IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 26, 2020, accepted January 17, 2021, date of publication February 1, 2021, date of current version February 9, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3055858

Empirical Investigation About the Factors Affecting the Cost Estimation in Global Software Development Context

JUNAID ALI KHAN, SAIF UR REHMAN KHAN, JAVED IQBAL, AND INAYAT UR REHMAN
Department of Computer Science, COMSATS University Islamabad, Islamabad 42000, Pakistan

Corresponding author: Junaid Ali Khan (junaidalikhan17@gmail.com)

The work of Junaid Ali Khan was supported by the COMSATS University Islamabad, Pakistan.

ABSTRACT Software organization always aims at developing a quality software product using the estimated development resources, effort, and time. Global Software Development (GSD) has emerged as an essential tool to ensure optimal utilization of resources, which is performed in globally distributed settings in various geographical locations. Global software engineering focuses on reducing the cost, increasing the development speed, and accessing skilled developers worldwide. Estimating the required amount of resources and effort in the distributed development environment remains a challenging task. Thus, there is a need to focus on cost estimation models in the GSD context. We nevertheless acknowledge that several cost estimation techniques have been reported. However, to the best of our knowledge, the existing cost estimation techniques/models lack considering the additional cost drivers required to compute the accurate cost estimation in the GSD context. Motivated by this, the current work aims at identifying the other cost drivers that affect the cost estimation in the context of GSD. To achieve the targeted objectives, current state-of-the-art related to existing cost estimation techniques of GSD is reported. We adopted SLR and Empirical approach to address the formulated research questions. The current study also identifies the missing factors that would help the practitioners improve the cost estimation models. The results indicate that previously conducted work ignores the additional elements necessary for the cost estimation in the GSD context. Moreover, the current work proposes a conceptual cost estimation model tailored to fit the GSD context.

INDEX TERMS Global software development, distributed development, cost estimation, systematic review.

I. INTRODUCTION

The Globalization of software companies is increasing rapidly. Many software industries are trying to adopt it due to the advantages that it provides. As technology advances and new communication mediums are introduced, the development's Globalization also emerged [1]. This emergence of Globalization increases global software development projects. However, the studies predict that the number of offshore projects will increase with time over time. Global Software Development (GSD) projects are expected to grow from 20 to 30% in countries, including India and China. Many western software industries are developing in Eastern Europe and Asia due to the lower labor rates in these countries. There are many other reasons for adopting GSD, such as improving time to market through time zone differences, using virtual teams with vast skills [1], [2].

However, there are different reasons for adopting GSD that fit the purpose, but this research mainly focuses on the lower cost, which is among the most crucial factors. Globalization's primary goal is to lower the cost of development; this is considered the primary justification for not developing locally. It can be misleading if we do not consider the challenges of this type of development, i.e., the difference in time zone and culture [3]. Thus, it can also take additional time and effort if we do not consider GSD's factors. One of the crucial issues is to estimate the effort and cost in GSD [4] to assess whether this will benefit us or attain significance through local production. There are many tools and techniques available for the estimation of the cost. Many models are developed before the GSD concept, so these techniques lack the factors and the cost drivers associated with this development [5]. We are uncertain about the applicability of these techniques in GSD. However,

The associate editor coordinating the review of this manuscript and approving it for publication was Pinjia Zhang.

if we identify the factors that influence GSD estimation, we can only crosscheck those with the existing techniques for the amplification.

The cost is the main driving factor for any project only if it is done correctly. For software development, effective investment is achieved when it is accurately estimated [3]. The basic idea of GSD is described in Figure 1.

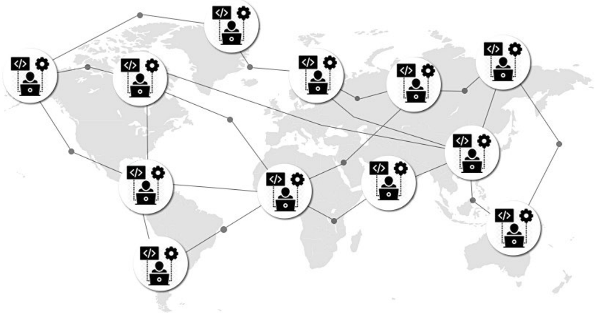


FIGURE 1. Basic Idea of Global Software Development.

Figure 1 represents the virtual teams working from remote locations and collaborating the communication infrastructures to achieve the end product. GSD's primary purpose is cost-effectiveness, in which the base organization hires developers from countries with low labor rates. GSD is widely used for the cost-effectiveness objective. However, many project management challenges are associated with planning, executing, and managing disperses resources [6].

Cost estimation is one of the primary issues for project managers. It is even more crucial in the GSD context, where the task allocation and predicting the resources are way more difficult due to its dispersed nature [7]. The current state-of-the-art lacks in considering additional cost drivers effective to enhance the estimation accuracy for GSD. Motivated by this, current work aims at identifying the cost drivers that affect the cost estimation process in GSD. Evaluating these cost drivers can improve GSD's cost estimation process and help practitioners handle these cost drivers through implementation. Thus, this research aims at addressing the following research objectives:

RO1: To extensively review the factors affecting cost estimation in the GSD context

RO1.1: To identify the key categories of cost drivers in GSD

RO1.2: To obtain the industrial perspective regarding the identified cost drivers

RO1.3: To identify the critical cost drivers that affect the cost estimation process in GSD

RO2: To identify the supporting metrics or techniques for cost estimation in the GSD context

RO2.1: To identify the cost estimation models from the practitioner's perspective

RO3: To analyze the shortcomings of the existing cost estimation techniques in GSD

TABLE 1. Sections Distribution.

Section(s)	Agenda
Section 2	Research Motivation
Section 3	Research Methodology
Section 4	Results and Findings
Section 5	Discussion (Summarized Results)
Section 6	Proposed Conceptual Model
Section 7	Future Directions
Section 8	Threat to Validity
Section 9	Conclusion

As the current state of the art lacks in the identification and categorization of the cost drivers of GSD, our research contributes to the identification of the hidden cost drivers that are not considered in the estimation process of GSD projects. Therefore, this research focuses on providing a thematic taxonomy of cost estimation's identified factors in the GSD context. Meanwhile, based on the results and analysis, a conceptual model is developed to assist practitioners in estimating in the GSD context.

The remaining of work is categorized as follows:

II. RESEARCH MOTIVATION

Many researchers focused on developing the techniques and models for cost estimation in the GSD context. The estimation techniques are high-level approaches adopted to estimate a project, i.e., automated, semi-automated, model-driven, or regression-based [4]. Simultaneously, the estimation models are more specified, corresponding to the particular mechanism for accurate estimation like COCOMO II based, cobra based, or machine learning-based [5]. Based on selected estimation techniques, the estimation models are created. The reported cost estimation models could assist the practitioners through improved estimates as they are amplified for GSD's need. The models include Cost overhead [8], SOCEM [9], and Analogy-based [10]. However, regardless of these various models, we are not achieving the desired results. The author [11] justifies that we're still lacking in considering the cost drivers that affect GSD's overall cost. The attention given to cost estimation in the GSD context is still limited.

The challenges in collocated and distributed development are different due to the characteristics that these development types exhibit. Moreover, project management challenges are of great importance because the challenges associated with project management can directly impact the overall project, and their negative consequences could lead to project failure. The initial challenges that we counter in any project are related to cost estimation. In this phase, we calculate the project profit by estimating the resources and the time required to complete the desired task. The existing studies do not explicitly mention the cost drivers that could impact a GSD project.

Jain and Suman [12] reported more GSD complications, i.e., geographic dispersion, time zone difference, competence level, and hidden cost. The author [12] termed it as hidden

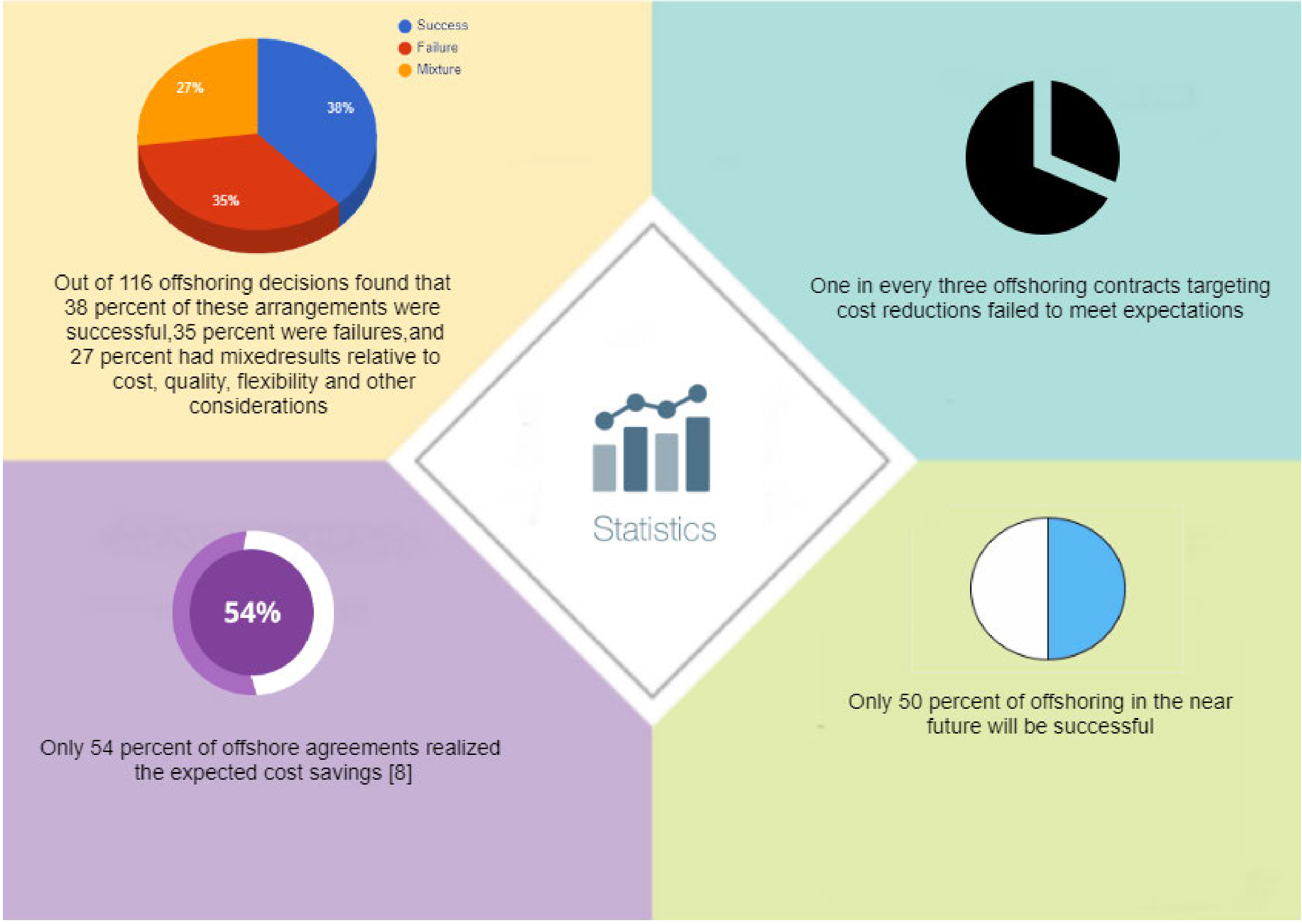


FIGURE 2. **Statistics regarding Cost of Offshoring Projects [5].**

cost drivers because these factors are not considered during the estimation but affect the project in terms of cost, time, or effort. So, it is suggested to identify the hidden cost drivers and consider them in cost estimation to achieve precise and accurate results. If neglected, then these cost drivers could be problematic for the project.

Prikladnicki *et al.* [13] reported that the companies adopted GSD to lower the cost. Still, they do not meet the expectations because the additional factors of GSD are not being considered. Figure 2 represents some statistics regarding the cost of Offshoring that how these additional factors could be misleading for the overall development if not properly analyzed and evaluated. The statistics are extracted from the existing literature [5].

The statistics depicted in Figure 2 provided us motivation to work in this area and to improve the aspects that are lacking to save the cost of Offshoring. In summary, little research has been carried out in the targeted context, and no study has been found that has identified the additional cost drivers of GSD systematically. However, it is of great importance to identify the factors that affect the cost of GSD projects. By classifying these cost drivers, we would assist the practitioners in accurately estimating the cost. Moreover, this classification of cost drivers could serve as a guideline for the project managers to estimate the GSD projects by considering the critical cost drivers.

III. RESEARCH METHODOLOGY

To achieve the targeted research objective, we adopted the Systematic Literature Review (SLR) technique to identify the cost drivers and conducted an empirical study to validate identified cost drivers. Notice that SLR is different from a literature review because it is performed planned and systematically. The results obtained through SLR are more comprehensive as compared to an informal literature review. SLR helps us identify the factors systematically in an unbiased manner, evaluate the specified results, and categorize the elements accordingly [14], [15]. To conduct the SLR efficiently, the guidelines presented by Kitchenhem [16] are followed. SLR consists of three main phases, “Planning the review,” “Conducting the review,” and “Reporting the review.” All these phases are briefly discussed in the next sections. Figure 3 illustrates the adopted research methodology.

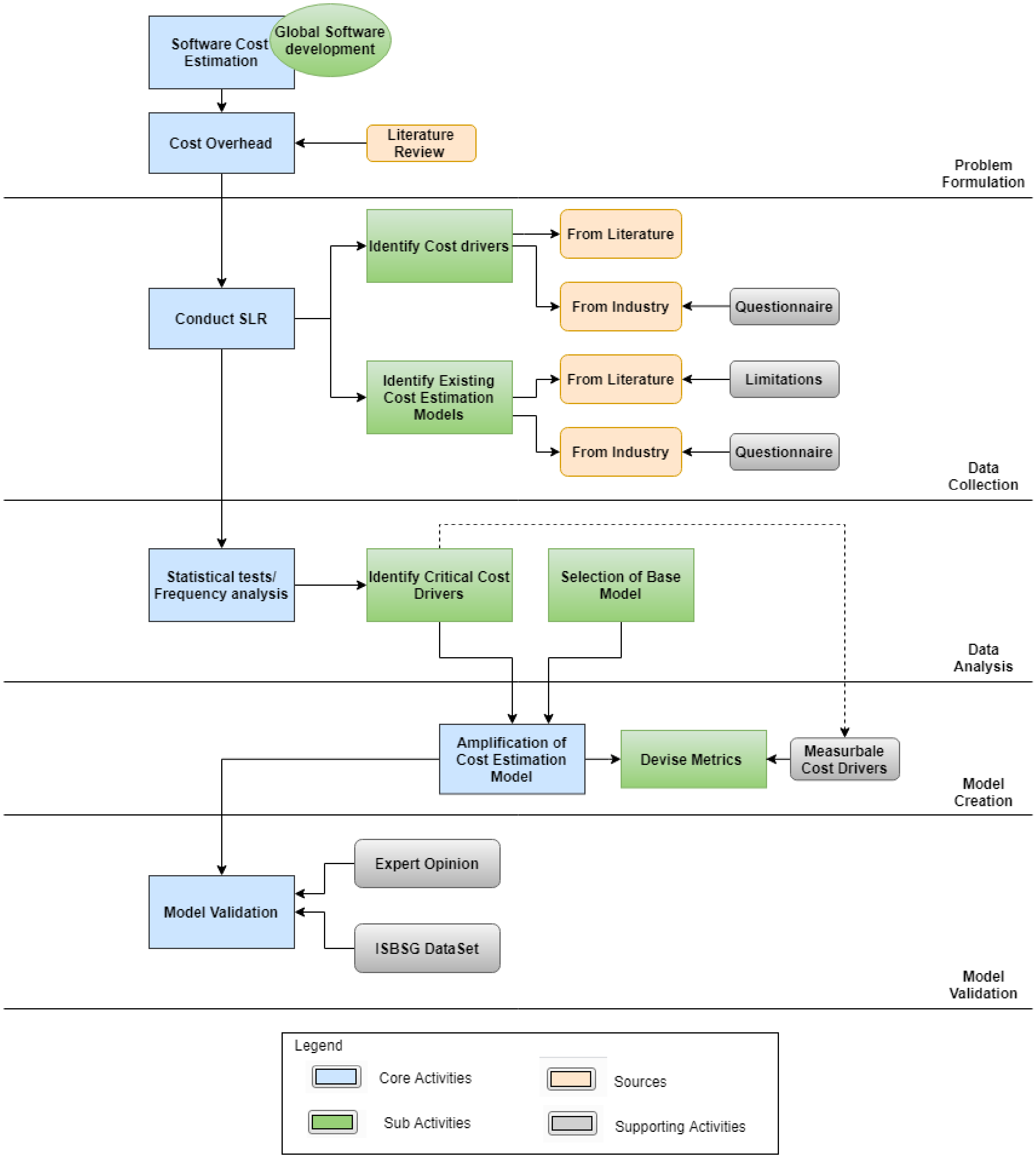


FIGURE 3. Adopted Research Methodology.

Figure 3 depicts the overall research flow of our study. Initially, we started with the problem formulation by conducting a general literature review. Cost overhead in the context of GSD is selected as a base problem. Then we performed a planned SLR to extract the cost drivers and GSD-Specific cost estimation models. The obtained results of SLR are then validated through an empirical study targeting the project managers of GSD. Through this multi- perspective, we highlighted the critical cost drivers by applying various statistical tests. Finally, we presented a conceptual model based on our findings. The ultimate aim and future direction of this research are to formalize the proposed model to assist the practitioners by considering cost estimation's additional challenges. The subsequent sections present three main phases of SLR.

A. PHASE 1: PLANNING THE REVIEW

This phase includes the primary steps that are required to conduct an SLR. To initiate an SLR, several pre-requisite activities are necessary to build its foundation. These steps are given below:

- Formation of the research question
- Selection of appropriate research repositories
- Developing a search string for the extraction of articles
- Defining inclusion and exclusion criteria for the article
- Defining quality criteria

The subsequent sections provide a detailed discussion of the activities mentioned above.

1) RESEARCH QUESTIONS

Based on the targeted objectives, a set of Research Questions (RQs) are formulated and described as:

RQ1: What are the cost drivers that could affect the cost estimation in GSD?

RQ1.1: How can the identified cost drivers be categorized?

RQ1.2: What are the critical cost drivers in each identified category?

RQ1.3: What are the additional factors, according to practitioners, that may affect the cost estimation in GSD?

RQ2: Are there any supporting cost estimation metrics/techniques to handle the identified cost drivers?

RQ2.1: What are the GSD specific cost estimation models according to the industrial perspective?

RQ3: What are the factors that current cost estimation techniques lack?

2) DATA REPOSITORIES

Search repositories are selected based on the previous experience and the recommendations provide by Chen *et al.* [17]. The digital libraries that are accessed for this research are IEEE Xplore Digital Library, ACM digital library, Science Direct, Google Scholar, Wiley Inter-Science, and Springer Link.

The search mechanism in data repositories differed, so search queries were tailored accordingly.

3) SEARCH STRING

The search strings that we applied are based on the primary and alternative keywords of our research questions. The keywords and their alternatives were chosen based on the available literature in the context of Software cost estimation and Global software development [4]–[7]. We categorized the search terms in our research into two groups; the first group includes the Cost Estimation terms. The second group comprises the terms related to the Global software development context.

Finally, we applied the content analysis technique on the data extracted to obtain the categories' frequency.

TABLE 2. Search String.

Group	Search String
Cost estimation	("Cost estimation" OR "cost prediction" OR "cost evaluation" OR "effort prediction" OR "effort estimation")
	AND
Global Software development	("Global software development" OR "global software engineering" OR "GSD" OR "GSE" OR "distributed development" OR "collaborative development" OR "Outsource" OR "Offshore")

TABLE 3. Quality Assessment Questions.

S.no	Checklist Question
Q1	Do the objectives of the proposed study discussed?
Q2	Is the research method clearly defined and documented?
Q3	Does the study explicitly focus on cost estimation for GSD?
Q4	Does the study address cost drivers of GSD?
Q5	Do the results of the proposed study answer the research question s and research objectives?

4) INCLUSION CRITERIA

The inclusion and exclusion criteria of our study were formed based on the guidelines presented by Kitchenham *et al.* [15] and Kitchenham [16]. Following were the inclusion criteria for primary study selection:

IC1: The selected primary study should be a journal, conference, or book chapter

IC2: The studies that focused on cost estimation in the GSD context

IC3: The studies that discuss the challenges or cost drivers affecting cost estimation in GSD

IC4: The studies that present GSD specific cost estimation models or techniques

IC5: The selected studies must be available full-text articles, specifically in the English language

IC6: The studies that were published between 2010 and 2020

5) EXCLUSION CRITERIA

Following exclusion criteria were applied for primary study selection:

EC1: The studies that do not answer the questions defined in Section 1

EC2: The studies that do not discuss the cost estimation process in the GSD context

EC3: The studies that do not highlight the challenges or cost drivers of cost estimation in GSD

EC4: The studies that were not written in English

EC5: The studies that are published before 2010

6) QUALITY ASSESSMENT OF SELECTED STUDIES

Quality assessment of the selected studies plays an essential role in quality research. The following questions were used to assess the quality of the studies:

TABLE 4. Total Selected Studies using Tollgate Approach.

E-databases	Papers extracted through Search Terms	Inclusion/\exclusion based on title and abstract	Inclusion/exclusion based on introduction and conclusions	Inclusion/exclusion the basis of the full text	Total selected articles for the primary study	Percentage of the final selected article (n=23)
IEEE Xplore	1187	125	22	7	7	30.44
ACM Digital Library	68	29	12	3	3	13.04
Science Direct	283	38	13	3	3	13.04
Google Scholar	1360	114	59	10	10	43.57
Springer Link	277	45	3	0	0	0
Wiley Inter Science	44	30	2	0	0	0
Total	3219	381	111	23	23	100

For the checklist questions mentioned above, the assessment we have done is as follows:

- The papers addressing the appropriate answer to the checklist question are given 1 point
- The papers addressing the partial answer to the checklist question are given 0.5 point
- The papers not addressing the desired checklist questions are given 0 points

B. PHASE 2: CONDUCTING THE REVIEW

In this phase of SLR, we conduct our review by applying the devised query and selecting the primary studies. Standard approaches are used for the extraction of the articles. Once the papers are extracted, they are synthesized according to the formulated criteria. The subsequent sections discuss these activities in detail.

1) PRIMARY STUDY SELECTION

During the primary studies selection process, various research articles were found, and the tollgate approach proposed by Afzal *et al.* [18] was used to refine the selection process. Tollgate approach is comprised of five phases and is shown in Fig. 4 and described in Table 4.

In the first phase of the tollgate approach, 3219 articles are extracted from all databases based on the keywords and search terms.

In the second phase, 381 articles were selected by applying the title and abstract inclusion and exclusion criteria.

In the third phase, only those articles were selected whose introduction and conclusion involves the answers to our proposed questions. The resulted articles were 111.

After the third phase, we included a secondary phase of duplication removal in which the articles with duplicated content were removed. After the elimination of duplicate articles, the resulted studies were 80.

Finally, after applying the inclusion and exclusion criteria based on full text, only those papers were selected that included cost drivers or presented some cost estimation model in the GSD context. The resultant primary studies were 23. Figure 4 contains the pictorial representation of the phases of the tollgate approach.

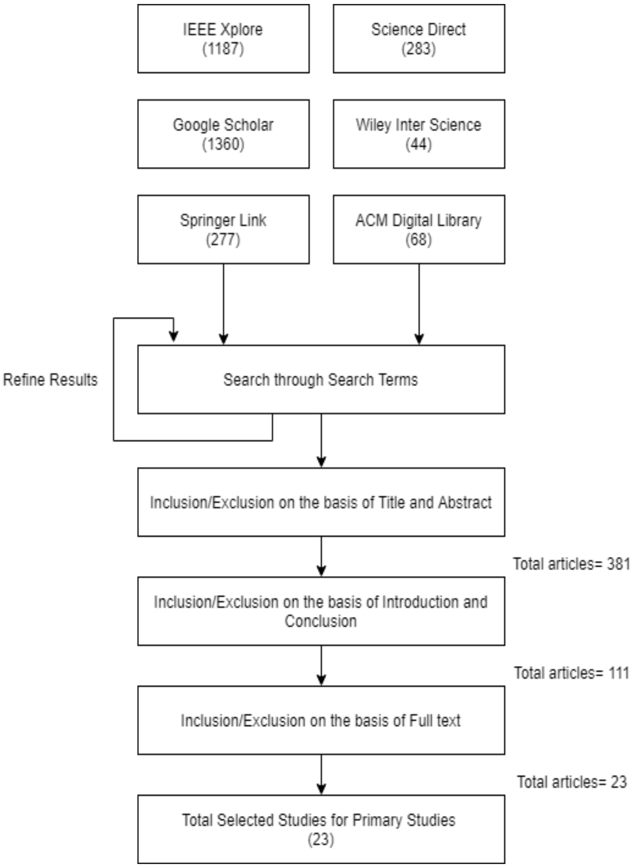

IEEE Xplore (1187)
Science Direct (283)
Google Scholar (1360)
Wiley Inter Science (44)
Springer Link (277)
ACM Digital Library (68)
Refine Results
Search through Search Terms
Inclusion/Exclusion on the basis of Title and Abstract
Total articles= 381
Inclusion/Exclusion on the basis of Introduction and Conclusion
Total articles= 111
Inclusion/Exclusion on the basis of Full text
Total articles= 23
Total Selected Studies for Primary Studies (23)

FIGURE 4. Tollgate Approach for Article Selection.

The result demonstrated that most of the studies related to software cost estimation (GSD) are extracted from Google scholar (43.57%) and IEEE (30.44%). Google scholar and IEEE are the most relevant and active digital libraries associated with cost drivers and GSD-specific cost estimation techniques. The lists of selected primary studies are presented in Table 6.

2) DATA EXTRACTION

The articles were selected based on the parameters, i.e., study id, author name, publication year, research method, study

TABLE 5. Quality Score of Selected Primary Studies.

Primary Study	Q1	Q2	Q3	Q4	Q5	Total Score
PS1	1	0.5	1	1	0.5	4
PS2	1	1	1	0.5	0.5	4
PS3	1	1	1	1	0.5	4.5
PS4	1	1	0.5	0.5	0.5	3.5
PS5	1	1	0.5	0.5	1	4
PS6	1	1	0	0.5	0.5	3
PS7	1	1	0	0.5	1	3.5
PS8	1	1	1	0.5	1	4.5
PS9	1	1	1	1	0.5	4.5
PS10	1	1	0.5	0.5	0.5	3.5
PS11	1	1	0.5	0	1	3.5
PS12	1	0.5	1	0.5	1	4
PS13	1	0.5	1	0.5	1	4
PS14	1	1	0.5	0.5	0.5	3.5
PS15	1	0.5	0.5	0.5	0.5	3
PS16	1	1	1	0.5	0.5	4
PS17	0.5	1	1	0	0.5	3
PS18	1	1	0.5	0	0.5	3
PS19	1	1	1	0	1	4
PS20	1	0.5	0.5	0	0.5	2.5
PS21	1	1	1	0	0.5	3.5
PS22	1	1	1	0	0.5	3.5
PS23	1	1	1	0	1	4

type, and limitations associated with each article. The selected list of articles is presented in Table 6. Moreover, the formulated research questions were mapped with the identified studies to ensure relevancy.

3) DATA SYNTHESIS

In this phase, the data extracted from the primary studies are formed and evaluated against the formulated research questions. Moreover, the articles were filtered through the levels of the tollgate approach for synthesizing the data, as shown in Figure 4. From a total of 23 articles, 41 cost drivers and 9 GSD-specific cost estimation models were identified.

C. PHASE 3: REPORTING THE REVIEW

In this phase, the selected primary studies are evaluated against the devised quality questions, and a list of primary studies is formulated. Along with the study type, the section also discusses the studies' temporal distribution to identify the trend. The detail of the phases is provided in the subsequent sections.

1) QUALITY ATTRIBUTES

All the selected studies were assessed and reviewed to cross-check their quality to be included in SLR Table 5 contains each primary study's quality score corresponding to the formulated quality questions presented previously in Table 2.

As the average quality score presented in Table 5 is greater than 2.5. This high-quality score depicted that the selected articles fulfilled the quality criteria and were most relevant to our topic. The chosen studies targeted cost estimation, specifically in the GSD context. Some primary studies focused on the cost drivers whereas, some presented models in the GSD context.

2) TEMPORAL DISTRIBUTION OF SELECTED STUDIES

Of the total 23 primary studies, 50% of research papers were published in the years (2010-2014), and the same percent of articles (50%) were published in the years (2015-2020). This represents the constant interest of the authors in this particular domain. The noticeable work has been carried out in the context of cost estimation of the GSD projects. Furthermore, we applied the snowballing technique for the extraction of relevant work through citations. The additional studies found were three that were included in our primary studies. The rationale behind the inclusion is that these studies were able to answer our formulated research questions.

IV. RESULTS AND FINDINGS

In this section, the results and findings of each formulated research question are discussed. Furthermore, analysis and empirical evaluation are performed for the legitimacy of the obtained results.

RQ1: What are the cost drivers that affect the cost estimation in global software development?

Cost drivers are the factors that influence global software development in a multi-dimensional way; that may be positive or negative. In GSD, these cost drivers are hidden in nature because of the distributed characteristic of this development type. Due to the dispersed nature of GSD, these cost drivers are often neglected and cause cost overhead later in the project. These factors should be considered during the estimation process for a realistic prediction of the effort and resources. A total of 17 out of 23 articles were targeting the cost drivers, while some articles specifically discussed GSD-specific cost estimation models. The existing studies discussed cost drivers but lacked empirical evidence to validate these cost drivers. Initially, we obtained a list of cost drivers from the literature, and the analyses have been performed. The resultant cost drivers are validated through the empirical study to overcome the shortcomings of the existing work. The cost drivers extracted through the literature are presented in Table 7. The labels include the cost driver name, its frequency, percentage, and the reference.

A total number of 41 cost drivers were identified that are further discussed in the following sections:

A. FACTORS HAVING A CRITICAL IMPACT ON COST

To analyze each cost driver's importance, we have adopted the criteria of frequency > 50% as critical cost drivers. The same criteria are followed in various similar studies [34], [34], [36]. Using the mentioned criteria, we identified eight critical cost drivers. These critical costs drivers are [CD1-CD3], [CD6], [CD8], [CD10], [CD25], and [CD31].

B. FACTORS WITH MODERATE IMPACT ON COST

We adopted the criteria to categorize the cost drivers with moderate impact as the frequency was between 25 and

TABLE 6. List of Selected Primary Studies using SLR.

S.no	Reference	Study type
[PS1]	[4] N. Ramasubbu and R. K. Balan, "Overcoming the challenges in cost estimation for distributed software projects," *Proc. - Int. Conf. Softw. Eng.*, pp. 91–101, 2012, doi: 10.1109/ICSE.2012.6227203.	Factor-based
[PS2]	[7] J. Koskenkyla, "Cost estimation in global software development - Review of estimation techniques," p. 109, 2012.	Factor-based
[PS3]	[5] D. Wickramaarachchi and R. Lai, "Effort estimation in global software development - a systematic review," *Comput. Sci. Inf. Syst.*, vol. 14, no. 2, pp. 393–421, 2017, doi: 10.2298/CSIS160229007W.	Factor-based
[PS4]	[19] M. Usman and R. Britto, "Effort estimation in co-located and globally distributed agile software development: A comparative study," *Proc. - 26th Int. Work. Softw. Meas. IWSM 2016 11th Int. Conf. Softw. Process Prod. Meas. Mensura 2016*, pp. 219–224, 2017, doi: 10.1109/IWSM-Mensura.2016.042.	Factor-based
[PS5]	[20] M. Muhairat, S. Aldaajeh, and R. E. Al-Qutaish, "The impact of global software development factors on effort estimation methods," *Eur. J. Sci. Res.*, vol. 46, no. 2, pp. 221–232, 2010.	Factor-based
[PS6]	[21] T. F. C. Tait and E. H. M. Huzita, "Software project management in distributed software development context," *ICEIS 2013 - Proc. 15th Int. Conf. Enterp. Inf. Syst.*, vol. 2, pp. 216–222, 2013, doi: 10.5220/0004442402160222.	Factor-based
[PS7]	[22] A. Lamersdorf and J. Munch, "Studying the Impact of Global Software Development Characteristics on Project Goals: A Causal Model~!2009-09-25~!2010-05-03~!2010-05-17~!," *Open Softw. Eng. J.*, vol. 4, no. 2, pp. 2–13, 2010, doi: 10.2174/1874107x01004020002.	Factor-based
[PS8]	[23] M. El Bajta, A. Idri, J. L. Fernández-Alemán, J. N. Ros, and A. Toval, "Software cost estimation for global software development: A systematic map and review study," *ENASE 2015 - Proc. 10th Int. Conf. Eval. Nov. Approaches to Softw. Eng.*, pp. 197–206, 2015, doi: 10.5220/0005371501970206.	Factor-based
[PS9]	[24] R. Britto, E. Mendes, and M. Usman, "Effort Estimation in Global Software Development : A Systematic Literature Review," *2014 IEEE 9th Int. Conf. Glob. Softw. Eng.*, pp. 135–144, 2014, doi: 10.1109/ICGSE.2014.11.	Factor-based
[PS10]	[25] S. Ramacharan and K. V. G. Rao, "Parametric Models for Effort Estimation for Global Software Development," *Lect. Notes Softw. Eng.*, vol. 1, no. 2, pp. 178–182, 2013, doi: 10.7763/lnse.2013.v1.40.	Model and Factor-based
[PS11]	[26] M. El Bajta *et al.*, "Software project management approaches for global software development: A systematic mapping study," *Tsinghua Sci. Technol.*, vol. 23, no. 6, pp. 690–714, 2018, doi: 10.26599/TST.2018.9010029.	Model and Factor-based
[PS12]	[27] S. Ramacharan and K. V. G. Rao, "Scheduling based cost estimation model: An effective empirical approach for GSD project," *IFIP Int. Conf. Wirel. Opt. Commun. Networks, WOCN*, vol. 2016-Novem, pp. 0–4, 2016, doi: 10.1109/WOCN.2016.7759881.	Model and Factor-based
[PS13]	[8] A. Lamersdorf, J. Münch, A. Fernández-Del Viso Torre, C. R. Sánchez, and D. Rombach, "Estimating the effort overhead in global software development," *Proc. - 5th Int. Conf. Glob. Softw. Eng. ICGSE 2010*, pp. 267–276, 2010, doi: 10.1109/ICGSE.2010.38.	Model and Factor-based
[PS14]	[12] R. Jain and U. Suman, "A Project Management Framework for Global Software Development," *ACM SIGSOFT Softw. Eng. Notes*, vol. 43, no. 1, pp. 1–10, 2018, doi: 10.1145/3178315.3178329.	Factor-based
[PS15]	[11] M. Niazi *et al.*, "Challenges of project management in global software development: A client-vendor analysis," *Inf. Softw. Technol.*, vol. 80, pp. 1–19, 2016, doi: 10.1016/j.infsof.2016.08.002.	Factor-based
[PS16]	[28] R. Britto, M. Usman, and E. Mendes, "Effort estimation in agile global software development context," *Lect. Notes Bus. Inf. Process.*, vol. 199, pp. 182–192, 2014, doi: 10.1007/978-3-319-14358-3_15.	Factor-based
[PS17]	[29] S. Ramacharan and K. Venu Gopala Rao, "Software effort estimation of GSD Projects Using Calibrated parametric estimation models," *ACM Int. Conf. Proceeding Ser.*, vol. 04-05-Marc, 2016, doi: 10.1145/2905055.2905177.	Model-Based & Factor-Based
[PS18]	[10] M. El Bajta, "Analogy-based software development effort estimation in global software development," *Proc. - 2015 IEEE 10th Int. Conf. Glob. Softw. Eng. Work. ICGSEW 2015*, pp. 51–54, 2015, doi: 10.1109/ICGSEW.2015.19.	Model-Based
[PS19]	[30] R. Madachy," Distributed global development parametric cost modeling," Software Process Dynamics and Agility, pp. 159-168, 2007	Model-Based
[PS20]	[9] J. Ahmad, A. W. Khan, and I. Qasim, "Software Outsourcing Cost Estimation Model (SOCEM). A Systematic Literature Review Protocol," vol. 2, no. 1, 2018.	Model-Based
[PS21]	[31] M. Humayun and C. Gang, "111-K0003," vol. 2, no. 3, 2012.	Model-Based
[PS22]	[32] S. Betz and J. Mki," Amplification of the COCOMO II regarding Offshore Software Projects," Offshoring of software development, p. 33, 2008	Model-Based
[PS23]	[33] P. Keil, D. Paulish, and R. Sangwan," Cost estimation for global software development," 2006, p. 10	Model-Based

50%. Through literature, we identified a total of 16 cost drivers with moderate impact on cost estimation. These cost drivers are [CD4], [CD5], [CD7], [CD9], [CD12], [CD15], [CD18-CD21], [CD24], [CD26], [CD30], [CD39].

C. FACTORS WITH LOW SIGNIFICANT IMPACT ON COST

We adopted the criteria of factors with frequency $< 25\%$ for the least significant cost drivers to be categorized as the least significant factors. These cost drivers have a very negligible impact on the cost. The same criteria are followed in a similar study [37]. These cost drivers could be neglected based on their low significance value. A total number of 17 cost drivers are included in this category. The cost drivers lying in this category are [CD11], [CD16], [CD17], [CD22], [CD23], [CD27], [CD28], [CD29], [CD32], [CD33], [CD34], [CD35], [CD36], [CD38], [CD40], [CD41].

TABLE 7. **Identified Cost Drivers of Cost Estimation (GSD).**

Sr.	Cost Driver	Reference	Frequency (n=17)	Percentage (%)
CD1	Time zone difference	[PS1],[PS2],[PS3],[PS4],[PS5],[PS7],[PS8],[PS9],[PS10],[PS12],[PS14],[PS15],[PS16],[PS17]	14	82.35
CD2	Language and Cultural Differences	[PS1],[PS2],[PS3],[PS4],[PS7],[PS8],[PS9],[PS10] ,[PS11],[PS12],[PS13],[PS14], [PS15],[PS16],[PS17]	15	88.24
CD3	Communication infrastructure and process	[PS1],[PS2],[PS5],[PS6],[PS7],[PS8],[PS9] ,[PS11],[PS13],[PS14],[PS15],[PS16]	12	70.59
CD4	Process model	[PS1],[PS2],[PS3],[PS8],[PS9],[PS11],[PS16]	7	40.18
CD5	Travel cost	[PS1],[PS3],[PS2],[PS7],[PS9],[PS16]	6	35.38
CD6	Competence level	[PS1],[PS2],[PS4],[PS7],[PS8],[PS9],[PS10],[PS11],[PS16],[PS17]	10	58.82
CD7	Requirements legibility	[PS1],[PS2],[PS4],[PS8],[PS9],[PS14],[PS15],[PS16]	8	47.15
CD8	Process compliance	[PS1],[PS2],[PS3],[PS4],[PS9],[PS13],[PS14],[PS15],[PS16]	9	52.94
CD9	Response delay	[PS1],[PS2],[PS3],[PS9],[PS5]	5	29.41
CD10	Team trust	[PS1],[PS2],[PS5],[PS7],[PS8],[PS9],[PS11],[PS14],[PS15]	9	52.94
CD11	Client Unawareness	[PS1],[PS5],[PS9]	3	17.65
CD12	Shared resources	[PS1],[PS5],[PS6],[PS7],[PS9]	5	29.41
CD13	Team structure	[PS1],[PS5],[PS8] ,[PS9],[PS11]	5	29.41
CD14	Work dispersion	[PS1],[PS2],[PS8],[PS9],[PS16]	5	29.41
CD15	Work pressure	[PS1],[PS5],[PS8],[PS9],[PS13]	5	29.41
CD16	Range of parallel sequential work handover	[PS1],[PS2],[PS9],[PS16]	4	23.53
CD17	Client Specific knowledge	[PS1],[PS2],[PS9],[PS16]	4	23.53
CD18	Lack of Client involvement	[PS1],[PS2],[PS4],[PS8],[PS9],[PS14],[PS16]	7	40.18
CD19	Design and technology newness	[PS1],[PS2],[PS8],[PS9],[PS16]	5	29.41
CD20	Team size	[PS1],[PS2],[PS8],[PS9],[PS16]	5	29.41
CD21	Project effort	[PS1],[PS2],[PS3],[PS9],[PS16]	5	29.41
CD22	Development productivity	[PS1],[PS2],[PS8],[PS9]	4	23.53
CD23	Defect density	[PS1],[PS2],[PS9],[PS16]	4	23.53
CD24	Rework	[PS1],[PS2],[PS3],[PS9],[PS16]	5	29.41
CD25	Project management effort	[PS1],[PS2],[PS3],[PS4],[PS6],[PS8],[PS9],[PS10],[PS16],[PS17]	10	58.82
CD26	Reuse	[PS1],[PS2],[PS8],[PS9],[PS16]	5	29.41
CD27	Code size	[PS1],[PS2],[PS8],[PS9]	4	23.53
CD28	Product Complexity	[PS8],[PS9],[PS13]	3	17.65
CD29	Platform Volatility	[PS8],[PS9]	2	11.76
CD30	Task allocation	[PS6],[PS8],[PS9],[PS14],[PS15]	5	29.41
CD31	Geographic distance	[PS2],[PS3],[PS4],[PS6],[PS7],[PS10],[PS12],[PS14],[PS15]	9	52.94
CD32	Social factors	[PS3],[PS12]	2	11.76
CD33	Product architecture	[PS3]	1	5.88
CD34	Unavailability of concerned personal	[PS5],[PS6]	2	11.76
CD35	Exchange rate fluctuation	[PS3],[PS14]	2	11.76
CD36	Unrealistic Milestones	[PS5],[PS9]	2	11.76
CD37	Training meeting/Sessions	[PS3],[PS6],[PS14]	3	17.65
CD38	Rules/Laws	[PS6],[PS7],[PS14]	3	17.65
CD39	Process maturity	[PS3],[PS4],[PS2],[PS7],[PS10],[PS13]	6	35.38
CD40	Organizational differences	[PS2],[PS7],[PS14]	3	17.65
CD41	Overoptimism	[PS4],[PS8]	2	11.76

RQ1.1: How can the identified cost drivers be categorized?

Most of the studies lack proper categories to define the factors, whereas some use a standard grouping scheme to represent it. Approximately 70% of the articles lack the categorization of the factors. The remaining 30% of the papers categorized the elements based on 4P'S (Process, project, personnel, and product) or PMBOK (Project management body of knowledge) Standard. PMBOK is a grouping of the practices of project management [12], [38]. These categories can be used to compensate for the additional cost drivers of GSD.

For this research, we selected 4P's for categorizing the cost drivers because the identified factors mapped with all the categories of 4P's, whereas if we talk about PMBOK, then it is more generalized and does not contain the specific sub-areas of cost estimation [38].

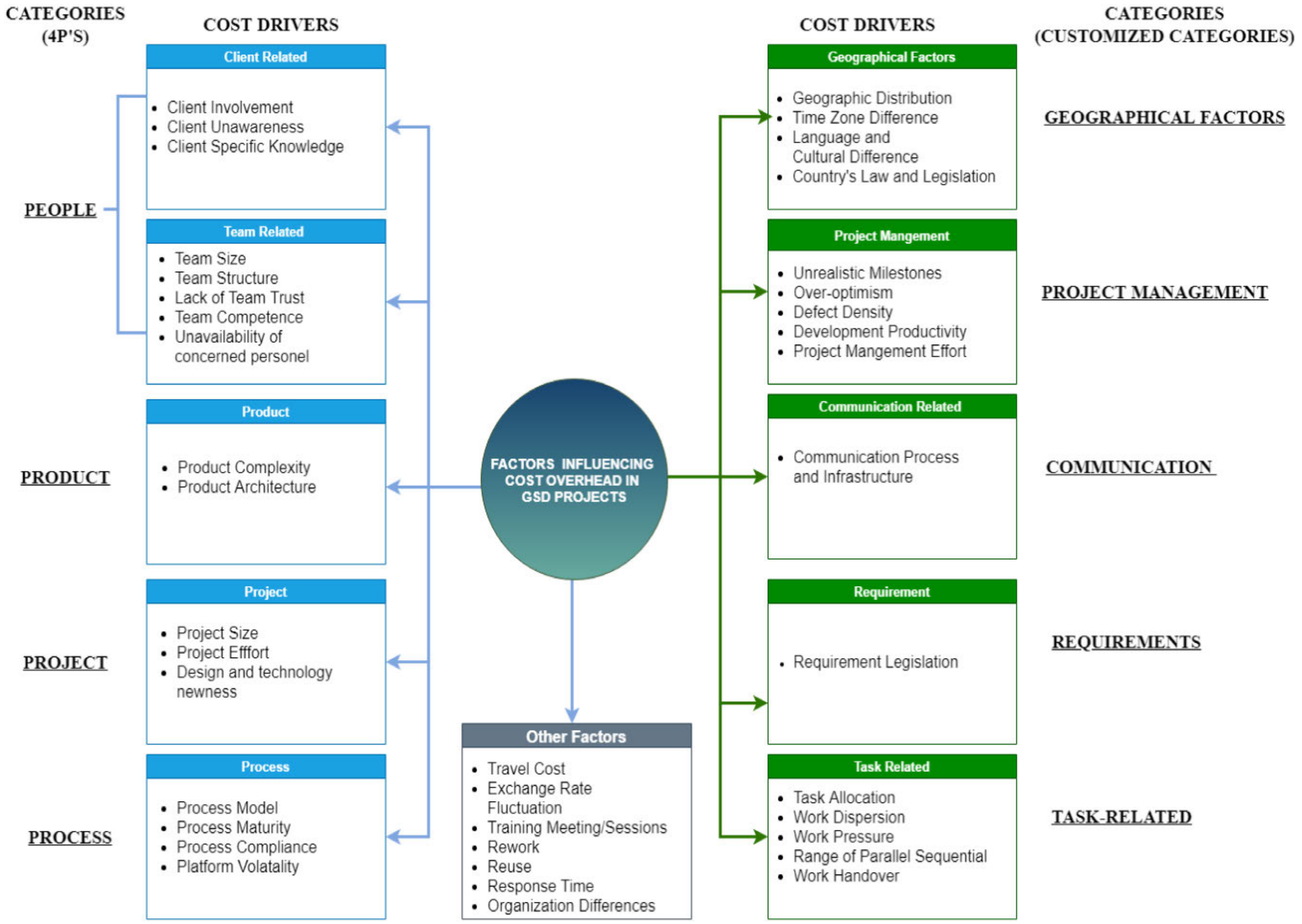


FIGURE 5. Thematic Taxonomy of Identified Cost Drivers.

Representation of cost drivers in 4P's could help the project managers understand better how knowledge areas require consideration of cost estimation factors to achieve better estimates.

D. EMPIRICAL EVALUATION OF COST DRIVERS

This section includes details regarding the design and execution of our empirical study. Moreover, it contains the analysis performed on the results obtained from the GSD organization. Finally, a comparison has been drawn between literature and industry for the identification of critical cost drivers.

1) SURVEY DESIGN

An online questionnaire was designed to acquire the industrial perspective regarding the obtained cost drivers from literature. The targeted respondents of the questionnaire were the project managers of global software organizations. A total number of 175 project managers were targeted, with all belonging to the different multinational companies. Moreover, the questionnaire was distributed in more than 20 countries for the legitimacy of the results. Five points Likert scale was used in the questionnaire with each identified cost driver (Strongly agree, Agree, Neutral, Disagree, Strongly disagree). The obtained responses were then converted into percentages and further refined with data analysis techniques. The demographics of targeted GSD organizations are represented in Figure 6.

2) DATA ANALYSIS

For data analysis, we performed a T-test and Spearman correlation test. All these statistical tests were used to ensure the result's legitimacy and distinguish the cost drives that either identified cost drivers are as critical as the literature indicates.

3) LEVENE'S TEST FOR RQ1

Levene's test is applied to check the homogeneity of variance between the results obtained from SLR and the Empirical study. The resultant values of variance and the percentages obtained from literature and industry are presented in Table 8.

4) T-TEST FOR RQ1

In addition to Levene's test, the t-test is also applied to check the mean differences between SLR and Empirical study data. The results of the t-test were t= −1.61 and p= 0.05,

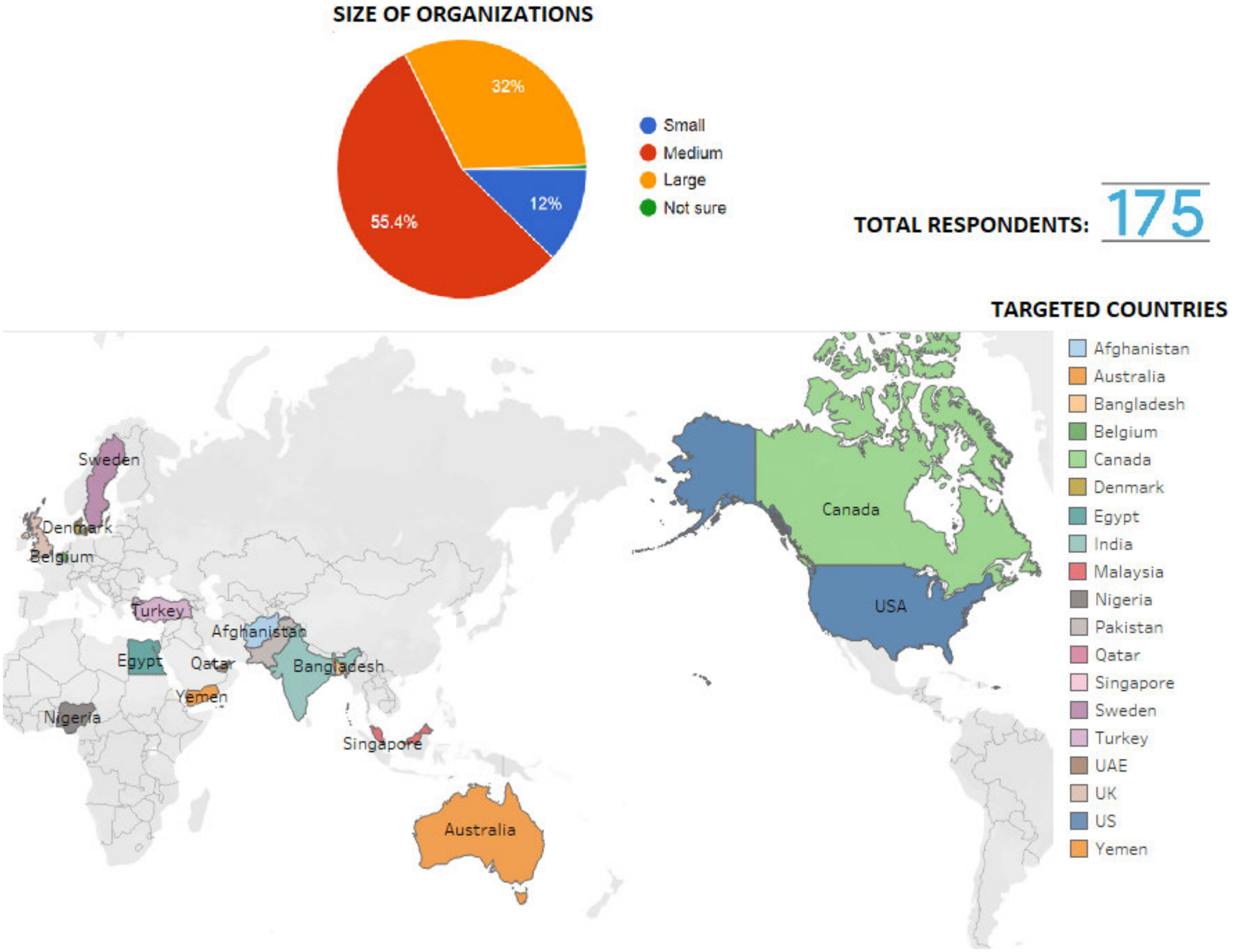


FIGURE 6. Demographics of Targeted Survey.

TABLE 8. Levene's Test for RQ1.

SUMMARY						
Groups	Count	Sum	Average	Variance		
Percentage (Literature)	24	1039.42	43.3091667	318.772825		
Percentage (Industry)	24	1226.81	51.1170833	241.407065		
ANOVA						
Source of *Variation*	SS	df	MS	F	P-value	F *crit*
Between Groups	731.562752	1	731.562752	2.61188509	0.11290425	4.05174869
Within Groups	12884.1375	46	280.089945			
Total	13615.7002	47				

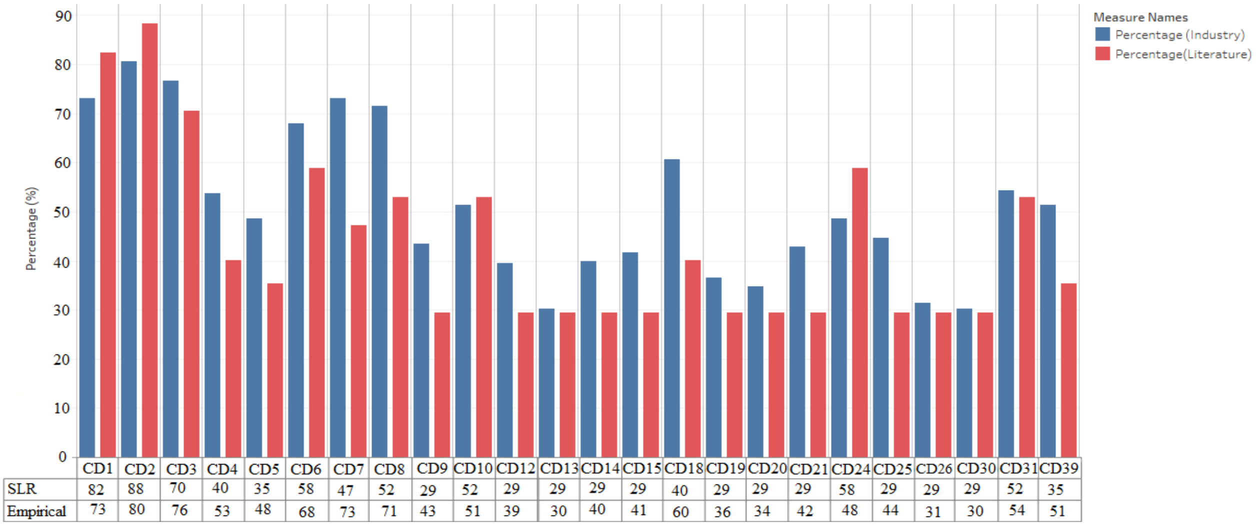
demonstrating that there is no significant difference between the rankings of SLR and Empirical study.

Figure 7 represents the comparison of the percentages of Cost drivers obtained from SLR and the Empirical Study. Only the cost drivers having a moderate or critical impact are considered in the comparison. The cost drivers having a low impact on estimation are neglected due to their low significant impact.

5) SPEARMAN TEST FOR RQ1

In addition to the t-test, the spearman correlation test is also applied. The comparison of the ranks obtained from SLR and Empirical study are presented in Table 9.

For the evaluation of the significance of the differences between the results of SLR and Empirical study, we performed spearman's rank-order correlation test. For the Spearman correlation test, the value of coefficient (Rs) closer



	CD1	CD2	CD3	CD4	CD5	CD6	CD7	CD8	CD9	CD10	CD12	CD13	CD14	CD15	CD18	CD19	CD20	CD21	CD24	CD25	CD26	CD30	CD31	CD39
SLR	82	88	70	40	35	58	47	52	29	52	29	29	29	29	40	29	29	29	58	29	29	29	52	35
Empirical	73	80	76	53	48	68	73	71	43	51	39	30	40	41	60	36	34	42	48	44	31	30	54	51

FIGURE 7. Comparison of Cost Drivers obtained from SLR and Empirical Study.

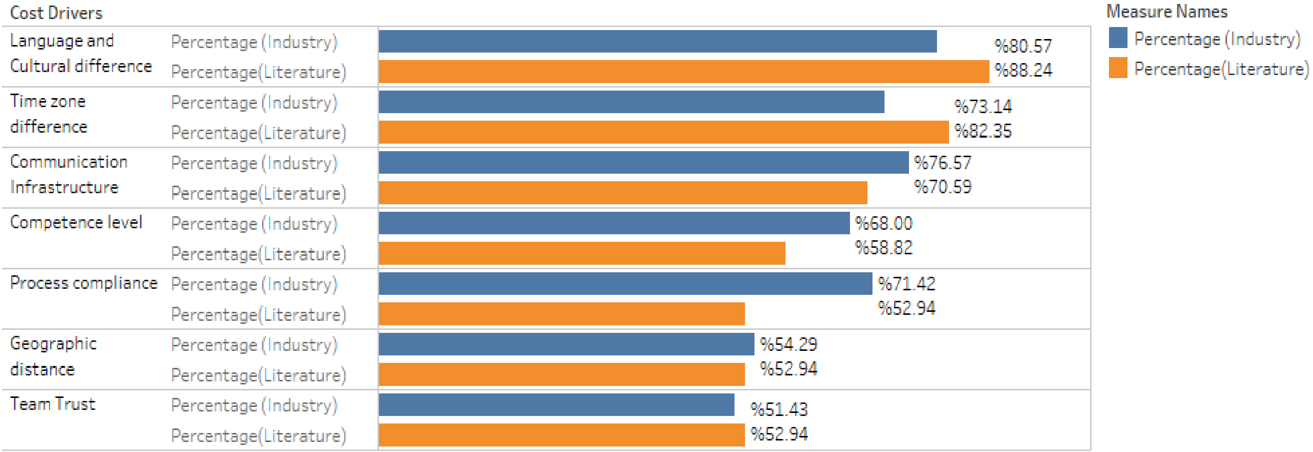


FIGURE 8. Critical Cost Drivers.

TABLE 9. T-test for RQ1.

	Percentage (Literature)	Percentage (Industry)
Mean	43.3091667	51.1170833
Variance	318.772825	241.407065
Observations	24	24
Pooled Variance	280.089945	
Hypothesized Mean Difference	0	
Df	46	
t Stat	-1.61613276	
P(T<=t) one-tail	0.05645213	
t Critical one-tail	1.67866041	
P(T<=t) two-tail	0.11290425	
t Critical two-tail	2.0128956	

to 1 represents the positive correlation, whereas the resulted values of Rs closer to −1 indicate a negative correlation. For our study, the Spearman coefficient was found to be 0.88727957, indicating a strong positive correlation between the rankings obtained from SLR and Empirical research results.

RQ1.2: What are the critical cost drivers in each identified category?

The identified factors were analyzed to extract essential drivers of cost. Those factors were considered critical, which have frequency $> 50\%$ in both literature and from an industrial perspective. The same criteria are followed by different studies [34]–[36]. Having frequency $> 50\%$ in literature and industry indicates that a highlighted cost driver is equally essential for the practitioners and must be considered in the estimation models. The extraction of critical drivers is of prime importance because these cost drivers could impact the other cost drivers. The critical cost drivers, along with their percentages (literature and industry), are shown in Figure 8.

CRITICAL COST DRIVERS (CCDS)

The critical cost drivers presented in Figure 8 are discussed below:

TABLE 10. Comparative Ranking of Identified Cost Drivers.

Sr.no	Cost Drivers	Percentage (Literature)	SLR Rank	Percentage (Industry)	Industry Rank
1	Time zone difference	82.35	2	73.14	3.5
2	Language and Cultural difference	88.24	1	80.57	1
3	Communication Infrastructure	70.59	3	76.57	2
4	Process Model	40.18	10.5	53.71	9
5	Travel cost	35.38	12.5	48.57	12
6	Competence level	58.82	4.5	68	6
7	Requirements legibility	47.15	9	73.14	3.5
8	Process compliance	52.94	7	71.42	5
9	Response delay	29.41	19	43.43	15
10	Team Trust	52.94	7	51.43	10
11	Shared resources	29.41	19	39.43	19
12	Team structure	29.41	19	30.29	23
13	Work dispersion	29.41	19	40	18
14	Work pressure	29.41	19	41.71	17
15	Client involvement	40.18	10.5	60.57	7
16	Design and technology newness	29.41	19	36.57	20
17	Team size	29.41	19	34.86	21
18	Project effort	29.41	19	42.85	16
19	Project management effort	58.82	4.5	48.56	13
20	Rework	29.41	19	44.57	14
21	Reuse	29.41	19	31.43	22
22	Task allocation	29.41	19	30.28	24
23	Geographic distance	52.94	7	54.29	8
24	Process maturity	35.38	12.5	51.42	11

CCD1: LANGUAGE AND CULTURAL DIFFERENCE

Language differences create communication misunderstandings among the team. These differences can result in several hidden cost drivers, i.e., rework, low quality of the work, or a company's bad image [23].

While cultural differences are the variations of values in different countries, these differences are an extra burden on management. They also associate hidden cost drivers for example, idle time in one site due to a public holiday in another country [5].

CCD2: GEOGRAPHIC DISTANCE

Geographic distance represents the residing physical distance between the teams or the individual developers. Geographic distance is symmetric, and it can create communication difficulties between the remote teams. Several hidden cost drivers can be introduced through geographic distance; travel time, review meetings, and so on [5].

CCD3: TIME ZONE DIFFERENCES

Time zone difference is one of the crucial cost drivers presented in several studies [5], [23]. Its impact is asymmetric and variations in time zone affect the total overlapping hours, directly affecting the communication among the virtual teams. Overlapping hours should be maintained to have effective communication. Less overlapping hours could result in communication difficulties. The hidden cost associated with the time zone is the idle time in which a developer could not proceed as he is waiting for a response from a remote team [5].

CCD4: TEAM TRUST

Team trust is one of GSD's social factors, and if not formed well, it can negatively impact the motivation, desire to work with, and other different issues in knowledge sharing. Geographic, cultural, and temporal distances can significantly impact the trust among the remote teams. Because specifically, in GSD, we lack informal and face-to-face meetings, so members have less opportunity to develop interpersonal relations and emotional bonds [7].

CCD5: COMMUNICATION INFRASTRUCTURE

Communication is the core of any Global Software Organization because we cannot have face-to-face meetings in GSD Organizations, but everything is managed through the communication channels. The occurrence of this cost driver in several studies [4], [7], [20], [21] highlights its importance. As effective communication plays a vital role in the success of a GSD project, GSD organizations should ensure the use of video and teleconferencing at each site. For proper use of these tools, training must be provided. By the effective use of these channels, we can complete a project in its allocated time and budget, and if we lack effective information sharing mechanisms, then it can lead to the delay and additional effort to the project [7].

CCD6: PROCESS COMPLIANCE

In a distributed environment, we have the possibility of introducing new (incompatible) processes and tools at different sites. Having multiple processes, methodologies, tools, and templates that do not integrate or interoperate with each other, can lead to rework or data loss during transference from one tool to another, which may decrease the quality [7].

CCD7: COMPETENCE LEVEL

Competence level is an important factor of a distributed environment. It lies under the category of "Skill

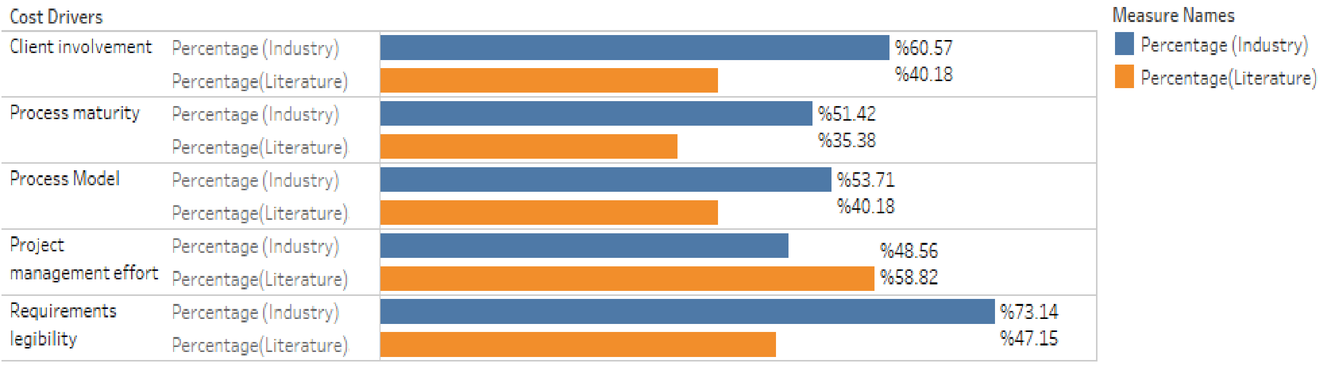


FIGURE 9. Variance in Cost Drivers.

management," in which we select a highly proficient and competent team from remote locations [39].

Along with the mentioned cost drivers, the industry considered some other cost drivers as critical, but these cost drivers lie in the moderate category of literature (25-50%). These cost drivers are shown in Figure 9. The cost drivers presented in Figure 9 with their percentages > 50% in industry depicts their importance. Client involvement is an essential factor for managing a project. Projects with no client involvement can lead to difficulties, and the project could exceed in terms of time or budget. Similarly, process maturity depicts that if our processes are mature and standardize to complete a specific project. In case if we lack formal processes, then our time and cost will be compensated.

Project management effort is not given the same importance from the industry as presented in the literature. We have analyzed that the practitioners consider the measurable factors that could be computed and be of their use. However, project management effort is regarded as an abstract term; neither could it be accurately measured. So, its percentage lies below 50% from an industrial perspective. The only reason for the variance of some cost drivers' percentages is the practitioners' mindset; they are diverted more toward the measurable factor, whereas literature considers the non-measurable factors. But in the end, these are only measurable cost drivers that are required by the cost estimation models.

RQ1.3: What are the additional challenges, according to practitioners, that may affect the cost estimation in GSD?

Along with the validation of the listed cost drivers of literature, the project managers, with their experiences, have shown interest in providing us some additional challenges that may impact the cost estimation in GSD. These additional cost drivers are shown in Figure 9.

The additional challenges depicted in Figure 10 are discussed below:

1) ORGANIZATION'S ADAPTABILITY TO NEW TECHNOLOGIES

Sometimes selecting a project on the latest technologies may impact the cost estimation due to the lack of knowledge

1) PRESSURE FROM HIGHER AUTHORITIES

Due to the pressure from higher authorities in the concern of "Winning a project," the project's cost and effort are accommodated.

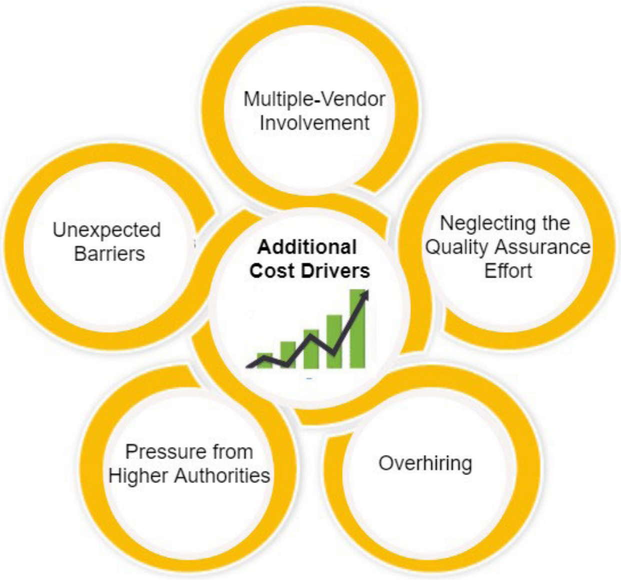

Multiple-Vendor Involvement
Neglecting the Quality Assurance Effort
Overhiring
Pressure from Higher Authorities
Unexpected Barriers
Additional Cost Drivers

FIGURE 10. Additional Challenges identified from Empirical Study.

1) MULTIPLE-VENDORS INVOLVEMENT

Cost estimation is affected when multiple vendors are involved with their different inter-company politics.

1) OVERHIRING

It does not happen commonly, but if a company over-hired the engineers that were not required for the current project, this could influence the cost. Choosing the right number does matters

1) NEGLECTING THE QA EFFORT

The effort required for quality assurance should not be ignored, or else it could affect the overall cost of a product.

1) UNEXPECTED BARRIERS

Referring to the current situation of Covid-19, companies should have some strategic plan and an allotted budget for the unexpected barriers to deal with them properly, or else these unexpected barriers can affect the development process.

RQ2: Are there any Metrics/Techniques defined to handle the identified cost drivers?

Although there are various cost estimation techniques for the estimation of a project, the author [7] reviewed the

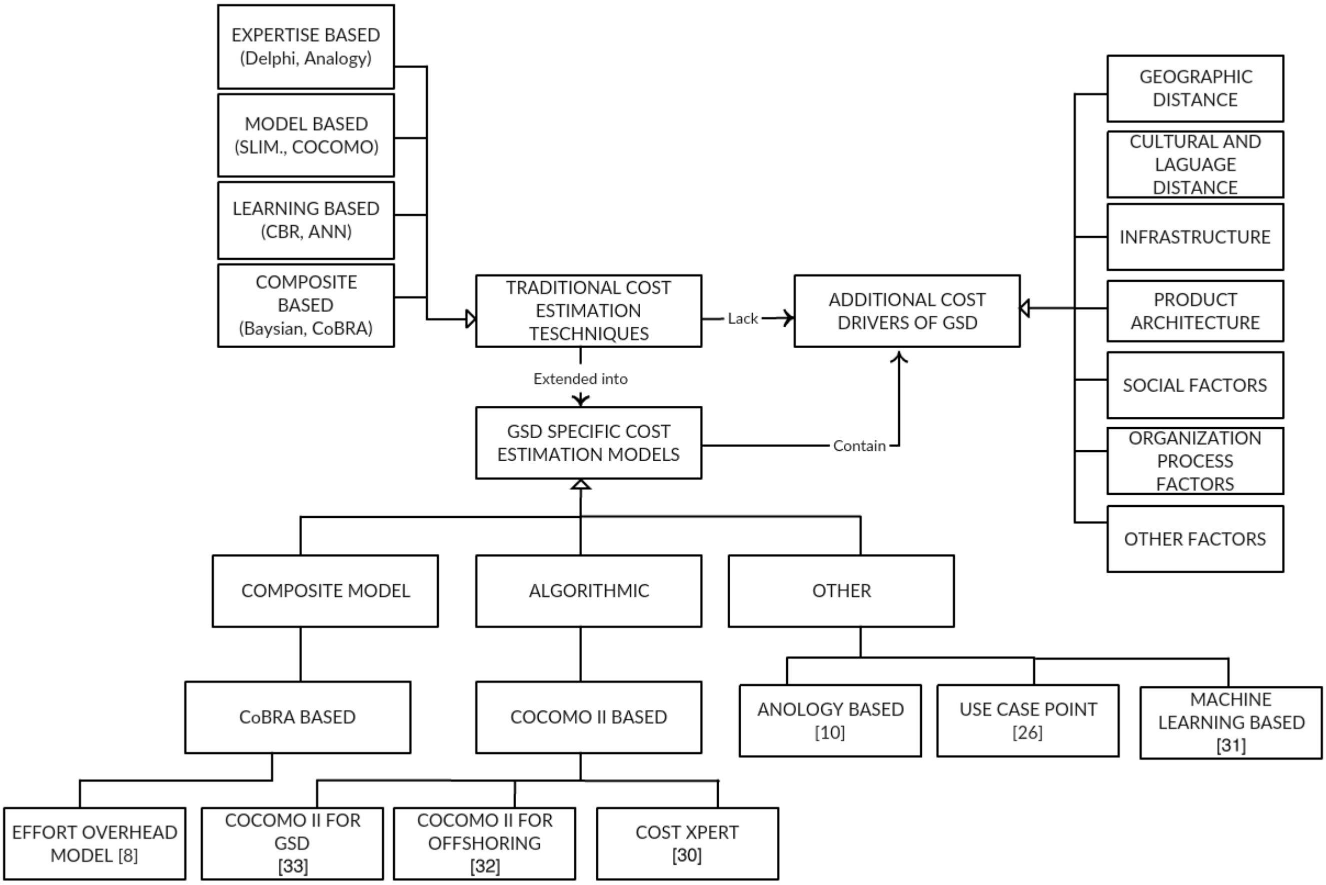

EXPERTISE BASED (Delphi, Analogy)
MODEL BASED (SLIM., COCOMO)
LEARNING BASED (CBR, ANN)
COMPOSITE BASED (Baysian, CoBRA)
TRADITIONAL COST ESTIMATION TESCHNIQUES
Lack
ADDITIONAL COST DRIVERS OF GSD
GEOGRAPHIC DISTANCE
CULTURAL AND LAGUAGE DISTANCE
INFRASTRUCTURE
PRODUCT ARCHITECTURE
SOCIAL FACTORS
ORGANIZATION PROCESS FACTORS
OTHER FACTORS
Extended into
GSD SPECIFIC COST ESTIMATION MODELS
Contain
COMPOSITE MODEL
ALGORITHMIC
OTHER
CoBRA BASED
COCOMO II BASED
ANOLOGY BASED [10]
USE CASE POINT [26]
MACHINE LEARNING BASED [31]
EFFORT OVERHEAD MODEL [8]
COCOMO II FOR GSD [33]
COCOMO II FOR OFFSHORING [32]
COST XPERT [30]

FIGURE 11. Conceptual Model of Existing GSD-Specific Cost Estimation Models.

algorithmic and non-algorithmic techniques for cost estimation, but these techniques do not fit for the GSD due to the additional challenges of global software development. The techniques that are used for GSD are amplified from the existing cost estimation by adding additional factors. In most of the research articles, the algorithmic approach is amplified, i.e., Keil *et al.* [33], Betz and Mäkiö [32], Madachy [30] amplified the COCOMO II model for the GSD context. Though, some other techniques are also introduced by analyzing the cost overhead. Lamersdorf *et al.* [8] introduced a model based on the cost overhead by amplifying COBRA's base model. Figure 11 represents a conceptual model of the existing GSD specific cost estimation techniques.

Figure 11 represents the phenomenon of traditional cost estimation techniques being amplified into GSD specific techniques. These techniques, their early stage and are not properly validated. These techniques lack many cost drivers that are identified in the literature [5]. So, there is a gap for the researchers that could be filled up if we consider the missing factors, i.e., time zone differences. Considering these factors is very important for efficient estimation, or our budget can overflow with these hidden factors.

RQ2.1: What are the GSD specific cost estimation models according to the industrial perspective?

This study has also taken the industrial view regarding cost estimation models being used for the GSD environment. For that, we have identified the models through a questionnaire used by practitioners to estimate the cost in the GSD context. It is observed that the companies do not rely on a specific cost estimation model but instead, they use multiple models for the estimation of a project. This is also discussed in the literature that we are taking benefit from multiple models by their combined usage. In our study, we have considered 175 project managers with all belonging to the different GSD companies.

TABLE 11. Usage of Cost Estimation Models (Industrial Perspective).

Models	No of Responses	Percentage (Usage)
Expert Judgment	107	61.1%
Analogy Based	77	44%
Pay as Go	53	30.3%
Algorithmic	27	15.4%
Hybrid	22	12.6%
Machine Learning	12	6.9%

Table 11 depicts that multiple organizations are using a combination of cost estimation models in the context of GSD. We still have a high frequency of expert judgment and analogy-based models, which means that we still lack an appropriate formal model. The GSD based organizations still prefer to use non-formal cost estimation models because the results of formal models in this context are still not satisfactory; they lack in considering the additional cost drivers of GSD. The limitations of existing cost estimation models are discussed in the forthcoming sections. Another fact is that the percentage of algorithmic-based estimation lies above all the formal models, so in the future, we can improve these types of models to enhance estimation accuracy.

RQ3: What are the factors that current cost estimation techniques lack?

There are different GSD-specific cost estimation models, but all these models have some shortcomings that are listed in Table 12. The review matrix contains all the desired information regarding each model and highlights its limitations.

Table 12 presented the detailed aspects regarding each cost estimation model and the specific limitations. We have then generalized the limitations of all these models and showed them in Figure 12.

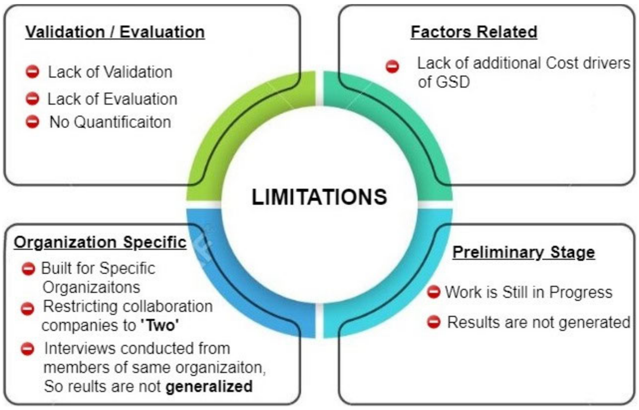

LIMITATIONS
Validation / Evaluation
Lack of Validation
Lack of Evaluation
No Quantificaiton
Factors Related
Lack of additional Cost drivers of GSD
Organization Specific
Built for Specific Organizaitons
Restricting collaboration companies to 'Two'
Interviews conducted from members of same organizaiton, So reults are not generalized
Preliminary Stage
Work is Still in Progress
Results are not generated

FIGURE 12. Limitations of GSD-Specific Cost Estimation Models.

V. DISCUSSIONS

The study aimed to identify the factors influencing cost overhead in the GSD context. The research's ultimate goal is to develop a cost estimation model based on the critical cost drivers to improve GSD's estimation process. In this section, we discussed the comments on our research questions.

Regarding RQ1, we have identified that:

- There are many additional hidden cost drivers for the GSD context that should be considered in the estimation process. Our research identified 41 initial cost drivers that were further refined according to the performed empirical study
  - There are many additional hidden cost drivers for the GSD context that should be considered in the estimation process. Our research identified 41 initial cost drivers that were further refined according to the performed empirical study
  - The identified factors have inter-relationship with each other, which means that they can also influence or affect each other
  - It is recognized that team culture, geographic distance, temporal distance, and communication infrastructure have the highest frequency, which makes them crucial.
  - There are also additional challenges associated with cost estimation that is not highlighted in the literature, i.e., pressure from higher authorities, unexpected barriers, multiple vendor involvement
  - There is a need to develop the metrics by considering these factors, i.e., Time differences may be calculated through the overlapping hours.
  - Finally, we identified two primary grouping schemes for the categorization of these factors. The 4P'S and PMBOK could be used to compensate for the additional elements. PMBOK is used for high-level classification, whereas 4P's could be used for the categorization of detailed low-level cost drivers.
  - The identified factors are presented in 4P's based taxonomy for better visualization and analysis of the knowledge areas

Regarding the RQ2, we have determined that:

- Most of the techniques that are used in GSD are amplified from the existing cost estimation techniques that were aimed to estimate the collocated projects
- 50% of the GSD specific techniques are amplified using the Algorithmic approach (COCOMO)
- From the industrial perspective, we have analyzed that the models being used to estimate GSD projects are mainly Analogy based or through Expert judgment. Hence, there is a need to develop a formal model in this context.
- It is also suggested to use some other techniques, e.g., machine learning and neural networks. Because each type of technique has its limitations. So, it's better to try some hybrid techniques.

Regarding the RQ3, we have identified that:

- The recognized techniques are at an early stage and lack calibration
- The techniques are not validated and consider the limited number of organizations [24]
- Some of the existing models are context-specific, which means that they are developed for some specific environment, so they could not be generalized for the GSD context
- Missing cost factors from the techniques can result in un-accurate results of the estimation. So, there is a gap for the researchers and practitioners to improve these models for the GSD context.

VI. PROPOSED CONCEPTUAL MODEL

Based on our research outcome, we have developed a conceptual model of cost estimation in the GSD context. As shown in Figure 13, the model consists of three main components. In phase 1, the critical cost drivers (CCD's) of the GSD context are selected along with a base cost estimation model. The rationale behind selecting these CCDs is to include the hidden cost associated with the GSD projects counted when

TABLE 12. **Review Matrix of Existing GSD Specific Cost Estimation Models.**

Author(s)	Model/Technique	Research Description	Finding(s)	Limitation(s)	Evaluation Measures
Betz and Makio [32]	COCOMO II based	• Proposed a model for estimating effort in GSD. • New effort multipliers are added to make it effective	• Most comprehensive model as compared to other • 11 New effort multipliers are added	• The model restricted the collaboration companies to two • No systematic approach is followed for the quantification of factors, which resulted in unclear numerical values	Not validated
Lamersdorf [8]	Cost Overhead Model for GSD (CoBRA Based)	• Presented a cost overhead model based on influencing factors and causal relationships	• The causal model can help in understanding the importance of the factors and their relationship	• All interviews were conducted within one organization • The model could not be generalized, as it is company-specific	Evaluation is done through the Goal Question Metric Paradigm
Mamoona Humayun [31]	Machine Learning-Based	• The paper presented an overview of different ML techniques for cost estimation in GSD	• The techniques address the distributed characteristics of GSD • Techniques are available according to the situation	• Handling cost drivers are not discussed. • The accuracy of these types of models is highly dependent on the kind of data on which they are trained • Unsuitable data can result in inaccurate estimates	Evaluated based on a Canadian dataset
Manal El Bajta [26]	Use Case Points	• UCP is a simple and relatively easy approach for estimation where application size is estimated through use case diagrams	• A simple and easy approach does not require a technical understanding of the concepts	• Detailed Use Cases are required for accurate estimation • The mentioned factors are application-specific and depend upon the type of application	In the stage of data extracting and no results have been published, so the model is not evaluated
Manal El Bajta [10]	Analogy based cost estimation model	• The model is related to Case-Based Reasoning where it is assumed that similar projects will have similar cost but in the case of global software development, first have to identify the similarity attributes to calculate the cost	• The model is useful only if we have experience with similar projects	• This type of estimation is problematic when data is not available or have not done any similar project in the past	The work is still in progress, so the approach is not evaluated
Mr.S.Ramacharan [27]	Scheduling based cost estimation model	• Scheduling and productivity parameters are identified, and cost estimation is based on the line of code	• The effectiveness of the model is evaluated by comparing it with other models, which is not done in any other model	• The variation of records is not considered; it is instead stated in the future work of the research	Comparison with other Models
Jamshed Ahmad [9]	SOCEM: Software Outsourcing Cost Estimation Model	• A model is presented to identify the challenges faced by the vendor organization regarding cost estimation in GSD	• The model works on the challenges of cost estimation	• An abstract level estimation model is presented. • No discussion about cost drivers and handling cost drivers through model • Organization Specific	Evaluation is left for future work.
Madachy [30]	Cost Xpert	• Proposed a model based on phase-sensitive effort multipliers. • The model considers the cost factors related to the people in different teams	• The model has proposed with industrial collaboration, and effort multipliers are phase sensitive	• No quantification and validation of the proposed model • Additional cost drivers of COCOMO II are not included, i.e., communication and collaboration	No validation

TABLE 12. ***(Continued.)*** **Review Matrix of Existing GSD Specific Cost Estimation Models.**

P.Keil [33]	COCOMO II Based	• Proposed a model to provide a decision-making framework to calculate the tradeoff between the estimation in collocated and GSD	• The model introduced factors related to multisite communication and multisite collocation.	• No quantification of the complexity factor • No systematic approach is adopted for factors extraction	No validation

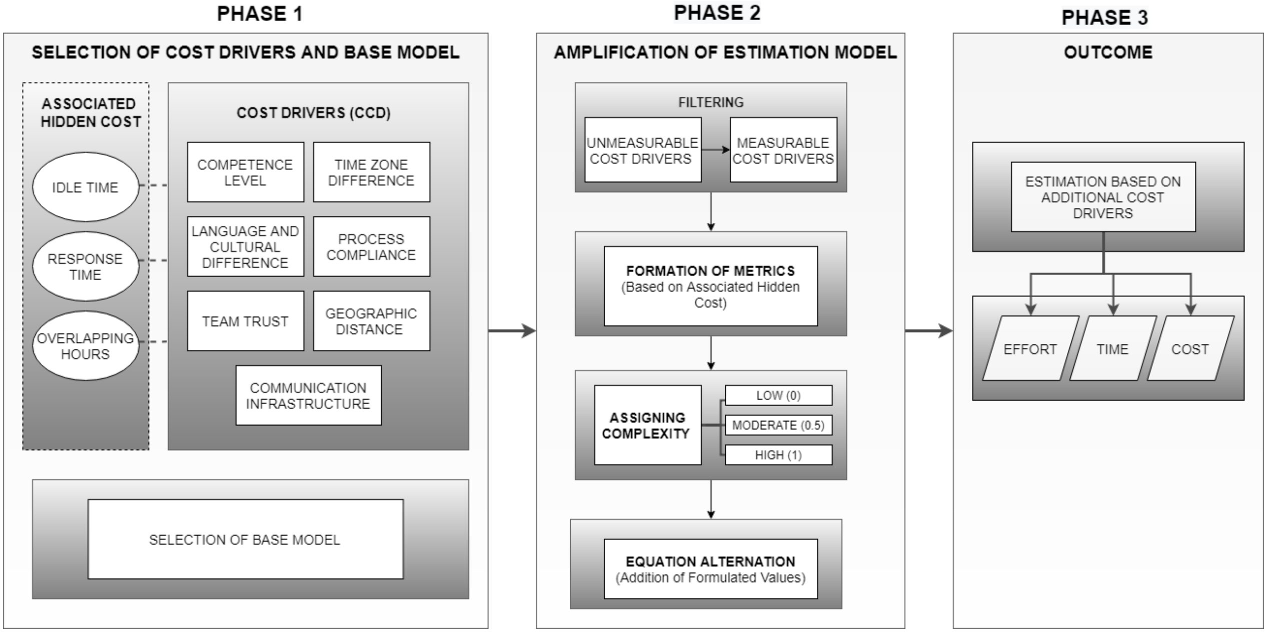


FIGURE 13. **Proposed Conceptual Model of Cost Estimation in GSD Context.**

estimation is performed. In phase 2, the amplification of the base model is performed where standardize modification activities are selected. Initially, the immeasurable cost drivers are converted into measurable cost drivers, then criteria are developed, and the formulation of metrics occurs. Once the metrics are formed, then these cost drivers are assigned values considering their level of occurrence. Finally, these values are added to the base model equation, and the results are generated. In phase 3, we estimate the additional cost drivers that were not considered in the traditional cost estimation model. In paper [32], a similar work is presented. The author extracted the effort multipliers of outsourcing (EMO) and accommodated them in COCOMO II. However, the extracted EMOs were not empirically validated. The presented model considered the factors and challenges of GSD, whereas the critical factors were not highlighted. Simultaneously, in our conceptual model, the extracted factors were empirically validated by 175 project managers. We have considered the moderate and critical impact cost drivers due to their high significance impact. Therefore, our model considers an exhaustive list of cost drivers covering all the aspects of the estimation process essential in the GSD context.

VII. FUTURE DIRECTIONS

The outcome of this research could be used to develop a model based on the identified cost drivers. The ultimate aim is to amplify the COCOMO II and CoBRA model based on the identified cost drivers as the percentage (usage) of algorithmic models from empirical study depicted a need for a formal model. The rationale behind the selection of algorithmic models is the availability of the literature and free tools. Therefore, we will try to quantify the cost drivers by defining the metrics, setting criteria for each cost driver, and assigning values accordingly so that these factors could be used for improvement in the estimation process. We can also consider the cost overhead for estimation, as mentioned in [8], which will allow us to calculate the inter-relationship between the factors.

VIII. THREAT TO VALIDITY

The selected primary studies might not have reported the reasons for a particular cost driver's occurrence, which could be an internal validity threat. The mitigation of this threat is challenging for us as the origin of the cost drivers was not formally identified in the selected primary studies.

Another possible threat related to our work is that the authors of the selected primary studies are from academia, so they may not have detailed knowledge and understanding of GSD's current trends. But we mitigated this threat by evaluating the cost drivers through practitioners. Through this, we covered both academic and industrial perspectives, which strengthen our research.

TABLE 13. Summary of Research Questions.

Research Questions	Discussion
RQ1: What are the cost drivers that could affect the cost estimation in GSD?	Identified a total of 41 cost drivers from literature. The specified cost drivers are Time zone difference, Language and Cultural Differences, Communication infrastructure and process, Process model, Travel cost, Competence level, Requirements legibility, Process Compliance, Response delay, Team trust, Client unawareness, Shared resources, team structure, work dispersion, work pressure
RQ1.1: How can the identified cost drivers be categorized?	Most of the primary studies did not categorize the factors, but some authors did it through 4P's or PMBOK categorization. For our research, we selected 4P's for the categorization of the identified cost drivers. It has been found that categorization through 4P's is more detailed regarding our context. In contrast, PMBOK contains high-level knowledge areas and does not explicitly target the low-level cost estimation areas.
RQ1.2: What are the critical cost drivers in each identified category?	A total of 8 critical cost drivers are identified based on criteria (frequency >50%) in both literature and empirical study. The identified CCD were time zone difference, language and cultural difference, competence level, process compliance, communication infrastructure, geographic distance, and team trust
RQ1.3: What are the additional challenges, according to practitioners, that may affect the cost estimation in GSD?	Five additional challenges were identified from the empirical study. The extracted challenges were organization's adaptability to new technologies, pressure from higher authorities, Multiple-vendors Involvement, Over-hiring, Neglecting the QA effort, Unexpected Barriers
RQ2: Are there any supporting cost estimation metrics/techniques to handle the identified cost drivers?	The existing GSD specific cost estimation models were tailored from the traditional models. The identified GSD-specific cost estimation models were the cost overhead model, COCOMO II for GSD, COCOMO II for Offshoring, analogy-based, machine-learning based, and a model based on use case points (UCP.)
RQ2.1: What are the GSD specific cost estimation models according to the industrial perspective?	The cost estimation models identified from the empirical study were mainly Analogy based and expert judgment. It has been analyzed that industries lack formal models in this context. Only 17% of companies were using formal estimation models. While all others were relying on traditional models (analogy and expert opinion)
RQ3: What are the factors that existing cost estimation techniques lack?	It was found that the existing cost estimation techniques were at the preliminary stage. The identified models were not validated or evaluated. Some of the models were context-specific, which means that these models could not be generalized. The models also lack the additional cost drivers of GSD.

IX. CONCLUSION

The Globalization of the software industries is rapidly increasing. This rapid increase in Globalization motivated us to identify the cost drivers that could influence GSD projects' cost overhead. To achieve the devised research objectives, we conducted SLR and a survey questionnaire to identify a total of 41 cost drivers. Subsequently, the cost drivers were categorized from low significance to high significance value. Afterward, 175 project managers of various GSD organizations were selected, particularly from more than 20 countries. The cost drivers were then validated from the industrial perspective. The critical cost drivers from literature and empirical studies were language and cultural differences, time zone differences, lack of proper communication infrastructure, the team's competence level, process compliance, geographic distance, and team trust. These critical cost drivers could be used as a guide for estimating the project in the GSD context. However, the empirical study highlighted some additional challenges of cost estimation that were not presented in the literature. These additional challenges were multiple-vendor involvement, neglecting the quality assurance effort, over-hiring, pressure from higher authorities, and some unexpected barrier, i.e., the current situation of Covid-19. These additional challenges should be considered during the cost estimation of a GSD project. Neglecting these challenges can associate the hidden cost that could influence the cost overhead.

Moreover, we identified the cost estimation models presented in literature and those used in the GSD context's software industries. We identified significant differences between the results obtained from the two studies. Most of the models are formal and mathematical based on COCOMO or other algorithmic models from a literature perspective. Whereas, in our empirical analysis, the most used models were found to be an analogy-based and expert judgment, which concludes that we still lack formal models of cost estimation for the GSD context. The existing formal models are at the preliminary stage of development, and these models still need to be improved so that the software industries could use them.

We believe that this paper's findings could be used to deal with the issues associated with the cost estimation of GSD projects. Dealing with these issues is vital to the progression and success of a GSD organization.

ACKNOWLEDGMENT

The authors are sincerely thankful to the Software Reliability Engineering Group (SREG) members at COMSATS University Islamabad, who provided them with their feedback and critical analysis of the research. Moreover, they would like to extend their appreciation to the project managers who participated in their survey and provided them with a timely response.

REFERENCES

[1] A. A. Khan, S. Basri, and P. D. D. Dominc, "A proposed framework for communication risks during RCM in GSD," *Procedia-Social Behav. Sci.*, vol. 129, pp. 496–503, May 2014, doi: 10.1016/j.sbspro.2014.03.706.

[2] E. Ó. Conchuir, P. J. Ågerfalk, H. H. Olsson, and B. Fitzgerald, "Global software development: Where are the benefits?" *Commun. ACM*, vol. 52, no. 8, pp. 127–131, Aug. 2009, doi: 10.1145/1536616.1536648.

[3] S. M. A. Suliman and G. Kadoda, "Factors that influence software project cost and schedule estimation," in *Proc. Sudan Conf. Comput. Sci. Inf. Technol. (SCCSIT)*, Nov. 2017, pp. 1–9, doi: 10.1109/SCCSIT.2017.8293053.

[4] N. Ramasubbu and R. K. Balan, ''Overcoming the challenges in cost estimation for distributed software projects,'' in *Proc. 34th Int. Conf. Softw. Eng. (ICSE)*, Jun. 2012, pp. 91–101, doi: 10.1109/ICSE.2012.6227203.

[5] D. Wickramaarachchi and R. Lai, ''Effort estimation in global software development–a systematic review,'' *Comput. Sci. Inf. Syst.*, vol. 14, no. 2, pp. 393–421, 2017, doi: 10.2298/CSIS160229007W.

[6] M. R. Riaz, ''PMCMG: Project management challenges model for global software development,'' King Fahd Univ. Petroleum Minerals, Dhahran, Saudi Arabia, Tech. Rep. 31261, 2013.

[7] J. Koskenkyla, *Cost Estimation in Global Software Development—Review of Estimation Techniques*. Aaltodoc Publications, 2012, p. 109.

[8] A. Lamersdorf, J. Munch, A. F.-D.-V. Torre, C. R. Sanchez, and D. Rombach, ''Estimating the effort overhead in global software development,'' in *Proc. 5th IEEE Int. Conf. Global Softw. Eng.*, Aug. 2010, pp. 267–276, doi: 10.1109/ICGSE.2010.38.

[9] J. Ahmad, A. W. Khan, and I. Qasim, ''Software outsourcing cost estimation model (SOCEM). A systematic literature review protocol,'' *Univ. Sindh J. Inf. Commun. Technol.*, vol. 2, no. 1, pp. 25–30, 2018.

[10] M. E. Bajta, ''Analogy-based software development effort estimation in global software development,'' in *Proc. IEEE 10th Int. Conf. Global Softw. Eng. Workshops*, Jul. 2015, pp. 51–54, doi: 10.1109/ICGSEW.2015.19.

[11] M. Niazi, S. Mahmood, M. Alshayeb, M. R. Riaz, K. Faisal, N. Cerpa, S. U. Khan, and I. Richardson, ''Challenges of project management in global software development: A client-vendor analysis,'' *Inf. Softw. Technol.*, vol. 80, pp. 1–19, Dec. 2016, doi: 10.1016/j.infsof.2016.08.002.

[12] R. Jain and U. Suman, ''A project management framework for global software development,'' *ACM SIGSOFT Softw. Eng. Notes*, vol. 43, no. 1, pp. 1–10, Mar. 2018, doi: 10.1145/3178315.3178329.

[13] R. Prikladnicki, J. Nicolas Audy, and R. Evaristo, ''A reference model for global software development: Findings from a case study,'' in *Proc. IEEE Int. Conf. Global Softw. Eng. (ICGSE)*, Oct. 2006, pp. 18–25, doi: 10.1109/ICGSE.2006.261212.

[14] *Guidelines for Performing Systematic Literature Reviews in Software Engineering*, School Comput. Sci. Math., Keele Univ., Keele, U.K., 2007.

[15] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, ''Systematic literature reviews in software engineering—A systematic literature review,'' *Inf. Softw. Technol.*, vol. 51, no. 1, pp. 7–15, Jan. 2009, doi: 10.1016/j.infsof.2008.09.009.

[16] B. Kitchenham, ''Procedures for performing systematic reviews,'' *Keele, U.K., Keele Univ.*, vol. 33, no. 2004, pp. 1–26, 2004.

[17] L. Chen, M. A. Babar, and H. N. Zhang, ''Towards an evidence-based understanding of electronic data sources,'' in *Proc. 14th Int. Conf. Eval. Assessment Softw. Eng. (EASE)*, 2010, pp. 1–4, doi: 10.14236/ewic/ease2010.17.

[18] W. Afzal, R. Torkar, and R. Feldt, ''A systematic review of search-based testing for non-functional system properties,'' *Inf. Softw. Technol.*, vol. 51, no. 6, pp. 957–976, 2009, doi: 10.1016/j.infsof.2008.12.005.

[19] M. Usman and R. Britto, ''Effort estimation in co-located and globally distributed agile software development: A comparative study,'' in *Proc. Joint Conf. Int. Workshop Softw. Meas. Int. Conf. Softw. Process Product Meas. (IWSM-MENSURA)*, Oct. 2016, pp. 219–224, doi: 10.1109/IWSM-Mensura.2016.042.

[20] M. Muhairat, S. Aldaajeh, and R. E. Al-Qutaish, ''The impact of global software development factors on effort estimation methods,'' *Eur. J. Sci. Res.*, vol. 46, no. 2, pp. 221–232, 2010.

[21] T. F. C. Tait and E. H. M. Huzita, ''Software project management in distributed software development context,'' in *Proc. 15th Int. Conf. Enterp. Inf. Syst.*, vol. 2, 2013, pp. 216–222, doi: 10.5220/0004442402160222.

[22] A. Lamersdorf and J. Munch, ''Studying the impact of global software development characteristics on project goals: A causal model !2009-09-25 !2010-05-03 !2010-05-17 !'' *Open Softw. Eng. J.*, vol. 4, no. 2, pp. 2–13, May 2010, doi: 10.2174/1874107x01004020002.

[23] M. El Bajta, A. Idri, J. L. Fernández-Alemán, J. N. Ros, and A. Toval, ''Software cost estimation for global software development—A systematic map and review study,'' in *Proc. 10th Int. Conf. Eval. Novel Approaches Softw. Eng.*, 2015, pp. 197–206, doi: 10.5220/0005371501970206.

[24] R. Britto, V. Freitas, E. Mendes, and M. Usman, ''Effort estimation in global software development: A systematic literature review,'' in *Proc. IEEE 9th Int. Conf. Global Softw. Eng.*, Aug. 2014, pp. 135–144, doi: 10.1109/ICGSE.2014.11.

[25] S. Ramacharan and K. V. G. Rao, ''Parametric models for effort estimation for global software development,'' *Lect. Notes Softw. Eng.*, vol. 1, no. 2, pp. 178–182, 2013, doi: 10.7763/lnse.2013.v1.40.

[26] M. El Bajta, A. Idri, J. N. Ros, J. L. Fernandez-Aleman, J. M. C. D. Gea, F. Garcia, and A. Toval, ''Software project management approaches for global software development: A systematic mapping study,'' *Tsinghua Sci. Technol.*, vol. 23, no. 6, pp. 690–714, Dec. 2018, doi: 10.26599/TST.2018.9010029.

[27] S. Ramacharan and K. V. G. Rao, ''Scheduling based cost estimation model: An effective empirical approach for GSD project,'' in *Proc. 13th Int. Conf. Wireless Opt. Commun. Netw. (WOCN)*, Jul. 2016, pp. 1–4, doi: 10.1109/WOCN.2016.7759881.

[28] R. Britto, M. Usman, and E. Mendes, ''Effort estimation in agile global software development context,'' in *Agile Methods. Large-Scale Development, Refactoring, Testing, and Estimation* (Lecture Notes in Business Information Processing) vol. 199. Springer, 2014, pp. 182–192, doi: 10.1007/978-3-319-14358-3_15.

[29] S. Ramacharan and K. V. G. Rao, ''Software effort estimation of GSD projects using calibrated parametric estimation models,'' in *Proc. 2nd Int. Conf. Inf. Commun. Technol. Competitive Strategies (ICTCS)*, 2016, pp. 1–8, doi: 10.1145/2905055.2905177.

[30] R. Madachy, ''Distributed global development parametric cost modeling,'' in *Software Process Dynamics and Agility (ICSP)* (Lecture Notes in Computer Science), vol. 4470. Berlin, Germany: Springer, 2007, pp. 159–168, doi: 10.1007/978-3-540-72426-1_14.

[31] M. Humayun and C. Gang, ''Estimating effort in global software development projects using machine learning techniques,'' *Int. J. Inf. Educ. Technol.*, vol. 2, no. 3, Jun. 2012.

[32] S. Betz and J. Mäkiö, ''Amplification of the COCOMO II regarding offshore software projects,'' in *Proc. 2nd IEEE Int. Conf. Global Softw. Eng. (ICGSE)*, Munich, Germany, 2007.

[33] P. Keil, D. J. Paulish, and R. S. Sangwan, ''Cost estimation for global software development,'' in *Proc. Int. Workshop Econ. Driven Softw. Eng. Res. (EDSER)*, 2006, p. 7, doi: 10.1145/1139113.1139117.

[34] A. A. Khan, J. Keung, S. Hussain, M. Niazi, and S. Kieffer, ''Systematic literature study for dimensional classification of success factors affecting process improvement in global software development: Client–vendor perspective,'' *IET Softw.*, vol. 12, no. 4, pp. 333–344, Aug. 2018, doi: 10.1049/iet-sen.2018.0010.

[35] A. A. Khan and M. A. Akbar, ''Systematic literature review and empirical investigation of motivators for requirements change management process in global software development,'' *J. Softw., Evol. Process*, vol. 32, no. 4, pp. 180–205, Apr. 2020, doi: 10.1002/smr.2242.

[36] A. A. Khan and J. Keung, ''Systematic review of success factors and barriers for software process improvement in global software development,'' *IET Softw.*, vol. 10, no. 5, pp. 125–135, Oct. 2016, doi: 10.1049/iet-sen.2015.0038.

[37] P. Sharma and A. L. Sangal, ''Investigating the factors which impact SPI implementation initiatives in software SMEs—A systematic map and review,'' *J. Softw., Evol. Process*, vol. 31, no. 7, Jul. 2019, doi: 10.1002/smr.2183.

[38] G. Jamali and M. Oveisi, ''A study on project management based on PMBOK and PRINCE2,'' *Mod. Appl. Sci.*, vol. 10, no. 6, p. 142, Apr. 2016, doi: 10.5539/mas.v10n6p142.

[39] R. Britto, ''Knowledge classification for supporting effort estimation in global software engineering projects,'' Blekinge Inst. Technol. Licentiate, Karlskrona, Sweden, Series 2015:04, 2015.

JUNAID ALI KHAN received the B.E. degree in software engineering from COMSATS University Islamabad, Wah Campus, Pakistan, in 2017, where he is currently pursuing the M.S. degree in software engineering. His research interests include software project management, software process improvement, and their application in GSD context.

SAIF UR REHMAN KHAN received the Ph.D. degree in software engineering/software testing from the University of Malaya, Kuala Lumpur, Malaysia, in 2018. He is currently serving with the Department of Computer Science, COMSATS University Islamabad (CUI), Islamabad, Pakistan. His research interests include software engineering, search-based software engineering, verification and validation, cyber-physical systems, model-based testing, software reuse, requirements engineering, and software project management. He has been in several experts' review panels, both locally and internationally. He was a recipient of the Best Paper Presentation Award from the Faculty of Computer Science and Information Technology, UM, in 2014, and the Paper Reviewing Award (Future Generation Computer Systems) in 2016.

JAVED IQBAL received the Ph.D. degree in computer sciences from the University of Malaya (UM), Kuala Lumpur, Malaysia, in 2016. He is currently an Assistant Professor with the Department of Computer Sciences, COMSATS University Islamabad, Pakistan. His research interests include software process improvement, requirements engineering, software development outsourcing, global software development, and machine learning.

INAYAT UR REHMAN received the Ph.D. degree in computer sciences from COMSATS University Islamabad, Pakistan, in 2017. He is currently an Assistant Professor with the Department of Computer Sciences, COMSATS University Islamabad. His research interests include computer assisted education, designing learning tools, computer animations for learning, role of human computer in designing learning tools, cognitive learning, and use of educational psychology for e-learning applications. Using his expertise of development side, he also designed learning tools and tutorials. He is also working on investigating impact of learning environments and sense of belongings in learnability.