The Case for Bayesian Meta-analysis as a Complement to Systematic Literature Reviews in Software Engineering: Evidence from Defect Prediction

Vinicius H. S. Durelli
Universidade Federal de São Carlos
São Paulo, Brazil
vinicius.durelli@ufscar.br

Andre T. Endo
Universidade Federal de São Carlos
São Paulo, Brazil
andreendo@ufscar.br

Fabiano C. Ferrari
Universidade Federal de São Carlos
São Paulo, Brazil
fcferrari@ufscar.br

Abstract

In Software Engineering (SE), systematic literature reviews (SLRs) are widely used to organize and synthesize research evidence. However, given the scant adoption of meta-analysis, evidence synthesis in SE still relies predominantly on narrative aggregation. This paper argues that Bayesian meta-analysis should be adopted more broadly as a complement to SLRs in SE. To ground this argument empirically, we investigate software defect prediction as an application context and examine the overall predictive effectiveness of machine learning models reported in the selected studies. Rather than conducting a new SLR, we relied on an existing SLR as the basis for our study and applied additional eligibility criteria to identify studies suitable for quantitative synthesis. From the selected primary studies, we extracted *model* × *dataset* observations and analyzed them using a Bayesian hierarchical model. The results suggest that defect prediction models achieve, on average, a moderate level of predictive effectiveness. Importantly, this result is expressed as a posterior distribution, which provides a pooled quantitative estimate together with an explicit representation of uncertainty. Because the result is obtained by pooling information across the selected studies, it provides a more coherent quantitative summary of the literature than a narrative account or a simple collection of descriptive statistics. The analysis also indicates that this effectiveness is not uniform across the included evidence, but varies across the empirical contexts represented in the selected studies. Our results show that Bayesian meta-analysis can yield insights that would remain inaccessible under a conventional SLR, including pooled quantitative estimates, probabilistic characterizations of uncertainty, and a fine-grained view of heterogeneity across empirical contexts. More broadly, the paper makes the case that, whenever SE researchers have already invested the effort required to identify, screen, and extract structured data from primary studies, they should take the further step of subjecting that corpus to a Bayesian meta-analysis when the underlying evidence permits.

Keywords

Bayesian Meta-analysis, Evidence-based Software Engineering, Systematic Literature Review, Software Defect Prediction

1 Introduction

Systematic literature reviews (SLRs) are carried out to identify, retrieve, and examine the body of evidence relevant to a given topic [11]. The collected evidence is then synthesized to generate an integrated understanding of the topic. Central to this process is the application of procedures that are transparent, rigorous, and reproducible, with the objective of mitigating subjectivity. Ideally, a systematic review should be conducted in such a way that independent researchers applying the same review protocol would identify a comparable set of primary studies and arrive at similar conclusions. Systematic reviews are often referred to as secondary studies, owing to the fact that they generate their outcomes by aggregating material from a set of primary studies [4, 11].

SLRs have transitioned from a sparsely used methodological approach [14] to a mainstream and well-established research methodology in Software Engineering (SE) [1]. Given that systematic reviews can take different forms [3], the corresponding approaches to evidence synthesis also vary considerably in rigor and analytical depth. In SE, secondary studies have predominantly taken the form of SLRs and mapping studies. These approaches have contributed to organizing the literature, identifying clusters of related studies, and revealing gaps where additional primary research is needed. However, despite the laudable purpose of systematic reviews to provide an objective and evidence-based synthesis, they have largely emphasized classification and narrative synthesis rather than quantitative integration of empirical findings.

Given the growing emphasis on evidence-based inquiry in SE [17], we posit that researchers should move beyond relying exclusively on SLRs as the primary mechanism for secondary synthesis. In particular, whenever the available primary evidence allows for quantitative integration, SLRs should be complemented by meta-analyses. Despite their well-established role in other disciplines, meta-analyses remain seldom used in the context of SE research [11], which may in part suggest that meta-analytic methods are not yet widely understood within the SE community [13].

Meta-analysis is a statistical methodology designed to aggregate and synthesize research results across multiple studies [7, 8]. By pooling empirical evidence, it can provide greater statistical power and inferential robustness than individual studies considered in isolation [11]. Moreover, it enables a form of quantitative aggregation that is methodologically more rigorous and formally superior to the predominantly narrative synthesis typically employed in SLRs and mapping studies.

As mentioned, in more established evidence-based disciplines (e.g., medicine and psychology), quantitative synthesis through meta-analysis has evolved into a standard practice for consolidating empirical findings [5, 6, 8, 18, 19], with SLRs typically providing the foundation upon which meta-analyses are performed. By comparison, SE has only marginally adopted such approaches. This is partly due to the methodological challenges posed by SE experiments, including small sample sizes, heterogeneous study designs (e.g.,

cross-over and repeated measures), and non-normal outcome distributions. As a result, SE researchers have only marginally adopted advanced meta-analytic methods, often relying on simplified or ad hoc aggregation strategies.

More recently, researchers have begun to address these limitations by proposing meta-analytic procedures tailored to the specific characteristics of SE experiments. Specifically, the study by Kitchenham and Madeyski [12] introduces a framework based on non-parametric effect sizes, i.e., the probability of superiority ($\hat{p}$) and Cliff's d, designed to improve robustness in the presence of small samples and non-normal data. The proposed approach also revisits variance estimation and provides guidance for aggregating results across families of experiments. Despite these advances, most approaches remain largely grounded in a frequentist paradigm and, as such, inherit several of its well-known limitations. In particular, uncertainty is typically summarized through point estimates and confidence intervals, which provide only a partial and sometimes unintuitive characterization of the variability in the aggregated evidence. By contrast, a Bayesian approach represents uncertainty using full probability distributions, allowing researchers to directly quantify how likely different outcomes are. This enables more intuitive interpretations (e.g., the probability that one technique outperforms another) and provides more robust estimates in the presence of small samples. Additionally, Bayesian methods enable the incorporation of prior knowledge (i.e., expert judgment or evidence from previous studies) into the analysis, thereby supporting a more cumulative and transparent process of evidence synthesis [7].

In this paper, we argue that a Bayesian approach to meta-analysis provides a principled and flexible alternative that is well suited to the challenges encountered in SE. By modeling effect sizes within a hierarchical probabilistic framework, Bayesian meta-analysis naturally accounts for between-study heterogeneity and enables the incorporation of prior knowledge when available. Furthermore, it yields full posterior distributions over parameters of interest, allowing for direct probabilistic statements about effect magnitudes and differences between techniques. To ground this argument in an empirical setting, we decided not to undertake an SLR. Instead, we chose an existing SLR in a domain that lends itself well to quantitative synthesis. We selected software defect prediction as the application context because its literature includes relatively standardized benchmark datasets and commonly reported performance metrics, thus providing favorable conditions for a Bayesian meta-analysis. It is worth noting that defect prediction is not the primary focus of this paper. Rather, we selected this body of evidence, which is relatively standardized in terms of benchmark datasets and performance metrics, to illustrate the kinds of insights that become accessible when the extracted data are subjected to a more elaborate quantitative synthesis.

In this context, this paper has two overarching goals. First, it argues that the adoption of Bayesian meta-analysis in SE would represent an important step toward more rigorous empirical research. Second, it illustrates the feasibility and analytical benefits of this framework through an empirical application in defect prediction. In doing so, the paper advances Bayesian meta-analysis as a promising alternative to current evidence synthesis practices in SE, which remain constrained by both the limited adoption of meta-analysis and the limitations of existing approaches. More specifically, we argue that, whenever researchers in SE have already undertaken the considerable effort required to identify, screen, and extract structured data from primary studies in the course of an SLR, they should, whenever the underlying evidence permits, take the further step of subjecting that corpus to a Bayesian meta-analysis. Its probabilistic and hierarchical formulation is better suited to the characteristics of SE data, including small samples, heterogeneous experimental designs, and non-normal distributions. More broadly, this study is intended to stimulate discussion on the role of probabilistic methods in SE and to encourage a shift toward a more robust and cumulative methodology for synthesizing evidence in SLRs.

The remainder of this paper is organized as follows. Section 2 describes the design of the Bayesian meta-analysis. Section 3 presents the results of the Bayesian meta-analysis, focusing on the posterior distribution of the overall expected F1-score and on the heterogeneity captured by the model. Section 4 discusses the implications of these findings in relation to the original SLR and to evidence synthesis practices in SE. Section 5 presents concluding remarks.

2 Meta-analysis Design

As mentioned, in more established fields, SLRs provide the foundation upon which meta-analyses are conducted. Accordingly, to obviate the need to conduct an SLR, we decided to rely on an already published SLR as the basis for our study. Specifically, to identify relevant primary studies for our meta-analysis, we relied on an existing SLR as a starting point. The selected SLR focuses on software defect prediction using artificial intelligence [16], a topic that lends itself to quantitative synthesis due to the prevalence of standardized evaluation datasets and performance metrics. This choice allows us to leverage an already curated body of literature without having to conduct a systematic search and screening process. We formulated a research question (RQ) that was not explicitly addressed in the original SLR: we set out to examine more the overall effectiveness of machine learning models for defect prediction. To investigate this systematically, we posed the following RQ:

> **RQ:** What is the overall predictive effectiveness of defect prediction models across the selected studies, as measured by F1-score?

Upon closer examination of the set of studies identified by Pachouly et al. [16], we observed that not all primary studies are suitable for inclusion in a meta-analysis. Consequently, we established a set of eligibility criteria to identify the subset of studies that meet the requirements of a quantitative synthesis. These criteria, detailed in the next subsection, are mostly designed to ensure comparability across studies, particularly with respect to the datasets employed, the evaluation procedures adopted, and the performance metrics reported. In doing so, we avoid aggregating fundamentally incomparable results, thereby supporting the validity of the resulting meta-analysis.

2.1 Eligibility Criteria

Although the selected SLR reports a total of 146 primary studies, not all primary studies are aligned with the objectives of our meta-analysis. Specifically, a substantial subset of studies focuses on

Table 1: Selected primary studies included in the meta-analysis

ID	Study	Focus	Caveats
S1	Ji et al. [10], *A new weighted naive Bayes method based on information diffusion for software defect prediction*	Weighted Naive Bayes for defect prediction	Reports dataset-level Recall, Precision, and F-measure on PROMISE datasets. Uses repeated cross-validation, but reports only mean values, with no variance estimates.
S2	Ali et al. [2], *Software Defect Prediction Using Variant based Ensemble Learning and Feature Selection Techniques*	Ensemble learning and feature selection	Reports dataset-level F-measure on NASA/PROMISE datasets. Uses a 70/30 holdout split rather than repeated cross-validation and does not report variance estimates.
S3	Naseem et al. [15], *Investigating Tree Family Machine Learning Techniques for a Predictive System to Unveil Software Defects*	Tree-family machine learning techniques	Reports dataset-level F-measure across multiple NASA/PROMISE datasets and classifiers. Uses 10-fold cross-validation, but reports only mean values.
S4	Ji et al. [9], *A Two-Stage Feature Weighting Method for Naive Bayes and Its Application in Software Defect Prediction*	Two-stage feature weighting for Naive Bayes	Reports dataset-level Recall, Precision, and F-measure on NASA/PROMISE datasets. Uses repeated 10×10 cross-validation, but reports only mean values, with no variance or confidence intervals.

ancillary aspects of defect prediction, such as dataset construction, class imbalance handling, feature selection, and data preprocessing techniques [16]. While these contributions are important to the broader defect prediction literature, they do not directly evaluate or compare predictive modeling approaches. Therefore, to be eligible, studies must investigate software defect prediction as a binary classification problem (i.e., defective vs. non-defective) and report quantitative performance results. More specifically, eligible studies are required to report performance at the dataset level (i.e., results for individual datasets and models), allowing the extraction of observations of the form *model* × *dataset*. To support a quantitative synthesis, studies must report at least one of the following metrics: recall, precision, or F-measure.[1] It is worth noting that preference was given to studies reporting the F-measure or combinations of recall and precision, as these metrics are consistent with the construct of prediction effectiveness adopted in this study. Furthermore, selected studies must use datasets from the PROMISE repository or closely related NASA datasets, ensuring a comparable experimental context across studies.

We explicitly exclude studies that report only aggregated results (e.g., averages across datasets) or lack extractable numeric values. These decisions were made to ensure that all included studies support a consistent unit of analysis, i.e., dataset-level observations, and enable quantitative synthesis. Similarly, studies reporting metrics that are not compatible with our extraction protocol were excluded, as their inclusion would introduce heterogeneity that could compromise the validity and interpretability of the meta-analytic model. After this additional filtering step, we ended up with the set of primary studies shown in Table 1.

2.2 Data Extraction

As shown in Table 1, the selected primary studies exhibit several methodological limitations, most notably the absence of variance estimates and the reliance on different evaluation protocols. While these factors would typically pose challenges for traditional meta-analytic approaches, they do not preclude a Bayesian meta-analysis.

From the four selected studies shown in Table 1, we extracted a total of 184 *model* × *dataset* observations, which constitute the unit of analysis in our Bayesian meta-analysis. Although multiple performance metrics (e.g., recall and precision) were extracted when available, the present analysis focuses on the F1-score as the primary measure of predictive effectiveness, given its ability to capture the balance between these metrics. In the interest of space, the complete set of observations is not included in the paper but is made available in the replication package.

2.3 Summary of Extracted Data

Table 2 shows a summary of the defect prediction models extracted from the selected studies, including the number of datasets, evaluation procedures, and mean F1-scores reported for each model. As shown, the studies include a mix of baseline machine learning techniques and proposed approaches, with the latter consistently achieving the highest performance within each study. In total, the meta-analysis considers a diverse set of models spanning multiple families of machine learning techniques, including probabilistic models (e.g., Naive Bayes), tree-based methods (e.g., Decision Trees, Random Forest, and their variants), instance-based learners (e.g., k-Nearest Neighbors), and neural approaches (e.g., Multilayer Perceptron).

Several models appear in more than one study, most notably Naive Bayes, Random Forest, and Random Tree, which are evaluated across different experimental settings and datasets. This overlap is particularly valuable for the meta-analysis, as it enables indirect comparisons across studies and contributes to a more connected body of evidence.

It is worth noting that S3 does not introduce a proposed approach, but instead evaluates a broad set of baseline models within the tree-family of classifiers. As such, the best-performing model in this study (Forest-PA) corresponds to a variant within this family rather than to a novel technique. This contrasts with the other studies, in which explicitly proposed approaches (e.g., NB-ID, the ensemble

[1]Throughout this paper, we use the terms F-measure and F1-score interchangeably, as the selected studies report the standard F1 formulation, i.e., the harmonic mean of precision and recall.

Table 2: Summary of defect prediction models extracted from the selected studies, including number of datasets, evaluation protocols, mean F1-scores, and model category. Models are categorized as baseline or proposed.

ID	Model	#Datasets	Evaluation	Mean F1	Category
S1	(NB-ID) Naive Bayes with Information Diffusion	10	10×10 cross-validation	**0.670**	proposed (best-performing)
S1	(NB) Naive Bayes	10	10×10 cross-validation	0.597	baseline
S2	(ProposedEnsemble) Ensemble Learning Model	4	70/30 holdout	**0.621**	proposed (best-performing)
S2	(DT) Decision Tree	4	70/30 holdout	0.473	baseline
S2	(RF) Random Forest	4	70/30 holdout	0.430	baseline
S2	(KNN) k-Nearest Neighbors	4	70/30 holdout	0.420	baseline
S2	(KStar) K* Instance-Based Learner	4	70/30 holdout	0.384	baseline
S2	(NB) Naive Bayes	4	70/30 holdout	0.348	baseline
S2	(MLP) Multilayer Perceptron	4	70/30 holdout	0.341	baseline
S2	(OneR) One Rule Classifier	4	70/30 holdout	0.305	baseline
S2	(PART) Partial Decision Tree	4	70/30 holdout	0.277	baseline
S2	(RBF) Radial Basis Function Network	4	70/30 holdout	0.257	baseline
S2	(SVM) Support Vector Machine	4	70/30 holdout	0.156	baseline
S3	(Forest-PA) Forest by Penalizing Attributes	10	10-fold cross-validation	**0.894**	baseline (best-performing)
S3	(DS) Decision Stump	10	10-fold cross-validation	0.889	baseline
S3	(J48) J48 Decision Tree	10	10-fold cross-validation	0.851	baseline
S3	(HT) Hoeffding Tree	10	10-fold cross-validation	0.850	baseline
S3	(CS-Forest) Cost-Sensitive Forest	10	10-fold cross-validation	0.842	baseline
S3	(LMT) Logistic Model Tree	10	10-fold cross-validation	0.839	baseline
S3	(RF) Random Forest	10	10-fold cross-validation	0.834	baseline
S3	(CDT) Categorical Decision Tree	10	10-fold cross-validation	0.834	baseline
S3	(REP-T) Reduced Error Pruning Tree	10	10-fold cross-validation	0.831	baseline
S3	(RT) Random Tree	10	10-fold cross-validation	0.810	baseline
S4	(TSWNB) Two-Stage Weighted Naive Bayes	5	10×10 cross-validation	**0.413**	proposed (best-performing)
S4	(RT) Random Tree	5	10×10 cross-validation	0.349	baseline
S4	(NB) Naive Bayes	5	10×10 cross-validation	0.340	baseline
S4	(LR) Logistic Regression	5	10×10 cross-validation	0.300	baseline

model, and TSWNB) outperform their respective baselines. The table also highlights the diversity of evaluation protocols employed, ranging from repeated cross-validation to holdout validation, which underscores the need for careful alignment of experimental settings when aggregating results.

Table 3 summarizes the datasets used across the selected studies. As shown in Table 3, there is some degree of overlap, particularly among datasets from the PROMISE and NASA repositories (e.g., KC1, JM1, and PC1), which are used in multiple studies. This overlap is beneficial for the purposes of meta-analysis, as it enhances comparability across studies while maintaining a consistent experimental context. At the same time, differences in dataset composition across studies introduce a degree of heterogeneity that is explicitly accounted for in the meta-analytic model.

2.4 Bayesian Hierarchical Modeling

The overall predictive effectiveness of defect prediction models was assessed using a Bayesian hierarchical model, implemented in PyMC.[2] Specifically, to answer our RQ, we modeled the F1-score using a Beta distribution rather than a Gaussian one.[3] This choice requires the response variable to lie strictly within the open interval $(0, 1)$. Accordingly, as part of the data preprocessing stage, and prior to specifying the statistical model, we excluded observations with missing F1 values and verified that all retained observations satisfied this requirement.

Because the extracted F1-scores are clustered within both studies and datasets, we modeled these sources of dependence explicitly through a hierarchical formulation. To this end, study and dataset identifiers were encoded as grouping variables, allowing the model to estimate separate random effects for each study and each dataset. This specification accounts for the non-independence of observations arising from shared study-level and dataset-level characteristics, while still enabling the estimation of an overall pooled level of predictive effectiveness. More formally, let y_i denote the F1-score associated with observation i, where each observation corresponds to a *model* × *dataset* result. We assume

$$y_i \sim \text{Beta}(\alpha_i, \beta_i),$$

with

$$\alpha_i = \theta_i \phi \qquad \text{and} \qquad \beta_i = (1 - \theta_i)\phi,$$

where $\theta_i \in (0, 1)$ is the expected F1-score for observation i, and $\phi > 0$ is a precision parameter controlling the dispersion of the Beta distribution.

The expected F1-score is linked to a global intercept and study- and dataset-level random effects through a logit link:

$$\text{logit}(\theta_i) = \mu + u_{s(i)} + v_{d(i)},$$

[2] A probabilistic programming framework for Bayesian inference (https://www.pymc.io/welcome.html).

[3] An inspection of the empirical distribution of the extracted F1-scores indicated a bounded and potentially asymmetric distribution, which further supported the use of a Beta likelihood rather than a Gaussian outcome model.

where μ denotes the overall average predictive effectiveness on the logit scale, $u_{s(i)}$ is the random effect associated with the study corresponding to observation i, and $v_{d(i)}$ is the random effect associated with the dataset corresponding to observation i. The random effects are assumed to follow zero-centered normal distributions:

$$u_s \sim \mathcal{N}(0, \sigma^2_{\text{study}}) \qquad \text{and} \qquad v_d \sim \mathcal{N}(0, \sigma^2_{\text{dataset}}).$$

Under this specification, μ captures the pooled average level of predictive effectiveness across the included evidence, whereas σ_{study} and σ_{dataset} quantify the extent of between-study and between-dataset heterogeneity, respectively.

Table 3: Datasets used in the selected primary studies included in the meta-analysis.

ID	#Datasets	Datasets
S1	10	CM1, Ivy-1.1, Lucene-2.2, Lucene-2.4, MW1, PC4, Poi-1.5, Poi-2.5, Velocity-1.6, Xalan-2.6
S2	4	JM1, KC1, PC4, PC5
S3	10	AR1, AR3, CM1, KC2, KC3, MW1, PC1, PC2, PC3, PC4
S4	5	CM1, MW1, PC1, PC3, PC4

3 Bayesian Meta-analysis Results

To facilitate interpretation, posterior draws of μ were transformed back to the original F1 scale using the inverse-logit function, yielding the posterior distribution of the overall expected F1-score shown in Figure 1. As shown in Figure 1 the posterior mean is centered at approximately 0.59, with a 95 % highest-density interval (HDI) ranging from about 0.33 to 0.81. This suggests that, after accounting for variation across studies and datasets, the overall predictive effectiveness of the included defect prediction models can be characterized as moderate (although the associated uncertainty is non-negligible). The width of the posterior interval suggests substantial heterogeneity in predictive performance across the underlying evidence base, which is in line with the methodological and empirical differences observed among the selected studies. Thus, with respect to our RQ, the analysis suggests that machine learning defect prediction models achieve, on average, a moderate level of predictive effectiveness when performance is summarized using the F1-score.

The uncertainty associated with the pooled F1-score is consistent with the posterior heterogeneity estimates obtained from the hierarchical model. As shown in Table 4, the posterior mean of σ_{study} was 1.178 (95% HDI [0.543, 1.944]), while that of σ_{dataset} was 1.280 (95% HDI [0.828, 1.740]). These results indicate that the predictive effectiveness of defect prediction models is not constant across the evidence base, but instead varies across studies and datasets. In particular, the slightly larger value of σ_{dataset} suggests that differences between datasets may contribute somewhat more to performance variation than differences between studies.[4]

To further examine dataset-level heterogeneity, we inspected the posterior random effects associated with each dataset, summarized in the caterpillar plot shown in Figure 2. In Figure 2, each point represents the posterior mean of the random effect associated with a given dataset, whereas the corresponding horizontal interval denotes its 95 % highest-density interval. The dashed vertical line at zero marks the pooled mean on the logit scale. Thus, datasets located to the right of zero are associated with higher-than-average predictive effectiveness, whereas datasets located to the left are associated with lower-than-average predictive effectiveness.

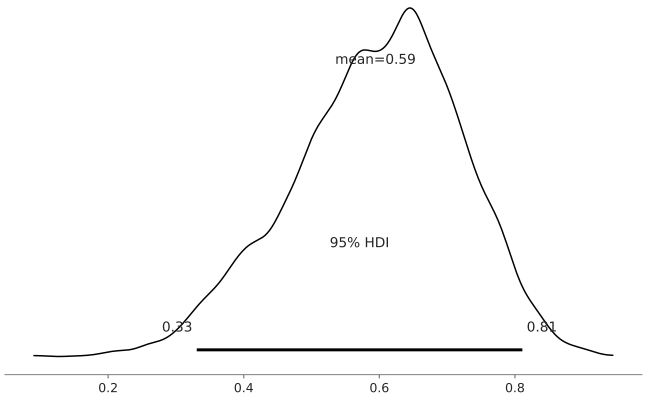

Figure 1: Posterior distribution of the overall F1-score.

Table 4: Posterior summaries of the study- and dataset-level heterogeneity parameters.

Parameter	Mean	SD	95% HDI
σ_{study}	1.178	0.384	[0.543, 1.944]
σ_{dataset}	1.280	0.243	[0.828, 1.740]

As mentioned, predictive effectiveness is not distributed uniformly across datasets: some datasets are associated with higher-than-average performance, whereas others are associated with lower-than-average performance. In particular, datasets such as PC2 and several of the Apache-related datasets appear on the positive side of the pooled mean, while datasets such as KC3, JM1, KC1, PC5, and PC4 appear on the negative side (Figure 2). Intervals lying largely away from zero indicate more pronounced dataset-specific deviations from the pooled mean, whereas intervals overlapping zero suggest datasets whose predictive effectiveness is not clearly distinguishable from the overall average. As shown in the caterpillar plot (Figure 2), these deviations are not evenly scattered around the pooled mean but are distributed across clearly positive and negative regions. This pattern supports the interpretation of σ_{dataset} as evidence of substantial dataset-level heterogeneity and suggests that the pooled posterior F1-score should be understood as an average over markedly different dataset contexts rather than as a homogeneous level of predictive effectiveness.

Key takeaway regarding our RQ: Our Bayesian meta-analysis indicates that defect prediction models achieve a *moderate* level of predictive effectiveness, while also exhibiting variation in performance across datasets.

[4]The heterogeneity parameters σ_{study} and σ_{dataset} are defined on the logit scale, since they represent the standard deviations of the corresponding random effects in the linear predictor. As such, they quantify relative heterogeneity and should not be interpreted as direct differences in F1-score.

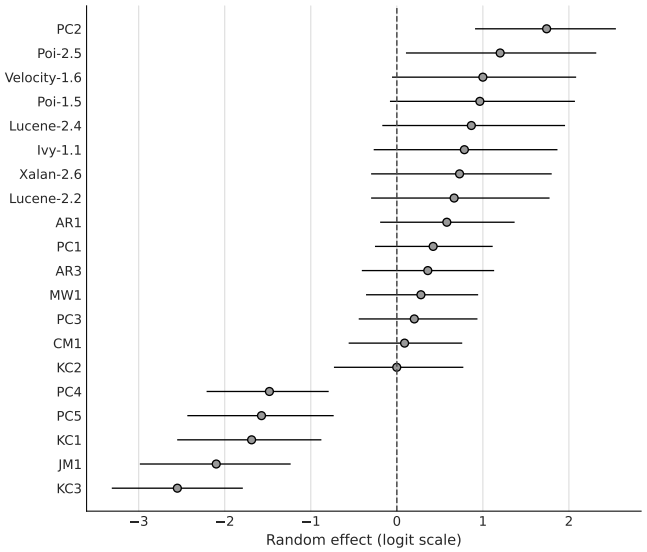

Figure 2: Summary of the posterior dataset-level random effects: points denote posterior means and horizontal lines denote 95 % HDIs, effects are shown relative to the pooled mean on the logit scale.

4 Discussion

A difference of our study, relative to the original SLR, lies in the granularity of the resulting evidence synthesis. Whereas the SLR provides a broad and valuable narrative overview of defect prediction research, it does not provide a pooled quantitative estimate of predictive effectiveness nor a formal account of variation across empirical contexts. By contrast, although our analysis is based on only four primary studies selected from the broader set of studies identified in the original SLR, we refine the unit of analysis to the level of *model* × *dataset* observations and subject this structured corpus to a Bayesian meta-analysis. This shift in granularity makes it possible to estimate the overall expected F1-score, quantify uncertainty through posterior distributions (rather than relying on narrative claims about performance), and explicitly model heterogeneity across both studies and datasets.

This finer-grained quantitative synthesis yields insights that are not accessible through the original SLR. Most notably, even with this deliberately scoped subset of four studies, the analysis allows us to conclude that defect prediction models achieve, on average, a moderate level of predictive effectiveness, while also showing that this performance is not uniform across datasets. Moreover, the posterior heterogeneity parameters indicate that predictive effectiveness varies substantially across the studies and datasets included in the analysis. Accordingly, our study shows how a Bayesian meta-analysis can turn a selected subset of an already curated SLR corpus into a quantitatively integrated body of evidence, thereby supporting conclusions about overall effectiveness, uncertainty, and context sensitivity that would otherwise remain implicit.

5 Concluding Remarks

SLRs have become a mainstream form of secondary research in SE. However, these studies rarely culminate in a formal quantitative synthesis of the extracted evidence. In many cases, the results of SLRs are summarized narratively or through descriptive tabulations, even when the underlying corpus contains empirical results that couldbe aggregated more systematically. As pointed out by Kitchenham et al. [11], if evidence-based SE is to become more useful to the software industry, as well as attain the maturity level of other practical sciences, SE researchers need to move beyond predominantly narrative forms of synthesis and conduct more meta-analyses capable of statistically aggregating empirical findings across primary studies. We argue that this gap is partly methodological. Traditional meta-analytic approaches often require assumptions and reported information that are not always available in SE studies, such as comparable effect-size definitions, variance estimates, and relatively homogeneous experimental designs. By contrast, Bayesian meta-analysis provides a more flexible framework for synthesizing the kinds of data commonly encountered in SE reviews [7]. Its probabilistic and hierarchical formulation can accommodate small samples, unbalanced evidence, heterogeneous datasets, and partially comparable study designs.Thus, when researchers have already assembled a body of primary evidence through an SLR, Bayesian meta-analysis offers a principled way to take the next step from descriptive synthesis to cumulative quantitative inference.

Our selection of four primary studies was not intended to support an exhaustive meta-analysis of the defect prediction literature. Rather, it reflects a scoping decision aligned with the methodological purpose of this paper. Including a broader set of studies would likely require a more complex modeling strategy, capable of accommodating additional outcome metrics, prediction settings, evaluation protocols, and forms of missing or partially reported information. While such an extension is both possible and desirable, it would shift the emphasis of the paper from giving an overview of the value of Bayesian meta-analysis to addressing the methodological complexity of defect prediction as a research domain. Accordingly, our goal is straightforward: to show that, when an SLR has already produced a structured body of evidence, Bayesian meta-analysis can provide a principled next step for quantitative synthesis. We hope that this paper encourages SE researchers to carry out meta-analyses as a complement to systematic literature reviews, particularly when the available evidence allows for quantitative synthesis.

Artifact Availability

The dataset and analysis scripts are publicly available in our Zenodo replication package: https://doi.org/10.5281/zenodo.20347487. The analysis script also includes posterior predictive checks, which compare the observed F1-scores with replicated data generated from the fitted posterior distribution, thereby providing a visual assessment of model adequacy.

Acknowledgements

This work was supported by the National Council for Scientific and Technological Development (CNPq) under grant nos. 444311/2024-6, 406941/2025-4 and 310967/2025-2.

References

[1] Nauman B. Ali and Muhammad Usman. 2018. Reliability of search in systematic reviews: Towards a quality assessment framework for the automated-search strategy. *Information and Software Technology* 99 (2018), 133–147.

[2] Umair Ali, Shabib Aftab, Ahmed Iqbal, Zahid Nawaz, Muhammad Salman Bashir, and Muhammad Anwaar Saeed. 2020. Software Defect Prediction Using Variant-Based Ensemble Learning and Feature Selection Techniques. *International Journal of Modern Education and Computer Science* 12, 5 (2020), 29–40.

[3] Andrew Booth, Anthea Sutton, Mark Clowes, and Marrissa Martyn-St James. 2026. *Systematic Approaches to a Successful Literature Review* (4th ed.). Sage Publications Ltd. 416 pages.

[4] David Budgen and Pearl Brereton. 2006. Performing systematic literature reviews in software engineering. In *Proceedings of the 28th International Conference on Software Engineering* (Shanghai, China) *(ICSE '06)*. ACM, New York, NY, USA, 1051–1052.

[5] Christopher Cox, Christina Bergmann, Emma Fowler, Tamar Keren-Portnoy, Andreas Roepstorff, Greg Bryant, and Riccardo Fusaroli. 2023. A systematic review and Bayesian meta-analysis of the acoustic features of infant-directed speech. *Nature Human Behaviour* 7, 1 (2023), 114–133.

[6] Jordan Garrett, Carly Chak, Tom Bullock, and Barry Giesbrecht. 2024. A systematic review and Bayesian meta-analysis provide evidence for an effect of acute physical activity on cognition in young adults. *Communications Psychology* 2, 1 (2024), 82.

[7] Robert Grant and Gian Luca Di Tanna. 2025. *Bayesian Meta-Analysis: A Practical Introduction.* CRC Press. 304 pages.

[8] Jessica Gurevitch, Julia Koricheva, Shinichi Nakagawa, and Gavin Stewart. 2018. Meta-analysis and the science of research synthesis. *Nature* 555, 7695 (2018), 175–182.

[9] Haijin Ji, Song Huang, Xuewei Lv, Yaning Wu, and Zhanwei Hui. 2018. A Two-Stage Feature Weighting Method for Naive Bayes and Its Application in Software Defect Prediction. *International Journal of Performability Engineering* 14, 7 (2018), 1468.

[10] Haijin Ji, Song Huang, Yaning Wu, Zhanwei Hui, and Changyou Zheng. 2019. A new weighted naive Bayes method based on information diffusion for software defect prediction. *Software Quality Journal* 27, 3 (2019), 923–968.

[11] Barbara Kitchenham, David Budgen, and Pearl Brereton. 2015. *Evidence-Based Software Engineering and Systematic Reviews.* Chapman and Hall/CRC Press. 434 pages.

[12] Barbara Kitchenham and Lech Madeyski. 2024. Recommendations for analysing and meta-analysing small sample size software engineering experiments. *Empirical Software Engineering* 29, 6 (2024), 137.

[13] Barbara Kitchenham, Lech Madeyski, and Pearl Brereton. 2020. Meta-analysis for families of experiments in software engineering: a systematic review and reproducibility and validity assessment. *Empirical Software Engineering* 25, 1 (2020), 353–401.

[14] Barbara Kitchenham, Rialette Pretorius, David Budgen, O. Pearl Brereton, Mark Turner, Mahmood Niazi, and Stephen Linkman. 2010. Systematic literature reviews in software engineering - A tertiary study. *Information and Software Technology* 52, 8 (2010), 792–805.

[15] Rashid Naseem, Bilal Khan, Arshad Ahmad, Ahmad Almogren, Saima Jabeen, Bashir Hayat, and Muhammad Arif Shah. 2020. Investigating Tree Family Machine Learning Techniques for a Predictive System to Unveil Software Defects. *Complexity* 2020, 1 (2020), 6688075.

[16] Jalaj Pachouly, Swati Ahirrao, Ketan Kotecha, Ganeshsree Selvachandran, and Ajith Abraham. 2022. A systematic literature review on software defect prediction using artificial intelligence: Datasets, Data Validation Methods, Approaches, and Tools. *Engineering Applications of Artificial Intelligence* 111 (2022), 104773.

[17] S. Lawrence Pfleeger and Barbara Kitchenham. 2025. Evidence-Based Software Engineering Guidelines Revisited . *IEEE Transactions on Software Engineering* 51, 03 (2025), 814–819.

[18] Ashish Shetty, Gayathri Delanerolle, Chunli Deng, Anish Thillainathan, Heitor Cavalini, Xiaojie Yang, Yassine Bouchareb, Amy Boyd, Peter Phiri, Jian Qing Shi, and Timothy Deer. 2024. A systematic review and bayesian meta-analysis of medical devices used in chronic pain management. *Scientific Reports* 14, 1 (2024), 13549.

[19] Steeven Ye, David Attali, Maria Ghazi, Arnaud Cachia, Mathieu Cassotti, and Grégoire Borst. 2026. Systematic review and meta-analysis of the evidence for an illusory truth effect and its determinants. *Nature Communications* 17, 1 (2026), 3270.