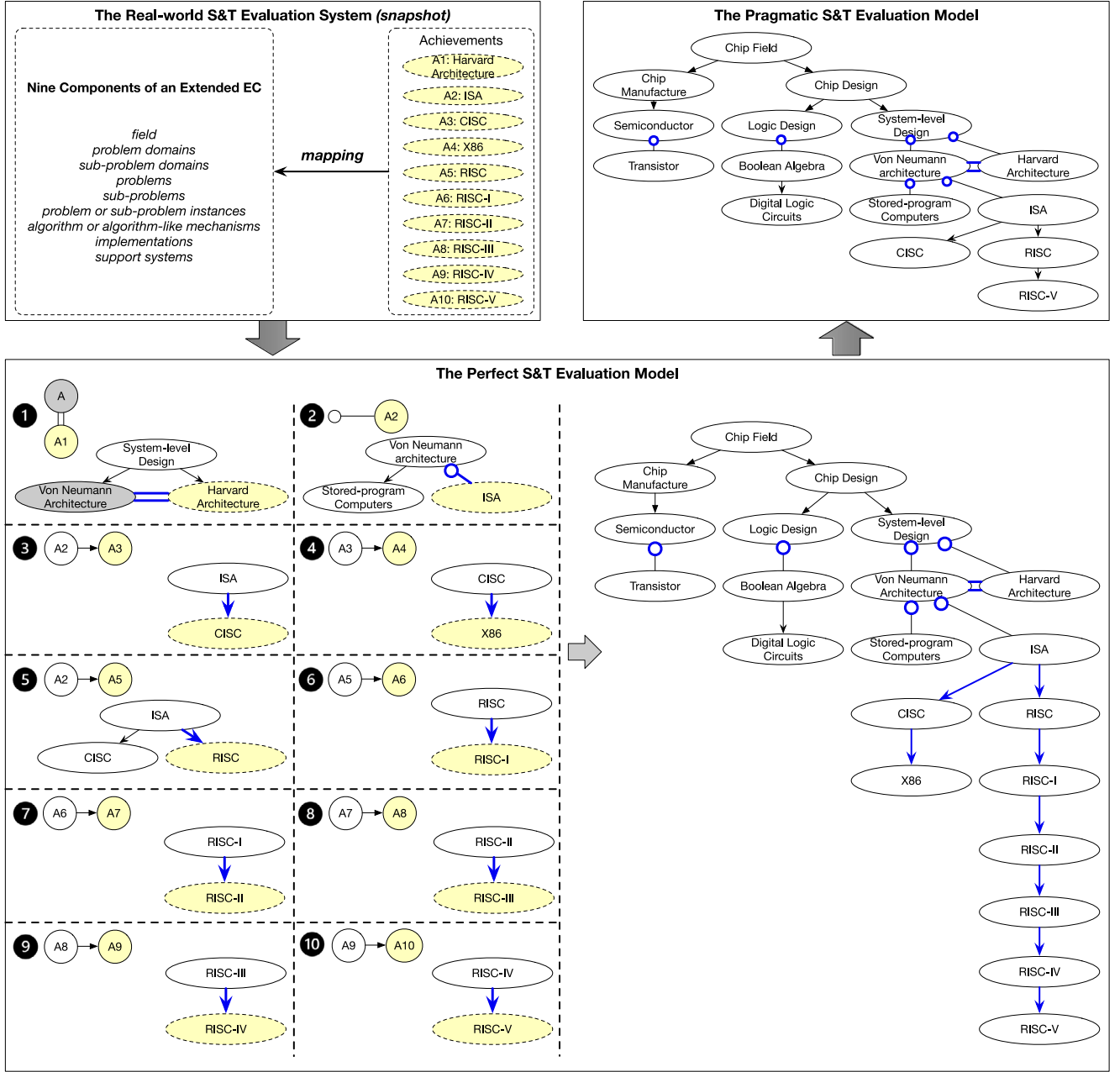


Fig. 5. Illustrating S&T Evaluatology with an example.

For instance, let us consider a scenario where a researcher proposes a new problem and provides a preliminary algorithm for solving that problem. In this case, the subject, a specific S&T achievement, comprises multiple components. These components include:

- problem: The specific problem being addressed or investigated.
- Algorithm: The preliminary algorithm proposed by the researcher to solve the given problem.

Third, based on their mapped extended EC components as well as their temporal and citation links, we establish two primary relationships: *pioneering and progressive* and two auxiliary relationships: *parallel* and *related but not connected* to illustrate the connections among different achievements. Section 3.3 will provide the details of four relationships.

Fourth, according to the theory of evaluatology, S&T evaluation involves applying a well-defined extended EC to the subject—a specific S&T achievement. This process allows for the creation of an EM or

ES. Within a relationship under an extended EC, evaluators can effectively compare different S&T achievements by carefully addressing the influence of confounding variables [6,7].

In the subsequent four steps, we will adhere to and implement the universal evaluation methodology proposed by Zhan et al. [6] to address the intricate S&T evaluation scenarios.

Fifth, we establish a real-world S&T ES, which encompasses the complete collection of S&T achievements. Moreover, each achievement will be decomposed into its respective components within an extended EC. In establishing a real-world S&T ES, it is crucial to characterize the real-world S&T ecosystems. In line with the aim of identifying the top N S&T achievements, the proposed real-world S&T ES in this article encompasses the entire collection of S&T achievements while ignoring the other components of the real-world S&T ecosystems.

Sixth, under the premise that all evaluated achievements belong to the same field, we assume the existence of a “perfect S&T EM” that can accurately trace the S&T evolution and development in terms of four

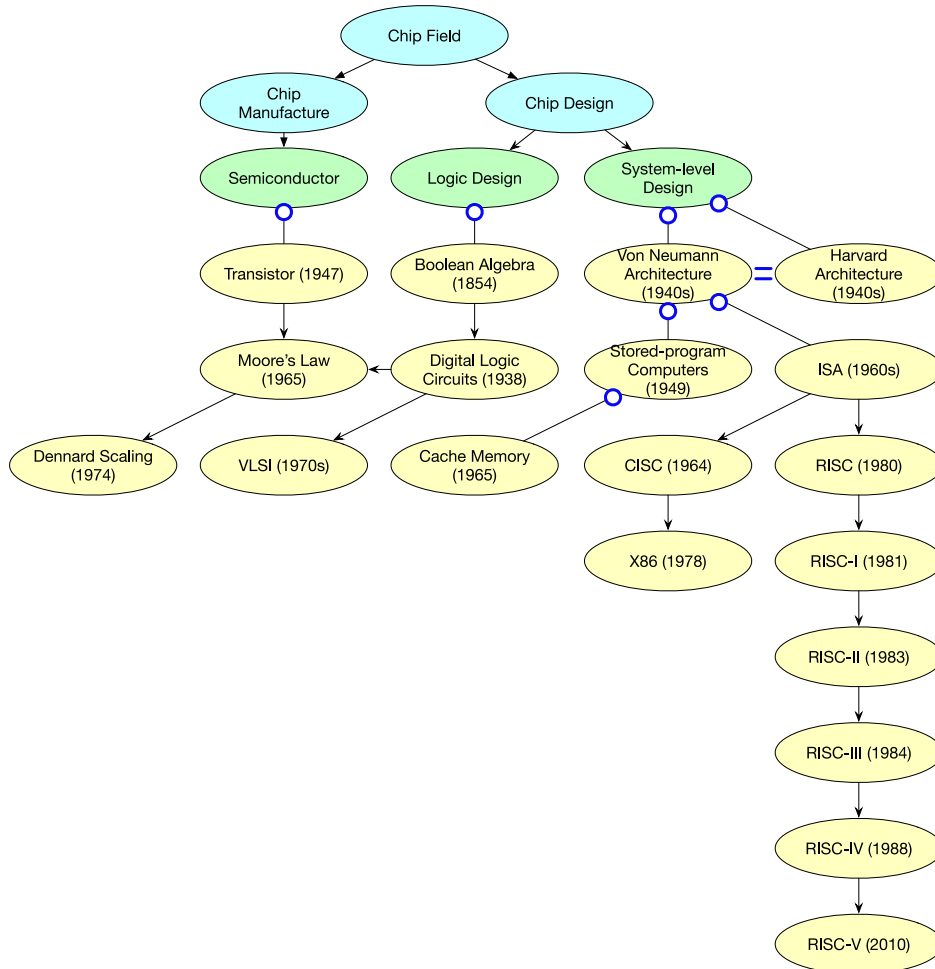


Fig. 6. The localized snapshot of a pragmatic EM in the field of chip technology.

relationships. That is to say, a “perfect S&T EM” can track the evolution of a real-world S&T ES from ES_i to ES_{i+1} in a rigorous manner. This model operates under the premise that only one change is made at a time. By implementing one change at a time, we ensure that only one achievement is added.

Seventh, as the perfect S&T EM contains huge states, we propose several simple rules to prune non-significant achievements to establish a pragmatic S&T EM that captures the fundamental S&T achievements. Essentially, the pragmatic S&T EM is a collection of top N achievements. The basic idea behind this process is that we compare achievements that have a pioneering or progressive relationship under the extended EC they involve. We will explain the simple rules in Section 3.6.

Fig. 5 illustrates S&T Evaluatology with an example, while Fig. 6 offers a localized snapshot of a pragmatic EM in the field of chip technology, showcasing individual achievements. Within the chip technology field are several critical problem domains like ‘Chips Manufacture’ and ‘Chips Design’. This localized snapshot highlights the diversity and complexity within the chip technology field.

3.2. The definition of an extended EC

In [6], Zhan et al. emphasized that “understanding the composition of the problem domain is crucial in identifying the problem that best represents the whole. Across different disciplines, a field often exhibits a hierarchical structure, where a significant problem domain can be broken down into several problems”, which provide the methodology to model an extended EC.

An extended EC consists of nine basic components [6,7], as shown in Fig. 4: (1) the field that can be broken down into several problem domains; (2) the set of problem domains, each of which can be broken down into various sub-problem domains; (3) the sub-problem domains, each of which can be decomposed into several problems; “(4) the set of a collective of equivalent problems, each of which can be broken down into multiple sub-problems; (5) the set of a collective of equivalent sub-problems; (6) the set of a collective of problems or sub-problem instances; (7) the algorithms or the algorithm-like mechanisms that tackle a problem or sub-problem; (8) the implementations of algorithms or the algorithm-like mechanisms; (9) the support systems that provide necessary resources and environments [6,7]”.

As depicted in Fig. 7, the essential steps of the methodology can be summarized as follows. The first and second steps are to define the field and compose it into different problem domains. If necessary, the third step is to decompose each problem domain into several sub-problem domains. The fourth step is to break down problem domains or sub-problem domains into the problems. If necessary, the fifth step is to decompose each problem into several sub-problems. The sixth step proposes the problem instances or sub-problem instances. The seventh step is to figure out the algorithms or algorithm-like mechanisms to solve the problem or sub-problem. The eighth step encompasses the implementation of algorithms or algorithm-like mechanisms. The last step is to define the support system.

For example, chip design is a problem domain in the chip field. The system-level design is a typical sub-problem domain in chip design. The computer architecture design is one of the problems of the system-level design. The Von Neumann architecture was the pioneering

work that defined the computer architecture design problem and proposed algorithm-like mechanisms to address it. Any specific processor that aligns with the Von Neumann architecture can be viewed as an implementation of this mechanism.

3.3. The formal definition of four relationships

In this section, based on their mapped extended EC components as well as their temporal and citation links, we propose two primary relationships and two auxiliary relationships to connect achievements, as shown in Fig. 8.

3.3.1. Two primary relationships

Two fundamental relationships contain a pioneering relationship and a progressive relationship.

Relationship one: A pioneering relationship. *Definition:* A pioneering relationship pertains to an achievement that opens up a new research direction in the form of establishing a new field, problem domain, sub-problem domain, problem, sub-problem, algorithm or algorithm-like mechanism, implementation, or support system within an extended EC. The pioneering relationship recognizes the pioneering nature of such achievements, which lay the foundation for future advancements and innovations.

Formal expression: Let A represent an achievement. The pioneering relationship for A can be formally expressed as:

$$P(A) = \begin{cases} 1 & \text{if } A \text{ opens up a new research direction in the} \\ & \text{form of establishing a new field, problem do-} \\ & \text{main, sub-problem domain, problem, sub-} \\ & \text{problem, algorithm or algorithm-like} \\ & \text{mechanism, implementation, or support} \\ & \text{system within an extended EC,} \\ 0 & \text{otherwise} \end{cases}$$

This binary expression indicates whether A qualifies as a pioneering achievement (1) or not (0). It is based solely on the novelty and originality of the achievement A , without any preceding work.

Examples: Pioneering relationships manifest across various industries and disciplines, highlighting achievements that are the first to propose a novel field, problem domain, sub-problem domain, problem, sub-problem, solution, or support system within an extended EC.

- **Chip:** The Instruction Set Architecture (ISA) represents the pioneering work that defined the instruction set design sub-problem within the computer architecture design problem and proposed corresponding mechanisms to address it. The Reduced Instruction Set Computer (RISC) and Complex Instruction Set Computers (CISC) are subsequent developments following ISA.
- **AI:** The first computational model of a neuron, the McCulloch-Pitts neuron [16], is a pioneering algorithm-like mechanism in the field of neural networks.

Relationship two: A progressive relationship. *Definition:* For the achievements that involve the same component of an extended EC, e.g., a problem or sub-problem, a progressive relationship indicates subsequent achievements are inspired by preceding ones, and the latter publicly acknowledges this influence through citations.

Formal expression: A progressive relationship between two achievements A_i and A_j is defined as:

$$\begin{aligned} S(A_i, A_j) &= 1 \iff (Q(A_i) = Q(A_j)) \\ &\quad \wedge (EC(A_i) \cap EC(A_j) \neq \emptyset) \\ &\quad \wedge ((T(A_{i-e}) < T(A_{j-b})) \vee (T(A_{i-b}) > T(A_{j-e}))) \\ &\quad \wedge ((A_i \in R(A_j)) \vee (A_j \in R(A_i))) \end{aligned} \quad (1)$$

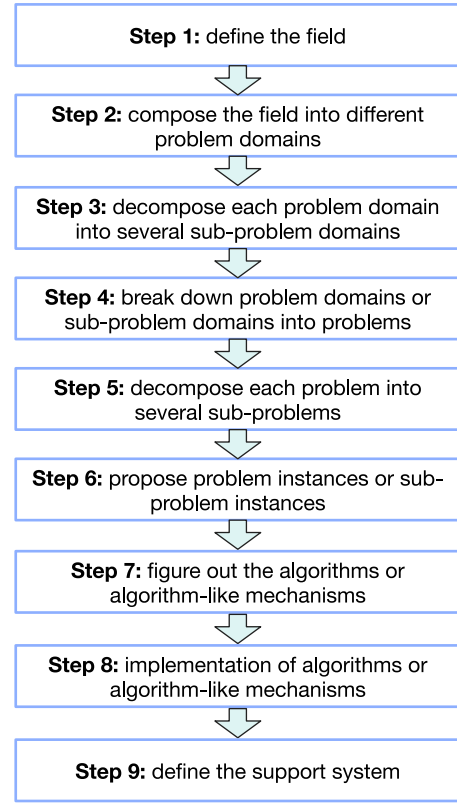


Fig. 7. Essential steps of S&T evaluatology.

where:

- $Q(a)$ is the key problem domain, sub-problem domain, problem, or sub-problem that achievement a addresses.
- $EC(a)$ denotes the EC involved in achievement a .
- $T(a_b)$ and $T(a_e)$ represent the begin time and end time of achievement a , respectively. Thus, $T(A_{i-e}) < T(A_{j-b})$ indicates that achievement A_i precedes achievement A_j in time. $T(A_{i-b}) > T(A_{j-e})$ indicates that achievement A_j precedes achievement A_i in time.
- $R(a)$ indicates the references of achievement a . Thus, $A_i \in R(A_j)$ indicates achievement A_j publicly acknowledge the influence of achievement A_i .

A many-to-one progressive relationship. *Definition:* A many-to-one progressive relationship is an instance of a progressive relationship, indicating multiple much preceding achievements inspire a single subsequent achievement.

Formal expression: A many-to-one progressive relationship between achievements $A_{i1}, A_{i2}, \dots, A_{in}$ and A_j is defined as:

$$S(A_{i1}, A_j) \wedge S(A_{i2}, A_j) \wedge \dots \wedge S(A_{in}, A_j) = 1 \quad (2)$$

where:

- $\{A_{i1}, A_{i2}, \dots, A_{in}\}$ are multiple preceding achievements.
- A_j is a single subsequent achievement.

An one-to-many progressive relationship. *Definition:* A one-to-many progressive relationship is an instance of a progressive relationship, indicating a single preceding achievement inspires multiple subsequent achievements.

Formal expression: A one-to-many progressive relationship between achievement A_i and $A_{j1}, A_{j2}, \dots, A_{jn}$ is defined as:

$$S(A_i, A_{j1}) \wedge S(A_i, A_{j2}) \wedge \dots \wedge S(A_i, A_{jn}) = 1 \quad (3)$$

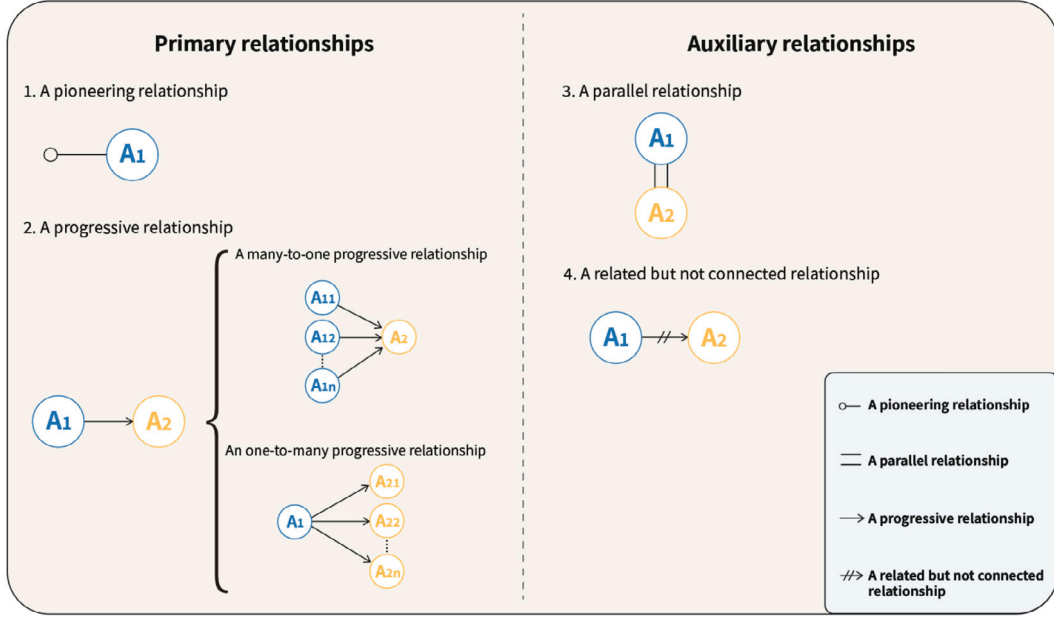


Fig. 8. Two fundamental relationships and two auxiliary relationships among the S&T achievements.

where:

- A_i is a single preceding achievement.
- $\{A_{j1}, A_{j2}, \dots, A_{jn}\}$ are multiple subsequent achievements.

Examples: Progressive relationships demonstrate how knowledge and technology evolve over time, with each new development building on the previous ones.

- **Chip:** The RISC-V instruction set architecture has its origins in and was developed from the original RISC design.
- **AI:** LeNet [17,18] is a pioneering convolutional neural network that inspired AlexNet [19], a milestone in the field of deep learning.
- **Open-sources systems:** OpenBLAS [20] is a progressive achievement of GotoBLAS2 [21].
- **Benchmarks:** The CH-benCHmark [22] exemplifies a many-to-one progressive relationship as it integrates aspects from both the TPC-C [23] and TPC-H [24] benchmarks. This benchmark is designed to evaluate a hybrid workload by combining the transactional operations characteristic of TPC-C with the complex querying features of TPC-H.

3.3.2. Two auxiliary relationships

Two auxiliary relationships contain a parallel relationship and a connected but not related relationship.

Relationship three: A parallel relationship. *Definition:* A parallel relationship indicates that the achievements that involve the same component of an extended EC, e.g., problem or sub-problem, are proposed simultaneously within a brief and shared timeframe.

Formal expression: For a set of achievements A with each achievement $A_i \in A$, a parallel relationship between two achievements A_i and A_j is defined as:

$$\begin{aligned}
 P(A_i, A_j) = 1 &\iff (Q(A_i) = Q(A_j)) \\
 &\quad \wedge (EC(A_i) \cap EC(A_j) \neq \emptyset) \\
 &\quad \wedge ([T(A_i.b), T(A_i.e)] \cap [T(A_j.b), T(A_j.e)] \neq \emptyset)
 \end{aligned} \tag{4}$$

where:

- $Q(a)$ is the key problem domain, sub-problem domain, problem, or sub-problem that achievement a addresses.
- $EC(a)$ denotes the EC involved in achievement a .
- $T(a.b)$ and $T(a.e)$ represent the begin time and end time of achievement a , respectively. The achievements A_i and A_j are considered to be in a parallel relationship if their time intervals overlap.

Examples: Parallel relationships occur across multiple fields where different approaches are employed simultaneously to address a common issue within a shared timeframe.

- **Chip:** The Von Neumann architecture and the Harvard architecture are two parallel works in computer system-level design in the 1940s.
- **AI:** BERT [25] and GPT [26] are two parallel works in the research of big models.
- **Open-sources systems:** Ubuntu [27] and CentOS [28] are two parallel works in open-source software.
- **Benchmarks:** BigDataBench [29] and BigBench [30] are two benchmarks specifically designed for evaluating big data systems, and they epitomize a parallel relationship as both were published within a year of each other, representing concurrent efforts in the problem domain of big data benchmarking.

Relationship four: A related but not connected relationship.. *Definition:* For the achievements that involve the same component of an extended EC, e.g., a problem or sub-problem, a related but not connected relationship suggests that these achievements are not proposed simultaneously within a brief and shared timeframe. Instead, they are related in some way, but there is no explicit public acknowledgment cited by the later achievements indicating inspiration or influence from the earlier ones.

Formal expression: A related but not connected relationship characterizes that two achievements are not parallel and have similar components inheriting the same high-level component of an extended EC but lack a citation.

This relationship carries three implications. First, two achievements have similar components inheriting the same high-level component of an extended EC. Second, they are not parallel in nature, meaning they are not proposed simultaneously. Third, though the two achievements

Algorithm 1 Identify four fundamental relationships among numerous S&T achievements

```

1: Input: IDs, TimeStamps_b, TimeStamps_e, References, EC, ProblemsQ
2: Output: PioneerRelationship, ParallelRelationship, ProgressiveRelationship, Related But Not Connected Relationship
3: Initialize PioneerRelationship, ParallelRelationship, ProgressiveRelationship, Related But Not Connected Relationship to empty sets
4: for each achievement i do
5:   if i opens up a new research direction in the form of establishing a new field, problem domain, sub-problem domain, problem, sub-problem, algorithm or algorithm-like mechanism, implementation, or support system within an extended EC. then
6:     Add i to PioneerRelationship
7:   end if
8: end for
9: for each pair of achievements (i, j) where i ≠ j do
10:  if ProblemsQ[i] = ProblemsQ[j] then
11:    if TIMEINTERVALSEXISTOVERLAP([TimeStamps_b[i], TimeStamps_e[i]], [TimeStamps_b[j], TimeStamps_e[j]]) AND EC[i] ∩ EC[j] ≠ ∅ then
12:      Add (i, j) to ParallelRelationship
13:    else if (TimeStamps_e[i] precedes TimeStamps_b[j] OR TimeStamps_e[j] precedes TimeStamps_b[i]) AND (i ∈ References[j] OR j ∈ References[i]) AND EC[i] ∩ EC[j] ≠ ∅ then
14:      Add (i, j) to ProgressiveRelationship
15:    end if
16:  end if
17: end for
18: for each consecutive pair of achievements (i, i + 1), sorted by TimeStamps_e do
19:  if ProblemsQ[i] = ProblemsQ[i + 1] AND Ai ∉ References[Ai+1] AND EC[i] ∩ EC[i + 1] ≠ ∅ AND TIMEINTERVALSNOOVERLAP([TimeStamps_b[i], TimeStamps_e[i]], [TimeStamps_b[i + 1], TimeStamps_e[i + 1]]) then
20:    Add (i, i + 1) to Related But Not Connected Relationship
21:  end if
22: end for
23: return PioneerRelationship, ParallelRelationship, ProgressiveRelationship, Related But Not Connected Relationship

```

have a chronological order, the later ones did not cite the earlier ones. While we cannot accurately disclose the underlying motivation, we emphasize the factual nature of these implications.

$$\begin{aligned}
C(A_i, A_{i+1}) = 1 &\iff (Q(A_i) = Q(A_{i+1})) \\
&\quad \wedge (EC(A_i) \cap EC(A_j) \neq \emptyset) \\
\wedge ([T(A_i, b), T(A_i, e)] \cap [T(A_{i+1}, b), T(A_{i+1}, e)] = \emptyset) &\quad (5) \\
&\quad \wedge ((A_i \notin R(A_{i+1})))
\end{aligned}$$

where:

- $Q(a)$ is the key problem domain, sub-problem domain, problem, or sub-problem that achievement a addresses.
- $EC(a)$ denotes the EC involved in achievement a .
- $T(a, b)$ and $T(a, e)$ represent the begin time and end time of achievement a , respectively.
- $R(a)$ indicates the references of achievement a . $A_i \notin R(A_{i+1})$ indicates achievement A_i does not in the reference list of achievement A_{i+1} .

Examples: related but not connected relationships trace the sequence of achievements that tackle similar issues across different timeframes. Although these developments may seem interconnected, they often evolve independently.

- **AI:** Condconv [31] and Dynamic Convolution [32] are two contemporary achievements for dynamical models with similar approaches.
- **Benchmarks:** TPC-C [23] and TPC-E [33], both developed to evaluate Online Transactional Processing (OLTP) databases, exemplify a related but not connected relationship. They sequentially advance the field of database benchmarking without direct influence from one another.

Fig. 8 illustrates the interplay among S&T achievements governed by four relationships: pioneering, progressive, parallel, and related but not connected. In S&T evaluatology, formalizing the four relationships is crucial for understanding and analyzing the interaction between various scientific achievements.

3.3.3. The algorithm to identify the four relationships

In this subsection, we present an algorithm designed to discern four significant types of relationships among a myriad of science and technology achievements: pioneering, progressive, parallel, and related but not connected relationships. The algorithm operates on a set of inputs comprising achievement IDs, timestamps, references (key references), evaluation conditions (EC), and the key problem domain, sub-problem domain, problem, or sub-problem Q addressed by each achievement. Subsequently, it outputs sets of achievement pairs categorized into pioneering, progressive, parallel, or related but not connected relationships.

Inputs:

- *IDs*: A list of achievement IDs or an optional list of pairs of achievement IDs for comparison.
- *TimeStamps_b*: Timestamps indicating the beginning time of achievements.
- *TimeStamps_e*: Timestamps indicating the end time of achievements.
- *References*: Key references or citations between achievements.
- *EC*: The involved EC components of each achievement.
- *ProblemsQ*: A compilation of key problem domain, sub-problem domain, problem, or sub-problem Q addressed by each achievement.

The algorithm proceeds as follows:

1. Identification of Pioneering Relationship:

- achievements that are the first to open up a new research direction by establishing a new field, problem domain, sub-problem domain, problem, sub-problem, algorithm or algorithm-like mechanism, implementation, or support system within an extended EC.

2. Identification of Parallel Relationship:

- achievements addressing the same problem domain, sub-problem domain, problem, or sub-problem are scrutinized.
- achievements occurring within overlapping time intervals are classified as having a parallel relationship.

3. Identification of Progressive Relationship:

- achievements sharing the same problem domain, sub-problem domain, problem, or sub-problem are paired.
- successive temporal order and mutual referencing between achievements, indicate a progressive relationship.

4. Identification of related but not connected Relationship:

- achievements within no-overlapping time intervals are evaluated.
- achievements addressing the same problem domain, sub-problem domain, problem, or sub-problem, without any mutual referencing, are considered to have a related but not connected relationship.

Outputs:

- *PioneerRelationship*: A set of achievement pairs in a Pioneering relationship.
- *ParallelRelationship*: A set of achievement pairs in a Parallel relationship.
- *ProgressiveRelationship*: A set of achievement pairs in a Progressive relationship.
- *RelatedButNotConnectedRelationship*: A set of achievement pairs in a related but not connected relationship.

This algorithm offers a systematic approach to unraveling the intricate interplay among S&T achievements, facilitating a deeper understanding of their underlying relationships.

3.4. Establishing the real-world S&T ES

This subsection presents how to model the real-world S&T ES (M_r), as depicted in Fig. 7. The proposed real-world S&T ES encompasses the entire collection of S&T achievements, each of which is mapped onto the several components of an extended EC. As the aim is to single out the top achievements, we ignore the other components of the S&T ecosystem, e.g., the mechanisms and policies within the S&T ecosystems that profoundly shape the evolution of these achievements.

Although this approach can identify all S&T achievements, the real-world S&T ES (M_r) is often susceptible to confounding factors. For instance, the communities tend to favor highly prestigious scientists, naturally drawing more attention to the research outcomes of well-known scientists. This bias stems from a real-world S&T ES (M_r)'s inability to track the developmental trajectory of S&T achievements and elucidate the relationships among these achievements.

To address these deficiencies in the real-world S&T ES (M_r), we will develop the perfect S&T EM (M_p) in Section 3.5, which systematically traces the evolution of S&T achievements and clarifies the interconnections among them.

3.5. Establishing the perfect S&T EM

The core objective of the perfect S&T EM is to track the evolution of S&T achievements. This model aims to capture these achievements' dynamic changes and progressions in terms of four relationships as they contribute to the S&T ecosystem. Doing so provides a full-picture understanding of the evolution of S&T within the real-world context.

Section 3.1 has offered a concise overview of the process for establishing a perfect S&T EM. This subsection will delve into the details,

comprehensively exploring the methodology.

The perfect S&T EM aims to track the evolution of the real-world S&T ES. A perfect S&T EM meticulously tracks the progression of a real-world S&T ES, from ES_i to ES_{i+1} , in a rigorous manner. This process ensures that only one achievement is added from ES_i to ES_{i+1} . This framework allows for an accurate description of the evolution of a field, starting from ES_0 and ultimately culminating in the development of a comprehensive real-world S&T ES.

In this framework, we also provide an auxiliary structure to depict the interconnected relationships among all the achievements. As we progress from ES_0 to ES_1 , from ES_i , then to ES_{i+1} , and ultimately towards a real-world S&T ES, we adhere to the principle of adding only one achievement at a time. When a new achievement is introduced in ES_{i+1} , we compare it to its counterpart in ES_i and determine the relationship based on the rules defined in Section 3.3. This approach ensures a systematic and logical evaluation of the evolving achievements within the S&T evaluation framework.

Meanwhile, as discussed in [6,7], the perfect S&T EM also implies exploring and understanding the entire spectrum of possibilities within a research field.

By embracing the concept of a perfect S&T EM, researchers can push the boundaries of knowledge and innovation. It encourages them to explore new avenues, challenge existing assumptions, and uncover hidden potentials. Fig. 7 shows a sample of a perfect S&T EM. The perfect S&T EM has almost entirely replicated the real-world S&T ES. Not only can it establish an extended EC, but it can also organize a roadmap of achievements' evolution by identifying relationships among achievements.

3.6. Establishing the pragmatic EM

Building upon the perfect S&T EM, we can establish the pragmatic evaluation model after filtering out non-significant achievements. The process of filtering out non-significant achievements is essentially the reverse of the process outlined in Section 3.5, which explains how an achievement is added from ES_i to ES_{i+1} . In the filtering process, we employ four rounds of filtering rules.

In the first round, our focus is to identify and filter out non-significant achievements from those that demonstrate progressive relationships. For the achievements that have progressive relationships, as they involve one or several same components of an extended EC, e.g., a problem domain or a problem, we compare achievements under the shared components of the extended EC and filter out those that are not significant.

In the second round, we will identify the achievements that exhibit parallel relationships or related but not connected relationships to the achievements preserved in the first round. Once we have compiled these achievements, we will proceed with an additional filtering process to eliminate any non-significant ones.

We categorize an achievement that exhibits a pioneering relationship as a pioneering achievement. In the third round, we will compare the pioneering achievements under the shared components of the extended EC and filter out achievements that are deemed non-significant.

In the fourth round, we will identify the achievements that exhibit parallel relationships or related but not connected relationships to the pioneering achievements preserved in the third round. Once we have compiled these achievements, we will proceed with an additional filtering process to eliminate any non-significant ones.

According to Zhan [34], an achievement can exert a positive change force over a counterpart by significantly enhancing the simplicity, user experience, cost-effectiveness, efficiency, or other fundamental features by several orders of magnitude. On the other hand, significant deviation from existing technology ecosystems can generate a negative change force. Additionally, when different usage patterns require users to incur significant learning costs, it can also result in a negative change force. This empirical law helps to explain why a certain achievement

dominates over the other one.

In theory, it is possible to quantitatively measure two achievements under the same extended EC from different dimensions, and each dimension is defined as X_i . To summarize these dimensions, we propose a simple rule of thumb. We differentiate between positive and negative signs and sum up the positive or negative values of $lg(X_i)$ (metrics from different dimensions), and the formula is shown in Eq. (6). This approach allows for a holistic assessment of the achievements, taking into account their various dimensions and providing a comprehensive understanding of their overall impact. By considering both positive and negative values, we can gain insights into the strengths and weaknesses of each achievement, enabling a more nuanced evaluation and comparison.

$$V = \sum_i lg(X_i) \quad (6)$$

4. The top N @X @Y methodology

As a typical case study, this section presents how to apply S&T evaluatology.

We propose Top N @X @Y, aiming to recognize the top N achievements within a specific period X in a particular field Y. Here, N represents the number of top achievements, X represents a specific period, and Y represents a particular field.

To optimize the effectiveness of evaluating science and technology, a standardized procedure has been devised as outlined below.

First, during a particular timeframe X, we create a real-world S&T ES that encompasses all achievements. Each achievement is decomposed into various components within the specific extended EC.

Second, based on the real-world S&T ES during a timeframe X, we construct a perfect S&T EM that traces the evolution of S&T achievements in the field of Y according to the four relationships.

Third, considering the total number of achievements (N), we assign different percentages that add up to 100% to the achievements that have pioneering relationships and progressive relationships.

Finally, following the four-round filtering process defined in Section 3.6, we filter out non-significant achievements to establish a pragmatic S&T EM that comprises the top N achievements during a timeframe X in the field of Y. Please note that the final step is iterative.

Following the aforementioned procedures, the top N achievements are obtained and can be presented in a tree form, as depicted in Fig. 6. Subsequently, we can proceed to rank these achievements along with their corresponding contributors and institutions.

We propose a simple rule to score each achievement, with higher scores leading to higher rankings. Initially, each selected achievement is assigned a score of 1.0 points. However, we give an extra score to each pioneering achievement. With each groundbreaking achievement paving the way for new research directions, we aggregate the cumulative scores of progressive achievements by applying a weight, which we call a pioneering weight, to the original score of the pioneering achievement.

Once the scores for each achievement are determined, we proceed to assess the contribution shares of each author and their respective institutions. The specific criteria for assessing the main academic contributors are as follows:

1. If the number of authors is three or fewer, the score is evenly distributed among all authors involved.
2. If there are more than three authors, and their contributions are stated to be equal, the score is evenly divided among all authors.
3. When there are more than three authors and their contributions are not stated as equal, the first author is assigned a first-author ratio, i.e., 0.3. In cases where multiple individuals share the first authorship, the first-author ratio is equally divided among them.

The corresponding author (or the last author in the absence of a designated corresponding author) receives a corresponding author ratio, i.e., 0.3. Similarly, if multiple individuals share the corresponding author role, the corresponding author ratio is evenly distributed among them. The remaining ratio is equally divided among the other authors.

As per the aforementioned rule, the score assigned to each achievement is subsequently distributed among the respective contributors based on their designated ratios. For every contributor, the corresponding institutions (which may be one or multiple) can be determined at the time of their contribution. In cases where a contributor is associated with multiple institutions, the score will be evenly divided among all the affiliated institutions.

5. A case study on the top 100 chip achievements

The chip industry plays a crucial role in driving technological advancements across various sectors, encompassing a vast ecosystem involved in software, hardware, and application development to harness their capabilities. Utilizing S&T evaluatology principles, the International Open Benchmark Council (BenchCouncil) has developed a well-defined extended EC to assess various aspects of chips comprehensively. The first level is the chip field, while the second level encompasses three problem domains: chip design, chip manufacturing, and chip packaging. At the third level, chip design involves several sub-problem domains, including system-level design, logic design, physical design, timing design, verification, and simulation. Chip manufacturing covers semiconductors, materials, and optics. Then, using the Top N @X @Y methodology, BenchCouncil has launched an ambitious initiative to systematically recognize and honor the most 100 groundbreaking and influential achievements in chip technology (Chip100) [8].

The current version of Chip100 uses the Top N @X @Y methodology, where N stands for 100, X spans from the 1940s (the advent of the first computer) to 2023, and Y indicates the chip field and the percentages of pioneering achievements and progressive achievements are 40% and 60%, respectively. For the ranking in Chip100, the pioneering weight is set as 0.2, the first-author ratio is 0.3, and the corresponding author ratio is 0.3.

The major influential accomplishments in chips are encompassed within Chip100. For example, as depicted in Fig. 6, the Instruction Set Architecture (ISA) was first introduced by Frederick Brooks in the 1960s. It defines a crucial sub-problem of computer architecture design (problem) of the system-level design (sub-problem domain) within the chip design problem domain: the challenge of designing the instruction set and proposing effective mechanisms. This concept led to the development of Complex Instruction Set Computers (CISC) and Reduced Instruction Set Computer (RISC). Subsequently, Instruction Set Architectures such as X86 and RISC-V emerged, drawing from the principles of CISC and RISC. This examination provides valuable insights into the connections among these achievements. So, Chip100 identified and evaluated significant achievements and researchers in the chip field that could not be discerned through the application of bibliometrics.

We use the data of Chip100, CSRankings, and the Highly cited Researchers from Elsevier to find the top 100 achievements, contributors, and institutions in the chip field.

CSRankings uses the metric of the number of publications at the top-tier conferences for gauging the academic influence of researchers or their affiliated institutions in computer science. The CSRankings database utilized by us extends across a timeline from 1970 to 2023, representing the most extensive timeframe available for CSRankings. The most matched areas include Computer Architecture and Design Automation.

The Highly Cited Researchers list is the typical metric based on citations. The main criteria for inclusion are “the authorship of multiple Highly Cited Papers™ within the past decade and being ranked

Table 2
Comparing Chip100 against CSRankings and highly cited researchers from Elsevier.

Methods	Top 20 achievements	Top 20 contributors	Top 20 institutions
Chip100 [8] (By the end of 2023)	Von Neumann Architecture, ISA, Stored-program computers, Cache memory, Boolean Algebra, Floating Point Unit, Formal Verification, Out-of-Order Execution, Stream Architecture, Amdahl's Law, Verilog, FPGA, Branch Predictor, CC-NUMA, ECC, EDA, Electrostatic Discharge, Harvard Architecture, Multi-Core Processors, NOC, SIMD Architecture, Single-Chip Multiprocessor, SOC, The Principle of Locality, and Virtual address translation	John von Neumann, Maurice Wilkes, Frederick Brooks, David A. Patterson, Gene Amdahl, George Boole, Robert Tomasulo, William Kahan, Phil Moorby, John L. Hennessy, Aart de Geus, Claude Shannon, Jen-Hsun Huang, John Gustafson, Lisa Su, Mark Hill, Michael J. Flynn, Michel Mardiguian, Richard Hamming Ross H. Freeman, Wayne Wolf, and William M. Johnson	Princeton University, IBM, Univ. of California - Berkeley, University of Cambridge, Stanford University, AMD, Intel, Massachusetts Institute of Technology, NVIDIA, Xilinx, University of Michigan, Gateway Design Automation, ARM, Bell Labs, Georgia Institute of Technology, Google, Harvard University, Motorola, Sandia National Laboratories, Synopsys, University of Paris South, University of Pennsylvania, and University of Washington
CSRankings [2] (By the end of 2023)	Achievements are predicated on the number of publications in top-tier conferences.	David T. Blaauw, Andrew B. Kahng, Srinivas Devadas, Josep Torrellas, Diana Marculescu, Mark Horowitz, Alberto L. Sangiovanni Vincentelli, Mahmut T. Kandemir, Jason Cong, Yuan Xie, Moinuddin K. Qureshi, Giovanni De Micheli, Sheldon X.D. Tan, Onur Mutlu, David Z. Pan, Yiran Chen, ohsen Imani, Zhiru Zhang, Xiaoyao Liang, and Margaret Martonosi	University of Michigan, Univ. of California - San Diego, Massachusetts Institute of Technology, Univ. of Illinois at Urbana-Champaign, Carnegie Mellon University, Stanford University, Univ. of California - Berkeley, Pennsylvania State University, Univ. of California - Los Angeles, Univ. of California - Santa Barbara, Georgia Institute of Technology, EPFL, Univ. of California - Riverside, ETH Zurich, University of Texas at Austin, Univ. of California - Irvine, Duke University, Shanghai Jiao Tong University, Cornell University, and Princeton University
Highly cited researchers [35] (2023)	Achievements are highly cited papers	There are a total of 98 highly cited researchers in the field of computer science, listed in no particular order. ^a	Chinese Academy of Sciences, Harvard University, Stanford University, National Institutes of Health, Tsinghua University, Massachusetts Institute of Technology, University of California San Diego, University of Pennsylvania, University of Oxford, Max Planck Society, University of California San Francisco, University College London, University of Hong Kong, Washington University, University of California Berkeley, Johns Hopkins University, Memorial Sloan Kettering Cancer Center, University of Cambridge, Yale University, University of California Los Angeles, and University of Washington Seattle (based on the summary of highly cited researchers from all research fields)

^a <https://clarivate.com/highly-cited-researchers/>.

in the top 1% based on citations in Web of Science™ [35]. Highly Cited Researchers™ represent a select group comprising only 0.1% of researchers in the world. The data of Highly Cited Researchers utilized by us was released in the year 2023, and hence, the timeframe is from 2013 to 2023, representing the most extensive timeframe available for this database. The matched area is Computer Science, as it cannot be narrowed down to focus solely on the chip field.

Table 2 outlines a compilation of the Top 20 outcomes from Chip100, CSRankings, and the Highly Cited Researchers list published by Elsevier. Throughout the remainder of this section, we will focus on analyzing the top five achievements, contributors, and institutions from various rankings to identify any notable distinctions.

First, we contrast the results of Chip100 with those from CSRankings. From Table 2, we can see that the results are totally different.

According to the analysis conducted by Chip100 (1940s–2023), The top five achievements include the Von Neumann Architecture, ISA, Stored-program computers, Cache memory, and Boolean Algebra. These achievements are crucial in driving the development of chips. Conversely, the achievements in CSRankings are solely based on the volume of publications in top-tier conferences.

Furthermore, the Top five institutions in the chip field encompass Princeton University (recognized for advancements like Von Neumann Architecture, The Principle of Locality, and Virtual address translation), IBM (recognized for advancements like ISA, CISC, Amdahl's Law, and Dennard Scaling Law), UC Berkeley (known for achievements

in Floating Point Unit design, RISC architecture, and RISC-V implementation), University of Cambridge (highlighted for innovations in Stored-program computers, Cache Memory, and Advanced RISC Machines), and Stanford University (acknowledged for progress in MIPS architecture, Superscalar processing, and Single-Chip Multiprocessor development).

In contrast, CSRankings only emphasizes the number of publications at top-tier computer science conferences. In the field of Computer Architecture and Design Automation, covering the period from 1970 to 2023, the top five research institutions include the University of Michigan, University of California-San Diego, Massachusetts Institute of Technology, University of Illinois at Urbana-Champaign, and Carnegie Mellon University.

Among the top five research institutions selected by CSRankings, only the University of Michigan (No. 11), Massachusetts Institute of Technology (No. 8), and Carnegie Mellon University (No. 24) are included within the Chip100 (1940s–2023), while the University of California-San Diego and the University of Illinois at Urbana-Champaign are not featured.

The Top five contributors in Chip100 are John von Neumann (recognized for Von Neumann Architecture), Maurice Wilkes (known for Stored-program computers and Cache Memory mechanism), Frederick Brooks (credited with ISA), David A. Patterson (recognized for the monograph “Computer Architecture: A Quantitative Approach”, RISC, and RISC-V), and Gene Amdahl (recognized for CISC and Amdahl's

Law). Contrasting with this viewpoint, the top five chip research contributors according to CSRankings by the end of 2023 are Onur Mutlu (117 contribution papers), Yuan Xie (116 contribution papers), Jason Cong (112 contribution papers), Alberto L. Sangiovanni-Vincentelli (108 contribution papers), and David Z. Pan (104 contribution papers). The noticeable disparity between these rankings is apparent, with none of the top five researchers in CSRankings being featured in the Chip100 list spanning from the 1940s to 2023.

Another well-known ranking is the Highly Cited Researchers published by Elsevier. The achievements are constrained to highly cited papers as viewed through the lens of the Highly Cited Researchers. The top institutions listed in Table 2 are determined based on a roster of highly cited researchers from all research fields. In 2023, a total of 7125 researchers were recognized as Highly Cited Researchers, including 98 in the field of computer science. It is challenging to conduct precise searches for top institutions or researchers within a specific and focused field, such as Chip.

The criteria for this recognition clearly prioritize the impact of papers from a bibliometric perspective, as indicated by their citation counts. As a result, none of the top five contributors listed in the Chip100 have been encompassed in the Highly Cited Researchers list. On the other hand, none of the 98 Highly Cited Researchers in the field of computer science have been included in Chip100 as well.

6. Conclusion

This article systematically reveals three severe bibliometrics limitations in recognizing top science and technology achievements and researchers. To address these shortcomings, we introduce science and technology evaluatology, which exemplifies the application of evaluatology in evaluating science and technology achievements. At the heart of this approach lies the concept of an extended evaluation condition, encompassing nine crucial components. We define four relationships that illustrate the connections among various achievements based on their mapped extended EC components, as well as their temporal and citation links: pioneering, progressive, parallel, and related but not connected. Within a pioneering or progressive relationship under an extended evaluation condition, evaluators can effectively compare these achievements by carefully addressing the influence of confounding variables. The case studies show the effectiveness of the proposed methodology compared with bibliometrics.

CRedit authorship contribution statement

Guoxin Kang: Contributions to the summary of the related work for CSRankings and c-score in Section 2.2, the whole-session discussion, the mathematical formulations of four relationships, the algorithm for identifying the four fundamental relationships in Section 3.3, the benchmark examples in Section 3.3, and the presentations of Figure 1, 2, 3 and Table 1. **Wanling Gao:** Contributions to the summary of the related work for H-index in Section 2.2, the whole-session discussion, partial revision of mathematical formulations and algorithms in Section 3.3, and the presentations of Figure 4, 5, 6, 7. **Lei Wang:** Contributions to the summary of the related work for CiteScore and SNIP in Section 2.2, the whole-session discussion, the writing of Section 4, Section 5, the chip examples in 3.3, and the presentations of Figure 5, 6. **Chunjie Luo:** Contributions to the part of SJR in related work, the AI examples in Section 3.3, and the discussion. **Hainan Ye:** Contributions to the data for Chip100 rankings and the discussion. **Qian He:** Contributions to the presentations of Figure 8 and the discussion. **Shaopeng Dai:** Contributions to the discussion. **Jianfeng Zhan:** Contributions to the proposal for the evaluatology-based science and technology evaluation methodology, including the extended EC, four relationships, real-world ES, perfect S&T EM, pragmatic S&T EM, and Top N@X @Y methodology. Contributions to the presentation of most texts, excluding figures and tables. Also contributions to other aspects of the work unless otherwise explicitly stated.

Funding

This research is supported by the Strategic Research Special Funding of the Bureau of Development and Planning, Chinese Academy of Sciences.

Declaration of competing interest

Jiafeng Zhan is the editor in chief, Lei Wang, Wanling Gao, Chunjie Luo are the founding editors of BenchCouncil Transactions on Benchmarks, Standards and Evaluations and were not involved in the editorial review or the decision to publish this article. The other authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- [1] BenchCouncil, BenchCouncil science and technology achievement evaluation, 2023, <https://www.benchcouncil.org/evaluation/>.
- [2] Emery berger. CSRankings, 2023, <https://csrankings.org/#/index?all&us>.
- [3] John P.A. Ioannidis, Jeroen Baas, Richard Klavans, Kevin W. Boyack, A standardized citation metrics author database annotated for scientific field, *PLoS Biol.* 17 (8) (2019) e3000384.
- [4] Jorge E. Hirsch, An index to quantify an individual's scientific research output, *Proc. Natl. Acad. Sci.* 102 (46) (2005) 16569–16572.
- [5] H-index, 2023, <https://en.wikipedia.org/wiki/H-index>.
- [6] Jianfeng Zhan, Lei Wang, Wanling Gao, Hongxiao Li, Chenxi Wang, Yunyou Huang, Yatao Li, Zhengxin Yang, Guoxin Kang, Chunjie Luo, et al., Evaluatology: The science and engineering of evaluation, *BenchCounc. Trans. Benchmark. Stand. Eval.* 4 (1) (2024) 100162.
- [7] Jianfeng Zhan, A short summary of evaluatology, *BenchCounc. Trans. Benchmark. Stand. Eval.* 4 (2) (2024).
- [8] Top 100 chips achievements, 2023, <https://www.benchcouncil.org/evaluation/>.
- [9] Digital Bibliography & Library Project, University of Trier, 2023, <https://dblp.org/>.
- [10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin, Attention is all you need, *Adv. Neural Inf. Process. Syst.* 30 (2017).
- [11] Accepted papers of CVPR 2022, 2022, <https://cvpr2022.thecvf.com/accepted-papers>.
- [12] Accepted papers of MICRO 2022, 2022, <https://microarch.org/micro55/index.php>.
- [13] Chris James, Lisa Colledge, Wim Meester, Norman Azoulay, Andrew Plume, Citescore Metr.: Creat. J. Metr. Scopus Cit. Index (2018) arXiv preprint arXiv: 1812.06871.
- [14] Henk F. Moed, Measuring contextual citation impact of scientific journals, *J. Informetr.* 4 (3) (2010) 265–277.
- [15] Kevin W. Boyack, Mapping knowledge domains: Characterizing PNAS, *Proc. Natl. Acad. Sci.* 101 (Suppl_1) (2004) 5192–5199.
- [16] Warren S. McCulloch, Walter Pitts, A logical calculus of the ideas immanent in nervous activity, *Bull. Math. Biophys.* 5 (1943) 115–133.
- [17] Yann LeCun, Bernhard Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne Hubbard, Lawrence D. Jackel, Backpropagation applied to handwritten zip code recognition, *Neural Comput.* 1 (4) (1989) 541–551.
- [18] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, Gradient-based learning applied to document recognition, *Proc. IEEE* 86 (11) (1998) 2278–2324.
- [19] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, Imagenet classification with deep convolutional neural networks, *Adv. Neural Inf. Process. Syst.* 25 (2012).
- [20] Xianyi Zhang, Qian Wang, Yunquan Zhang, OpenBLAS: a high performance blas library on loongson 3a cpu, *J. Softw.* 22 (Zk2) (2012) 208–216.
- [21] Kazushige Goto, Robert A. van de Geijn, Anatomy of high-performance matrix multiplication, *ACM Trans. Math. Softw.* 34 (3) (2008) 1–25.
- [22] Richard Cole, Florian Funke, Leo Giakoumakis, Wey Guy, Alfons Kemper, Stefan Krompass, Harumi Kuno, Raghunath Nambiar, Thomas Neumann, Meikel Poess, et al., The mixed workload CH-benchmark, in: *Proceedings of the Fourth International Workshop on Testing Database Systems*, 2011, pp. 1–6.
- [23] TPC-c benchmark, 2010, <http://www.tpc.org/tpcc/>.
- [24] TPC-H benchmark, 2010, <http://www.tpc.org/tpch/>.
- [25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.
- [26] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, Improving language understanding by generative pre-training, OpenAI (2018).
- [27] Ubuntu. <https://ubuntu.com/>, Initial release in 2004.

- [28] CentOS. <https://www.centos.org/>, Initial release in 2004.
- [29] Lei Wang, Jianfeng Zhan, Chunjie Luo, Yuqing Zhu, Qiang Yang, Yongqiang He, Wanling Gao, Zhen Jia, Yingjie Shi, Shujie Zhang, et al., Bigdatabench: A big data benchmark suite from internet services, in: 2014 IEEE 20th International Symposium on High Performance Computer Architecture, HPCA, 2014, pp. 488–499.
- [30] Ahmad Ghazal, Tilmann Rabl, Mingqing Hu, Francois Raab, Meikel Poess, Alain Crolotte, Hans-Arno Jacobsen, Bigbench: Towards an industry standard benchmark for big data analytics, in: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, 2013, pp. 1197–1208.
- [31] Brandon Yang, Gabriel Bender, Quoc.V Le, Jiquan Ngiam, Condconv: Conditionally parameterized convolutions for efficient inference, Adv. Neural Inf. Process. Syst. 32 (2019).
- [32] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, Zicheng Liu, Dynamic convolution: Attention over convolution kernels, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11030–11039.
- [33] Tpc Benchmark™ E. Citeseer, Transaction Processing Performance Council, 2010.
- [34] Jianfeng Zhan, Three laws of technology rise or fall, BenchCounc. Trans. Benchmark. Stand. Eval. 2 (1) (2022) 100034.
- [35] Highly cited researchers, 2023, <https://clarivate.com/highly-cited-researchers/>.



Research Article

Analyzing the obstacles to the establishment of sustainable supply chain in the textile industry of Bangladesh

Md. Hasibul Hasan Hemal^a, Farjana Parvin^a, Alberuni Aziz^{b,*}

^a Department of Industrial Engineering and Management, Khulna University of Engineering & Technology, Khulna, 9203, Bangladesh

^b Department of Textile Engineering, Khulna University of Engineering & Technology, Khulna, 9203, Bangladesh

ARTICLE INFO

Keywords:

Textile industry
Sustainable supply chain
Sustainability
MCDM
DEMATEL
Fuzzy TOPSIS

ABSTRACT

Bangladesh's textile sector plays a crucial role in its economy by creating jobs and significantly contributing to export revenue. However, this industry faces challenges, including contaminated water sources and the release of airborne pollutants due to its high-water usage, chemical dyes, and manufacturing processes. Therefore, establishing a sustainable supply chain is essential. This study aims to identify the critical obstacles to establishing a sustainable supply chain. Multi-Criteria Decision Making (MCDM) techniques, such as DEMATEL, help reveal the relationships between different components and determine the relative importance of each in the decision-making model. Meanwhile, Fuzzy TOPSIS proves reliable in situations of uncertainty, allowing for effective ranking of the barriers. The findings indicate that the most pressing barriers include resistance to change and the adoption of innovation, financial constraints or high costs, and a lack of support and commitment from top management. This assessment helps pinpoint crucial obstacles that must be addressed to achieve sustainability in the textile sector. By effectively identifying and eliminating these barriers, this study aims to assist those involved in the industry in their pursuit of a more sustainable future.

1. Introduction

The textile industry significantly boosts the nation's economy by generating export revenue and job opportunities. Known for low labor costs, it produces various textiles like fabrics and clothing, driving global trade and growth [1]. Bangladesh, now the world's twelfth-largest clothing producer, derives approximately 77 % of its foreign exchange and 50 % of its industrial workforce from this sector [2]. The textile industry contributes 81 % to the country's GDP and is its top export earner, with around 5600 factories in operation [3]. The textile sector has a dark side, particularly the release of contaminated water from industrial sources, which poses serious environmental threats and harms living organisms [4]. It ranks just after the oil industry as one of the most polluting industries, negatively impacting all aspects of sustainability: environmental, economic, and social [5]. The industry's supply chain contributes to waste, pollution, and resource depletion, consuming significant amounts of energy, chemicals, and water throughout a product's life cycle. To promote environmental sustainability, clothing designers and supply chains must adopt ecologically and socially responsible design principles [6]. A sustainable supply chain in the

textile sector is crucial for minimizing environmental harm, promoting ethical practices, and ensuring long-term profitability. It fosters transparency, reduces waste, and meets consumer demand for eco-friendly products [7]. In today's business climate, prioritizing sustainable supply chain management can provide a competitive edge [8]. While many studies focus on performance and enablers for establishing sustainable supply chains, few address the obstacles to their long-term viability [9–11]. This study aims to identify these critical barriers in Bangladesh's textile sector, which is vital to the country's economy.

The goal of this study is to identify the current issues faced by the textile sector. The barriers identified are sourced from existing literature through an extensive review and are organized in a coherent sequence with the assistance of experts. This organization aims to help researchers gain a better understanding of the field. Various Multi-Criteria Decision Making (MCDM) tools are then employed. The Decision-Making Trial and Evaluation Laboratory (DEMATEL) method is used to determine how these barriers are interconnected, while the Fuzzy Technique for Order Preference by Similarities to Ideal Solution (TOPSIS) method prioritizes the barriers. The structure of the paper is as follows: The introduction provides a foundation for the research. Section 2 outlines

* Corresponding author at: Department of Textile Engineering, Khulna University of Engineering & Technology, Khulna, 9203, Bangladesh.

E-mail addresses: hasibulhasan1802@gmail.com (Md.H.H. Hemal), farjanamousumi17@iem.kuet.ac.bd (F. Parvin), alberunitekuet@gmail.com (A. Aziz).

<https://doi.org/10.1016/j.tbench.2024.100185>

Received 27 August 2024; Received in revised form 1 November 2024; Accepted 14 December 2024

Available online 15 December 2024

2772-4859/© 2024 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Table 1

Selected barriers and their sources.

Sl no.	Denoted by	Barriers	Sources
1	A	Consumer desire for lower prices	[11–13]
2	B	Lack of government support	
3	C	Organizational culture resistance to change	[14–16]
4	D	Lack of green materials, processes and technology	[13,17,18]
5	E	Lack of commitment and support by the top management level	[19–22]
6	F	Lack of training and education about sustainability	[18,23–25]
7	G	Monetary constraints or high costs	[26,27,28,29]
8	H	Resistance to change and adopt innovation	[14,30–32]

the research methodology, which includes the Fuzzy TOPSIS and DEMATEL methodologies. Section 3 presents a discussion that primarily highlights the findings from the applied research methods. Finally, Section 4 includes the conclusions and recommendations derived from the investigation, as well as considerations of constraints and future scope.

This study's originality lies in its data collection from experts in sustainability within Bangladesh's textile industry. It employs two distinct Multi-Criteria Decision-Making (MCDM) tools, each with its own mechanism, addressing different criteria for analysis. As a result, barriers are prioritized in two unique ways. The findings from these methods are compared, with any relevant circumstances discussed in detail. The research follows a two-phase approach: Phase 1 involves a preliminary assessment to identify the barriers hindering Sustainable Supply Chain Management, while Phase 2 focuses on determining the primary barriers.

2. Methodology

2.1. Identification of barriers

There are two categories of barriers to implementing sustainable supply chain management (SSCM): internal and external [12]. Through literature review and expert opinions, eight key barriers were identified (Table 1). External barriers include insufficient regulations, unreliable metrics for performance evaluation, and low market demand for sustainable products [13]. Internal barriers involve organizational challenges such as financial constraints, lack of knowledge and awareness, and insufficient support from senior management [14].

2.2. Dimensions of a sustainable supply chain

Sustainability in engineering encompasses social, environmental, and economic issues. It involves balancing economic development, environmental stewardship, and social equity [33]. The triple bottom line theory advocates for businesses to enhance the economy, society, and environment for long-term benefits [34]. In Bangladesh's textile sector, addressing challenges like fair labor standards and worker empowerment is essential for social sustainability [35]. Economic sustainability in this sector requires balancing expansion, resource efficiency, and financial stability, focusing on productivity, innovation, and market growth while managing debts and production stability [36–38].

Ensuring environmental sustainability in Bangladesh's textile industry is essential for reducing ecological impact. Key strategies include adopting eco-friendly production processes, decreasing water and energy consumption, and implementing waste management policies. As global consumers increasingly favor eco-friendly products, these environmental concerns are increasingly linked to trade [39–40]. Barriers to sustainability are analyzed using DEMATEL and Fuzzy TOPSIS, clarifying relationships and ranking elements [41–43]. Both methods will

Table 2

Pairwise comparison matrix of an expert's opinion.

	A	B	C	D	E	F	G	H
A	0	1	2	2	2	3	2	2
B	2	0	2	1	2	2	3	2
C	2	2	0	1	2	3	2	2
D	1	2	2	0	3	2	2	2
E	1	1	3	2	0	2	1	2
F	2	2	2	3	2	0	1	2
G	2	3	3	2	1	1	0	2
H	2	2	2	3	3	2	3	0

Table 3

Normalized direct-relation matrix.

	A	B	C	D	E	F	G	H
A	0.000	0.072	0.140	0.092	0.160	0.116	0.164	0.124
B	0.092	0.000	0.136	0.132	0.108	0.124	0.148	0.124
C	0.120	0.112	0.000	0.096	0.148	0.160	0.136	0.156
D	0.128	0.120	0.120	0.000	0.128	0.124	0.164	0.148
E	0.096	0.120	0.136	0.124	0.000	0.136	0.108	0.140
F	0.108	0.092	0.128	0.108	0.120	0.000	0.112	0.156
G	0.124	0.104	0.132	0.148	0.144	0.112	0.000	0.152
H	0.120	0.124	0.144	0.116	0.136	0.132	0.128	0.000

serve as benchmarks for future studies [44]. This approach aligns with Ronald Fisher's Design of Experiments (DOE), allowing for thorough analysis and minimizing confounding variables [45].

2.3. DEMATEL

The DEMATEL method determines causal dependencies among predefined factors, helping to identify critical barriers that need immediate attention [46–49]. It relies on expert judgment rather than sample size and effectively analyzes relationships in complex systems [50]. By creating a visual representation of interrelated elements, DEMATEL clarifies interconnectedness and aids in complex decision-making [51]. The approach follows criteria from Zhan et al. in Evaluatology, changing one factor at a time to ensure accurate results. This method ultimately establishes cause-effect relationships among controlled factors [44].

The following steps are to be followed to carry out a full-fledged DEMATEL analysis:

- **Step 1: Expert opinions are gathered:** A questionnaire is developed based on selected barriers and distributed to experts for their input, which is then documented. Responses are assigned numerical values on a scale from 0 (No influence) to 4 (Very high impact). Pairwise matrices (Table 2) are created from expert feedback, leading to a combined matrix using a specific formula:

$$A = [A_{ij}]_{n \times n} = \frac{1}{H} \sum_{k=1}^H [X_{ij}^k]_{n \times n} \quad (1)$$

In the above formula, H is the number of experts, and n is the number of barriers. Each expert provides the impact of barrier i on barrier j. The impacts are presented in the matrix $X^k = [X_{ij}^k]_{n \times n}$.

- **Step 2: Normalized primary direct matrix is computed:** This normalized primary direct matrix (Table 3) is also known as initial influence matrix, D. The following formula is used for this step:

$$D = \frac{A}{S} \quad (2)$$

Table 4

Total relationship matrix.

	A	B	C	D	E	F	G	H
A	0.782	0.813	1.032	0.890	1.058	0.988	1.062	1.078
B	0.865	0.743	1.026	0.921	1.013	0.992	1.048	1.076
C	0.933	0.888	0.962	0.940	1.099	1.074	1.092	1.158
D	0.947	0.900	1.076	0.859	1.092	1.051	1.124	1.160
E	0.862	0.845	1.021	0.908	0.909	0.996	1.010	1.082
F	0.844	0.795	0.982	0.866	0.983	0.844	0.980	1.059
G	0.932	0.877	1.072	0.975	1.090	1.029	0.968	1.148
H	0.913	0.878	1.064	0.935	1.066	1.029	1.064	0.998

Table 5

Cause and effect determination.

Indicated as	Barriers	R _i	C _i	R _i -C _i	Identity
A	Consumer desire for lower prices	7.703	7.077	0.626	Cause
B	Lack of government support	7.683	6.738	0.945	Cause
C	Organizational culture resistance to change	8.146	8.235	-0.089	Effect
D	Lack of green materials, processes and technology	8.208	7.293	0.915	Cause
E	Lack of commitment and support by the top management	7.633	8.310	-0.677	Effect
F	Lack of training and education about sustainability	7.353	8.004	-0.651	Effect
G	Monetary constraints or high costs	8.091	8.348	-0.257	Effect
H	Resistance to change and adopt innovation	7.946	8.758	-0.812	Effect

Table 6

Rank of the barriers by DEMATEL method.

Indicated as	Barriers	R _i	C _i	R _i +C _i	Rank
A	Consumer desire for lower prices	7.703	7.077	14.780	7
B	Lack of government support	7.683	6.738	14.421	8
C	Organizational culture resistance to change	8.146	8.235	16.381	3
D	Lack of green materials, processes and technology	8.208	7.293	15.501	5
E	Lack of commitment and support by the top management level	7.633	8.310	15.943	4
F	Lack of training and education about sustainability	7.353	8.004	15.357	6
G	Monetary constraints or high costs	8.091	8.348	16.439	2
H	Resistance to change and adopt innovation	7.946	8.758	16.704	1

Where,

$$S = \max \left(\max \sum_{j=1}^n a_{ij}, \max \sum_{i=1}^n a_{ij} \right) \quad (3)$$

- **Step 3: Direct/Indirect influence matrix is calculated:** The interrelationships between the matrix elements are demonstrated in this matrix through both direct and indirect effects. I denoted identity matrix. T, the total relation matrix (Table 4), which is computed using:

Table 7

Cause and effects by threshold value.

	A	B	C	D	E	F	G	H
A	0.782	0.813	1.032	0.890	1.058	0.988	1.062	1.078
B	0.865	0.743	1.026	0.921	1.013	0.992	1.048	1.076
C	0.933	0.888	0.962	0.940	1.099	1.074	1.092	1.158
D	0.947	0.900	1.076	0.859	1.092	1.051	1.124	1.160
E	0.862	0.845	1.021	0.908	0.909	0.996	1.010	1.082
F	0.844	0.795	0.982	0.866	0.983	0.844	0.980	1.059
G	0.932	0.877	1.072	0.975	1.090	1.029	0.968	1.148
H	0.913	0.878	1.064	0.935	1.066	1.029	1.064	0.998

$$T = D(I - D)^{-1} \quad (4)$$

- **Step 4: Ri and Ci matrices are calculated.** Using the following formulas, the Ri and Ci values are determined:

$$Ri = \left(\sum_{i=1}^n t_{ij} \right)_{1 \times n} \quad (5)$$

$$Ci = \left(\sum_{j=1}^n t_{ij} \right)_{n \times 1} \quad (6)$$

$$T = [t_{ij}]_{n \times n} \quad (7)$$

Using the achieved values, we can determine Ri+Ci (the Total impacts provided and accepted by a barrier) and Ri-Ci (the overall effect contributed to the system by a barrier). If Ri-Ci is positive, it is a cause; otherwise, it is an effect.

- **Step 5: Assessing based on threshold (Alpha) value:** Determining the threshold value aids cause-and-effect identification and is optional. It is calculated by finding the mean of all values in the overall influence matrix, which is 0.981 in this case. Barriers C (1.032), E (1.058), F (0.988), G (1.062), and H (1.078) exceed this threshold, indicating that barrier A impacts them. Table 7 shows the cause and effects on the basis of the threshold value.

2.4. Fuzzy TOPSIS

Many real-life decisions depend on ambiguous evaluation data [52]. TOPSIS is a decision-making technique that aids in selecting the best option among various choices [53]. Fuzzy set theory addresses uncertainties from imprecision, enhancing decision quality [54–55]. Decision makers often use vague terms like "good" or "poor," leading to fuzziness in attribute weighting [56]. Triangular fuzzy numbers represent these linguistic expressions. The TOPSIS approach, introduced by Hwang and Yoon in 1981, selects options that are far from the negative-ideal solution and close to the positive-ideal one, based on precise attribute values and weights [57].

Table 8

Linguistic variables representing the significance weight of each criterion.

Linguistic Variable	Fuzzy Number
Extremely Low	(0,0,1)
Very Low	(0,1,3)
Low	(1,3,5)
Medium	(3,5,7)
High	(5,7,9)
Very High	(7,9,10)
Extremely High	(9,10,10)

Table 9
Integrated matrix.

Barriers	Social			Economic			Environmental		
A	3.7778	5.2778	6.6667	3.8889	5.5000	7.1667	2.8889	4.2222	5.8333
B	4.5556	5.8889	7.2222	2.0000	3.5000	5.3333	3.8333	5.2222	6.5556
C	5.6667	7.1111	8.2222	3.5000	5.0000	6.6667	2.0556	3.3889	5.2222
D	2.4444	3.7778	5.3333	3.3333	4.7222	6.2222	4.1667	5.8889	7.4444
E	3.1111	4.2778	5.6111	3.5556	5.1667	6.8333	2.2222	3.7778	5.5556
F	3.1111	4.2778	5.6111	4.8889	6.3333	7.5000	4.3889	5.9444	7.2778
G	3.1111	4.4444	5.9444	2.1667	3.3889	4.8889	3.1111	4.4444	5.9444
H	3.8333	5.4444	6.7778	4.5556	6.0556	7.3333	2.1111	3.3889	5.0556

Table 10
Normalized matrix.

Barriers	Social			Economic			Environmental		
A	0.6471	0.4632	0.3667	0.6286	0.4444	0.3411	0.8462	0.5789	0.4190
B	0.5366	0.4151	0.3385	1.2222	0.6984	0.4583	0.6377	0.4681	0.3729
C	0.4314	0.3438	0.2973	0.6984	0.4889	0.3667	1.1892	0.7213	0.4681
D	1.0000	0.6471	0.4583	0.7333	0.5176	0.3929	0.5867	0.4151	0.3284
E	0.7857	0.5714	0.4356	0.6875	0.4731	0.3577	1.1000	0.6471	0.4400
F	0.7857	0.5714	0.4356	0.5000	0.3860	0.3259	0.5570	0.4112	0.3359
G	0.7857	0.5500	0.4112	1.1282	0.7213	0.5000	0.7857	0.5500	0.4112
H	0.6377	0.4490	0.3607	0.5366	0.4037	0.3333	1.1579	0.7213	0.4835

The steps to be followed in order to completion of the Fuzzy TOPSIS method:

- **Step 1: Utilize the opinions of experts and use linguistic factors to assess the importance of attribute weights and ratings for various possibilities:** A questionnaire with 24 questions focused on social, economic, and environmental criteria, along with eight barriers, was distributed to textile industry experts. Their responses were recorded, and triangular fuzzy numbers (Table 8) were used to assign weightage based on their ratings.
- **Step 2: The ratings of barriers and weights of criteria are combined:** The criteria weights and barrier ratings are combined (Table 9) using the following calculation:

$$\tilde{w}_j = \frac{1}{t} [\tilde{w}_j^1 + \tilde{w}_j^2 + \dots + \tilde{w}_j^t] \quad (8)$$

$$\tilde{a}_{ij} = \frac{1}{t} [\tilde{a}_{ij}^1 + \tilde{a}_{ij}^2 + \dots + \tilde{a}_{ij}^t] \quad (9)$$

In the abovementioned equations, 't' is the number of decision-makers. The aggregated ratings a_{ij} of barriers x_j for attribute G_i and the average weight \tilde{w}_i of attribute G_i can be determined. It is generally assumed that each expert has the same knowledge base. However, that is inaccurate, as we know that not everyone has the same expertise in a specific domain.

- **Step 3: Normalize the complex fuzzy decision matrix:** Here, the complex fuzzy decision matrix $\tilde{A} = (\tilde{a}_{ij})_{s \times n} = [a_{lij}, a_{mij}, a_{uij}]_{s \times n}$ is

normalized (Table 10) into a corresponding matrix in the form of $\tilde{R}^{(\mathcal{A})} = (\tilde{r}_{ij}^{(\mathcal{A})})_{s \times n}$, Where:

$$\tilde{r}_{ij} = \left(\frac{a_{lij}}{a_{ui}^*}, \frac{a_{mij}}{a_{ui}^*}, \frac{a_{uij}}{a_{ui}^*} \right), i \in B \quad (10)$$

$$\tilde{r}_{ij} = \left(\frac{a_{li}}{a_{uij}}, \frac{a_{li}}{a_{uij}}, \frac{a_{li}}{a_{uij}} \right), i \in C \quad (11)$$

And,

$$a_{ui}^* = \max a_{uij}, \quad i \in B \quad (12)$$

$$a_{li}^- = \max a_{lij}, \quad i \in C \quad (13)$$

In the formulas, B represents benefit criteria and C represents cost criteria. Benefit criteria are desirable characteristics to optimize, with higher values being more advantageous. The goal is to maximize these advantages. In contrast, cost criteria indicate elements to minimize, with lower values being preferable. The objective here is to reduce expenses associated with each criterion.

- **Step 4: Develop the weighted normalized fuzzy decision matrix:** Utilizing the formula, the weighted normalized fuzzy decision matrix \tilde{V} (Table 11) is calculated:

$$\tilde{V} = [\tilde{v}_{ij}]_{m \times n} \quad (14)$$

Table 11
Weighted normalized matrix.

Barriers	Social			Economic			Environmental		
A	0.4982	0.4307	0.3667	0.4840	0.4133	0.3411	0.6515	0.5384	0.4190
B	0.4132	0.3860	0.3385	0.9411	0.6495	0.4583	0.4910	0.4353	0.3729
C	0.3322	0.3197	0.2973	0.5378	0.4547	0.3667	0.9157	0.6708	0.4681
D	0.7700	0.6018	0.4583	0.5647	0.4814	0.3929	0.4517	0.3860	0.3284
E	0.6050	0.5314	0.4356	0.5294	0.4400	0.3577	0.8470	0.6018	0.4400
F	0.6050	0.5314	0.4356	0.3850	0.3589	0.3259	0.4289	0.3824	0.3359
G	0.6050	0.5115	0.4112	0.8687	0.6708	0.5000	0.6050	0.5115	0.4112
H	0.4910	0.4176	0.3607	0.4132	0.3754	0.3333	0.8916	0.6708	0.4835

Table 12
FPIS (A*) & FNIS(A-).

	Social			Economic		Environmental			
A*	0.7700	0.6018	0.4583	0.8687	0.6708	0.5000	0.8916	0.6708	0.4835
A-	0.3322	0.3197	0.2973	0.3850	0.3589	0.3259	0.4289	0.3824	0.3359

Where,

$$\tilde{v}_{ij} = \tilde{w}_i \times \tilde{r}_{ij} \quad (15)$$

- **Step 5: Calculate the Fuzzy Positive Ideal Solution (FPIS) and Fuzzy Negative Ideal Solution (FNIS):** The FPIS (A*) and the FNIS (A-) are computed using the following sets of formulas in Table 12:

$$A^* = \{\tilde{v}_1^*, \tilde{v}_2^*, \dots, \tilde{v}_s^*\} \quad (16)$$

$$A^- = \{\tilde{v}_1^-, \tilde{v}_2^-, \dots, \tilde{v}_s^-\} \quad (17)$$

To simplify the calculation, FPIS and FNIS can be written as $\tilde{v}_i^* = [1, 1, 1]$ and $\tilde{v}_i^- = [0, 0, 0]$

- **Step 6: Calculate the distance of each barrier from A* and A-:** From the following equations, the distance of each barrier from FPIS and FNIS is calculated:

$$d_j^* = \sum_{i=1}^s d(\tilde{v}_{ij}, \tilde{v}_i^*), \quad (18)$$

$$d_j^- = \sum_{i=1}^s d(\tilde{v}_{ij}, \tilde{v}_i^-), \quad (19)$$

- **Step 7: Calculate the closeness coefficient of each barrier.** The closeness coefficient of each barrier is determined using the following formula:

$$CC_i = \frac{d_j^-}{d_j^* + d_j^-} \quad (20)$$

- **Step 8: Rank the barriers:** The barriers are then ranked using the closeness coefficient. According to this method, the barrier with the highest closeness coefficient will be ranked as the number one barrier. [55].

3. Discussion

This study identified and prioritized barriers within the textile sector of Bangladesh. Firstly, a DEMATEL analysis was conducted using a questionnaire distributed to experts. The objective of this study was to examine the interconnections among the barriers and evaluate their relative significance in influencing workers' decision-making processes. The primary goal was to determine the causes and effects of the identified barriers. In this analysis, if the value of Ri-Ci was negative, the barrier was classified as an effect; if it was positive, it was categorized as a cause. Among the eight identified barriers, three were classified as causes, while the remaining five were deemed effects, as shown in Table 5. Cause-based barriers are considered more critical than those based solely on their impact. Conversely, effect-based barriers are generally seen as dependable. According to the analysis, the barriers identified as causes were: consumer demand for low prices (A), lack of government support (B), and lack of green materials, processes, and technology (D). These three barriers were found to be the underlying

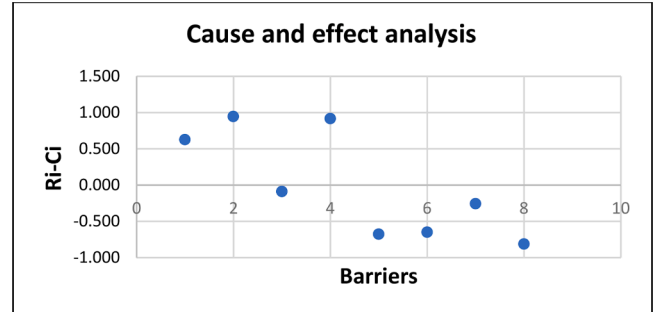


Fig. 1. Cause-effect digraph.

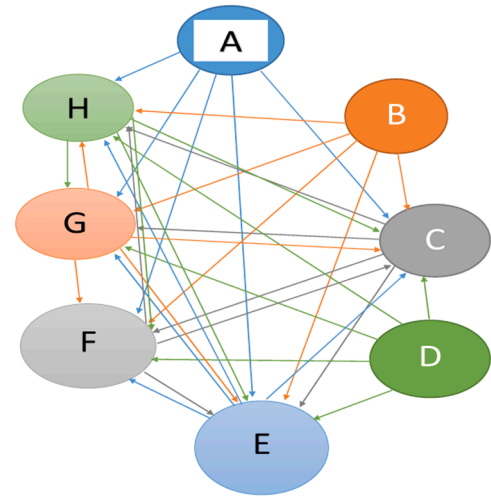


Fig. 2. Impact diagram.

Table 13
Barriers ranked by Fuzzy TOPSIS.

Barriers	Name of barriers	CCi	Rank
A	Consumer desire for lower prices	0.6865	7
B	Lack of government support	0.7262	5
C	Organizational culture resistance to change	0.7218	6
D	Lack of green materials, processes and technology	0.7384	4
E	Lack of commitment and support by the top management level	0.7993	2
F	Lack of training and education about sustainability	0.6203	8
G	Monetary constraints or high costs	0.8496	1
H	Resistance to change and adopt innovation	0.7398	3

reasons for the other five barriers all are evident in Fig. 1.

After identifying the causes and effects, we ranked the barriers based on their significance using the Ri + Ci method (as shown in Table 6). It was determined that resistance to change and adoption of innovation (H) ranked first, while monetary constraints or high costs (G) and resistance to change within organizational culture (C) ranked second and third, respectively. Subsequently, by utilizing the threshold value, we assessed the impact of each barrier on the others which is shown in Fig. 2.

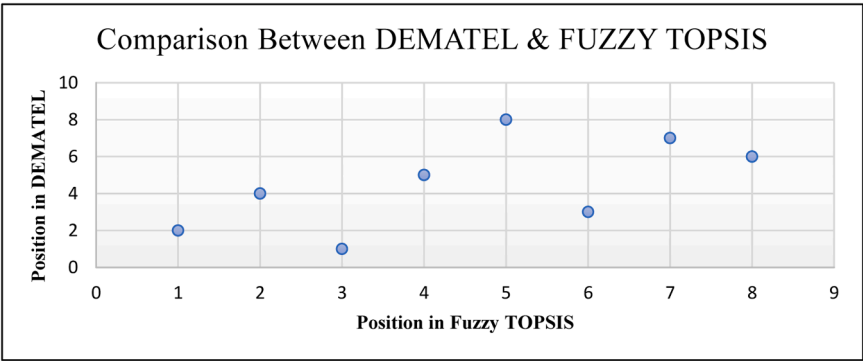


Fig. 3. Rank of barriers through DEMATEL and Fuzzy TOPSIS.

Table 14
Most Critical Barriers as recognized by previous papers.

References	Country	Name of the barrier
[4]	India	Communication gap among stakeholders
[58]	India	lack of effective governmental policies
[59]	Bangladesh	Insufficient financial incentives
[60]	Pakistan	Lack of ability to design a green product

After the DEMATEL analysis, the Fuzzy TOPSIS method was applied, leading to a ranking of barriers based on the closeness coefficient (Table 13). The analysis identified the most critical barrier as Monetary constraints (G), followed by Lack of support from top management (E), and Resistance to change (H). The results from DEMATEL and Fuzzy TOPSIS showed slight differences, as depicted in Fig. 3.

Although the results are similar, they differ because DEMATEL focuses on identifying key elements through cause-and-effect relationships, while Fuzzy TOPSIS uses fuzzy logic to address ambiguities in decision criteria. DEMATEL emphasizes interrelationships, whereas Fuzzy TOPSIS manages uncertainty.

Table 14 shows that different studies have identified various critical barriers. This variation is due to factors like differing production scales, labor costs, environmental practices, and technological advancements in the countries involved. If these factors are similar, the methodologies used may differ. For example, Rashid et al.’s study on Bangladesh presents different findings due to its distinct approach compared to this study.

4. Conclusion

This paper examines the barriers to sustainable supply chains in Bangladesh’s textile industry, identifying eight key hurdles based on recent studies. Utilizing the MCDM technique, expert opinions were gathered to assess the criticality and influence of these barriers. Two methods, DEMATEL and Fuzzy TOPSIS, were employed. DEMATEL analyzed the interrelations between barriers and classified them into cause-and-effect groups, considering the insights of 18 industry experts. Although companies may face varied issues, common barriers emerged due to the industry’s nature. Three barriers were identified as primary causes, illustrating how a small factor can have complex effects. The barriers were ranked, with resistance to change and innovation (H) in the top spot, followed by monetary constraints (G) and organizational culture resistance (C). Fuzzy TOPSIS further highlighted that monetary constraints (G), lack of top management support (E), and resistance to change (H) are the main obstacles.

4.1. Limitations and future scope of the work

This research has several theoretical and methodological constraints.

Firstly, only eight barriers are selected for analysis, limiting accuracy and length. A broader questionnaire may lead to reduced engagement from experts, resulting in random responses. Additionally, treating all expert opinions equally overlooks varying levels of knowledge, and biases may affect their scaling of barriers. The study is focused solely on SSCM procedures within the textile industry, making conclusions inapplicable to other sectors. Future research could validate these findings and explore other MCDM tools like AHP, grey theory, or ANP for comparative analysis.

Funding

This research received no external funding.

CRediT authorship contribution statement

Md. Hasibul Hasan Hemal: Writing – review & editing, Writing – original draft, Methodology, Data curation, Conceptualization. **Farjana Parvin:** Writing – review & editing, Validation, Supervision, Methodology. **Alberuni Aziz:** Writing – review & editing, Validation, Formal analysis, Data curation.

Declaration of competing interest

The authors declare no competing financial interests or personal relationships influencing this work.

Data availability

Data will be made available on request.

References

[1] S. Imran, M.A. Mujtaba, M.M. Zafar, A. Hussain, A. Mehmood, U.E. Farwa, T. Korakianitis, M.A. Kalam, H. Fayaz, C.A. Saleel, Assessing the potential of GHG emissions for the textile sector: a baseline study, *Heliyon*. 9 (2023) e22404, <https://doi.org/10.1016/j.heliyon.2023.e22404>.
[2] M.M. Islam, K. Mahmud, O. Faruk, M.S. Billah, Textile dyeing industries in Bangladesh for sustainable development, *Int. J. Environ. Sci. Dev.* (2011) 428–436, <https://doi.org/10.7763/ijesd.2011.v2.164>.
[3] S. Afrin, H.R. Shuvo, B. Sultana, F. Islam, A.A. Rus'd, S. Begum, M.N. Hossain, The degradation of textile industry dyes using the effective bacterial consortium, *Heliyon*. 7 (2021) e08102, <https://doi.org/10.1016/j.heliyon.2021.e08102>.
[4] A.S.M.M. Hasan, M. Rokonzaman, R.A. Tuhin, S.Md. Salimullah, M. Ullah, T. H. Sakib, P. Thollander, Drivers and Barriers to industrial energy efficiency in textile Industries of Bangladesh, *Energies* 12 (2019) 1775, <https://doi.org/10.3390/en12091775>.
[5] A. Vishwakarma, G.S. Dangayach, M.L. Meena, S. Gupta, Analysing barriers of sustainable supply chain in apparel & textile sector: a hybrid ISM-MICMAC and DEMATEL approach, *Cleaner Logist. Supply Chain* 5 (2022) 100073, <https://doi.org/10.1016/j.clscn.2022.100073>.
[6] A. Becker, T. Gries, 26 Sustainability in the textile industry. De Gruyter eBooks, 2023, pp. 473–480, <https://doi.org/10.1515/9783110670776-026>.
[7] L. Shen, L. Olfat, K. Govindan, R. Khodaverdi, A. Diabat, A fuzzy multi-criteria approach for evaluating green supplier’s performance in the green supply chain

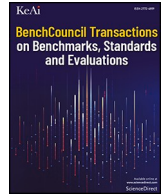
- with linguistic preferences, *Resour., Conserv. Recycl.* 74 (2013) 170–179, <https://doi.org/10.1016/j.resconrec.2012.09.006>.
- [8] S. Seuring, M. Müller, From a literature review to a conceptual framework for sustainable supply chain management, *J. Clean. Prod.* 16 (2008) 1699–1710, <https://doi.org/10.1016/j.jclepro.2008.04.020>.
 - [9] A.A. Hervani, M.M. Helms, J. Sarkis, Performance measurement for green supply chain management, *Benchmarking* 12 (2005) 330–353, <https://doi.org/10.1108/14635770510609015>.
 - [10] Mohd.N. Faisal, Analysing the barriers to corporate social responsibility in supply chains: an interpretive structural modelling approach, *Int. J. Logist.* 13 (2010) 179–195, <https://doi.org/10.1080/13675560903264968>.
 - [11] M.Y. Tay, A.A. Rahman, Y.A. Aziz, S. Sidek, A Review on Drivers and Barriers towards Sustainable Supply Chain Practices, *Int. J. Soc. Sci. Humanity* 5 (2015) 892–897, <https://doi.org/10.7763/ijssh.2015.v5.575>.
 - [12] R.J. Orsato, Competitive Environmental Strategies: when Does it Pay to Be Green? *Calif. Manage. Rev.* 48 (2006) 127–143, <https://doi.org/10.2307/41166341>.
 - [13] A. Sajjad, G. Eweje, D. Tappin, Managerial perspectives on drivers for and barriers to sustainable supply chain management implementation: evidence from New Zealand, *Bus. Strategy. Environ.* 29 (2019) 592–604, <https://doi.org/10.1002/bse.2389>.
 - [14] V. Midha and A. Mukhopadhyay, Recent trends in traditional and technical textiles, 2021. <https://doi.org/10.1007/978-981-15-9995-8>.
 - [15] A.M. Kitsis, L.J. Chen, Do stakeholder pressures influence green supply chain Practices? Exploring the mediating role of top management commitment, *J. Clean. Prod.* 316 (2021) 128258, <https://doi.org/10.1016/j.jclepro.2021.128258>.
 - [16] A. Lisa Allison, E. Ambrose-Dempster, T. Domenech Aparisi, M. Bawn, M. Casas, The environmental dangers of employing single-use face masks as part of a COVID-19 exit strategy, *UCL Press J.* (2020), <https://doi.org/10.14324/111.444/000031.v1>.
 - [17] A. Paulraj, L.J. Chen, C. Blome, Motives and performance Outcomes of Sustainable Supply chain Management Practices: a Multi-theoretical perspective, *J. Bus. Ethics* 145 (2015) 239–258, <https://doi.org/10.1007/s10551-015-2857-0>.
 - [18] D. Delmonico, C.J.C. Jabbour, S.C.F. Pereira, A.B.L. De Sousa Jabbour, D.W. S. Renwick, A.M.T. Thomé, Unveiling barriers to sustainable public procurement in emerging economies: evidence from a leading sustainable supply chain initiative in Latin America, *Resour., Conserv. Recycl.* 134 (2018) 70–79, <https://doi.org/10.1016/j.resconrec.2018.02.033>.
 - [19] M. Movahedipour, J. Zeng, M. Yang, X. Wu, An ISM approach for the barrier analysis in sustainable supply chain management, *Manage. Decis.* 55 (2017) 1824–1850, <https://doi.org/10.1108/md-12-2016-0898>.
 - [20] A. Majumdar, S. Sinha, Modeling the barriers of green supply chain management in small and medium enterprises, *Manage. Environ. Qual.* 29 (2018) 1110–1122, <https://doi.org/10.1108/meq-12-2017-0176>.
 - [21] K. Govindan, M. Kaliyan, D. Kannan, A.N. Haq, Barriers analysis for green supply chain management implementation in Indian industries using analytic hierarchy process, *Int. J. Prod. Econ.* 147 (2014) 555–568, <https://doi.org/10.1016/j.ijpe.2013.08.018>.
 - [22] G. Soni, S. Prakash, H. Kumar, S.P. Singh, V. Jain, S.S. Dhami, An interpretive structural modeling of drivers and barriers of sustainable supply chain management, *Manage. Environ. Qual.* 31 (2020) 1071–1090, <https://doi.org/10.1108/meq-09-2019-0202>.
 - [23] R.-J. Lin, R.-H. Chen, T.-H. Nguyen, Green supply chain management performance in automobile manufacturing industry under uncertainty, *Procedia: Soc. Behav. Sci.* 25 (2011) 233–245, <https://doi.org/10.1016/j.sbspro.2011.10.544>.
 - [24] H.T.S. Caldera, C. Desha, L. Dawes, Evaluating the enablers and barriers for successful implementation of sustainable business practice in 'lean' SMEs, *J. Clean. Prod.* 218 (2019) 575–590, <https://doi.org/10.1016/j.jclepro.2019.01.239>.
 - [25] A. Sajjad, G. Eweje, D. Tappin, Sustainable supply chain management: motivators and barriers, *Bus. Strategy. Environ.* 24 (2015) 643–655, <https://doi.org/10.1002/bse.1898>.
 - [26] A.A. Teixeira, C.J.C. Jabbour, A.B.L. De Sousa Jabbour, H. Latan, J.H.C. De Oliveira, Green training and green supply chain management: evidence from Brazilian firms, *J. Clean. Prod.* 116 (2016) 170–176, <https://doi.org/10.1016/j.jclepro.2015.12.061>.
 - [27] A.K.M.A.H. Asif, An overview of sustainability on apparel manufacturing industry in Bangladesh, *Sci. J. Energy Eng.* 5 (1) (2017), <https://doi.org/10.11648/j.sjee.20170501.11>.
 - [28] S.A. Zaabi, N.A. Dhaheri, A. Diabat, Analysis of interaction between the barriers for the implementation of sustainable supply chain management, *Int. J. Adv. Manuf. Technol.* 68 (2013) 895–905, <https://doi.org/10.1007/s00170-013-4951-8>.
 - [29] N. Bhanot, P.V. Rao, S.G. Deshmukh, Enablers and barriers of sustainable manufacturing: results from a survey of researchers and industry professionals, *Procedia CIRP* 29 (2015) 562–567, <https://doi.org/10.1016/j.procir.2015.01.036>.
 - [30] A. Esfahbodi, Y. Zhang, G. Watson, Sustainable supply chain management in emerging economies: trade-offs between environmental and cost performance, *Int. J. Prod. Econ.* 181 (2016) 350–366, <https://doi.org/10.1016/j.ijpe.2016.02.013>.
 - [31] E.R.G. Pedersen, K.R. Andersen, Sustainability innovators and anchor draggers: a global expert study on sustainable fashion, *J. Fashion Market. Manage.* 19 (2015) 315–327, <https://doi.org/10.1108/jfmm-08-2014-0059>.
 - [32] J. Sarkis, M.M. Helms, A.A. Hervani, Reverse logistics and social sustainability, *Corporate Soc.-Responsib. Environ. Manage.* 17 (2010) 337–354, <https://doi.org/10.1002/csr.220>.
 - [33] K.K. Muduli, S. Luthra, S.K. Mangla, C.J.C. Jabbour, S. Aich, J.C.F. De Guimarães, Environmental management and the "soft side" of organisations: discovering the most relevant behavioural factors in green supply chains, *Bus. Strategy. Environ.* 29 (2020) 1647–1665, <https://doi.org/10.1002/bse.2459>.
 - [34] S.A.R. Khan, Z. Yu, H. Golpira, A. Sharif, A. Mardani, A state-of-the-art review and meta-analysis on sustainable supply chain management: future research directions, *J. Clean. Prod.* 278 (2021) 123357, <https://doi.org/10.1016/j.jclepro.2020.123357>.
 - [35] R. Stewart, N. Bey, C. Boks, Exploration of the barriers to implementing different types of sustainability approaches, *Procedia CIRP* 48 (2016) 22–27, <https://doi.org/10.1016/j.procir.2016.04.063>.
 - [36] S.K. Sikdar, Sustainable development and sustainability metrics, *AIChE Journal* 49 (2003) 1928–1932, <https://doi.org/10.1002/aic.690490802>.
 - [37] A. Yıldızbaşı, C. Öztürk, D. Efendioğlu, S. Bulkan, Assessing the social sustainable supply chain indicators using an integrated fuzzy multi-criteria decision-making method: a case study of Turkey, *Environ. Dev. Sustain.* 23 (2020) 4285–4320, <https://doi.org/10.1007/s10668-020-00774-2>.
 - [38] R.L.H. Chiu, Social equity in housing in the Hong Kong Special Administrative Region: a social sustainability perspective, *Sustain. Dev.* 10 (2002) 155–162, <https://doi.org/10.1002/sd.186>.
 - [39] J.M. Harris, Sustainability and sustainable development, *Int. Soc. Ecol. Econ.* 1 (2003) 1–12.
 - [40] T. Al Khidir, S. Zailani, Going green in supply chain towards environmental sustainability, *Global J. Environ. Res.* (2009).
 - [41] M.N. Faisal, Sustainable supply chains: a study of interaction among the enablers, *Bus. Process Manage. J.* 16 (2010) 508–529, <https://doi.org/10.1108/14637151011049476>.
 - [42] P. Bohdanowicz, P. Zientara, E. Novotna, International hotel chains and environmental protection: an analysis of Hilton's 'swe care' programme (Europe, 2006–2008), *J. Sustain. Tour.* 19 (2011) 797–816, <https://doi.org/10.1080/09669582.2010.549566>.
 - [43] L. Preuss, Addressing sustainable development through public procurement: the case of local government, *Supply Chain Manage.* 14 (2009) 213–223, <https://doi.org/10.1108/13598540910954557>.
 - [44] J. Zhan, L. Wang, W. Gao, H. Li, C. Wang, Y. Huang, et al., Evaluatolgy: The science and engineering of evaluation, *BenchCouncil Transactions on Benchmarks Standards and Evaluations* (2024) 100162, <https://doi.org/10.1016/j.tbench.2024.100162> [Internet]Mar 1 Available from.
 - [45] Fisher R. The design of experiments [Internet]. 1935. Available from: <http://ci.niil.ac.jp/ncid/BA23310168>.
 - [46] F.E. Bowen, P.D. Cousins, R.C. Lamming, A.C. Farukt, The role of supply management capabilities in green supply, *Prod. Oper. Manage.* 10 (2001) 174–189, <https://doi.org/10.1111/j.1937-5956.2001.tb00077.x>.
 - [47] H. Chen, S. Liu, X. Wanyan, L. Pang, Y. Dang, K. Zhu, X. Yu, Influencing factors of novice pilot SA based on DEMATEL-AISM method: from pilots' view, *Heliyon* 9 (2023) e13425, <https://doi.org/10.1016/j.heliyon.2023.e13425>.
 - [48] A. Alsugair, K. Al-Gahtani, N. Alsanabani, G.M. Hommadi, M.A. Alosan, An integrated DEMATEL and system dynamic model for project cost prediction, *Heliyon*. (2024) e26166, <https://doi.org/10.1016/j.heliyon.2024.e26166>.
 - [49] M. Nilashi, S. Samad, A.A. Manaf, H. Ahmadi, T.A. Rashid, A. Munshi, W. Almukadi, O. Ibrahim, O.H. Ahmed, Factors influencing medical tourism adoption in Malaysia: a DEMATEL-Fuzzy TOPSIS approach, *Comput. Ind. Eng.* 137 (2019) 106005, <https://doi.org/10.1016/j.cie.2019.106005>.
 - [50] J. Wu, H. Wang, L. Yao, Z. Kang, Q. Zhang, Comprehensive evaluation of voltage stability based on EW-AHP and Fuzzy-TOPSIS, *Heliyon* 5 (2019) e02410, <https://doi.org/10.1016/j.heliyon.2019.e02410>.
 - [51] A.Khan Fahmi, N.T. Abdeljawad, M.A. Alqudah, Natural gas based on combined fuzzy TOPSIS technique and entropy, *Heliyon* 10 (2024) e23391, <https://doi.org/10.1016/j.heliyon.2023.e23391>.
 - [52] Q. Hung DO, V.T. Tran, T. Tran, Evaluating lecturer performance in Vietnam: an application of Fuzzy AHP and Fuzzy TOPSIS methods, *Heliyon*. (2024) e30772, <https://doi.org/10.1016/j.heliyon.2024.e30772>.
 - [53] Z.A. Eldukair, B.M. Ayyub, Multi-attribute fuzzy decisions in construction strategies, *Fuzzy. Sets. Syst.* 46 (1992) 155–165, [https://doi.org/10.1016/0165-0114\(92\)90128-q](https://doi.org/10.1016/0165-0114(92)90128-q).
 - [54] M. Yazdi, F. Khan, R. Abbassi, R. Rusli, Improved DEMATEL methodology for effective safety management decision-making, *Saf. Sci.* 127 (2020) 104705, <https://doi.org/10.1016/j.ssci.2020.104705>.
 - [55] Z.A. Eldukair, B.M. Ayyub, Multi-attribute fuzzy decisions in construction strategies, *Fuzzy. Sets. Syst.* 46 (1992) 155–165, [https://doi.org/10.1016/0165-0114\(92\)90128-q](https://doi.org/10.1016/0165-0114(92)90128-q).
 - [56] S. Feng, L.D. Xu, Decision support for fuzzy comprehensive evaluation of urban development, *Fuzzy. Sets. Syst.* 105 (1999) 1–12, [https://doi.org/10.1016/s0165-0114\(97\)00229-7](https://doi.org/10.1016/s0165-0114(97)00229-7).
 - [57] C.-T. Chen, Extensions of the TOPSIS for group decision-making under fuzzy environment, *Fuzzy. Sets. Syst.* 114 (2000) 1–9, [https://doi.org/10.1016/s0165-0114\(97\)00377-1](https://doi.org/10.1016/s0165-0114(97)00377-1).
 - [58] R. Raut, B.B. Gardas, B. Narkhede, Ranking the barriers of sustainable textile and apparel supply chains, *Benchmarking* 26 (2019) 371–394, <https://doi.org/10.1108/bij-12-2017-0340>.
 - [59] M.R. Rashid, S.K. Ghosh, Md.F.B. Alam, M.F. Rahman, A fuzzy multi-criteria model with pareto analysis for prioritizing sustainable supply chain barriers in the textile industry: evidence from an emerging economy, *Sustain. Oper. Comput.* 5 (2024) 29–40, <https://doi.org/10.1016/j.susoc.2023.11.002>.
 - [60] D. Janguo, Y.A. Solangi, Sustainability in Pakistan's textile industry: analyzing barriers and strategies for green supply chain management implementation, *Environ. Sci. Pollut. Res. Int.* 30 (2023) 58109–58127, <https://doi.org/10.1007/s11356-023-26687-x>.

Md. Hasibul Hasan Hemal received B.Sc. Engineering degree in Industrial and Production Engineering from Khulna University of Engineering & Technology. His research interests include systems Engineering, Data Analytics in Industrial Engineering, Simulation Modeling, Logistics Optimization, Inventory Management, Decision Analysis, and Supply Chain Management.

Farjana Parvin received B.Sc. and M.Sc. Engineering degree in Industrial and Production Engineering from Khulna University of Engineering & Technology. Her research interests include Human Factors Engineering, Product Design and Development, Project Management, Industrial Management, Supply Chain Management and Operations Management.

She is an assistant professor at the Department of Industrial Engineering and Management at Khulna University of Engineering & Technology, Khulna, Bangladesh.

Alberuni Aziz received B.Sc. Engineering degree in Textile Engineering and M.Sc. Engineering degree in Industrial Engineering and Management from Khulna University of Engineering & Technology. Currently he is pursuing PhD at Khulna University of Engineering & Technology. His research interests include Garments Production Optimization, Textile Composite Materials, Industry 4.0, Supply Chain Management and TPM. He is an Assistant Professor of Department of Textile Engineering at Khulna University of Engineering & Technology, Khulna, Bangladesh.



Full Length Article

Exploring the Orca Predation Algorithm for Economic Dispatch Optimization in Power Systems

Vivi Aida Fitria^{a,b}, Arif Nur Afandi^{a,*}, Aripriharta^a 

^a Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Indonesia

^b Informatics, Institut Teknologi dan Bisnis Asia Malang, Indonesia

ARTICLE INFO

Keywords:

Economic Dispatch
Orca Predation Algorithm
Metaheuristic algorithm
Bio-Inspired Algorithm
Optimization
Power System

ABSTRACT

The Economic Dispatch problem is essential for minimizing generation costs while satisfying power demand in electrical systems. This research looks into the Orca Predation Algorithm, an optimization method based on biology that can solve the Economic Dispatch problem for systems with 6, 13, or 15 producing units. The idea behind Orca Predation Algorithm came from the way orcas hunt for food. It solves problems that other optimization methods and bio-inspired algorithms have, like too much population diversity and too early convergence. This research shows that Orca Predation Algorithm consistently does better than other bio-inspired algorithms like Particle Swarm Optimization, Whale Optimization Algorithm, Grey Wolf Optimizer, the Bat Algorithm, Genetic Algorithm and Ant Colony Optimization in terms of minimum cost, average cost, and solution stability. The sensitivity analysis of the parameters regulating the exploration-exploitation balance in Orca Predation Algorithm demonstrated substantial performance enhancements. By changing these parameters, the best prices came in at \$15,275.93 for the 6-unit system, \$17,932.49 for the 13-unit system, and \$32,256.97 for the 15-unit system. These prices are lower than those in the previous parameter setting. Although Orca Predation Algorithm demonstrates greater performance, it necessitates extended computing time, which future research could mitigate by exploring parallelization or hybrid methodologies. This paper shows that Orca Predation Algorithm is a reliable tool for optimizing Economic Dispatch problems. It gives useful information to power system engineers who are looking for effective and scalable optimization methods for modern power systems.

1. Introduction

The economic dispatch (ED) problem is a critical issue in power system operation, with the primary aim of optimally distributing power among generating units to minimize overall generation costs [1]. This is accomplished while satisfying the load demand, with each generating unit subject to minimum and maximum power limits that must be observed [2,3]. The ED problem, being a convex optimization issue, necessitates a rapid and efficient resolution, particularly in extensive systems comprising numerous generating units [4,5]. Conventional optimization methods, including linear programming and quadratic programming, frequently fail to address the increasing complexity of power systems. This complexity originates from causes including system expansion, the incorporation of variable renewable energy sources, and the existence of numerous local optima in extensive power systems. Moreover, traditional methods encounter challenges in delivering effective global optimization, especially in systems necessitating swift

flexibility due to variable energy demands and resources [6]. Bio-inspired algorithms have lately garnered interest as viable solutions for addressing complicated optimization issues, including the ED problem [7]. These algorithms are especially beneficial for navigating extensive search areas and circumventing local optima, obstacles frequently encountered by conventional approaches. Researchers have thoroughly investigated a variety of bio-inspired algorithms for economic dispatch optimization, including Particle Swarm Optimization (PSO) [8,9,10], Bat Algorithm [11], Whale Optimization Algorithm (WOA) [12], Grey Wolf Optimization (GWO) [13], Genetic Algorithm (GA) [14] and Ant Colony Optimization (ACO) [15]. Many of these algorithms frequently encounter premature convergence as a result of insufficient population variety, which in turn limits their ability to fully explore the entire solution space.

Yuxin Jiang developed the Orca Predation Algorithm (OPA) in 2022, a unique bio-inspired optimization technique that demonstrates significant potential in addressing challenges of previous algorithms [16]. The

* Corresponding Author: Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Jl. Semarang 5 Malang 65145 Jawa Timur Indonesia
E-mail address: an.afandi@um.ac.id (A.N. Afandi).

<https://doi.org/10.1016/j.tbench.2024.100187>

Received 28 May 2024; Received in revised form 27 December 2024; Accepted 30 December 2024

Available online 3 January 2025

2772-4859/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

predatory tactics of orcas inspire OPA, which strikes a balance between exploration and exploitation to maintain population diversity and prevent early convergence. However, the application of OPA to the ED problems has not been studied, and this research aims to evaluate its performance in comparison to other bio-inspired algorithms such as PSO, Bat Algorithm, WOA, GWO, GA and ACO. The originality of this research lies in adapting OPA to the specific restrictions of the ED problem, which emphasizes the optimal distribution of generating units while minimizing costs. OPA is implemented for its capacity to address intricate optimization challenges and derive efficient solutions via adaptive pursuit and assault phases. This study entails the meticulous adjustment of the parameters (p_1 and p_2) of OPA to enhance its efficacy in addressing ED problems. Parameter p_1 governs the individual's subsequent phase, either driving or encircling, while p_2 adjusts the strength of the attack. A sensitivity study of these parameters is performed to assess their influence on solution quality and computational efficiency, ensuring that the method can accommodate many circumstances with little modification.

This work offers a thorough comparison of OPA with other prevalent biology-inspired optimization methods, emphasizing its benefits regarding solution stability and convergence speed. By comparing OPA to other methods already used, this study aims to show how reliable and effective it is as a strong optimization tool for solving economic dispatch problems in power systems. This work seeks to provide significant insights into the relevance and efficacy of OPA in treating ED issues while tackling critical concerns such as preserving solution diversity and reducing computational expenses. This study's results will offer practical direction for power system engineers in applying OPA to real-world ED scenarios, hence enhancing the development of efficient optimization solutions for contemporary power systems.

2. Related Work

2.1. Particle Swarm Optimization

Particle Swarm Optimization (PSO) is a global optimization technique developed by Kennedy and Eberhart in 1995 [17]. PSO was inspired by the behavior of avian flocks and schools of fish [18]. Each particle PSO possesses a variable velocity that navigates the search space based on its prior performance [19]. The particles have a propensity to migrate towards more advantageous search regions throughout the search process. Throughout the search process [20]. PSO has been employed in numerous intricate optimization challenges, including the resolution of economic dispatch issues. The findings demonstrate that PSO is proficient in circumventing local minima and achieving convergence to the global optimum, a crucial aspect for economic dispatch problems characterized by intricate and non-linear objective functions alongside various restrictions [21]. Nonetheless, PSO may occasionally encounter premature convergence or sluggish convergence rates, particularly if the parameters are inadequately calibrated. This may impact the precision and dependability of the derived solutions [22]. To enhance the efficacy of PSO in economic dispatch problems, certain research integrate it with alternative optimization methodologies, such as Ant Colony Optimization (ACO), to augment its performance and convergence rate. This hybrid methodology can yield superior answers, yet introduces complication to its execution [23].

2.2. Bat Algorithm

The Bat algorithm is a bio-inspired metaheuristic algorithm introduced by Xin-She Yang in 2010 [24]. Bats' foraging behavior, which involves locating prey and evading obstacles, serves as the basis for the Bat algorithm [25]. The fundamental principle underlying the development of the Bat algorithm is echolocation, which enables bats to identify food sources and obstacles while estimating their distances [26]. The second hypothesis posits that bats can modify their flying

patterns in response to experience and environmental feedback, facilitating the exploration of the solution space [27]. The third aspect pertains to the parameters utilized in the Bat algorithm, specifically the loudness and emission rate of the bat's pulse, which serve to equilibrate exploration and exploitation throughout the search process [28]. Prior studies have employed the Bat algorithm to address economic dispatch issues. Included are works that present the application of Ant Lion Optimization (ALOA) and Bat Algorithm (BA) to address economic dispatch issues in power systems. This study evaluated the efficacy of both algorithms on a small-scale three-generator system and a large-scale six-generator system, utilizing the IEEE-30 bus reliability test system. Nonetheless, the results indicate that ALOA exhibits a superior convergence rate compared to the Bat Algorithm (BA) [29].

2.3. Whale Optimization Algorithm

The Whale Optimization Algorithm (WOA) draws its optimization technique from the hunting strategies of whales in the ocean [30]. Specifically, it mimics the behavior of whales when they locate their prey, creating a bubble trap that constricts the prey's movement [31]. Upon ensnaring its prey, the whale will promptly consume it. This algorithm provides an effective method for identifying optimal solutions within intricate search spaces by integrating exploration and exploitation phases [32]. The WOA algorithm uses three main methods to find the best solution: first, it copies the way whales circle their prey, moving their positions based on the best solution found; second, it copies the way whales search for food, moving around randomly and exploring the solution space; and third, the bubble net attack mechanism copies the way whales hunt together, working together to completely surround and catch their prey [33]. Prior studies have demonstrated the efficacy of the WOA algorithm in addressing economic dispatch issues. The study demonstrated that WOA can yield optimal or near-optimal solutions by taking into account fuel expenses and pollutants. Superior convergence characteristics set it apart from traditional methods such as PSO [29]. However, if not adequately designed, WOA, akin to other meta-heuristics, has sluggish convergence, limited precision, and a propensity to become trapped in local optima [34].

2.4. Grey Wolf Optimization

The predatory behavior of wolves in their natural habitat serves as the basis for the Grey Wolf Optimizer (GWO) algorithm. The gray wolf is regarded as an apex predator, characterized by a pronounced social dominance structure. The foremost leaders are designated as alpha, the second tier as beta, the third tier as delta, and the final tier as omega [35]. The GWO algorithm is characterized by three primary phases: first, the wolves encircle the prey according to the positions of the alpha, beta, and delta wolves; second, the wolves adjust their positions by adhering to the alpha, beta, and delta wolves, emulating the actions of tracking, pursuing, and nearing the prey; and third, the wolves launch an attack [36]. The GWO algorithm does a much better job of solving the economic dispatch problem than other meta-heuristic approaches like Biogeography-based Optimization (BBO), Lambda Iteration (LI), Hopfield model approaches (HM), Cuckoo Search (CS), Firefly algorithms, Artificial Bee Colony (ABC), Neural Networks trained by Artificial Bee Colony (ABCNN), Quadratic Programming (QP), and General Algebraic Modeling System (GAMS) [37]. Despite GWO demonstrating superior efficiency relative to other metaheuristic algorithms, it remains computationally demanding for extensive systems or when confronted with intricate limitations [38].

2.5. Genetic Algorithm

The Genetic Algorithm (GA) is an evolutionary computer technique developed by John Holland in the 1970s. Genetic algorithms use selection, crossover, and mutation, along with other ideas from natural

selection and genetics, to come up with new solutions and make the ones we already have better. In a genetic algorithm, each solution, referred to as a 'chromosome,' undergoes iterative evolution to enhance its quality according to a designated fitness function [39]. GA have been extensively utilized for diverse complex optimization challenges, including ED issues, owing to their capacity to manage non-linear and non-convex objective functions [14]. Previous research has shown that genetic algorithms are very good at solving economic dispatch problems by finding the best balance between exploring and exploiting in the search space. Research has shown that genetic algorithms can get optimal or near-optimal solutions for extensive power systems, encompassing fuel cost reduction and emission limitations [40]. Nonetheless, genetic algorithms face issues including premature convergence and sensitivity to parameter configurations, such as population size and mutation rate, which can influence their efficacy and dependability. Recent advancements aim to amalgamate GA with other optimization methodologies, such as PSO, to enhance convergence velocity and solution efficacy [41].

2.6. Ant Colony Optimization

Marco Dorigo developed Ant Colony Optimization (ACO), a nature-inspired metaheuristic, in the early 1990s, based on the foraging behavior of ants. In ACO, artificial ants emulate the foraging and pheromone trail-following behaviors of actual ants, enabling them to identify the optimal route between a food source and their colony. Each ant looks at the pheromone trail and heuristic data to figure out what the best solution might be. The pheromone trail is changed over and over to reinforce the best options [42]. Numerous optimization challenges, including economic dispatch issues, have effectively utilized ACO due to its capacity to adjust to dynamic and intricate search environments. Research indicates that ACO excels in emergency department circumstances by effectively balancing exploration and exploitation, particularly in small to medium-sized power systems [15]. Efforts to integrate ACO with other algorithms, including GA and WOA, have demonstrated potential enhancements in efficiency and accuracy. Nonetheless, these hybrid methodologies frequently elevate computational complexity, rendering ACO less advantageous for extensive ED challenges that necessitate high precision and rapid convergence [6].

3. Research Method

3.1. Economic Dispatch Problem

The economic dispatch problem is a crucial concern in power systems, to optimize the distribution of power generation among different power plants to satisfy the demand at the most cost-effective rate [43]. The aim is to reduce the overall generation cost while complying with the operational limits of each facility. The generating cost function is typically expressed as a second-order polynomial of the power output (1) :

$$F_T = \sum_{i=1}^n F(P_i) = \sum_{i=1}^n (a_i P_i^2 + b_i P_i + c_i) \quad (1)$$

where F_T denotes the total generation cost in \$/hour, $F(P_i)$ signifies the generation cost of the i -th plant, P_i represents the power output of the i -th plant in MW, and a_i , b_i , c_i are cost coefficients derived from the operational characteristics and fuel type of the plant. The generator limits characterize the inequality conditions in the ED problem formularization.

$$P_{i,min} \leq P_i \leq P_{i,max} \text{ for } i = 1, 2, \dots, n$$

The ideal power flow in the power system is impacted by the transmission line losses. These losses can be expressed quantitatively as (2)

Table 1
Dataset of 6 unit system

Variable	Range	Unit
$P_{i,min}$	50-100	MW
$P_{i,max}$	120-500	MW
a	0.007-0.0095	-
b	7-12	-
c	190-240	-

Table 2
Dataset of 13 unit system

Variable	Range	Unit
$P_{i,min}$	0-60	MW
$P_{i,max}$	120-680	MW
a	0.00028 -0.00324	-
b	7.74-8.6	-
c	126-550	-

Table 3
Dataset of 15 unit system

Variable	Range	Unit
$P_{i,min}$	20-150	MW
$P_{i,max}$	55-470	MW
a	0.000183-0.00513	-
b	8.8-13.1	-
c	173-671	-

$$P_{Loss} = \sum_{i=1}^n \sum_{j=1}^n P_i B_{ij} P_j + \sum_{i=1}^n B_{0i} P_i + B_{00} \quad (2)$$

In this context, P_{Loss} denotes the total transmission loss in megawatts (MW), while B_{ij} , B_{0i} , and B_{00} are coefficients contingent upon the system setup and network topology. The B coefficients need to be established for a changeable system demand. The prerequisites for electrical equality in ED are shown in (3). [44]

$$P_D = \sum_{i=1}^n P_i - P_{Loss} \quad (3)$$

where P_D is the total system demand measured in megawatts (MW). This study used a data set comprising three test scenarios: 6 units system with a total load demand of 1263 MW, 13 units system with a total load demand of 1800 MW, and 15 units system with a total load demand of 2630 MW. The Tables 1, 2 and 3 displays the data sets used in this study from Hardiansyah's research as well as those from Zakian and Keveh. [45,46]:

3.2. Orca Predation Algorithm

The Orca Predation Algorithm (OPA) was initially proposed by [16]. The OPA is a bio-inspired metaheuristic optimization method that emulates the hunting behavior of orcas. The purpose is to imitate the hunting strategies of orcas, known for their advanced and highly coordinated hunting techniques, to address complex optimization problems. The algorithm consists of two distinct stages: driving and encircling. In the first step of OPA, important parameters are set, such as the number of populations (N), dimensions (D), maximum iterations, selection probability p_1 , p_2 , lower bound (lb), and upper bound (ub) of the design variables. The orca group's initial position is randomly created within the specified limits [lb, ub]. The objective function determines the second phase of fitness value assessment for each orca. The orca exhibiting the highest fitness value is designated as x_{best} , representing the optimal current solution. The final step involves updating the position

Table 4
Parameters Setting.

Algorithm Name	Parameters	Max_iter
OPA	$a, b, d \in [0, 1]; e \in [0, 2]; F = 2; q = 0.9; p_1 = 0.4; p_2 = 0.05; g_1 \in [0, 2]; g_2 \in [-2.5, 2.5]$ [16]	40
PSO	$c_1 = 2; c_2 = 2; \omega t = 0.9; \omega f = 0.2$ [47]	40
Bat-Algorithm	$f_{min} = 0; f_{min} = 2; A_0 = [0, 2]; r_0 = [0, 1]; \alpha = 0.9; \gamma = 0.9$ [48]	40
WOA	$a \in [2, 0]; r \in [0, 1]; b = 1$ [49]	40
GWO	$A \in [2, 0]$ [47]	40
GA	Crossover rate = 0.8; Mutation rate = 0.1 [47]	40
ACO	$\alpha = 1; \beta = 2; \text{Evaporation rate} = 0.2$ [50]	40

throughout the pursuit phase. At this juncture, the orca chooses between "driving" or "encircling" its prey based on the selection probability (p_1). This decision dictates whether the orca will prioritize exploration (seeking novel solutions) or exploitation (refining the existing optimal solution). Orcas utilize sonar to modify their location by recalibrating their position in accordance with Equations :

$$v_{chase\ 1,i}^t = a (dx_{best}^t - F(bM^t + cx_i^t)) \quad (4)$$

$$v_{chase\ 2,i}^t = ex_{best}^t - x_i^t \quad (5)$$

$$M = \frac{\sum_{i=1}^N x_i^t}{N} \quad (6)$$

$$c = 1 - b \quad (7)$$

$$\begin{cases} x_{chase\ 1,i}^t = x_i^t + v_{chase\ 1,i}^t & \text{if } rand > q \\ x_{chase\ 2,i}^t = x_i^t + v_{chase\ 2,i}^t & \text{if } rand \leq q \end{cases} \quad (8)$$

where t is the number of cycles, v^t is the chasing speed, M is the average position of the orca group, x^t is the position of the orca, a, b and d are random numbers between $[0, 1]$, e is a random number between $[0, 2]$, $F = 2$ and q is a number between $[0, 1]$. While the equation when the orca encircles the prey is:

$$x_{chase\ 3,i}^t = x_{j_1,i}^t + u(x_{j_2,i}^t - x_{j_3,i}^t) \quad (9)$$

$$u = 2(rand - 0.5) \frac{maxiter - t}{maxiter} \quad (10)$$

where $max\ iter$ denotes the maximum number of iterations, and j_1, j_2, j_3 represent three distinct orcas selected at random such that $j_1 \neq j_2 \neq j_3$. The position of the i -th whale after selecting the third chase method at time t is denoted as $x_{chase\ 3,i}^t$.

In this procedure, orcas ascertain the location of their prey and modify their position to enhance the efficacy of the solution. The fourth phase involves updating the position during the attack phase. During this phase, orcas enhance their positioning to effectively assault prey. Equations (11)-(13) guide the revision of the position. During the assault, certain orcas may attain the boundaries of the search zone. If they surpass the limit, their position will revert to the lower limit (lb). [16]

$$v_{attack\ 1,i}^t = \frac{(x_{first}^t + x_{second}^t + x_{third}^t + x_{four}^t)}{4 - x_{chase,j}^t} \quad (11)$$

$$v_{attack\ 2,i}^t = \frac{(x_{chase,j_1}^t + x_{chase,j_2}^t + x_{chase,j_3}^t)}{3 - x_i^t} \quad (12)$$

$$x_{attack,i}^t = x_{chase,i}^t + g_1 v_{attack\ 1,i}^t + g_2 v_{attack\ 2,i}^t \quad (13)$$

where $v_{attack\ 1,i}^t$ represents the speed vector of the i -th orca during the hunting phase at time t , $v_{attack\ 2,i}^t$ indicates the speed vector of the i -th orca returning to the cage at time t , $x_{first}^t, x_{second}^t, x_{third}^t, x_{four}^t$ denote the four orcas positioned optimally in succession, j_1, j_2, j_3 signify three

randomly selected orcas from a total of N in the pursuit phase, ensuring $j_1 \neq j_2 \neq j_3$. $x_{attack,i}^t$ indicates the position of the i -th orca at time t following the attack phase, g_1 is a random number within the range $[0, 2]$ and g_2 is a random number within the interval $[-2.5, 2.5]$. The fifth step involves the establishment of a new population. Following the assault phase, we establish a new pod of orcas. The orcas' placements are revised according to the outcomes of the pursuit and assault phases. This stage ensures the diversity of the population while maintaining the identified optimal solution. The final step is the loop's termination. The algorithm verifies if the maximum iteration count ($maxiter$) has been attained or if the optimal solution has been identified. If the termination condition remains unfulfilled, we repeat the procedure from Step 2.

3.3. Scenarios

This study uses the OPA to optimize the ED issue across three distinct power generation systems: 6, 13, and 15 generating units. The aim of each scenario is to reduce the overall generation cost while satisfying all power demand and operational restrictions. Each scenario assesses OPA's capacity to manage escalating system complexity and scale while benchmarking its performance against other bio-inspired optimization methods, including PSO, Bat Algorithm, WOA, GWO, GA and ACO.

3.2.1. Scenario Setup

Multiple constraints structure ED problem solving around test data: the first constraint mandates that each generating unit has a specified minimum and maximum generation capacity, the second constraint requires the aggregate power output of all units to align with system demand, and the objective is to minimize the total cost function, which is characterized as a quadratic function for each generator.

3.3.2. Parameter Tuning and Sensitivity Analysis

The parameter configurations for all algorithms included in this study were derived from prior research sources to guarantee appropriateness and pertinence. The subsequent Table 4 encapsulates the parameters.

The subsequent experiment aimed to optimize the settings of the OPA algorithm to enhance its performance. The grid search technique fine-tunes the essential parameters p_1 and p_2 of the OPA to optimize the unique characteristics of the ED issue using the OPA algorithm. These two criteria are crucial for balancing the exploration and exploitation stages. The parameter p_1 determines the chance of the orca transitioning into the driving or encircling phase, whereas p_2 governs the attack phase. These parameters control the balance between exploring and exploiting during the optimization process. Making sure they are set up correctly is very important to avoid premature convergence and make sure that the whole solution space is explored. Grid search is conducted by examining multiple combinations of p_1 and p_2 variables. This research grid search process has many key steps: initially, setting the parameter ranges p_1 and p_2 , where $p_1 \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and $p_2 \in \{0.01, 0.03, 0.05, 0.07, 0.1\}$, resulting in a 5×5 grid combination. Afterwards, the OPA algorithm is run 40 times in the ED scenario to test each parameter combination and find the fitness value (generation cost) for each setting. The performance is judged by writing down the lowest

Table 5

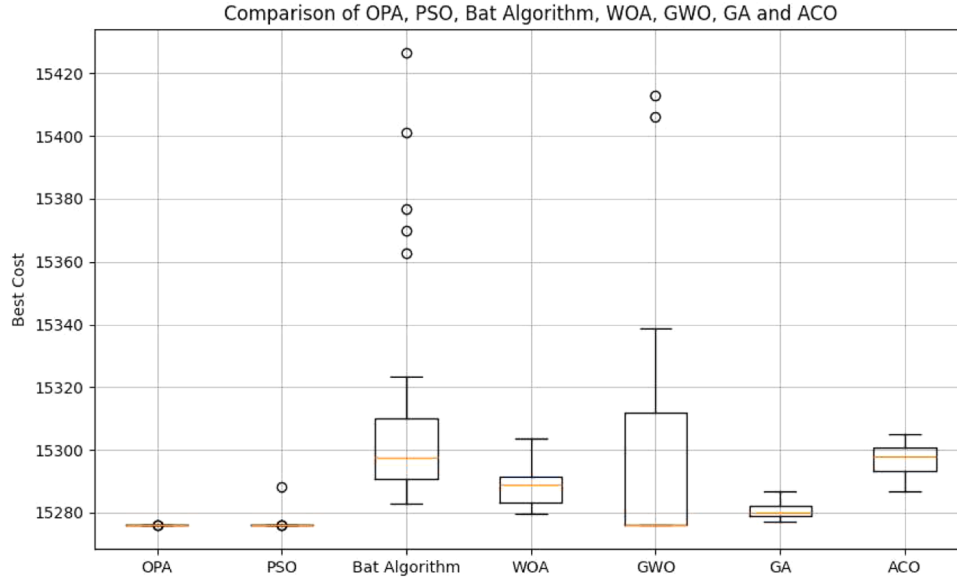
Actual output from the system's six generators

Algorithm	P1	P2	P3	P4	P5	P6
PSO	446.82995889	171.19809142	264.10692378	125.1393930	172.10194858	83.62368431
Bat Algorithm	443.68052541	188.20549581	269.86504536	108.6707606	159.0956169	93.48255591
WOA	105.2252307	172.17792636	278.86546458	130.10435371	173.48050005	83.44119756
GWO	446.56908253	169.21357222	264.21317023	125.86927672	172.14265957	84.99223872
GA	447.5046163	170.79042194	266.01133413	123.22171545	169.17913187	86.29278031
ACO	446.3378674	162.65821351	262.98062072	150.	162.65821351	78.36508487
OPA	446.64128102	171.26899505	264.14790971	125.18833083	172.11574993	83.63773346

Table 6

Economic Dispatch on 6 Generators

Algorithm	Minimum Cost	Mean	Std	Computation Time
PSO	15275.930594364689	15277.54602959566	3.352045776164392	0.05616219043731689
Bat Algorithm	15283.57585596129	15311.85874126833	37.97378990142223	0.11467342376708985
WOA	15276.354615304797	15284.83732050414	7.475160434950923	0.10255167484283448
GWO	15275.988847142546	15296.143454969084	31.57586518826314	0.195526856642503
GA	15276.129209510731	15281.089977499496	3.320602022291739	0.04007517496744792
ACO	15283.094249517922	15294.727679662261	5.124485757627291	0.3876126050949097
OPA	15275.930461640604	15275.933494696532	0.003482677781155	0.5794726530710856

**Fig. 1.** Box Plot Comparison on 6 Generators

cost that can be reached for each set of parameters. The test is run several times to lessen the effect of stochastic variations in the algorithm. The aim of the grid search is to determine the parameter configurations that yield the optimal balance between convergence rate and solution quality. As part of the tuning process, sensitivity analysis is used to see how changes in p_1 and p_2 affect the optimal cost, how the solution converges, and how stable it is in general. The sensitivity of parameters p_1 and p_2 is illustrated by contour plots, depicting the effect of parameter modification on the optimal cost attained. These contour plots show how stable the OPA algorithm is with different parameter settings. The best parameter setting is shown by the global minimum on the curve. Grid search carefully examines all parameter combinations, guaranteeing that every alternative is evaluated and no solution is overlooked. This research employs the grid search method because of its advantages in parallelization and implementation [51].

3.3.3. Evaluation Metrics

In all scenarios, the efficacy of OPA and other comparative algorithms is assessed using several critical metrics: the minimum total

generation cost attained by each algorithm, the mean generation cost across multiple tests to evaluate solution consistency, the standard deviation to gauge result variation and algorithm stability, and the average computation time. To compare OPA to current methods and determine how well it can adapt to growing system sizes, this study will use these criteria in all situations.

4. Result and Discussion

Experiments have been carried out to evaluate the performance of the OPA, PSO, Bat Algorithm, WOA and GWO on three well-known test systems: the 6-unit, 13-unit, and 15-unit systems.

4.1. 6-unit system

The system comprises six thermal units, 26 buses, and 46 transmission lines. The load demand is 1263 megawatts. Table 5 shows the actual power output of each unit based on the five optimization strategies used. This constitutes a crucial component of the optimal ED

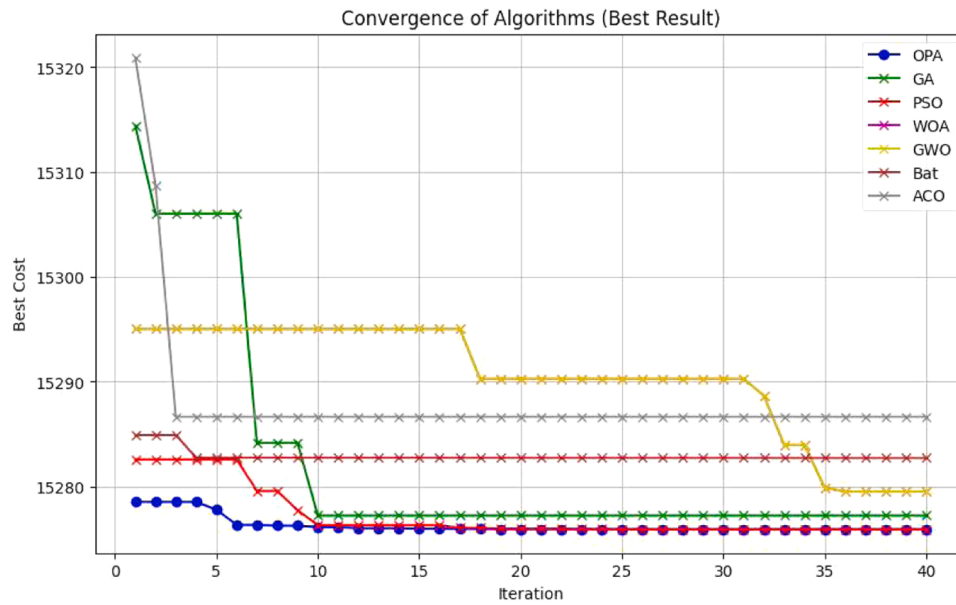


Fig. 2. Convergence curves on 6 Generators

calculation. Table 6 presents the minimum system cost, average cost, standard deviation, and computation time for each method following 30 trials. Fig. 1 depicts the minimum cost associated with each evaluated strategy across 40 iterations.

Table 6, which encompasses six power plants, indicates that OPA exhibits superior performance, with the lowest cost of 15275.9305 and the most stable average cost of 15275.9335 in comparison to GA, PSO, ACO, the Bat Algorithm, WOA, and GWO. Moreover, OPA's minimal standard deviation (0.0035) signifies that its optimization outcomes are highly consistent. This is markedly superior to other, more variable algorithms, such as the Bat Algorithm (37.9738) and GWO (31.5759). GA and PSO exhibited similar performance, with GA achieving a minimal cost of 15276.13 and PSO reaching a nearly identical minimum cost of 15275.9306, along with slightly higher standard deviations of 3.3206 and 3.3520, respectively. Although OPA's computation time is slightly increased at 0.579 seconds, its dependability in producing the ideal solution establishes it as the preeminent approach for the ED problem.

Fig. 1's boxplot demonstrates that OPA exhibits the most constrained cost distribution, lacking outliers and hence confirming its reliability. PSO and GA exhibit similarly tight distributions, albeit with slightly greater variability than OPA. Conversely, ACO exhibits a broader spectrum with a greater number of outliers, signifying less consistent performance. The Bat Algorithm and GWO display the broadest distribution with numerous outliers, indicating their instability, while WOA shows improved performance but remains less stable than OPA, PSO, and GA. The convergence curve in Fig. 2 shows that OPA is clearly better because it reaches stability and the lowest possible cost (about \$15,275) in just 10 rounds. PSO demonstrated rapid convergence, with a final outcome nearly identical to OPA. GA demonstrated rapid and consistent convergence, but with a somewhat elevated final cost. ACO necessitates

additional iterations to attain stability, resulting in a higher optimal cost compared to OPA, PSO, and GA. The Bat Algorithm and GWO exhibited negligible enhancement across iterations, with optimal costs persisting at elevated levels. Concurrently, WOA exhibited incremental enhancement but plateaued at a suboptimal cost of around 15284. This confirms that OPA is the leading algorithm, producing the most optimal solution and demonstrating efficiency and stability that significantly exceed those of alternative algorithms.

The experimental results for the OPA algorithm utilize the identical parameter values p_1 and p_2 as those in Yuxin's study. The forthcoming discussion will present the results of the sensitivity graph for parameters p_1 and p_2 following the adjustment of certain parameters.

The experimental results from the sensitivity graph of parameters p_1 and p_2 for 6 system units using the OPA algorithm show that the best cost changes a lot depending on the combinations of parameters p_1 and p_2 , especially when the values are low. In the lower left quadrant of the graph, particularly within the range of p_1 from 0.1 to 0.4 and p_2 from 0.01 to 0.08, a notable alteration in the optimal cost is observed. Conversely, when p_1 approaches 0.9, the fluctuation in the optimal cost diminishes, as evidenced by the more uniform coloration of the graph. This indicates that elevated levels of p_1 (approaching 1) and diminished p_2 yield stability in the outcomes, albeit with minimal variation in the optimal cost. Furthermore, experimental outcomes with the p_1 and p_2 parameter values from Yuxin's study yielded an optimal cost of 15275.930461640604. Further optimization yielded a superior cost of 15275.930398908155, using parameter values $p_1 = 0.9$ and $p_2 = 0.01$. The sensitivity graph indicates that parameter adjustment substantially influences the optimization outcomes. The combination of elevated p_1 (0.9) and diminished p_2 (0.01) demonstrated superior optimality compared to the parameter values employed in Yuxin's article. This

Table 7

Actual output from the system's 13 generators

Algorithm	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13
PSO	782	86	223	72	85	75	108	89	76	42	41	58	57
Bat Algorithm	510	288	175	72	93	70	97	103	165	60	40	57	64
WOA	658	228	225	101	114	108	99	97	92	40	39	55	56
GWO	499	257	253	100	101	101	101	99	101	40	40	55	55
GA	505	194	247	136	102	104	87	96	89	51	50	59	73
ACO	528	264	306	120	60	60	105	60	60	60	60	60	55
OPA	505	255	254	97	98	99	100	98	98	40	40	55	55

Table 8
Economic Dispatch on 13 Generators

Algorithm	Minimum Cost	Mean	Std	Computation Time
PSO	17981.852673781043	18010.25471739617	12.105198962784101	0.0828967014948527
Bat Algorithm	17972.28367998929	18006.380485156496	18.17393100098788	0.11570893923441569
WOA	17935.78738096655	17975.3459921142	15.979580254725192	0.06328445275624593
GWO	17932.597705266875	17932.841638157162	0.19262156142012746	0.1848301410675049
GA	17963.61906345404	17981.56961907532	9.25872245439762	0.05540952682495117
ACO	17978.631481258828	17998.38803106319	9.697338387501473	0.8677584489186605
OPA	17932.495398652336	17932.556635945206	0.058578778625625744	1.039674949645996

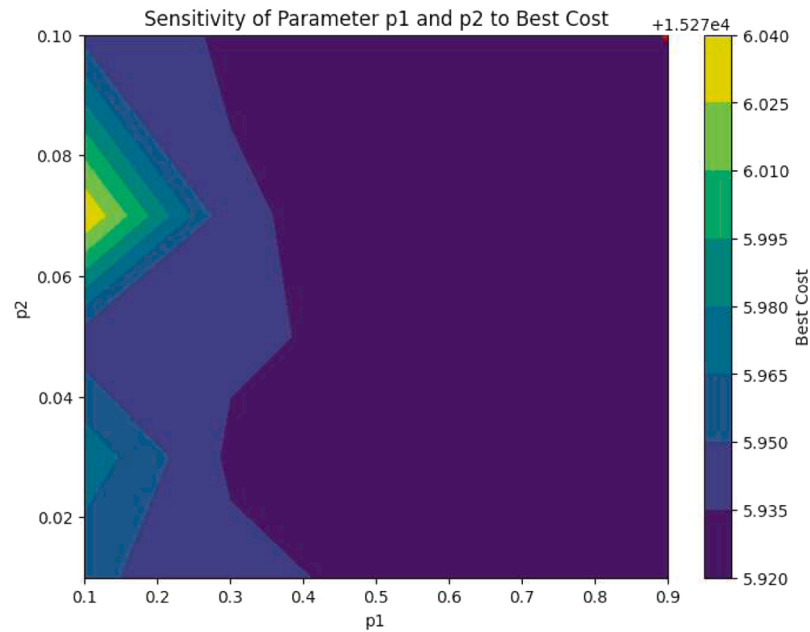


Fig. 3. Parameter Sensitivity Graph on 6 System Units

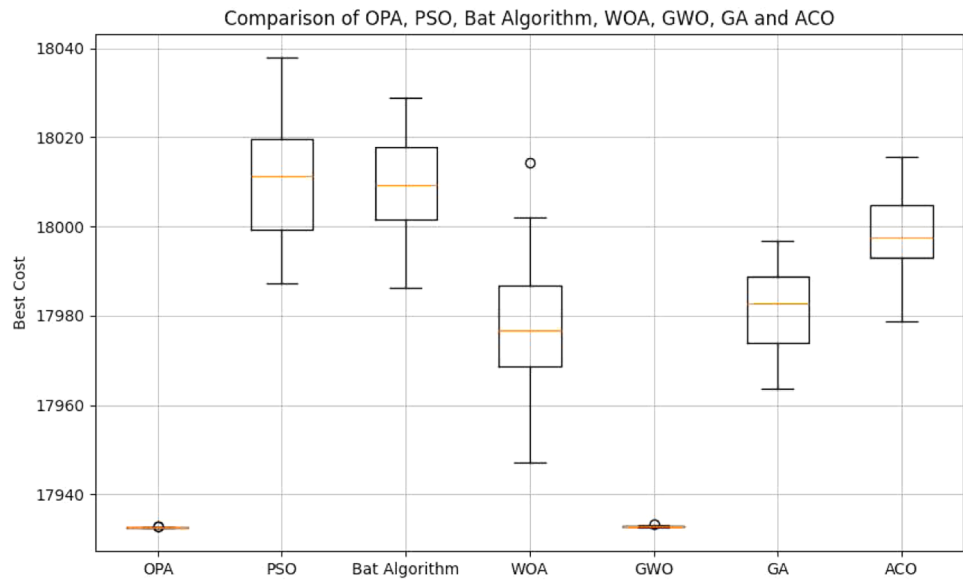


Fig. 4. Box Plot Comparison on 13 Generators

study demonstrates that parameter optimization in the OPA method can markedly enhance performance, particularly in the economic dispatch problem involving a system with six producing units. Through appropriate parameter tuning, a reduced optimal cost and enhanced stability can be achieved, demonstrating significant potential for improving the optimization outcomes of the OPA method.

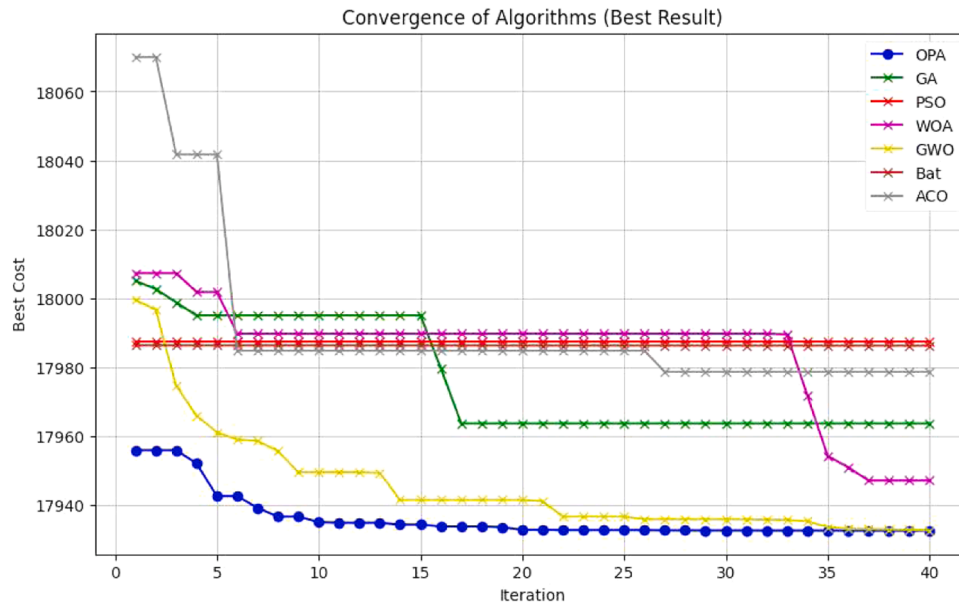


Fig. 5. Convergence curves on 13 Generators

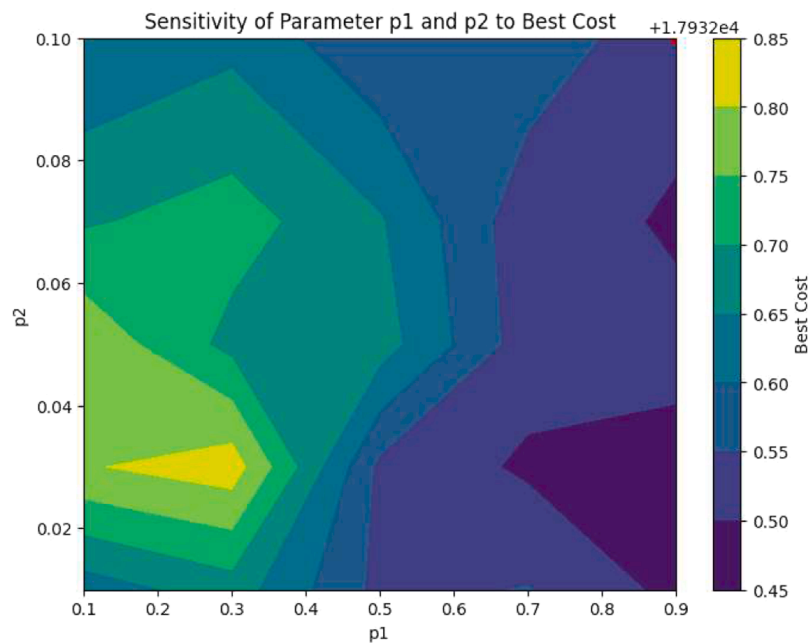


Fig. 6. Parameter Sensitivity Graph on 13 System Units

4.2. 13-unit system

The 13 generators in the 13-unit system collectively require a total load of 1800 MW. Table 7 displays the power output of each unit utilizing the five optimization procedures. Table 8 displays the minimum cost of the system following 30 trials, the average cost derived from these trials, the standard deviation, and the computation time for each technique. Fig. 3 illustrates the box plot for each algorithm assessed over 40 iterations. Fig. 4 illustrates their convergence curves.

The study found that OPA was the best way to solve the ED problem because it had the lowest cost (17932.4954), the most stable average cost (17932.5566), and the smallest standard deviation (0.0586). With a slightly higher average cost (17932.8416) and higher standard deviation (0.1926), GWO emerged as a strong contender. On the other hand, GA showed better results with a minimum cost of 17963.6191, even though

its standard deviation was higher (0.2587). ACO demonstrates advantageous characteristics relative to PSO and the Bat Algorithm; yet, it displays reduced convergence speed and increased variability. Regarding computational duration, GA is the most rapid at 0.0554 seconds, followed by WOA at 0.0633 seconds. The computing length of OPA is 1.0397 seconds, a duration that aligns with its exceptional accuracy and stability. A boxplot study shows that OPA is more consistent with the most tightly shaped cost distribution that doesn't have any outliers, but GWO is also stable. Convergence research Fig. 5 shows that OPA gets to the best cost (~17932) in just 10 iterations, which is faster than other methods. GWO and GA are also competitive, but not as reliable, options. These findings show that OPA is the most dependable and efficient method for addressing the ED problem involving 13 generators.

The sensitivity graph Fig. 6 of parameters p_1 and p_2 for ED over 13

Table 9

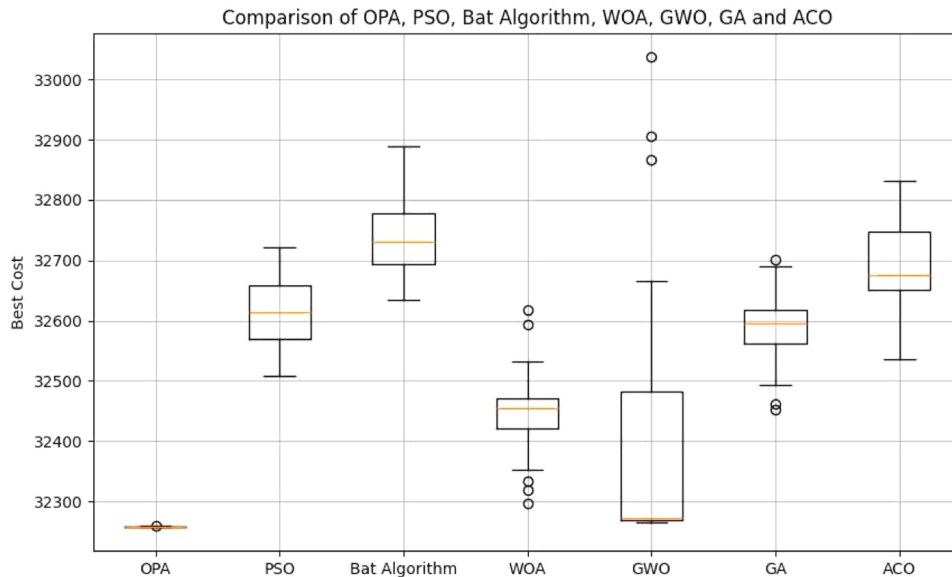
Actual output from the system's 15 generators

Algorithm	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15
PSO	637	413	127	126	156	357	363	156	57	66	35	55	37	15	23
Bat algorithm	289	408	116	91	416	444	455	89	113	57	47	28	29	16	22
WOA	434	446	87	127	172	446	437	298	154	143	62	65	78	15	26
GWO	424	455	130	130	283	458	465	60	25	25	51	54	25	15	15
GA	429	366	120	119	222	458	455	71	90	90	49	70	28	16	39
ACO	381	444	130	130	169	444	444	60	60	46	80	80	46	55	55
OPA	454	454	129	129	271	459	464	60	25	25	41	58	25	15	15

Table 10

Economic Dispatch on 15 Generators

Algorithm	Minimum Cost	Mean	Std	Computation Time
PSO	32479.77291159878	32623.175772738516	64.57281947957246	0.1150459369023641
Bat Algorithm	32517.27926269168	32736.693165221626	92.45938874166197	0.10075624783833821
WOA	32274.77527602057	32417.488311981335	90.06626186968653	0.04204538663228353
GWO	32260.96554107122	32373.93983597075	246.29498176071982	0.20730963548024495
GA	32452.48462298998	32591.145301781602	59.11278287536433	0.05900542736053467
ACO	32536.403120208663	32688.53318748755	76.27128490028328	0.9935892422993978
OPA	32257.018569117114	32257.602241102897	0.3277700800721914	0.9067840178807577

**Fig. 7.** Box Plot Comparison on 15 Generators

system units shows that the best cost changes depending on how p_1 and p_2 are combined. The experimental findings indicate that utilizing the parameter settings from Yuxin's work yields a maximum cost of 17932.495398652336, produced by the OPA method. Subsequent optimization yields a superior cost of 17932.485745827937, utilizing parameter values $p_1 = 0.9$ and $p_2 = 0.01$. The sensitivity graph indicates that within the range of p_1 from 0.2 to 0.4 and p_2 from 0.03 to 0.05, there is a notable variation in the optimal cost, with the most favorable outcomes occurring within these parameters. Simultaneously, as p_1 ascends to 0.9 and p_2 descends to 0.01, it is evident that the optimal cost diminishes, signifying that this parameter combination yields more stable and superior outcomes. The results of this 13-unit system demonstrate that the OPA algorithm can enhance its performance for solving economic dispatch problems in larger systems.

4.3. 15-unit system

The system consists of fifteen thermal units, and the specific parameters can be referenced in [16]. This test setting encompasses all the

nonlinear features and practical limitations associated with the ED problem. The load demand is 2630 MW. Fig. 8

The investigation reveals that OPA is the most efficient method for the ED issue involving 15 generators Tables 9 and 10, boasting the lowest cost (32257.0186) and the most consistent average cost (32257.6022) with a minimal standard deviation (0.3278), thereby confirming its exceptional stability. GWO emerged as a viable alternative with a marginally elevated cost (32260.9655) but increased variability (246.2950), while WOA showed commendable performance (32274.7753), albeit with a lower consistency (90.0663). GA yielded competitive findings (32452.4846), although its standard deviation (59.1128) constrained its reliability. ACO outperforms PSO and the Bat Algorithm; nonetheless, its slower convergence and elevated standard deviation suggest reduced stability. On the other hand, PSO (32623.1758) and the Bat Algorithm (32736.6932) have high costs that change a lot, which proves that they are unstable. Regarding calculation time, WOA is the most rapid at 0.0420 seconds, followed by GA at 0.0590 seconds, whereas OPA requires more time at 0.9068 seconds due to its meticulous optimization procedure. Notwithstanding the increased

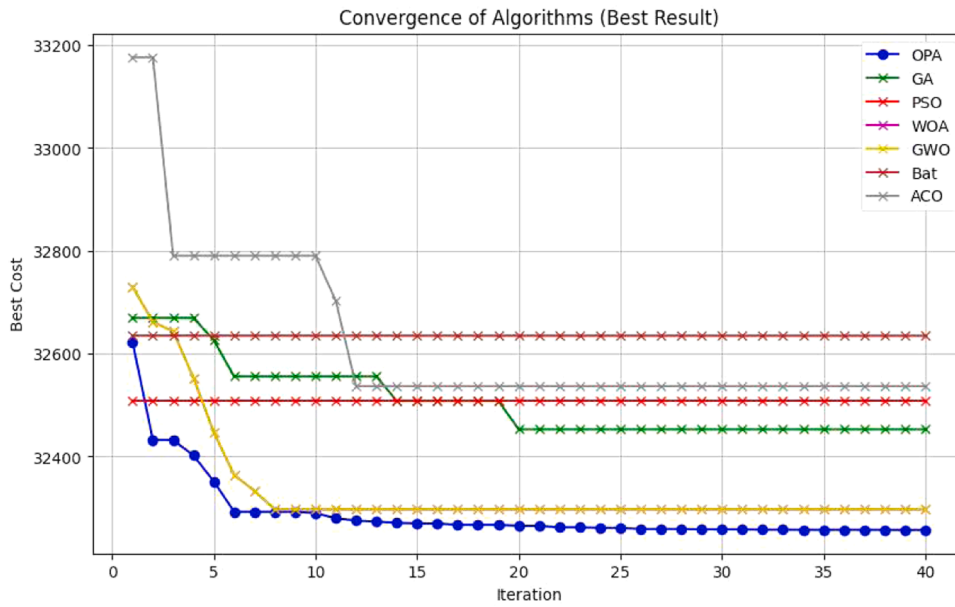


Fig. 8. Convergence curves on 15 Generators

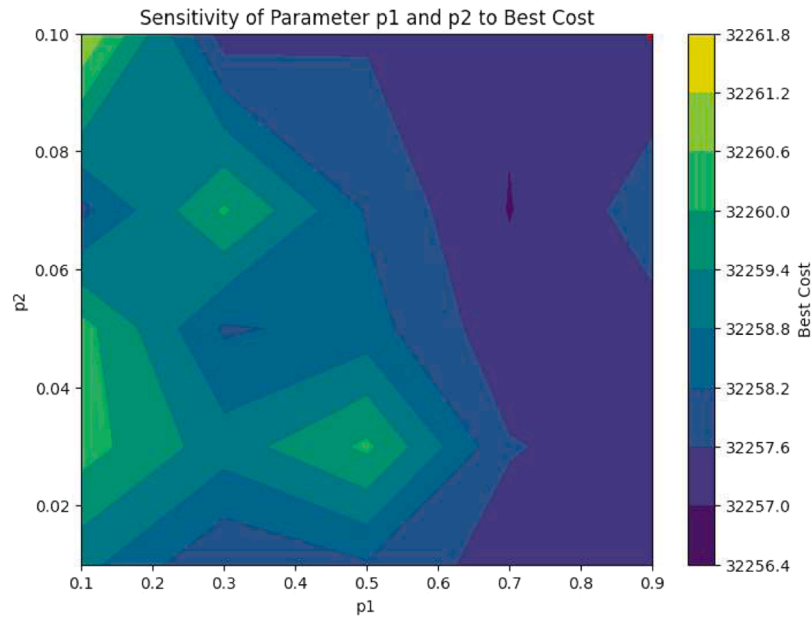


Fig. 9. Parameter Sensitivity Graph on 15 System Units

computing time, OPA typically yields improved outcomes, validating the trade-off. The boxplot study reinforces OPA's superiority, demonstrating the smallest cost distribution without outliers; hence, it affirms its exceptional consistency (Fig. 7). GWO exhibited stability with increased variability, whereas WOA, GA, and ACO demonstrated broader dispersion (Fig. 7). The PSO and Bat Algorithm exhibit the broadest distributions, indicating a deficiency in consistency. The convergence graph illustrates the efficacy of OPA, achieving optimal cost (~ 32257) after 10 rounds (Fig. 8). GWO approaches convergence at approximately 32260 after approximately 15 iterations; however, it exhibits diminished consistency. WOA stabilized at approximately 32274 across 20 iterations, indicating moderate performance. GA and ACO exhibited slower convergence, stabilizing at around 32452 and 32536, respectively. The PSO and Bat Algorithm did not attain competitive prices, stabilizing at elevated values with negligible

enhancement. These results show that OPA is the most reliable and effective way to solve the ED problem with 15 generators. It is better than all the other options in terms of correctness, stability, and convergence rate.

The sensitivity graph Fig. 9 of parameters p_1 and p_2 for ED over 15 system units shows that changing the parameters has a big effect on the optimal cost. Yuxin's essay's parameter values yielded an optimal cost of 32257.018569117114. Subsequent optimization yielded a reduced optimal cost of 32256.967447080402, achieved with parameter values $p_1 = 0.7$ and $p_2 = 0.07$. The graph shows that the optimal values for p_1 and p_2 are within the range of 0.7 and 0.07, respectively, allowing for cost minimization beyond the parameters presented in Yuxin's work. This proves that changing some parameters makes OPA algorithms work much better and more efficiently, especially for more complicated ED problems, as this 15-unit system shows.

5. Conclusion

This research shows that the Orca Predation Algorithm (OPA) consistently does a better job than other methods, such as PSO, the Bat Algorithm, WOA, GWO, GA, and ACO, when it comes to solving ED problems for systems with 6, 13, and 15 units in all possible configurations. OPA achieved the lowest cost, the most constant average cost, and the least fluctuation, confirming its remarkable stability and reliability. In the 6-unit system, OPA got the best price of \$15,275.9305 very precisely, beating out its competitors because it had the lowest standard deviation and the fastest convergence. In the 13-unit system, OPA maintained the lead with an ideal cost of 17932.4954 and consistent performance, although GWO presented a viable alternative, albeit with slightly greater variability. In the 15-unit system, OPA worked amazingly well, getting the best results with the lowest cost (32257.0186) and the most consistent costs. This solidified its reputation as the most reliable algorithm. While GWO exhibited commendable convergence speed, its greater variability and certain outliers underscored the superior consistency of OPA. WOA and GA yielded competitive outcomes in certain instances; nonetheless, they exhibited inferior stability and precision compared to OPA. Simultaneously, ACO demonstrated superior performance compared to PSO and the Bat Algorithm, but with delayed convergence and increased variability. The PSO and Bat Algorithm consistently produced bad results with wide variation, proving that they are not useful for solving ED problems. Although OPA had outstanding performance, its computational duration surpassed that of alternative approaches. The remarkable precision and stability of OPA justify this trade-off. Researchers may find ways to get around this problem in the future by finding ways to speed up OPA's calculations through parallelization or hybrid methods. Also, looking into how OPA can be used in bigger, more complicated systems with cost functions that aren't convex can show how useful it is for real-world power system optimization.

Funding statement

This work received no external funding.

CRediT authorship contribution statement

Vivi Aida Fitria: Writing – original draft, Investigation, Formal analysis, Data curation, Conceptualization. **Arif Nur Afandi:** Validation, Methodology. **Aripriharta:** Writing – review & editing, Validation, Supervision.

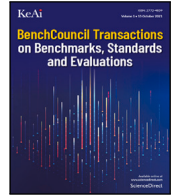
Declaration of competing interest

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

References

- [1] S.K. Rangu, P.R. Lolla, K.R. Dhenuvakonda, A.R. Singh, Recent trends in power management strategies for optimal operation of distributed energy resources in microgrids: A comprehensive review, *Int. J. Energy Res.* 44 (13) (Oct. 2020) 9889–9911, <https://doi.org/10.1002/er.5649>.
- [2] L. Al-Bahrani, M. Seyedmahmoudian, B. Horan, A. Stojcevski, Solving the Real Power Limitations in the Dynamic Economic Dispatch of Large-Scale Thermal Power Units under the Effects of Valve-Point Loading and Ramp-Rate Limitations, *Sustainability* 13 (3) (Jan. 2021) 1274, <https://doi.org/10.3390/su13031274>.
- [3] A.N. Afandi, I. Fadlika, A. Andoko, Comparing Performances of Evolutionary Algorithms on the Emission Dispatch and Economic Dispatch Problem, *TELKOMNIKA (Telecommunication Comput. Electron. Control)* 13 (4) (Dec. 2015) 1187, <https://doi.org/10.12928/telekomnika.v13i4.3166>.
- [4] H. Zhong, X. Yan, Z. Tan, Real-Time Distributed Economic Dispatch Adapted to General Convex Cost Functions: A Secant Approximation-Based Method, *IEEE Trans. Smart Grid* 12 (3) (May 2021) 2089–2101, <https://doi.org/10.1109/TSG.2020.3049054>.
- [5] Janter Pangaduan Simanjuntak, Khaled Ali Al-attab, Eka Daryanto, Bisrul Hapis Tambunan, Eswanto, Bioenergy as an Alternative Energy Source: Progress and Development to Meet the Energy Mix in Indonesia, *J. Adv. Res. Fluid Mech. Therm. Sci.* 97 (1) (Aug. 2022) 85–104, <https://doi.org/10.37934/arfmts.97.1.85104>.
- [6] F. Marzbani, A. Abdelfatah, Economic Dispatch Optimization Strategies and Problem Formulation: A Comprehensive Review, *Energies* 17 (3) (Jan. 2024) 550, <https://doi.org/10.3390/en17030550>.
- [7] P. Vasant, A. Banik, J. J. Thomas, J. A. Marmolejo-Saucedo, U. Fiore, and G.-W. Weber, “Bio-inspired approaches for a combined economic emission dispatch problem,” in *Human-Assisted Intelligent Computing*, IOP Publishing, 2023, pp. 3–13–38. <https://doi.org/10.1088/978-0-7503-4801-0ch3>.
- [8] B.M. Hussein, Evolutionary algorithm solution for economic dispatch problems, *Int. J. Electr. Comput. Eng.* 12 (3) (Jun. 2022) 2963, <https://doi.org/10.11591/ijece.v12i3.pp2963-2970>.
- [9] J. Zhang, J. Zhang, F. Zhang, M. Chi, L. Wan, An Improved Symbiosis Particle Swarm Optimization for Solving Economic Load Dispatch Problem, *J. Electr. Comput. Eng.* 2021 (Jan. 2021) 1–11, <https://doi.org/10.1155/2021/8869477>.
- [10] A.E. Prasetya, T. Wrahatnolo, Economic Dispatch Pada Pembangkit Termal Pln Apb Iv Jawa Timur Menggunakan Metode Particle Swarm Optimization (Pso), *J. Tek. Elektro* 9 (1) (2020) 885–892, <https://doi.org/10.26740/jte.v9n1.p825p>.
- [11] F. Tariq, S. Alelyani, G. Abbas, A. Qahmash, M.R. Hussain, Solving Renewables-Integrated Economic Load Dispatch Problem by Variant of Metaheuristic Bat-Inspired Algorithm, *Energies* 13 (23) (Nov. 2020) 6225, <https://doi.org/10.3390/en13236225>.
- [12] H.J. Touma, Study of The Economic Dispatch Problem on IEEE 30-Bus System using Whale Optimization Algorithm, *Int. J. Eng. Technol. Sci.* 3 (1) (Jun. 2016) 11–18, <https://doi.org/10.15282/ijets.5.2016.1.2.1041>.
- [13] S. Hosseini-Hemati, S. Derafshi Beigvand, H. Abdi, A. Rastgou, Society-based Grey Wolf Optimizer for large scale Combined Heat and Power Economic Dispatch problem considering power losses, *Appl. Soft Comput.* 117 (Mar. 2022) 108351, <https://doi.org/10.1016/j.asoc.2021.108351>.
- [14] W.-C. Yeh, et al., New genetic algorithm for economic dispatch of stand-alone three-modular microgrid in DongAo Island, *Appl. Energy* 263 (Apr. 2020) 114508, <https://doi.org/10.1016/j.apenergy.2020.114508>.
- [15] A. Srivastava, S. Singh, Implementation of Ant Colony Optimization in Economic Load Dispatch Problem, in: 2020 7th International Conference on Signal Processing and Integrated Networks (SPIN), IEEE, Feb. 2020, pp. 1018–1024, <https://doi.org/10.1109/SPIN48934.2020.9071407>.
- [16] Y. Jiang, Q. Wu, S. Zhu, L. Zhang, Orca predation algorithm: A novel bio-inspired algorithm for global optimization problems, *Expert Syst. Appl.* 188 (Feb. 2022) 116026, <https://doi.org/10.1016/j.eswa.2021.116026>.
- [17] A.G. Gad, Particle Swarm Optimization Algorithm and Its Applications: A Systematic Review, *Arch. Comput. Methods Eng.* 2022 295 29 (5) (Apr. 2022) 2531–2561, <https://doi.org/10.1007/S11831-021-09694-4>.
- [18] Y.R. Nugraha, A.P. Wibawa, I.A.E. Zaeni, Particle Swarm Optimization – Support Vector Machine (PSO-SVM) Algorithm for Journal Rank Classification, in: 2019 2nd International Conference of Computer and Informatics Engineering (IC2IE), IEEE, Sep. 2019, pp. 69–73, <https://doi.org/10.1109/IC2IE47452.2019.8940822>.
- [19] A.B. Putra Utama, A.P. Wibawa, M. Muladi, A. Nafalski, PSO based Hyperparameter tuning of CNN Multivariate Time- Series Analysis, *J. Online Inform.* 7 (2) (Dec. 2022) 193–202, <https://doi.org/10.15575/join.v7i2.858>.
- [20] W.M.A. Wan Abdul Razak, N. Rosmin, A.H. Musta'amal, S.M. Hussin, D. Mat Said, Aripriharta, Power loss minimization by optimal allocation and sizing of STATCOM via particle swarm optimization, in: E3S Web Conf 516, Apr. 2024, p. 05003, <https://doi.org/10.1051/e3sconf/202451605003>.
- [21] S. Tiwari, N. S. Pal, M. A. Ansari, D. Yadav, and N. Singh, “Economic Load Dispatch Using PSO,” 2020, pp. 51–64. https://doi.org/10.1007/978-981-15-2329-8_6.
- [22] K. Chaitanya, D.V.L.N. Somayajulu, P.R. Krishna, Memory-based approaches for eliminating premature convergence in particle swarm optimization, *Appl. Intell.* 51 (7) (Jul. 2021) 4575–4608, <https://doi.org/10.1007/s10489-020-02045-z>.
- [23] H. Suyono, E. Subekti, H. Purnomo, T. Nurwati, R.N. Hasanah, Economic Dispatch of 500 kV Java-Bali Power System using Hybrid Particle Swarm-Ant Colony Optimization Method, in: 2020 12th International Conference on Electrical Engineering (ICEENG), IEEE, Jul. 2020, pp. 5–10, <https://doi.org/10.1109/ICEENG45378.2020.9171771>.
- [24] E. Osaba, X.-S. Yang, and J. Del Ser, “Traveling salesman problem: a perspective review of recent research and new results with bio-inspired metaheuristics,” in *Nature-Inspired Computation and Swarm Intelligence*, Elsevier, 2020, pp. 135–164. <https://doi.org/10.1016/B978-0-12-819714-1.00020-8>.
- [25] Z.A.A. Alyasseri, et al., Recent advances of bat-inspired algorithm, its versions and applications, *Neural Comput. Appl.* 34 (19) (Oct. 2022) 16387–16422, <https://doi.org/10.1007/s00521-022-07662-y>.
- [26] A. Kaur, Y. Kumar, Recent Developments in Bat Algorithm: A Mini Review, *J. Phys. Conf. Ser.* 1950 (1) (Aug. 2021) 012055, <https://doi.org/10.1088/1742-6596/1950/1/012055>.
- [27] C.A. Diebold, A. Salles, C.F. Moss, Adaptive Echolocation and Flight Behaviors in Bats Can Inspire Technology Innovations for Sonar Tracking and Interception, *Sensors* 20 (10) (May 2020) 2958, <https://doi.org/10.3390/s20102958>.
- [28] W. Younas, et al., Improving Convergence Speed of Bat Algorithm Using Multiple Pulse Emissions along Multiple Directions, *Sensors* 22 (23) (Dec. 2022) 9513, <https://doi.org/10.3390/s22239513>.
- [29] R. Ali Abttan, A. Hasan Tawafan, S.Jaafar Ismael, Economic dispatch by optimization techniques, *Int. J. Electr. Comput. Eng.* 12 (3) (Jun. 2022) 2228, <https://doi.org/10.11591/ijece.v12i3.pp2228-2241>.
- [30] M.H. Nadimi-Shahraki, H. Zamani, Z.Ashgari Varzaneh, S. Mirjalili, A Systematic Review of the Whale Optimization Algorithm: Theoretical Foundation,

- Improvements, and Hybridizations, *Arch. Comput. Methods Eng.* 30 (7) (Sep. 2023) 4113–4159, <https://doi.org/10.1007/s11831-023-09928-7>.
- [31] M. Li, G. Xu, Y. Fu, T. Zhang, L. Du, Improved whale optimization algorithm based on variable spiral position update strategy and adaptive inertia weight, *J. Intell. Fuzzy Syst.* 42 (3) (Feb. 2022) 1501–1517, <https://doi.org/10.3233/JIFS-210842>.
- [32] S. Chakraborty, S. Sharma, A.K. Saha, A. Saha, A novel improved whale optimization algorithm to solve numerical optimization and real-world applications, *Artif. Intell. Rev.* 55 (6) (Aug. 2022) 4605–4716, <https://doi.org/10.1007/s10462-021-10114-z>.
- [33] L. Abualigah, et al., Whale optimization algorithm: analysis and full survey. *Metaheuristic Optimization Algorithms*, Elsevier, 2024, pp. 105–115, <https://doi.org/10.1016/B978-0-443-13925-3.00015-7>.
- [34] C. Tang, W. Sun, M. Xue, X. Zhang, H. Tang, W. Wu, A hybrid whale optimization algorithm with artificial bee colony, *Soft Comput* 26 (5) (Mar. 2022) 2075–2097, <https://doi.org/10.1007/s00500-021-06623-2>.
- [35] Q.A. Sias, I. Fadlika, I.D. Wahyono, A.Nur Afandi, Quasi Z-Source Inverter as MPPT on Renewable Energy using Grey Wolf Technique, in: 2018 5th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), IEEE, Oct. 2018, pp. 362–366, <https://doi.org/10.1109/EECSI.2018.8752686>.
- [36] I. Sharma, V. Kumar, S. Sharma, A Comprehensive Survey on Grey Wolf Optimization, *Recent Adv. Comput. Sci. Commun.* 15 (3) (Mar. 2022), <https://doi.org/10.2174/2666255813999201007165454>.
- [37] L.I. Wong, M.H. Sulaiman, M.R. Mohamed, M.S. Hong, Grey Wolf Optimizer for solving economic dispatch problems, in: 2014 IEEE International Conference on Power and Energy (PECon), IEEE, Dec. 2014, pp. 150–154, <https://doi.org/10.1109/PECON.2014.7062431>.
- [38] Y. Liu, A. As'arry, M.K. Hassan, A.A. Hairuddin, H. Mohamad, Review of the grey wolf optimization algorithm: variants and applications, *Neural Comput. Appl.* 36 (6) (Feb. 2024) 2713–2735, <https://doi.org/10.1007/s00521-023-09202-8>.
- [39] B. Alhijawi, A. Awajan, Genetic algorithms: theory, genetic operators, solutions, and applications, *Evol. Intell.* 17 (3) (Jun. 2024) 1245–1256, <https://doi.org/10.1007/s12065-023-00822-6>.
- [40] K.B. Sahay, A. Sonkar, A. Kumar, Economic Load Dispatch Using Genetic Algorithm Optimization Technique, in: 2018 International Conference and Utility Exhibition on Green Energy for Sustainable Development (ICUE), IEEE, Oct. 2018, pp. 1–5, <https://doi.org/10.23919/ICUE-GESD.2018.8635729>.
- [41] S. Katoch, S.S. Chauhan, V. Kumar, A review on genetic algorithm: past, present, and future, *Multimed. Tools Appl.* 80 (5) (Feb. 2021) 8091–8126, <https://doi.org/10.1007/s11042-020-10139-6>.
- [42] N. Nayar, S. Gautam, P. Singh, and G. Mehta, "Ant Colony Optimization: A Review of Literature and Application in Feature Selection," 2021, pp. 285–297. doi: [10.1007/978-981-33-4305-4_22](https://doi.org/10.1007/978-981-33-4305-4_22).
- [43] A.B. Kunya, A.S. Abubakar, S.S. Yusuf, Review of economic dispatch in multi-area power system: State-of-the-art and future prospective, *Electr. Power Syst. Res.* 217 (Apr. 2023) 109089, <https://doi.org/10.1016/j.epsr.2022.109089>.
- [44] M. Dashtdar, et al., Solving the environmental/economic dispatch problem using the hybrid FA-GA multi-objective algorithm, *Energy Reports* 8 (Nov. 2022) 13766–13779, <https://doi.org/10.1016/j.egy.2022.10.054>.
- [45] H. Hardiansyah, A Modified Particle Swarm Optimization Technique for Economic Load Dispatch with Valve-Point Effect, *Int. J. Intell. Syst. Appl.* 5 (7) (Jun. 2013) 32–41, <https://doi.org/10.5815/ijisa.2013.07.05>.
- [46] P. Zakian, A. Kaveh, Economic dispatch of power systems using an adaptive charged system search algorithm, *Appl. Soft Comput.* 73 (Dec. 2018) 607–622, <https://doi.org/10.1016/j.asoc.2018.09.008>.
- [47] J. Pokala and B. Lalitha, "A Novel Intrusion Detection System for RPL Based IoT Networks with Bio-Inspired Feature Selection and Ensemble Classifier." Apr. 28, 2021. [10.21203/rs.3.rs-442429/v1](https://doi.org/10.21203/rs.3.rs-442429/v1).
- [48] X.-B. Meng, X.Z. Gao, Y. Liu, H. Zhang, A novel bat algorithm with habitat selection and Doppler effect in echoes for optimization, *Expert Syst. Appl.* 42 (17–18) (Oct. 2015) 6350–6364, <https://doi.org/10.1016/j.eswa.2015.04.026>.
- [49] S. Chakraborty, A. Kumar Saha, S. Sharma, S. Mirjalili, R. Chakraborty, A novel enhanced whale optimization algorithm for global optimization, *Comput. Ind. Eng.* 153 (Mar. 2021) 107086, <https://doi.org/10.1016/j.cie.2020.107086>.
- [50] S. Shafaghi, M. Shokouhifar, R. Sabbaghi-Nadooshan, Swarm Intelligence Low Power Routing in Network-on-Chips, *Int. J. Energy, Inf. Commun.* 7 (2) (Apr. 2016) 21–40, <https://doi.org/10.14257/ijeic.2016.7.2.03>.
- [51] F. Abbas, et al., Optimizing Machine Learning Algorithms for Landslide Susceptibility Mapping along the Karakoram Highway, Gilgit Baltistan, Pakistan: A Comparative Study of Baseline, Bayesian, and Metaheuristic Hyperparameter Optimization Techniques, *Sensors* 23 (15) (Aug. 2023) 6843, <https://doi.org/10.3390/s23156843>.



Research Article

MultiPoint: Enabling scalable pre-silicon performance evaluation for multi-task workloads[☆]

Chenji Han^{a,b}^{*}, Xinyu Li^{a,b}, Feng Xue^{a,b}, Weitong Wang^{a,b}, Yuxuan Wu^c,
Wenxiang Wang^{b,c}, Fuxin Zhang^a

^a SKLP, Institute of Computing Technology, CAS, Beijing, China

^b University of Chinese Academy of Sciences, Beijing, China

^c Loongson Technology, Beijing, China

ARTICLE INFO

Keywords:

Performance modeling
Performance evaluation
Multi-task workloads

ABSTRACT

With the core numbers integrated within single processors growing and the fast development of cloud computing, performance evaluation for multi-core systems is increasingly crucial. It is typically conducted by executing multi-task workloads, exemplified by SPEC CPU Rate, to measure metrics like system's throughput. In response, several sampling-based methods have been developed for their pre-silicon performance evaluation. Nevertheless, these methods involve directly capturing multi-task checkpoints, which presents scalability issues of significant storage and time overheads. Therefore, enabling more scalable performance evaluation remains a critical problem.

In this work, we propose MultiPoint to enable scalable pre-silicon performance evaluation for multi-task workloads. It is noted that in the multi-task workloads of interest, each task executes independently without inter-task communication. Therefore, MultiPoint is motivated to construct the required multi-task checkpoints by recovering multiple single-task checkpoints across different cores and guarantee their smooth execution through address remapping and shuffling. We implemented MultiPoint on the Emulator Accelerator and assessed its evaluation accuracy against its post-silicon Loongson 3A6000 processor. Using SPEC CPU 2017 as the benchmark, MultiPoint achieved the estimation errors of 6.20%, 5.45%, and 6.99% for Rate 2, Rate 4, and Rate 8, respectively, achieving comparable accuracy compared to direct multi-task checkpointing but in a more scalable manner with substantially 86.0% lower storage and 93.7% less time overheads.

1. Introduction

Background. Pre-silicon performance evaluation is becoming increasingly critical, considering the continuous rise in fabrication costs and prolonged verification periods. For the single-task workloads, many representative sampling-based methods [1–8], such as SimPoint [1–5], have been developed. These methods involve profiling and clustering program's code signatures and eventually selecting the simulation points, as detailed in Section 2.1. In our practice, SimPoint achieved an average performance estimation error of 1.85% for SPEC CPU 2017 Rate 1 [9] with a speedup of 477 times. Besides, with the core numbers integrated within single processors growing and the fast development of cloud computing, performance evaluation for multi-core systems is becoming increasingly crucial. It is typically conducted by concurrently

executing multiple workloads [10,11], exemplified by SPEC CPU 2017 Rate, to measure metrics like system's throughput [12,13].

Research Problem. In response, several SimPoint-like method [14–19] have been developed for pre-silicon performance evaluations of multi-task workloads. These methods typically concatenate code signatures of concurrently executed programs to cluster and select representative simulation points. The multi-task checkpoints of selected simulation points are then captured, containing the architecture-level status of register values and the memory content of these tasks, which could be recovered on the target multi-core designs to conduct performance evaluations. However, the direct checkpointing of multi-task workloads involved in these methods presents **scalability issues** of significant storage and time overheads.

[☆] The authors would like to thank the helpful discussions with Ruiyang Wu, Yuxiao Chen, and Hongze Tan. This work was supported by the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. XDC05020100).

^{*} Corresponding author at: University of Chinese Academy of Sciences, Beijing, China.
E-mail address: hanchenji16@mails.ucas.ac.cn (C. Han).

<https://doi.org/10.1016/j.tbench.2025.100189>

Received 16 October 2024; Received in revised form 27 December 2024; Accepted 23 January 2025

Available online 13 February 2025

2772-4859/© 2025 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Firstly, checkpoints of multi-task workloads can only be employed to evaluate multi-core system with specific core numbers. For instance, the checkpoints captured on a four-core system can only be used to evaluate four-core designs. Therefore, any change in the assessed core numbers would necessitate recapturing the corresponding multi-task checkpoints, leading to a waste of computational resources as well as extra storage and time overheads.

Secondly, directly captured multi-task checkpoints require large storage overheads. The total size of a multi-task checkpoint is approximately equal to the combined size of checkpoints for the involved individual tasks. Besides, the content in multi-task checkpoints could be duplicated with certain single-task checkpoints. The storage overheads would be exacerbated as the core number increases.

Key Idea. It is noticed that in the multi-task workloads of interest, such as SPEC CPU Rate, each task executes independently without inter-task communication. It motivates us to restore multiple single-task checkpoints on different cores to construct the required multi-task checkpoints, which essentially involves concurrently running multiple single-core operating systems on the multi-core system with the shared memory. Consequently, the scalability issues of substantial storage and time overheads the direct multi-task checkpointing could be eliminated.

Requirements. However, composing the required multi-task checkpoint by multiple single-task checkpoints presents several requirements, as discussed below.

Firstly, it is required to guarantee multiple single-task checkpoints to smoothly execute at a multi-core system with shared memory. In the single-task checkpoint, the operating system manages memory within a fixed range. When multiple single-task checkpoints are restored simultaneously without special handles, they would inadvertently attempt to use the same memory regions. This overlap in memory spaces can prevent the tasks from executing normally.

Secondly, it is necessary to make the memory address characteristics similar to that in realistic multi-task workload executions, which are scattered across the memory and interleaved with each other, as shown in Fig. 3. This is because, differences in memory address characteristics could result in varying impacts on certain μ Arch structures, like the last-level cache, thus compromising the accuracy of performance evaluation, as discussed in Section 6.3.

Our Work. Corresponding to these requirements, we propose MultiPoint to enable scalable pre-silicon performance evaluation for multi-task workloads. MultiPoint is capable of composing the required multi-task workloads by simultaneously recovering multiple single-task workloads across different cores and ensuring their smooth concurrent execution through physical address remapping and shuffling. Specifically, MultiPoint introduces a checkpoint loader and proposes a synchronization mechanism to support the concurrent recovery of multiple single-task checkpoints at the Emulator Accelerator and enable their simultaneous initiation of execution. Besides, MultiPoint introduces a software-transparent address transform layer to support the concurrent smooth execution of multiple single-task checkpoints. To ensure the normal execution of these single-task checkpoints, MultiPoint remaps the memory requests from different cores to separate memory regions to avoid their interference. Furthermore, to mirror the actual memory access address characteristics of multi-task workloads, MultiPoint shuffles their memory addresses to make them scatter across the memory space and interleave with each other.

To sum up, the contributions of this work include:

1. We proposed MultiPoint to compose multi-task checkpoints through multiple single-task checkpoints, which enables scalable pre-silicon performance evaluation for multi-task workloads.
2. We implemented the evaluation routine of MultiPoint on the Emulator Accelerator and assessed its performance evaluation accuracy against its post-silicon Loongson 3A6000 commercial processor [20].

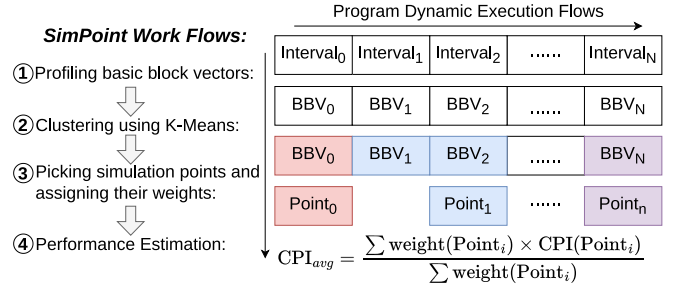


Fig. 1. Procedures of the representative sampling method, using SimPoint as an illustrative example.

3. We evaluated MultiPoint on the SPEC CPU 2017 Rate benchmark. MultiPoint achieved the score estimation errors of 6.20%, 5.45%, and 6.99% for Rate 2, Rate 4, and Rate 8, respectively, achieving comparable estimation accuracy compared to the direct multi-task checkpointing but in a more scalable manner with 86.0% lower storage and 93.7% less time overheads.

The remaining part of this work is organized as follows. Sections 2 and 3 give the background and motivation of MultiPoint. Section 4 details the method of MultiPoint. Section 5 introduces the experiment environment. Section 6 discusses the evaluation results. Section 7 lists the related works. Section 8 concludes this work.

2. Background

2.1. Evaluation for single-task workload

For pre-silicon performance evaluation of single-task workload, many representative sampling-based methods [1–8], such as SimPoint [1–5], are well-developed and widely employed in both the academia [21] and industries [22]. The motivation behind representative sampling is that programs' execution is composed of several recurring phases, instead of being chaos. Fig. 1 illustrates general procedures of the representative sampling, using SimPoint as the example. Specifically, SimPoint divides the program's dynamic execution flows into non-overlapping intervals with fixed lengths. For each interval, SimPoint profiles its frequency vector of basic blocks (BBVs) (①), which is a sequence of consecutive instructions with only one entrance and one exit. After profiling, the K-Means algorithm is leveraged to cluster these program intervals (②), after setting parameters of the maximum allowed cluster numbers $\max K$, projected dimension \dim , and Bayesian information criterion threshold BIC. As a result, program intervals closest to the centroid of each cluster are selected as the simulation points to represent the average performance of each cluster. Besides, the simulation points are assigned weights according to the number of program intervals that they represent (③). Finally, the program's performance could be extrapolated by the weighted average of performance of these representative simulation points' performance (④). In our practice, SimPoint achieved an average performance estimation error of 1.85% for SPEC CPU 2017 benchmark [9] with a speedup of 477 times.

2.2. Evaluation for multi-task workload

Sampling-like methods [14–19] have also been developed for homogeneous and heterogeneous multi-task workloads. Specifically, as demonstrated in Fig. 2, these methods involve concatenating BBVs of concurrently executed tasks (①) and utilizing the resultant concatenated vectors as program's code signatures (②). Following the SimPoint-like procedures in Fig. 1, these methods cluster these signatures and select the representative simulation points. Next, their corresponding multi-task checkpoints are captured (③). However, the direct checkpointing of multi-task workloads involved in these methods brings scalability issues of significant storage and time overheads.

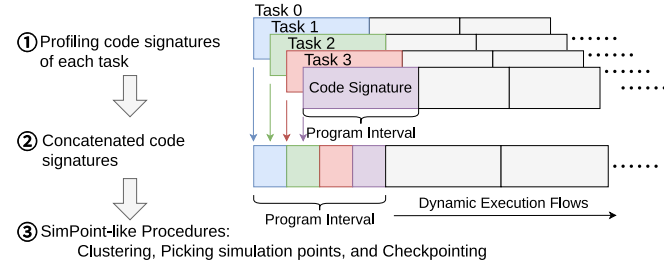


Fig. 2. Procedures of the SimPoint-like method for multi-task workload evaluations, using 4 tasks as the example.

3. Key idea of MultiPoint

It is noticed that, in scenarios of multi-task workloads such as SPEC CPU Rate, each task executes independently without inter-task communication. It motivates us to utilize multiple single-task checkpoints to agilely compose required checkpoints of multi-task workloads, which essentially involves concurrently running multiple single-core operating systems on the multi-core system with the shared memory. Correspondingly, such an approach must satisfy the following two requirements to ensure the correctness of program execution and maintain the evaluation accuracy.

3.1. Motivations of address isolation

Firstly, it is noted that in the single-task checkpoint, the operating system manages memory within a fixed range and is typically immutable during runtime. When multiple single-task checkpoints are restored simultaneously without special handles, they would inadvertently attempt to use the same memory regions. Consequently, these concurrently executed tasks would interfere with each other, preventing them from executing successfully.

Requirement 1. Physical addresses of memory requests from different cores should be isolated from each other.

Address isolation can be achieved by assigning different memory offsets to requests originating from different cores.

3.2. Motivations of address interleaving

Secondly, it is necessary to make the memory address characteristics of concurrently executed multiple single-task checkpoints similar to that in realistic multi-task workload executions, which are scattered across the memory and interleaved with each other. This is because differences in memory address characteristics could result in varying impacts on certain μ Arch structures. For example, in homogeneous workloads such as SPEC CPU 2017, if only the address isolation is implemented, the low bits of address accessed by different single-task checkpoints for semantically identical memory are identical, and only their highest address bits are distinct. This would result in significant cache set conflicts in the shared last-level cache.

Specifically, Fig. 3 presents the probability density distributions of physical address usage by different cores of four memory-intensive programs in SPEC CPU 2017 Rate 4. Other programs in SPEC CPU 2017 behave similarly and are thus not presented here. In Fig. 3, the program's physical address usage is collected via the Linux kernel interface, and the probability density distributions are calculated by the Gaussian KDE method [23]. Fig. 3 illustrates that memory addresses utilized by different cores are interleaved with each other rather than being distinctly isolated. It is noted that the specific distributions of memory usage by multi-task workloads can vary across

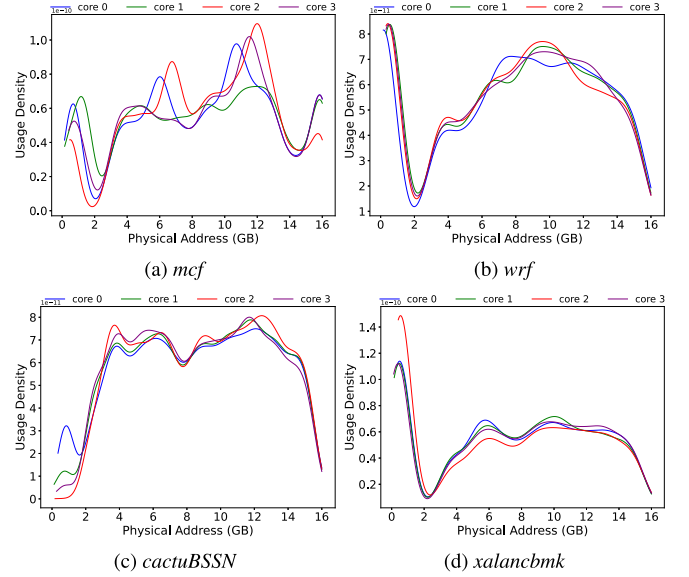


Fig. 3. Probability density distributions of physical address usage by different cores of four memory-intensive programs in SPEC CPU 2017 Rate 4. Instead of being isolated, their memory addresses are interleaved with each other.

different executions based on the real-time situation of the memory fragmentation [24]. Despite these variations, physical addresses allocated for different tasks are typically interleaved as a result of interleaved processing of their memory allocation requests.

Requirement 2. Physical addresses of memory requests from different cores should be interleaved with each other.

Address interleaving can be achieved by employing software-transparent address shuffling algorithm.

4. Method of MultiPoint

In this work, we propose MultiPoint to enable scalable performance evaluation for multi-task workloads. As illustrated in Fig. 4, MultiPoint is composed of three critical procedures as follows. **Checkpoint Recovery:** Multiple checkpoints of different or identical single-task workloads are concurrently recovered across different processor cores (①). Besides, a sync mechanism is implemented among the processor cores to guarantee their simultaneous initiation of evaluation. **Address Remapping:** To guarantee the smooth execution of these concurrently executed single-task checkpoints, MultiPoint remaps memory requests from different processor cores to different memory regions to prevent them from interfering with each other (②). **Address Shuffling:** To maintain the performance evaluation accuracy of constructing multi-task checkpoints via multiple single-task checkpoints, MultiPoint shuffles memory requests from different cores to make them scatter and interleave in the memory space (③). Collectively, MultiPoint introduces a software-transparent address transform mechanism to support the concurrent smooth execution of multiple single-task checkpoints. Consequently, the scalability issues of substantial storage and time overheads the direct multi-task checkpointing could be avoided.

4.1. Checkpoint recovery

MultiPoint is designed to concurrently recover multiple single-task checkpoints across different processor cores and employs a sync mechanism to guarantee their simultaneous initiation of evaluation. The detailed recovery process of a single-task checkpoint by the proposed checkpoint loader is illustrated in Fig. 5. The checkpoint consists of

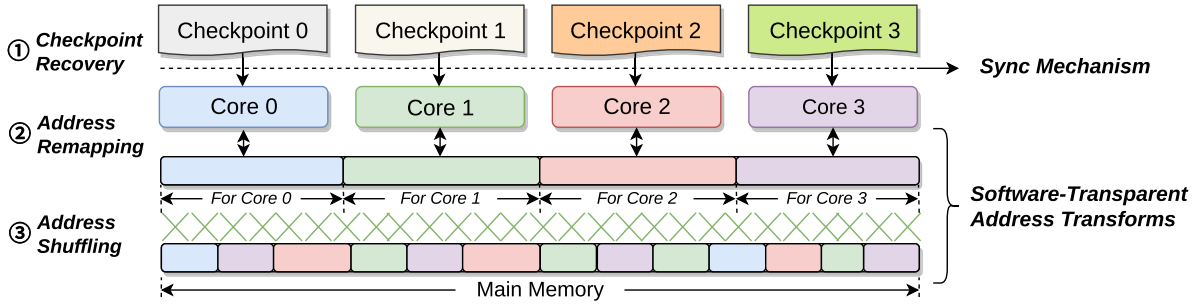


Fig. 4. Procedures of MultiPoint for performance evaluation of multi-task workload.

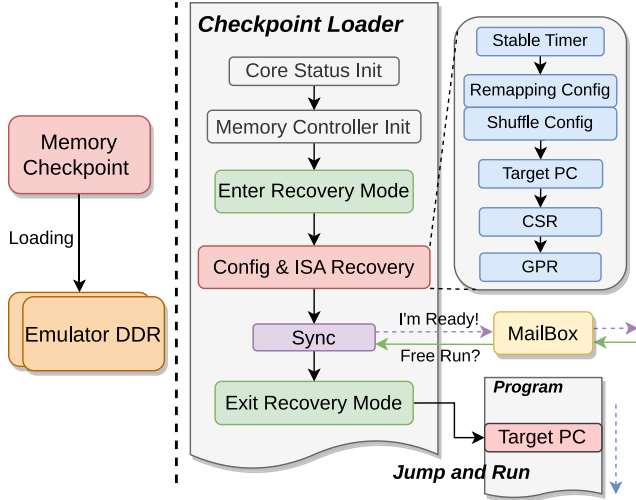


Fig. 5. Procedures of single-task checkpoint recovery by the checkpoint loader.

the complete architecture-level status required by the program's normal execution, including the memory contents and the values of various registers specified by the Instruction Set Architecture, such as general-purpose registers and control status registers. Specifically, for the recovery of memory content, MultiPoint utilizes the programming interface provided by the Emulator Accelerator to load the full contents of the main memory into the corresponding positions in the DDR system of the Emulator Accelerator.

For the recovery of register, MultiPoint introduces a checkpoint loader, which is firmware that is responsible for initializing the DDR system and preparing the execution environment before handing over control to the operating system in the checkpoint to be recovered. Specifically, in the checkpoint loader, after completing the system initialization, the processor core enters the recovery mode, in which all register values can be modified via instructions, regardless of their writable properties defined in the Instruction Set Architecture (ISA). Subsequently, the checkpoint loader begins executing instructions related to status configuration and register recovery.

The stable timer register, which supplies the wall clock time for Linux, is restored. Inaccurate restoration of this register can lead Linux to perceive the current time as earlier or significantly later than the actual last recorded time during subsequent system checks. This discrepancy can induce kernel panic and disrupt the normal operation of programs. Following this, remapping and shuffling settings are configured for the subsequent run-time physical transform mechanism, as detailed in Section 4.2. Next, the program counter (PC) for the first instruction in the checkpoint execution is logged. ISA registers, including control status registers (CSRs) and general purpose registers (GPRs) [25], are then restored by respective instructions. It is noted that since the recovery process is instruction-based, which requires

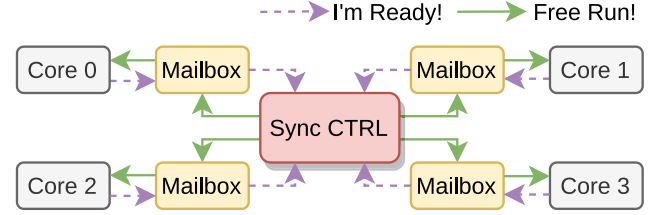


Fig. 6. Sync mechanism for simultaneous multiple single-task checkpoint recovery.

modification of GPR values, GPRs must be restored at the last step. As the speed of checkpoint recovery may vary across different cores, each processor core, upon completing its checkpoint recovery, would sync with other cores, ensuring all cores have finished their recovery. Once synchronization is achieved, a special instruction is executed, enabling all processor cores to exit recovery mode and simultaneously jump to the target PC and initialize their execution. It is required that this special instruction does not modify the values of any registers visible to the ISA. Collectively, when the processor cores begin their execution from the first instruction of the checkpoints, the values in memory, various registers, and the necessary operating system state have all been carefully restored, allowing the program to commence the normal execution.

During the synchronization procedure, processor cores communicate through their mailboxes, which is a hardware mechanism of the asynchronous inter-core communication. Specifically, as shown in Fig. 6, once a processor core completes its checkpoint recovery, it registers its readiness in its own mailbox and waits for a free run signal to trigger its execution. The Sync controller would monitor all cores' mailboxes. Upon detecting all the cores have been prepared, the Sync controller would send the free run message to all cores' mailboxes, thus simultaneously initiating executions of the multiple multi-task workloads.

4.2. Address remapping and shuffling

MultiPoint proposes a software-transparent physical address transform mechanism to support the concurrent and smooth execution of multiple single-task checkpoints. As illustrated in Fig. 7, after the virtual address translation through the Translation Lookaside Buffer (TLB), the cached memory read or write requests issued by the processor core are remapped and then shuffled. For the uncached memory access requests, which typically originate from Linux kernel interactions with I/O devices, their physical addresses are not translated to ensure accurate access to peripherals. Typically, the only peripheral in the pre-silicon performance evaluation is the serial port, which is used for program output printing. It is noted that above address transforms in MultiPoint are hardware-only and software-transparent, thus requiring no special software modifications and imposing no additional requirements for checkpoint capture.

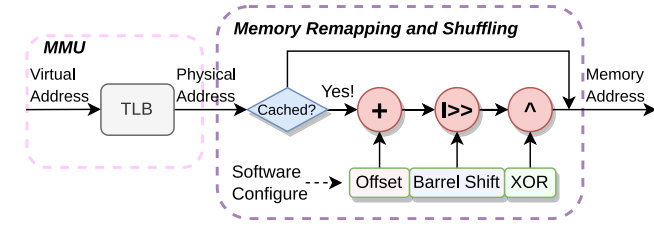


Fig. 7. Mechanism of physical address transforms in MultiPoint. The detailed procedures are presented in Algorithm 1.

Algorithm 1 Procedures of address transforms in MultiPoint

```

1: function LBS(addr, from_bit, to_bit, lbs_bits)
2:   origin  $\leftarrow$  addr[from_bit : to_bit]
3:   shuffled  $\leftarrow$  LEFTBARRELSHIFT(origin, lbs_bits)
4:   addr[from_bit : to_bit]  $\leftarrow$  shuffled
5:   return addr
6: end function
7: function XOR(addr, from_bit, to_bit)
8:   first_slice  $\leftarrow$  addr[start : start + to_bit - from_bit]
9:   addr[from_bit, to_bit]  $\oplus$  = first_slice
10:  return addr
11: end function
12: function REMAPPINGANDSHUFFLING(paddr, core_id)
13:   # Address Remapping
14:   paddr  $\leftarrow$  paddr + core_id  $\times$  (1  $\ll$  task_mem_bit)
15:   # Address Left Barrel Shift
16:   slices  $\leftarrow$  MAX([task_mem_bit - start) / copies], 1)
17:   for slice from 0 to slices do
18:     from_bit  $\leftarrow$  start + slice  $\times$  copies
19:     to_bit  $\leftarrow$  MIN(from_bit + copies, task_mem_bit)
20:     lbs_bits  $\leftarrow$  core_id + slice
21:     paddr  $\leftarrow$  LBS(paddr, from_bit, to_bit, lbs_bits)
22:   end for
23:   # Address Xor
24:   paddr  $\leftarrow$  XOR(paddr, to_bit, total_mem_bit)
25:   return paddr
26: end function

```

The detailed procedures of address remapping and shuffling in MultiPoint are illustrated in Algorithm 1. Specifically, for address remapping, the physical addresses *paddr* of memory requests issued by different cores are added offsets that are multiplied by *core_id* and the memory size for each task (line 14), ensuring that physical addresses across different cores do not overlap with each other. For the address shuffling, the remapped addresses *paddr* starting from the *start* bit are divided into several slices, each containing bit numbers of *copies*. In this algorithm, contiguous physical addresses within the 2^{start} -aligned ranges would retain their continuity after the address shuffling. The *slice* number is calculated by dividing the shuffled address bits by *copies* (line 16). For single-task checkpoint with 4 GB memory (*task_mem_bit* being 32), the shuffled address bits are correspondingly 32 - *start*. Each slice in *paddr* is conducted the left barrel shift (line 21), with the number of shifted bits determined by the sum of the *core_id* and the *slice* index (line 20). The remaining *paddr* bits starting from the boundary of left barrel shift to the end of effective physical address bits are XORed (line 24) with the corresponding bits stating from the *start* bit of *paddr*. Collectively, the physical addresses are remapped and shuffled in such a software-transparent and hardware-friendly manner.

Fig. 8 illustrates the impact of physical address remapping and shuffling when restoring four identical single-task checkpoints. Specifically, Fig. 8 depicts the distributions of physical addresses allocated to different cores. The physical addresses are randomly generated addresses

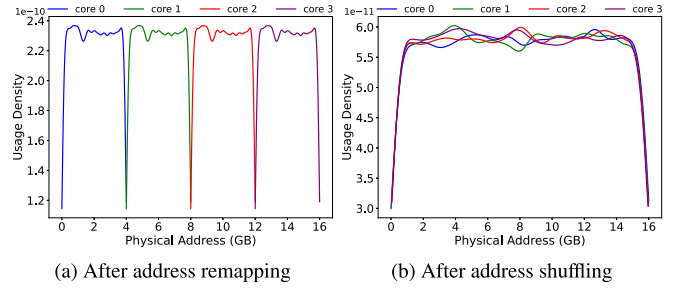


Fig. 8. Illustration of address distributions after remapping and shuffling for randomly generated addresses within [0, 4 GB].

within the range [0, 4 GB]. After address remapping, different cores operate duplicated data within distinct memory regions, as shown in Fig. 8(a). Subsequent address shuffling results in the physical addresses being scattered and interleaved across the memory among different cores, as presented in Fig. 8(b).

It is important to note that the address shuffling algorithm used in MultiPoint is empirically determined, because barrel shift and xor are hardware-friendly operations. Alternate algorithms can also be utilized, provided they ensure that the physical addresses from different cores are effectively interleaved and the guarantee the transforms are one-to-one mappings.

5. Methodology

5.1. Benchmarks

In this work, the SPEC CPU 2017 Rate [9] is employed as the benchmark and is compiled using GCC-13 at the Loongnix system with Linux 4.19. The simulation points of SPEC CPU 2017 are selected by SimPoint, utilizing parameters from previous studies [7,22] with the program interval length *N* being 100 million, the maximum allowed cluster numbers *maxK* being 30, the dimension of random linear projection *dim* being 15, and the Bayesian information criterion threshold *BIC* being 0.95. The checkpoints for these simulation points are captured by the modified system-level QEMU [26,27] with the equipped memory of 4 GB. When dumping the full-system checkpoint, physical pages containing all zero content would be suppressed. Besides, the ISA utilized in this work is the LoongArch [25].

It is noted that MultiPoint is not limited to evaluations of homogeneous workloads. Heterogeneous workloads can also be evaluated by restoring distinct single-task checkpoints at different cores to compose the required multi-task checkpoints determined by the sampling methods. We want to emphasize that MultiPoint is proposed to provide a more scalable alternative to the direct multi-task checkpointing. It is orthogonal to studies on how to sample and select the simulation points for multi-task workloads.

5.2. Metric and evaluation platform

We implemented the evaluation routine of MultiPoint on the Emulator Accelerator and assessed its performance evaluation accuracy of multi-task workloads against its post-silicon Loongson 3A6000 four-core processor [20], whose μ Arch specifications are presented in Table 1. The Emulator Accelerator is a commercial hardware platform designed to accelerate and verify complex chip designs through accurate simulation and real-time debugging. The absolute score error is utilized for evaluating performance estimation accuracy, with the definition given below:

$$\text{Error} = \frac{|Score_{MultiPoint} - Score_{3A6000}|}{Score_{3A6000}} \quad (1)$$

Table 1
Specifications of Loongson 3A6000 processor [20].

Components	Features
Core	4 LA664 Cores, with SMT2
Issue width	6 Insts per Cycle
Function unit	4 Fix, 4 Vec, 4 Mem
Reorder buffer	256 Entries
L1 Cache	64 kB DCache and ICache
L2 Cache	256 kB
Last level cache	16 MB
Main memory	2 Channel, DDR4-3200

Table 2
Memory bandwidth and latency of Emu.(Emulator Accelerator) and its post-silicon 3A6000 processor.

Benchmark	Emu.	3A6000	Error
Stream copy (MB/s)	36 397	35 977	1.17%
Stream scale (MB/s)	26 247	26 402	0.59%
Stream add (MB/s)	28 477	28 163	1.11%
stream triad (MB/s)	29 611	29 674	0.21%
Memory Latency (ns)	91.2	91.8	0.65%

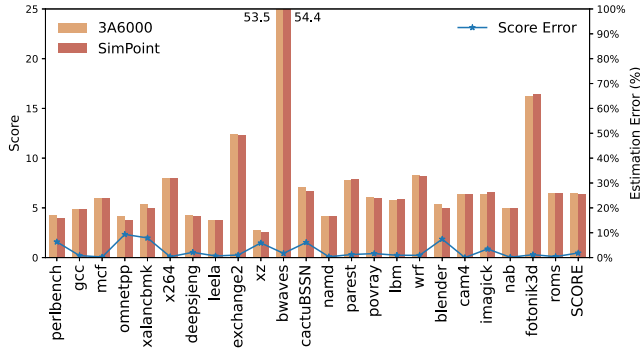


Fig. 9. Scores and estimation errors of SimPoint for each program in SPEC CPU 2017 Rate 1, where the SCORE is the geometric mean of individual program scores.

In SPEC CPU 2017 Rate, the score is calculated by multiplying the Rate numbers and ratios of the evaluated processor's runtime for each benchmark against a reference processor's runtime, which reflects system's throughput and scalability. The final SCORE is calculated by computing the geometric mean of the individual program scores. Due to potential variations in execution speeds among different cores, the runtime reported under Rate N mode is determined by the time taken by the slowest-running core.

The Emulator Accelerator is equipped with 32 GB memory, i.e., `addr_bit` in Algorithm 1 being 35, which enables the concurrent execution of up to eight checkpoint copies. Besides, the parameter `start` is assigned the value of 24, which indicates that contiguous physical addresses within the 16 MB-aligned space would retain their continuity after the address remapping and shuffling. Different parameter values of `start` are discussed in Section 6.3.

To ensure the reliability of the experimental results, we calibrated the Emulator Accelerator against its post-silicon 3A6000 processor. Given that the Emulator Accelerator and the 3A6000 processor share identical core logic, our calibration efforts were concentrated on the memory system. Specifically, we aligned the parameters of the memory controllers between these two systems. The calibration outcomes for their memory system are detailed in Table 2. We evaluated the alignment of their memory systems using the `stream` and `lat_mem_rd` benchmarks to assess memory bandwidth and latency, respectively. The results, as shown in Table 2, demonstrate that the memory systems of the Emulator Accelerator and its post-silicon 3A6000 processor are fundamentally aligned. Furthermore, the estimation accuracy of SimPoint on SPEC CPU 2017 Rate 1 is validated on the calibrated Emulator

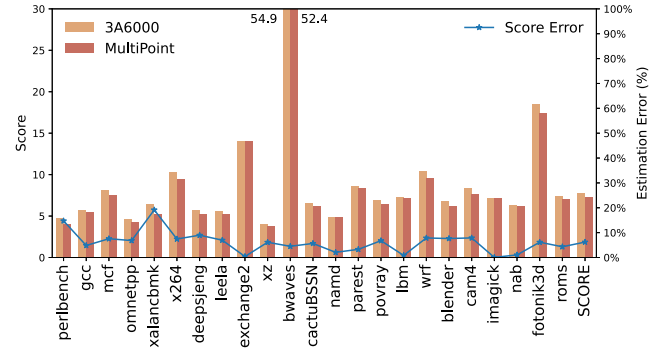


Fig. 10. Scores and errors of SPEC CPU 2017 Rate 2.

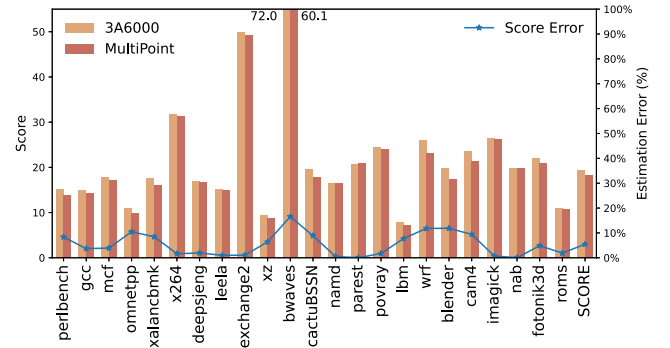


Fig. 11. Scores and errors of SPEC CPU 2017 Rate 4.

Accelerator platform. The corresponding scores and estimation errors are presented in Fig. 9, where SimPoint yields the total score error of only 1.85%, which is consistent with previous studies [8,28].

6. Evaluation

6.1. Analysis of performance evaluation accuracy

We employed the SPEC CPU 2017 Rate 2, Rate 4, and Rate 8 benchmarks to evaluate MultiPoint and assessed its score estimation accuracy against the post-silicon 3A6000 commercial processors. For Rate 2, these workloads are executed on two logic threads of one processor core. For Rate 4, these workloads are executed on four logic threads of four processor cores. For Rate 8, these workloads are executed on eight logic threads of four processor cores. The scores and estimation errors of these multi-task workloads are presented in Figs. 10, 11, and 12, respectively. In these figures, SCORE is the final SPEC CPU score. Specifically, for these multi-task workloads, MultiPoint achieved score errors of 6.20%, 5.45%, and 6.99%, respectively.

It is observed that for most programs in these multi-task workloads, MultiPoint could achieve estimation errors within 10%; while for several programs, such as *xalancbmk*, *wrf*, and *cam4* in Rate 8, MultiPoint yields relatively large estimation errors, with the value of 22.79%, 25.02%, and 15.11%, respectively. Nevertheless, studies [7,29] have illustrated that SimPoint-based method could exhibit stable relative errors across different μ Arch designs. It is important to note that, in pre-silicon μ Arch performance evaluations, the consistency of estimation errors across different μ Arch designs holds greater significance than the magnitude of the error itself [15].

Evaluations across different μ Arch designs. To evaluate the cross- μ Arch consistency of MultiPoint, programs *xalancbmk*, *wrf*, and *cam4* are evaluated at five distinct μ Arch designs and their estimation errors are demonstrated in Fig. 13. These programs are presented because their estimation errors are the highest in the Rate 8 evaluations. Besides

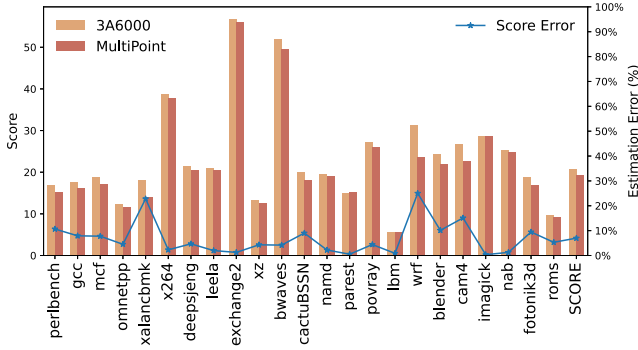


Fig. 12. Scores and errors of SPEC CPU 2017 Rate 8.

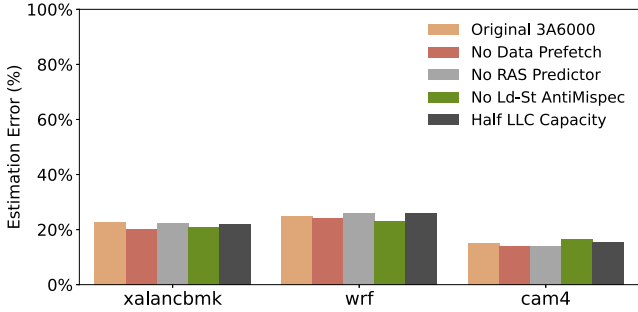
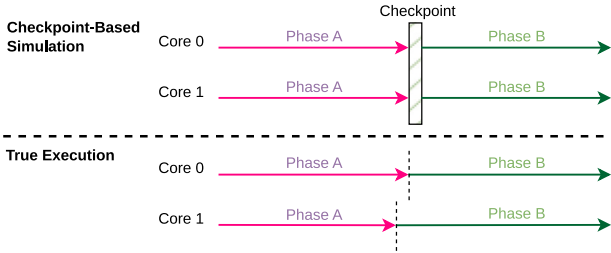
Fig. 13. Score estimation errors across five different μ Arch designs of MultiPoint for programs *xalancbmk*, *wrf*, and *cam4* in SPEC CPU 2017 Rate 8, whose estimation errors are the highest as shown in Fig. 12.

Fig. 14. Illustration of the discrepancy between checkpoint-based simulation and true execution for multi-task workloads.

of the original 3A6000, the introduced μ Arch designs involve various structure changes of (a) turning off data prefetching, (b) turning off branch predictor for return branch, (c) turning off load-store memory dependence prediction, and (d) reducing half of the LLC capacity, respectively. The post-silicon 3A6000 processor is configured these μ Arch changes through firmware modifications. As shown in Fig. 13, MultiPoint yields relatively stable estimation errors for these three programs across these five different μ Arch designs, with the standard variance being 1.04%, 1.13%, and 0.97%, respectively. Besides, the estimation errors of these programs exhibit same-sign bias across different μ Arch changes. To sum up, these consistent biases in estimation errors could enable designers to make correct trade-off decisions in the design space explorations of multi-core processors.

Error Comparisons with single-task workload. It is observed that for many programs, estimation errors in multi-task workloads of Rate 2, 4, and 8 are higher than those in single-task workloads of Rate 1. For example, the estimation errors for *wrf* in Rate 1, 2, 4, and 8 workloads are 0.91%, 7.85%, 11.80%, and 25.02%, respectively. This discrepancy arises from the inherent limitations of checkpoint-based methods, as discussed in previous studies [30–33]. Fig. 14 illustrates

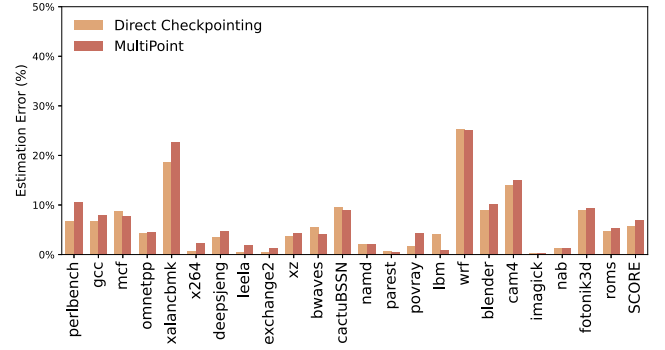


Fig. 15. Comparison of score errors of SPEC CPU 2017 Rate 8 for MultiPoint and direct multi-task checkpointing.

Table 3

Summary on estimation errors of MultiPoint and direct multi-task checkpointing for SPEC CPU 2017 Rate 2/4/8.

Error	MultiPoint	Direct checkpointing
Rate 2	6.20%	5.23%
Rate 4	5.45%	4.98%
Rate 8	6.99%	5.79%
Average	6.21%	5.33%

the discrepancies between checkpoint-based simulation and true execution for multi-task workloads. In realistic executions of multi-task workloads, variability in the execution speeds of individual tasks can occur due to differences in memory response order and memory access latency across cores. Therefore, to make inter-task relative progress independent of specific μ Arch implementations, current methods [34, 35] tried to enforce identical execution speeds across all cores. This approach introduces divergences between the actual runtime state and the initial checkpoint state. Such discrepancies lead to variations in memory access timing and patterns, resulting in different performance behaviors on μ Arch structures, such as the last-level cache and memory controller, during subsequent executions. Consequently, there can be significant performance differences between checkpoint-based simulations and true executions. In contrast, for single-task workloads, where inter-workload relative execution speed is not a concern, the initial state of the checkpoint aligns consistently with the realistic runtime state. As a result, checkpoint-based evaluation methods exhibit lower estimation errors for single-task workloads compared to multi-task workloads. Nevertheless, MultiPoint is still effective in pre-silicon performance evaluations for multi-task workloads, considering that it could yield stable estimation errors across different μ Arch designs, as shown in Fig. 13.

6.2. Comparisons with direct checkpointing

In this subsection, MultiPoint is evaluated against the direct multi-task checkpointing, as introduced in Section 2.2, in terms of their estimation accuracy and overheads.

Estimation Accuracy. The comparisons of score errors of each program in SPEC CPU 2017 Rate 8 for MultiPoint and direct multi-task checkpointing are depicted in Fig. 15, where MultiPoint achieves comparable estimation errors compared to the direct multi-task checkpointing. Specifically, the total score estimation errors of direct checkpointing and MultiPoint are 6.99% and 5.79%, respectively. The comparisons of estimation errors for Rate 2 and Rate 8 exhibit similar tendencies are hence not presented. The summary on estimation errors of MultiPoint and direct multi-task checkpointing for SPEC CPU 2017 Rate is given in Table 3, where their average total score errors are 6.21% and 5.33%, respectively. As demonstrated in Table 3, MultiPoint yields same-level estimation accuracy compared to the direct multi-task

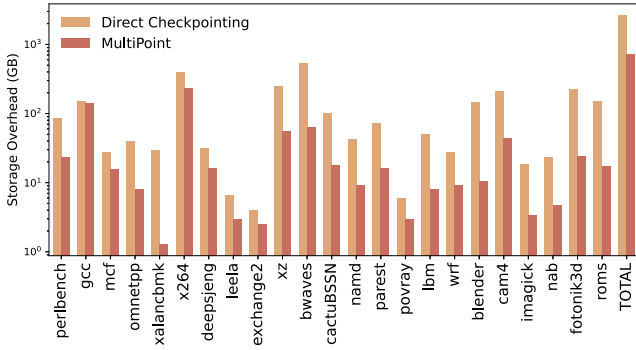


Fig. 16. Storage overheads of SPEC CPU 2017 Rate 8 for MultiPoint and direct multi-task checkpointing, where the y-axis is logarithmically scaled for better visualization.

Table 4

Summary on storage overheads of MultiPoint and direct multi-task checkpointing for SPEC CPU 2017 Rate 2/4/8.

Storage (TB)	MultiPoint	Direct checkpointing
Rate 2	0.71	0.98
Rate 4	0.71	1.51
Rate 8	0.71	2.58
Total	0.71	5.07

checkpointing. The estimation error discrepancies between MultiPoint and direct multi-task checkpointing mainly originate from two issues, as discussed below.

Firstly, while MultiPoint incorporates an address shuffling mechanism, its memory address usage is inevitably not identical that of realistic multi-task workloads. The differences in memory access patterns can result in varying performance behaviors on physical-address-aware μ Arch structures, such as caches, hardware data prefetchers, and memory controllers. Consequently, compared to direct multi-task checkpointing, MultiPoint demonstrates larger but still comparable estimation errors for most programs, as illustrated in Fig. 15. Besides, it is noted that MultiPoint's estimation errors could be improved if more sophisticated address shuffling algorithm is introduced.

Secondly, it is noted that the execution paths of profiling the code signatures and capturing the checkpoints of multi-task workloads are inevitably different due to non-deterministic events, such as random variations in task scheduling and memory accesses [30,32]. Therefore, the captured checkpoints of multi-task workloads do not strictly correspond to their simulation points. In contrast, single-task workloads exhibit more deterministic and repeatable execution paths. MultiPoint constructs the required multi-task checkpoints through multiple single-task checkpoints, which mitigates the misalignment between multi-task simulation points and their checkpoints. Besides, due to the statistical nature [29] of K-Means clustering in SimPoint, the simulation points selected by direct multi-task checkpoint is not identical to that constructed by MultiPoint, thereby introducing random discrepancies in their performance estimations. Collectively, compared to the direct checkpointing, MultiPoint can occasionally produce lower estimation errors for certain programs, such as the *lbm*, *bwaves*, and *mcf*.

Storage Overheads. The comparison of storage overheads of each program in SPEC CPU 2017 Rate 8 for MultiPoint and direct multi-task checkpointing are presented in Fig. 16, where the y-axis is logarithmically scaled for better visualization. Specifically, their total storage requirements are 0.71 TB and 2.58 TB, respectively. It is noted that MultiPoint significantly reduces the storage overheads by 72.5%. This is attributed to that MultiPoint composes the multi-task checkpoint through the combination of multiple single-task checkpoints, therefore its storage overheads of SPEC CPU 2017 Rate 8 is the same as Rate 1, which is much smaller. In contrast, direct multi-task checkpointing requires dumping the real run-time multi-task checkpoint, thus

Table 5

Summary on time overheads of checkpoint capture by MultiPoint and direct multi-task checkpointing for workloads of SPEC CPU 2017 Rate 2, 4, and 8.

Time (h)	MultiPoint	Direct checkpointing
Rate 2	9.0	19.4
Rate 4	9.0	37.5
Rate 8	9.0	85.9
Total	9.0	142.8

necessitating much more storage overheads.

Moreover, when evaluating different Rate N workloads, direct checkpointing requires distinct checkpoints for different rate number, resulting in redundant storage overheads. In contrast, for MultiPoint, when evaluating different Rate N workloads, there is no necessity to recapture respective checkpoints, thus avoiding the extra computing resources, timing, and storage overheads. For SPEC CPU Rate benchmark, the storage overhead of MultiPoint is always that of the Rate 1 workload, irrespective of how many different Rate numbers to be evaluated. The summary on storage overheads of MultiPoint and direct multi-task checkpointing for SPEC CPU 2017 Rate is presented in Table 4, where their total storage requirements are 0.71 TB and 5.07 TB, respectively. For these workloads, MultiPoint substantially decreases the storage requirements by 86.0%, while their estimation accuracy is comparable.

Time Overheads. Table 5 summarizes the time overheads for checkpoint capture using MultiPoint and direct multi-task checkpointing for workloads of SPEC CPU 2017 Rate 2, 4, and 8. It is noted that checkpoints are captured concurrently, with the total checkpoint capture time determined by the runtime of the longest-running program. Specifically, the total time overheads are 9.0 h and 142.8 h for MultiPoint and direct multi-task checkpointing, respectively. For these workloads, MultiPoint reduces the time overheads by 93.7% while maintaining comparable performance evaluation accuracy. This significant reduction is attributed to the property that MultiPoint's checkpoint capture time is independent of the Rate numbers being evaluated. By constructing the required multi-task checkpoints from multiple single-task checkpoints, MultiPoint eliminates the need for additional checkpoint capture time for workloads of Rate 2, 4, and 8, once the Rate 1 checkpoints have been captured. In contrast, for direct multi-task checkpointing, the capture time for Rate N increases linearly with the Rate number under the QEMU *icount* mode [26].

To sum up, MultiPoint achieves larger yet still comparable estimation accuracy compared to the direct multi-task checkpointing. However, MultiPoint achieves agile evaluations with substantially 86.0% fewer storage and 93.7% less timing overheads. Compared to direct multi-task checkpointing, MultiPoint enables more scalable performance evaluations for multi-task workloads.

6.3. Discussions of address shuffling

In this subsection, we analyzed the benefits and parameter sensitivity of introduced address shuffling on the accuracy of performance evaluation for multi-task workloads.

Benefits of address shuffling. Fig. 17 illustrates the estimation errors of MultiPoint with and without the address shuffling on the SPEC CPU 2017 Rate 8. The comparisons of estimation errors for Rate 2 and Rate 8 exhibit similar tendencies and are hence not presented. For MultiPoint without implementing address shuffling, only the address remapping is conducted to guarantee the normal concurrent execution of evaluated multi-task workloads. Specifically, as depicted in Fig. 17, incorporating address shuffling could markedly diminish the evaluation errors, significantly reducing total score errors from 43.60% to 6.99%. The error reductions are especially pronounced for programs like *bwaves*, *povray*, and *parast*, whose errors are significantly decreased by 76.33%, 73.30%, and 67.41%, respectively. It

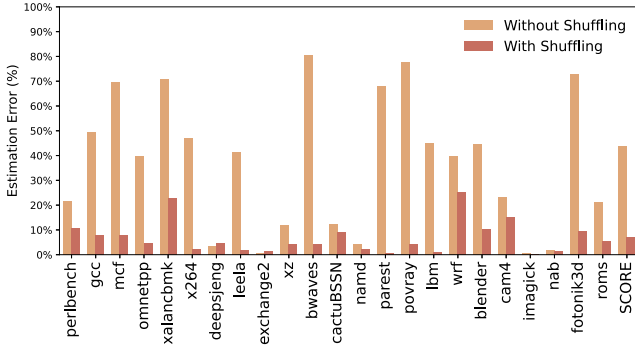


Fig. 17. Comparison of score estimation errors of SPEC CPU 2017 Rate 8 for MultiPoint **with** and **without** address shuffling.

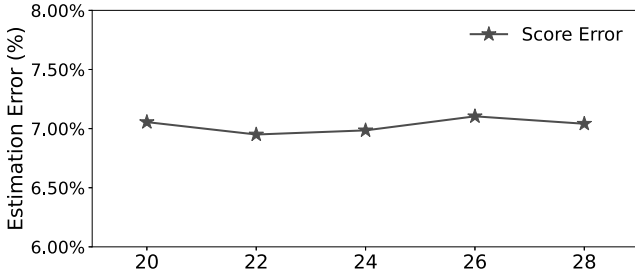


Fig. 18. Score estimation errors of SPEC CPU 2017 Rate 8 under different values of parameter *start* in Algorithm 1, which determines the granularity of address shuffling.

is noted that these programs own the common feature of extensive memory access demands. In contrast, for the program *deepsjeng* and *exchange2*, which exhibit moderate memory access demands, their estimation errors remained unaffected by the address shuffling.

The substantial decreases in estimation errors after introducing address shuffling are because that different physical memory address characteristics can impact the performance of certain μ Arch structures that are sensitive to the physical addresses. For instance, under homogeneous workloads, like SPEC CPU 2017 Rate, if address shuffling is not implemented, the low bits of address accessed by different single-task checkpoints for semantically identical memory are identical, with only their highest bits differing. This would result in significant cache set conflicts in the shared last-level cache. In contrast, in the actual multi-task workload executions, the physical addresses of memory requests issued by different cores are interleaved with each other. This allows the memory requests to be more evenly distributed across different cache sets, thereby resulting in fewer set conflicts. Fewer cache set conflicts would bring fewer cache misses and mitigate the performance degradation associated with the serial operations required by the cache coherence protocol for memory requests within the same cache set.

Sensitivities of Shuffling Algorithm. Fig. 18 presents the final score estimation errors of SPEC CPU 2017 Rate 8 under different values of the parameter *start* in Algorithm 1. The parameter *start* determines the granularity of address shuffling, where contiguous physical addresses within 2^{start} -aligned ranges maintain their continuity after shuffling. As shown in Fig. 18, the final score estimation errors exhibit low sensitivity to the specific value of *start*. In this work, the default value of *start* is set to 24 because its barrel shift bits, calculated as $32 - 24 = 8$, are divisible by 2, 4, and 8, which are the evaluated Rate numbers. This selection simplifies boundary condition handling in RTL coding, considering that different *start* values yield comparable results.

Collectively, Figs. 17 and 18 demonstrate that the address shuffling mechanism significantly influences the accuracy of performance

evaluation, while the specific values of *start* have only a minor impact on estimation accuracy. In essence, the coarse-grained isolation of memory usage illustrated in Fig. 8(a) fundamentally differs from the interleaved characteristics of real multi-task workloads depicted in Fig. 3. However, when the physical addresses of different tasks become interleaved, the granularity of interleaving has only a limited effect on the final performance outcomes. In this study, the address shuffling algorithm in MultiPoint is empirically selected, as barrel shift and XOR operations are hardware-friendly. Despite this, current method is effective in producing comparable evaluation accuracy to direct multi-task checkpointing. Exploring improved shuffling algorithms is a promising direction for further reducing the estimation error gap between MultiPoint and direct multi-task checkpointing.

7. Related work

Evaluation for Single-Task Workloads: There are mainly two categories of sampling methods that are widely used for single-task workload, or the single-threaded programs. **Representative Sampling** [1–8]: This method leverages the recurrent phases exhibited in the program’s dynamic execution and selectively chooses several intervals to represent and reconstruct the complete execution behavior of the program. **Systematic Sampling** [36–38]: This method involves systematically extracting many short program intervals to collectively represent the behavior of the entire program. This approach could statistically ensure a broad coverage of the program’s behavior, making it possible to estimate the overall performance more accurately.

Evaluation for Multi-Task Workloads: The multi-task workloads are primarily categorized into two types: homogeneous workloads, where different cores execute the same program, as exemplified by SPEC CPU 2017 Rate; and heterogeneous workloads, where different cores execute distinct workloads. For **homogeneous workloads**, Perelman et al. [15] introduced the parallel SimPoint method for efficient performance evaluation. For **heterogeneous workloads**, Jacobvitz et al. [10] provided a rigorous definition for benchmark evaluation purposes. Velázquez et al. [11] proposed a method involving random combinations of task loads to construct representative heterogeneous workloads. Eyerman et al. [12,39] developed several system-level performance evaluation metrics for multi-task workloads. Van et al. [19], NamKung et al. [14], Tawk et al. [17] introduced a series of sampling methods for heterogeneous workloads aimed at reducing the number of phase combinations needed for simulation, thereby enhancing simulation efficiency. Prieto et al. [40,41] proposed extracting core loops from programs and established a statistical model to validate the consistency.

Evaluation for Multi-Threaded Workloads: Because of the synchronization and communication among threads, the dynamic instruction counts of multi-threaded workloads can fluctuate significantly with each execution [42], and the execution paths are unpredictable [43]. Besides, the relative execution relationship among threads is microarchitecture-dependent [30,31]. These factors introduce many challenges to the performance evaluation of the multi-threaded workloads. Currently, pre-silicon performance evaluation methods for multi-threaded workloads are divided into three categories: **Timing-based Sampling** [30,31] methods periodically switch processor cores between fast-forward and detailed simulation states until the program execution concludes. **Communication primitive-based Sampling** [32, 44,45] involves segmenting the program based on statements like barriers, loops, and tasks, thereby obtaining simulation points that are naturally independent of the program’s execution path. **Work-based Sampling** [33–35] divides the program according to the total effective works so as to acquire simulation points that are independent of the dynamic instruction counts of the program.

8. Conclusion

In this work, we propose MultiPoint to enable pre-silicon performance evaluation for multi-task workloads. The key idea of MultiPoint is to construct the required multi-task workloads by combining multiple single-task workloads. To guarantee the smooth concurrent execution of these single-task checkpoints, MultiPoint introduces the mechanisms of address remapping. Furthermore, to maintain the accuracy of performance evaluation, MultiPoint shuffles memory requests from different cores to make them scatter and interleave in the memory space. MultiPoint is evaluated on the SPEC CPU 2017 and yields score estimation errors of 6.20%, 5.45%, and 6.99% for Rate 2, Rate 4, and Rate 8, respectively, achieving comparable performance evaluation accuracy compared to direct multi-task checkpointing but in a more scalable manner with substantially 86.0% lower storage and 93.7% less time overheads.

CRedit authorship contribution statement

Chenji Han: Writing – original draft, Methodology, Investigation. **Xinyu Li:** Writing – review & editing, Methodology. **Feng Xue:** Writing – original draft, Methodology. **Weitong Wang:** Writing – review & editing, Methodology. **Yuxuan Wu:** Writing – review & editing, Methodology. **Wenxiang Wang:** Writing – review & editing, Methodology. **Fuxin Zhang:** Writing – review & editing, Methodology, Funding acquisition.

Declaration of competing interest

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

References

- [1] Timothy Sherwood, Erez Perelman, Greg Hamerly, Brad Calder, Automatically characterizing large scale program behavior, in: ASPLOS X, 2002.
- [2] Jose Renau, Fangping Liu, Hongzhang Shan, Sang Wook Stephen Do, Enabling reduced simpoint size through LiveCache and detail warmup, *BenchCouncil Trans. Benchmarks Stand. Eval.* 2 (4) (2022) 100082.
- [3] Harish Patil, Alexander Isaev, Wim Heirman, Alen Sabu, Ali Hajiabadi, Trevor E Carlson, ELFies: executable region checkpoints for performance analysis and simulation, in: 2021 IEEE/ACM International Symposium on Code Generation and Optimization, CGO, IEEE, 2021, pp. 126–136.
- [4] Odysseas Chatzopoulos, Maria Trakosa, George Papadimitriou, Wing Shek Wong, Dimitris Gizopoulos, SimPoint-based microarchitectural hotspot & energy-efficiency analysis of RISC-v OoO CPUs, in: 2024 IEEE International Symposium on Performance Analysis of Systems and Software, ISPASS, IEEE, 2024, pp. 1–12.
- [5] Charles Yount, Harish Patil, Mohammad S Islam, Aditya Srikanth, Graph-matching-based simulation-region selection for multiple binaries, in: 2015 IEEE International Symposium on Performance Analysis of Systems and Software, ISPASS, IEEE, 2015, pp. 52–61.
- [6] Kenneth Hoste, Lieven Eeckhout, Microarchitecture-independent workload characterization, *IEEE Micro* 27 (3) (2007) 63–72.
- [7] Weihua Zhang, Jiaxin Li, Yi Li, Haibo Chen, Multilevel phase analysis, *ACM Trans. Embed. Comput. Syst.* 14 (2015) 31:1–31:29.
- [8] Hongwei Cui, Yujie Cui, Honglan Zhan, Shuhao Liang, Xianhua Liu, Chun Yang, Xu Cheng, MBAPIS: Multi-level behavior analysis guided program interval selection for microarchitecture studies, in: 2023 32nd International Conference on Parallel Architectures and Compilation Techniques, PACT, IEEE, 2023, pp. 297–308.
- [9] James Bucek, Klaus-Dieter Lange, Jákím v. Kistowski, SPEC CPU2017: Next-generation compute benchmark, in: Companion of the 2018 ACM/SPEC International Conference on Performance Engineering, ICPE '18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 41–42.
- [10] Adam N. Jacobvitz, Andrew D. Hilton, Daniel J. Sorin, Multi-program benchmark definition, in: 2015 IEEE International Symposium on Performance Analysis of Systems and Software, ISPASS, IEEE, 2015, pp. 72–82.
- [11] Ricardo A. Velásquez, Pierre Michaud, André Seznec, Selecting benchmark combinations for the evaluation of multicore throughput, in: 2013 IEEE International Symposium on Performance Analysis of Systems and Software, ISPASS, 2013, pp. 173–182.
- [12] Stijn Eyerman, Pierre Michaud, Wouter Rogiest, Multiprogram throughput metrics: A systematic approach, *ACM Trans. Archit. Code Optim.* (TACO) 11 (3) (2014) 1–26.
- [13] Jovan Stojkovic, Chunao Liu, Muhammad Shahbaz, Josep Torrellas, μ Manycore: A cloud-native CPU for tail at scale, in: Proceedings of the 50th Annual International Symposium on Computer Architecture, 2023, pp. 1–15.
- [14] Jeffrey Namkung, Dohyun Kim, Rajesh Gupta, Igor Kozintsev, Jean-Yves Bouget, Carole Dulong, Phase guided sampling for efficient parallel application simulation, in: Proceedings of the 4th International Conference on Hardware/Software Codesign and System Synthesis, 2006, pp. 187–192.
- [15] Erez Perelman, Marzia Polito, J-Y Bouguet, Jack Sampson, Brad Calder, Carole Dulong, Detecting phases in parallel applications on shared memory architectures, in: Proceedings 20th IEEE International Parallel & Distributed Processing Symposium, IEEE, 2006, pp. 10–pp.
- [16] Alvaro Wong, Dolores Rexachs, Emilio Luque, Extraction of parallel application signatures for performance prediction, in: 2010 IEEE 12th International Conference on High Performance Computing and Communications, HPCC, IEEE, 2010, pp. 223–230.
- [17] Melhem Tawk, Khaled Z. Ibrahim, Smail Niar, Multi-granularity sampling for simulating concurrent heterogeneous applications, in: Proceedings of the 2008 International Conference on Compilers, Architectures and Synthesis for Embedded Systems, 2008, pp. 217–226.
- [18] Melhem Tawk, Khaled Z. Ibrahim, Smail Niar, Adaptive sampling for efficient mpoc architecture simulation, in: 2007 15th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, IEEE, 2007, pp. 186–192.
- [19] M. Van Biesbrouck, Lieven Eeckhout, Brad Calder, Considering all starting points for simultaneous multithreading simulation, in: 2006 IEEE International Symposium on Performance Analysis of Systems and Software, IEEE, 2006, pp. 143–153.
- [20] Loongson, Loongson LS3a6000 processor, 2024, URL <https://www.loongson.cn/EN/product/show?id=11>.
- [21] Yinan Xu, Zihao Yu, Dan Tang, Guokai Chen, Lu Chen, Lingrui Gou, Yue Jin, Qianruo Li, Xin Li, Zuojun Li, Jiawei Lin, Tong Liu, Zhigang Liu, Jiazhan Tan, Huaqiang Wang, Huizhe Wang, Kaifan Wang, Chuanqi Zhang, Fawang Zhang, Linjuan Zhang, Zifei Zhang, Yangyang Zhao, Yaoyang Zhou, Yike Zhou, Jiangrui Zou, Ye Cai, Dandan Huan, Zulong Li, Jiye Zhao, Zihao Chen, Wei He, Qiyuan Quan, Xingwu Liu, Sa Wang, Kan Shi, Ninghui Sun, Yungang Bao, Towards Developing High Performance RISC-V Processors Using Agile Methodology, in: 2022 55th IEEE/ACM International Symposium on Microarchitecture, MICRO, 2022, pp. 1178–1199.
- [22] Brian Grayson, Jeff Rupley, Gerald D. Zuraski, Eric Quinell, Daniel A. Jiménez, Tarun Nakra, P. W. Kitchen, Ryan Hensley, Edward Brekelbaum, Vikas Sinha, Ankit Ghiya, Evolution of the samsung exynos CPU microarchitecture, in: 2020 ACM/ IEEE 47th Annual International Symposium on Computer Architecture, ISCA, 2020, pp. 40–51.
- [23] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R.J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E.A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, SciPy 1.0 Contributors, SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, *Nature Methods* 17 (2020) 261–272, <http://dx.doi.org/10.1038/s41592-019-0686-2>.
- [24] Mel Gorman, Physical page allocation, 2013, URL <https://www.kernel.org/doc/gorman/html/understand/understand009.html>. (Accessed 24 December 2024).
- [25] Loongson Technology, LoongArch documentation, 2022, URL <https://loongson.github.io/LoongArch-Documentation/>.
- [26] Fabrice Bellard, QEMU, a fast and portable dynamic translator., in: USENIX Annual Technical Conference, FREENIX Track, Vol. 41, California, USA, 2005, p. 46.
- [27] Xinyu Li, Plugin for checkpoint dumping in QEMU, 2024, URL https://github.com/rrwhx/qemu_plugins_loongarch.
- [28] Björn Gottschall, Silvio Campelo de Santana, Magnus Jahre, Balancing accuracy and evaluation overhead in simulation point selection, in: 2023 IEEE International Symposium on Workload Characterization, IISWC, IEEE, 2023, pp. 43–53.
- [29] Erez Perelman, Greg Hamerly, Brad Calder, Picking statistically valid and early simulation points, in: 2003 12th International Conference on Parallel Architectures and Compilation Techniques, 2003, pp. 244–255.
- [30] Ehsan K. Ardestani, Jose Renau, ESEC: A fast multicore simulator using time-based sampling, in: 2013 IEEE 19th International Symposium on High Performance Computer Architecture, HPCA, 2013, pp. 448–459.
- [31] Trevor E. Carlson, Wim Heirman, Lieven Eeckhout, Sampled simulation of multithreaded applications, in: 2013 IEEE International Symposium on Performance Analysis of Systems and Software, ISPASS, 2013, pp. 2–12.

- [32] Trevor E Carlson, Wim Heirman, Kenzo Van Craeynest, Lieven Eeckhout, Barrierpoint: Sampled simulation of multi-threaded applications, in: 2014 IEEE International Symposium on Performance Analysis of Systems and Software, ISPASS, IEEE, 2014, pp. 2–12.
- [33] Changxi Liu, Alen Sabu, Akanksha Chaudhari, Qingxuan Kang, Trevor E. Carlson, Pac-Sim: Simulation of multi-threaded workloads using intelligent, live sampling, *ACM Trans. Arch. Code Optim.* (2024).
- [34] Alen Sabu, Harish Patil, Wim Heirman, Trevor E. Carlson, LoopPoint: Checkpoint-driven sampled simulation for multi-threaded applications, in: 2022 IEEE International Symposium on High-Performance Computer Architecture, HPCA, 2022, pp. 604–618.
- [35] Alen Sabu, Changxi Liu, Trevor E. Carlson, Viper: Utilizing hierarchical program structure to accelerate multi-core simulation, *IEEE Access* (2024).
- [36] Sina Hassani, Gabriel Southern, Jose Renau, LiveSim: Going live with microarchitecture simulation, in: 2016 IEEE International Symposium on High Performance Computer Architecture, HPCA, 2016, pp. 606–617.
- [37] Roland E. Wunderlich, Thomas F. Wenisch, Babak Falsafi, James C. Hoe, SMARTS: accelerating microarchitecture simulation via rigorous statistical sampling, in: 30th Annual International Symposium on Computer Architecture, 2003. Proceedings, 2003, pp. 84–95.
- [38] Uday Kumar Reddy Vengalam, Anshujit Sharma, Michael C. Huang, LoopIn: A loop-based simulation sampling mechanism, in: 2022 IEEE International Symposium on Performance Analysis of Systems and Software, ISPASS, 2022, pp. 224–226.
- [39] Stijn Eyerman, Lieven Eeckhout, System-level performance metrics for multiprogram workloads, *IEEE Micro* 28 (3) (2008) 42–53.
- [40] Pablo Prieto, Pablo Abad, Jose Angel Herrero, Jose Angel Gregorio, Valentin Puente, SPECcast: A methodology for fast performance evaluation with SPEC cpu 2017 multiprogrammed workloads, in: Proceedings of the 49th International Conference on Parallel Processing, 2020, pp. 1–11.
- [41] Pablo Prieto, Pablo Abad, Jose Angel Gregorio, Valentin Puente, Fast, accurate processor evaluation through heterogeneous, sample-based benchmarking, *IEEE Trans. Parallel Distrib. Syst.* 32 (12) (2021) 2983–2995.
- [42] Alaa R. Alameldeen, David A. Wood, IPC considered harmful for multiprocessor workloads, *IEEE Micro* 26 (4) (2006) 8–17.
- [43] Weihua Zhang, Xiaofeng Ji, Bo Song, Shiqiang Yu, Haiibo Chen, Tao Li, Pen-Chung Yew, Wenyun Zhao, Varcatcher: A framework for tackling performance variability of parallel workloads on multi-core, *IEEE Trans. Parallel Distrib. Syst.* 28 (4) (2016) 1215–1228.
- [44] Miguel Tairum Cruz, Sascha Bischoff, Roxana Rusitoru, Shifting the barrier: extending the boundaries of the barrierpoint methodology, in: 2018 IEEE International Symposium on Performance Analysis of Systems and Software, ISPASS, IEEE, 2018, pp. 120–122.
- [45] Thomas Grass, Trevor E Carlson, Alejandro Rico, German Ceballos, Eduard Ayguade, Marc Casas, Miquel Moreto, Sampled simulation of task-based programs, *IEEE Trans. Comput.* 68 (2) (2018) 255–269.