Pontifícia Universidade Católica
do Rio de Janeiro

Marcelo Costalonga Cardoso

Can Machine Learning Replace a Reviewer in the Selection of Studies for Systematic Literature Review Updates?

Dissertação de Mestrado

Dissertation presented to the Programa de Pós–graduação em Informática of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Informática.

Advisor: Prof. Marcos Kalinowski

Rio de Janeiro
April 2024

Pontifícia Universidade Católica
do Rio de Janeiro

Marcelo Costalonga Cardoso

Can Machine Learning Replace a Reviewer in the Selection of Studies for Systematic Literature Review Updates?

Dissertation presented to the Programa de Pós–graduação em Informática of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Informática. Approved by the Examination Committee:

Prof. Marcos Kalinowski
Advisor
Departamento de Informática – PUC-Rio

Prof. Markus Endler
PUC-Rio

Prof. Maria Teresa Baldassarre
Uniba

Rio de Janeiro, April 29th, 2024

Marcelo Costalonga Cardoso

Graduated in computer engineering by the Pontifical Catholic University of Rio de Janeiro (PUC-Rio) at 2021.

Bibliographic data

Cardoso, Marcelo Costalonga

Can Machine Learning Replace a Reviewer in the Selection of Studies for Systematic Literature Review Updates? / Marcelo Costalonga Cardoso; advisor: Marcos Kalinowski. – 2024.

50 f: il. color. ; 30 cm

Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2024.

Inclui bibliografia

1. Informática – Teses. 2. Revisão Sistemática da Literatura. 3. Aprendizado de Máquina. 4. Classificação de Texto. 5. Seleção Automática de Estudos. I. Kalinowski, Marcos. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

To my parents,
my sister, Patrícia,
my girlfriend, Marina,
and my friend, Joana Dias,
whose endless support made this journey possible.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

I would like to express my deepest gratitude to my adviser, Professor Marcos Kalinowski, and Dr. Bianca Napoleão for their guidance and partnership in carrying out this work.

I would also like to extend my appreciation to the Graduate Program in Computer Science from the Department of Informatics at PUC-Rio for providing the resources and environment necessary for the completion of this dissertation.

I am also grateful to my friends, Vinícius Segura and Clarissa Gandour, for being an inspiration and encouraging me to pursue my dreams and academic goals.

A special thanks to my friend and colleague, Guilherme Lacerda, for his invaluable support, flexibility, and understanding during crucial moments of this journey.

Finally, I would like to thank my family, my girlfriend, Marina, and my friend, Joana Dias, for their endless love and support. This journey would not have been possible without their encouragement.

Abstract

Cardoso, Marcelo Costalonga; Kalinowski, Marcos (Advisor). **Can Machine Learning Replace a Reviewer in the Selection of Studies for Systematic Literature Review Updates?**. Rio de Janeiro, 2024. 50p. Dissertação de Mestrado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

[Context] The importance of systematic literature reviews (SLRs) to find and synthesize new evidence for Software Engineering (SE) is well known, yet performing and keeping SLRs up-to-date is still a big challenge. One of the most exhaustive activities during an SLR is the study selection because of the large number of studies to be analyzed. Furthermore, to avoid bias, study selection should be conducted by more than one reviewer. [Objective] This dissertation aims to evaluate the use of machine learning (ML) text classification models to support the study selection in SLR updates and verify if such models can replace an additional reviewer. [Method] We reproduce the study selection of an SLR update performed by three experienced researchers, applying the ML models to the same dataset they used. We used two supervised ML algorithms with different configurations (Random Forest and Support Vector Machines) to train the models based on the original SLR. We calculated the study selection effectiveness of the ML models in terms of precision, recall, and f-measure. We also compared the level of similarity and agreement between the studies selected by the ML models and the original reviewers by performing a Kappa Analysis and Euclidean Distance Analysis. [Results] In our investigation, the ML models achieved an f-score of 0.33 for study selection, which is insufficient for conducting the task in an automated way. However, we found that such models could reduce the study selection effort by 33.9% without loss of evidence (keeping a 100% recall), discarding studies with a low probability of being included. In addition, the ML models achieved a moderate average kappa level of agreement of 0.42 with the reviewers. [Conclusion] The results indicate that ML is not ready to replace study selection by human reviewers and may also not be used to replace the need for an additional reviewer. However, there is potential for reducing the study selection effort of SLR updates.

Keywords

Systematic Literature Review; Machine Learning; Text Classification; Automatic Study Selection.

Resumo

Cardoso, Marcelo Costalonga; Kalinowski, Marcos. **Machine Learning pode substituir um revisor na seleção de estudos de atualizações de revisões sistemáticas da literatura?**. Rio de Janeiro, 2024. 50p. Dissertação de Mestrado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

[Contexto] A importância das revisões sistemáticas da literatura (RSLs) para encontrar e sintetizar novas evidências para Engenharia de Software (ES) é bem conhecida, mas realizar e manter as RSLs atualizadas ainda é um grande desafio. Uma das atividades mais exaustivas durante uma RSL é a seleção de estudos, devido ao grande número de estudos a serem analisados. Além disso, para evitar viés, a seleção de estudos deve ser conduzida por mais de um revisor. [Objetivo] Esta dissertação tem como objetivo avaliar o uso de modelos de classificação de texto de *machine learning* (ML) para apoiar a seleção de estudos em atualizações de RSL e verificar se tais modelos podem substituir um revisor adicional. [Método] Reproduzimos a seleção de estudos de uma atualização de RSL realizada por três pesquisadores experientes, aplicando os modelos de ML ao mesmo conjunto de dados que eles utilizaram. Utilizamos dois algoritmos de ML supervisionado com configurações diferentes (Random Forest e Support Vector Machines) para treinar os modelos com base na RSL original. Calculamos a eficácia da seleção de estudos dos modelos de ML em termos de precisão, recall e f-measure. Também comparamos o nível de semelhança e concordância entre os estudos selecionados pelos modelos de ML e os revisores originais, realizando uma análise de Kappa e da Distância Euclidiana. [Resultados] Em nossa investigação, os modelos de ML alcançaram um f-score de 0.33 para a seleção de estudos, o que é insuficiente para conduzir a tarefa de forma automatizada. No entanto, descobrimos que tais modelos poderiam reduzir o esforço de seleção de estudos em 33.9% sem perda de evidências (mantendo um recall de 100%), descartando estudos com baixa probabilidade de inclusão. Além disso, os modelos de ML alcançaram em média um nível de concordância moderado com os revisores, com um valor médio de 0.42 para o coeficiente de Kappa. [Conclusões] Os resultados indicam que o ML não está pronto para substituir a seleção de estudos por revisores humanos e também pode não ser usado para substituir a necessidade de um

revisor adicional. No entanto, há potencial para reduzir o esforço de seleção de estudos das atualizações de RSL.

Palavras-chave

Revisão Sistemática da Literatura; Aprendizado de Máquina; Classificação de Texto; Seleção Automática de Estudos.

Table of contents

List of figures

List of tables

List of Abreviations

AI – Artificial Intelligence

Anova-F – Analysis of Variance

Chi2 – Chi-squared

CSLR – Continuous Systematic Literature Review

DT – Decision Tree

EBSE – Evidence-Based Software Engineering

ED – Euclidean Distance

FN – False Negatives

FP – False Positives

FR – Final Results

FS – Feature Selection

KNN – K-Nearest Neighbor

LLM – Large Language Model

ML – Machine Learning

NB – Naive Bayes

NLP – Natural Language Processing

RF – Random Forest

RQ – Research Question

R1 – Reviewer 1

R2 – Reviewer 2

R3 – Reviewer 3

SE – Software Engineering

SLR – Systematic Literature Review

SVM – Support Vector Machine

TN – True Negatives

TP – True Positives

VTM – Visual Text Mining

Software gets slower faster than hardware gets faster.

Niklaus Wirth, .

1
Introduction

1.1
Motivation

In the context of Evidence-Based Software Engineering (EBSE), Systematic Literature Reviews (SLR) are the main instrument to identify, synthesize and summarize current evidence on a research topic or phenomenon of interest (KITCHENHAM; BUDGEN; BRERETON, 2015). Since the introduction of SLR in the Software Engineering (SE) field in 2004 (KITCHENHAM, 2004), especially over the last years, the number of SLR has been increased substantially (MENDES et al., 2020; aO et al., 2021).

As stated by Mendes *et al.* (MENDES et al., 2020), several SLRs in SE are potentially outdated. Only 20 SLRs were updated between 2006 and 2018 in a scenario of over 400 published SLRs in SE (MENDES et al., 2020). Outdated SLRs could lead researchers to make obsolete decisions or conclusions about a research topic (WATANABE et al., 2020). Despite the several initiatives in SE to keep SLRs updated (e.g. processes (DIESTE; LÓPEZ; RAMOS, 2008; MENDES et al., 2020); guidelines (WOHLIN et al., 2020; FELIZARDO et al., 2016); and experience reports (GARCéS et al., 2017; FELIZARDO et al., 2020)) there is a lack of investigation on automation tools to support the SLR update activities.

Performing an SLR update demands significant effort and time for reasons such as (i) the rapid increase in the number of available evidence (ZHANG et al., 2018; STOL; FITZGERALD, 2015), which hampers and slows down the identification of relevant evidence; and (ii) the lack of detailed protocol documentation and data availability (AMPATZOGLOU et al., 2019; ZHOU et al., 2015), which makes the SLR update process even more difficult because most of the tacit knowledge from the original SLR conduction is lost (FELIZARDO et al., 2020; FABBRI et al., 2013). These two reasons impact directly on the activity of selecting new studies during SLR updates, since they are crucial to determining if new evidence should or not be considered. On top of that, SLR updates require the expertise of multiple reviewers in the area of study. Ideally, these tasks should be performed by more than one reviewer (preferably an odd number) to ensure proper application of agreement criteria during the study selection process. However, due to the labor-intensive and time-consuming nature of this activity, there is often a shortage of available

experts. As a result, the correct application of agreement criteria can sometimes be overlooked.

1.2 Objective

Considering the potential benefits of exploring Machine Learning (ML) algorithms in the SLR and SLR updates context (NAPOLEÃO; PETRILLO; HALLÉ, 2021; WATANABE et al., 2020), our study **goal** is to present an empirical investigation on the adoption of ML Models to support the selection of studies for SLR updates in three perspectives: (i) effectiveness of the ML Models in selecting relevant evidence; (ii) effort reduction; and (iii) agreement level during the selection activity. To achieve our study goal, we investigate the potential of two supervised ML Models: Support Vector Machines (SVM) and Random Forest (RF) comparing their results with a rigorous manual selection process for an ongoing SLR update performed by three experienced researchers in SE.

1.3 Document Structure

The remainder of this document is structured as follows. Chapter 2 presents some background information about the process of SLRs and SLR Updates, as well as related work relevant to our problem. Chapter 3 presents the research method describing our research goals and questions, investigation strategy, case selection and the data acquisition. Chapter 4 presents the results. Chapter 5 presents discussions and threats to validity. Last, contributions and future work are addressed in Chapter 6.

2
Background and Related Work

2.1
Introduction

In this chapter, we provide a comprehensive overview of important concepts and contributions that set the groundwork of our study. Section 2.2 introduces the methodology of conducting SLRs, emphasizing their important role in synthesizing empirical evidence in SE. Section 2.3 addresses the challenges and importance of keeping SLRs up-to-date due to the rapid growth of new scientific evidence, as well as to perform the update. Section 2.4 reviews significant contributions related to the automation of the study selection process in SLRs, particularly through the application of Artificial Intelligence (AI), highlighting its potential alongside with its current challenges.

2.2
Systematic Literature Review

SLRs are a form of secondary study which provide a means of evaluating, interpreting and aggregating relevant evidence around one or more research questions, summarizing its benefits and limitations and also identifying current research areas in need of further investigation (KITCHENHAM, 2004; KITCHENHAM; CHARTERS, 2007).

There are three main phases during the process of an SLR (planning, conducting and documenting) and each phase consists of a different number of activities (BRERETON et al., 2007). During the first phase (planning), the reviewers specify the research questions and establish a protocol for the study, to minimize bias and define how the SLR should be conducted. During the second phase (conducting), the reviewers execute the search strategy and then analyze each result by applying inclusion and exclusion criteria to select primary studies. Subsequently, the reviewers assess study quality, extract the required data and synthesize it to answer the review questions. Once the SLR is completed, during the third phase (documenting), the reviewers write a report to document the decisions made throughout the process, and finally, they validate the report. The Figure 2.1 illustrates all tasks that are executed during the process of an SLR.

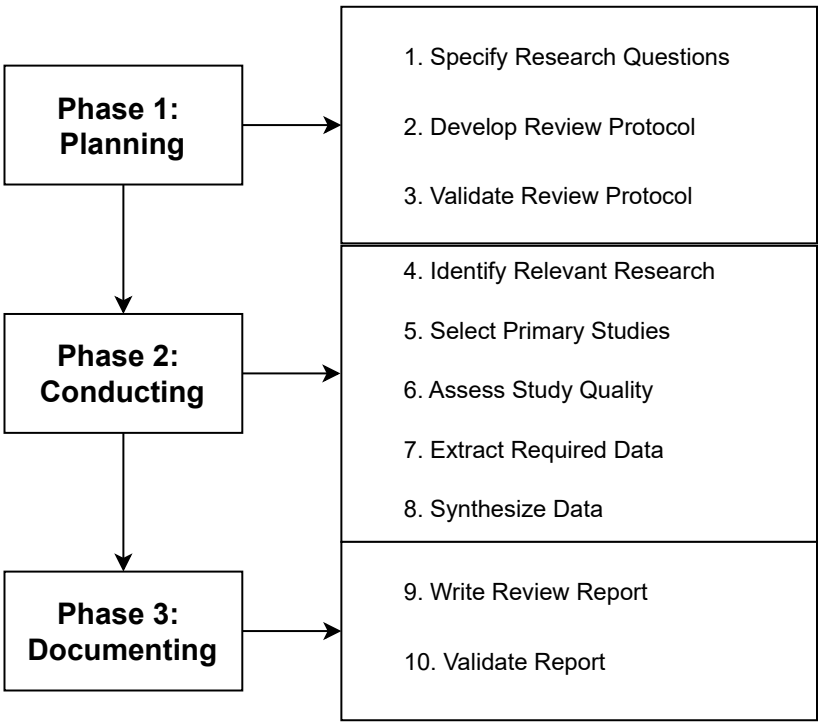

Figure 2.1: SLR process proposed by Brereton and Kitchenham *et al.* (BRERETON et al., 2007)

2.3 Systematic Literature Review Updates

Considering the fast speed new empirical evidences appear in SE, SLRs can become outdated in a short period of time. In order to assert all contributions provided by an SLR, researches need to perform an SLR update from time to time. The process to perform an SLR update is very similar to the process to perform an SLR, as described in the previously in Section 2.2, its structure consists of basically the same phases of an SLR, as illustrated in Figure 2.1.

Mendes *et al.* (MENDES et al., 2020) provide guidelines to determine if and when an SLR should be updated. Their work introduces a framework consisting of up to eight questions to assess whether an SLR is current or outdated, and whether updating it is worthwhile. The framework is structured in three steps: the first step is to assess the currency of the SLR, which involves evaluating if the SLR still addresses a relevant question, had good accessibility, and used valid methods; the second step is to identify new relevant information, focusing on finding new methods, studies, and other pertinent information; the third step is to assess the impact of updating the review, determining whether incorporating new information will change the conclusions or credibility of the review. Figure 2.2 illustrates this framework, including all the steps and

questions necessary to decide whether an SLR should be updated.

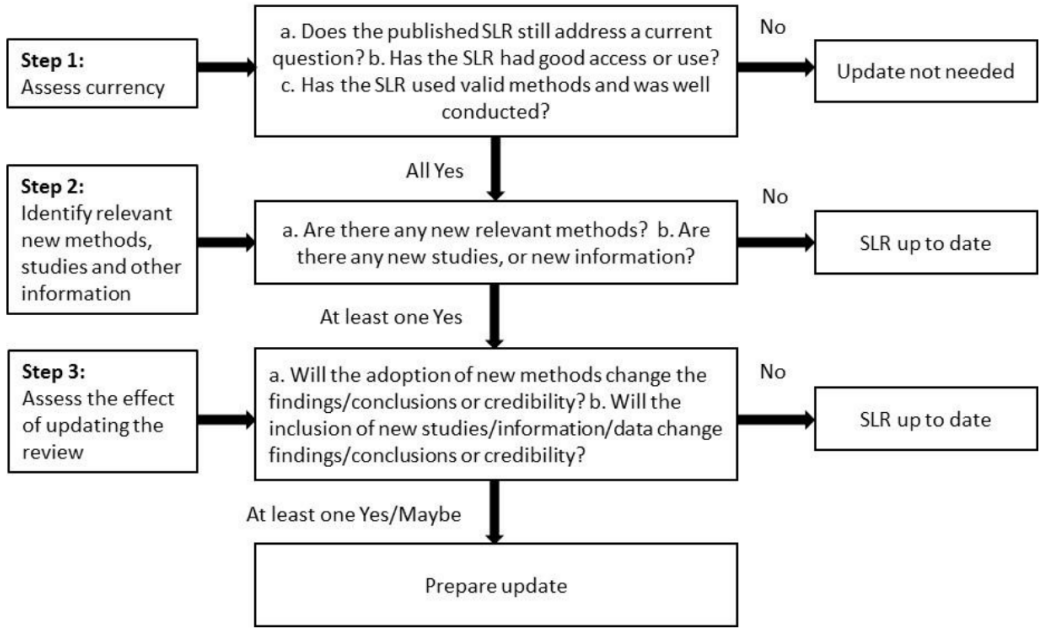

Figure 2.2: Decision framework to assess SLRs for updating proposed by Mendes *et al.*. (MENDES et al., 2020)

While the process to perform an SLR update can be very similar to the process that was performed during its original SLR, the two can differ in some aspects. For instance, to perform an SLR update, there's no need to execute the phase 1 (planning) again, since the research questions and review protocol were already defined in the original SLR. Another difference between the two is the search strategy for primary studies. Although using the same search strategy used to find the primary studies during the process of the original SLR can be a reliable approach, recent studies have presented alternative search strategies that have been proven to be more efficient when updating SLRs. (WOHLIN et al., 2022).

As stated by Felizardo *et al.* (FELIZARDO et al., 2016), an efficient strategy to find new evidence for an SLR update is to apply forward snowballing on the original SLR. This method, which utilizes the list of primary studies included in the original SLR as a seed set to find new citations, has proven to significantly reduce the workload by limiting the number of studies to be reviewed, although it can risk overlooking at some studies. They concluded that the forward snowballing technique has a precision high enough to be considered a valuable strategy for updating SLRs. Figure 2.3 illustrates the process of selecting new studies for an SLR update by using forward snowballing.

The work of Wohlin *et al.* (WOHLIN et al., 2020) proposes a series of guidelines designed to optimize the search strategies to find new evidences

when performing SLR updates, by investigating even further the benefits of using the forward snowballing technique. They performed a comparative analysis involving an original SLR and its subsequent updates in the context of SE, and only executed a single forward snowballing iteration using Google Scholar. They concluded that this approach is the most cost-effective strategy for maintaining current and accurate SLRs up-to-date. Their contributions also highlight the critical role of collaborative efforts in the activity of selecting studies, emphasizing how having multiple researchers working together on this activity can significantly diminish potential bias.

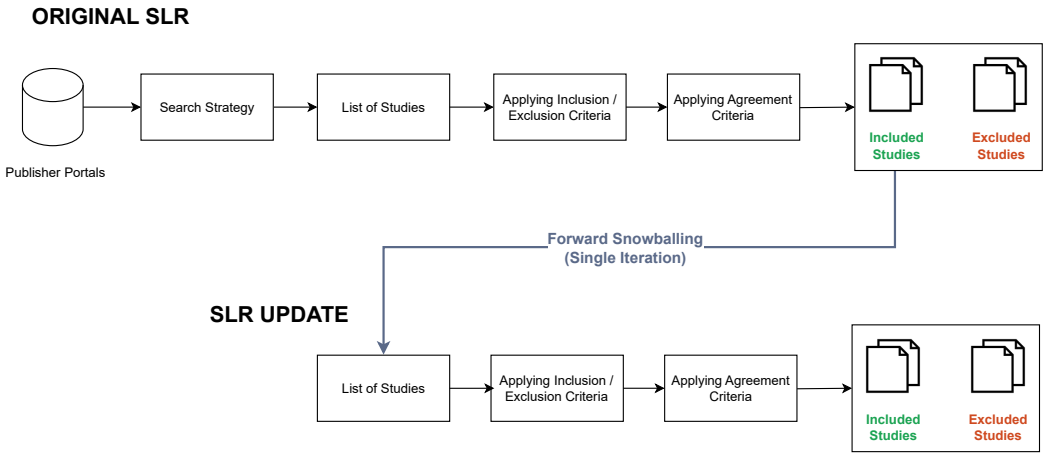

Figure 2.3: Study selection for an SLR update using a single iteration of forward snowballing.

2.4 Automatic Selection of Studies

Based on the contributions of the mapping study performed by Napoleão *et al.* (NAPOLEÃO; PETRILLO; HALLÉ, 2021) a variety of tools and text classification approaches have been evaluated to automate the activities of searching and selecting evidences for secondary studies. However, only two studies were applied in the context of SLRs updates in SE ((FELIZARDO et al., 2014); (WATANABE et al., 2020)).

Felizardo *et al.* (FELIZARDO et al., 2014) also propose an alternative to support the selection of studies for SLR updates, based on Visual Text Mining (VTM) techniques. The authors propose a tool called Revis which links new evidence with the original SLR's evidence using the K-Nearest Neighbor (KNN) Edges Connection technique. The tool output is presented in two distinct visualizations, a content map and an Edge Bundles diagram, to support knowledge discovery through visual processing and interactive data

exploration. The results showed an increase in the number of studies correctly included compared to the traditional manual approach.

The work of Watanabe *et al.* (WATANABE et al., 2020) also evaluated the use of text classification (text mining combined with ML Models) to support the study selection activity for SLR updates in SE. They performed an evaluation with 8 SLRs from different research domains in a cross-validation procedure using Decision Tree (DT) and SVM as ML classification algorithms. The results achieved on average an *F-score* of 0.92, *Recall* of 0.93 and *Precision* of 0.92. Unlike the approach proposed in (WATANABE et al., 2020), our study evaluates the ML Models SVM and RF using a detailed database of a solid ongoing SLR update conducted by renowned researchers in the field of EBSE. Furthermore, we perform a Euclidean Distance analysis and a *Kappa* analysis (COHEN et al., 2010; KITCHENHAM; BUDGEN; BRERETON, 2015) to evaluate the similarity and the agreement level between our ML Models with the expert reviewers.

Octaviano *et al.* (OCTAVIANO et al., 2022) propose a semi-automated strategy called SCAS-AI to support the initial selection task in SLRs in SE. SCAS-AI incorporates fuzzy logic and genetic algorithms to refine the selection process, aiming to reduce the human effort and potential bias when selecting studies. They conducted a quasi-experiment with eight SLRs demonstrated that SCAS-AI could reduce the effort required by 39.1% compared to the original strategy, while maintaining a low error rate with false negatives at 0.3% and false positives at 3.3%. These results underline the effectiveness of integrating AI techniques to support critical tasks in SLR processes, potentially revolutionizing how studies are selected and reviewed.

The work of Napoleão et al. (NAPOLEÃO et al., 2022) addresses the challenges of keeping SLRs up-to-date due to the fast-paced nature of research and limited update efforts. Their work propose a model of Continuous Systematic Literature Review (CSLR). Inspired by the Living Systematic Review (LSR) in medicine and the concept of Continuous Integration practices from DevOps (in the context of software development), CSLR integrates open science principles to ensure SLRs are continuously updated, facilitating access to the latest research findings. This model promises to mitigate the risks associated with outdated reviews and supports the ongoing advancement of Evidence-Based Software Engineering (EBSE).

Bolaños *et al.* (BOLANOS et al., 2024) performed a secondary study to evaluate the current use of AI techniques in the automation of SLRs, specifically during the screening and extraction activities of the second phase. They evaluated 21 tools that incorporate AI features, particularly examining

the impact of advanced Large Language Models (LLMs) during the research process. They concluded that AI shows great potential in enhancing the automation of SLRs during the studies selection and data extraction activities, specially considering the recent emerge of new tools based on LLMs. However, they also highlighted the current challenges of using LLM tools, such as ensuring the accuracy and ethical use of AI-generated.

2.5 Concluding Remarks

In this chapter, we have provided a comprehensive overview of the methodologies and technological advancements related to the process of conducting SLRs in SE. These advancements underscore a step towards more automated and efficient processes of performing SLR and SLR updates, highlighting the potential of reducing the human effort by adopting the use o AI-based solutions to support these activities. Chapter 3 will build upon this groundwork, detailing the specific research methods employed in this study and will describe how the research questions were formulated, how the investigation strategy was designed, and how the case selection process was executed, all aimed at achieving the goals of this study.

3
Research Method

3.1
Introduction

In this chapter, we present the key aspects of our research method. We begin by introducing our proposition and then describe the small-scale evaluation (WOHLIN; RAINER, 2022) conducted to assess it. Our research method is divided as follows. In Section 3.2 we present how our research questions were formulated according to GQM practices. In Section 3.3 we detail our proposed solution, a pipeline developed to train and configure the investigated ML models, providing details of its architecture. In Section 3.4 we describe our case selection and data acquisition process used in our small-scale evaluation to train and test our ML models, as well as the contributions made by the team assessment.

3.2
Research Goals and Questions

Our goal is to evaluate the adoption of ML models to support researchers during the activity of selecting studies for SLR updates. We translated our goal into three different research questions (RQs).

1. **RQ1:** *How effective are ML Models in selecting studies for SLR updates?*

   We represent the effectiveness of the ML models in supporting the selection of studies activity using metrics such as *Recall*, *Precision* and *F-measure* (NAPOLEÃO; PETRILLO; HALLÉ, 2021; WATANABE et al., 2020). Our ML automated analysis considers only title and abstract of the studies to make its predictions. The results made by our ML models are compared with the included studies selected manually for the SLR update under evaluation.

2. **RQ2:** *How much effort can ML Models reduce during the study selection activity for SLR updates?*

   We calculate the effort reduction by the relation of the number of studies that will need to have their title, abstract and keywords manually analyzed without the support of ML Models versus the number of studies to be analyzed after the use of the ML solution.

3. **RQ3:** *Can Machine Learning replace a reviewer in the selection of studies for SLRs?*

   We performed an agreement and similarity analysis using the ML Model with the highest *F-score*. Firstly, we used the Cohen's Kappa coefficient to measure the level of agreement between the ML Model and reviewers (COHEN et al., 2010; KITCHENHAM; BUDGEN; BRERETON, 2015). Then, we used the Euclidean Distance to measure the similarity between the ML Model, reviewers and final result (considering the studies that were actually included and excluded for the SLR update).

The Figure 3.1 illustrates our goal, research questions and metrics used to answer each question altogether.

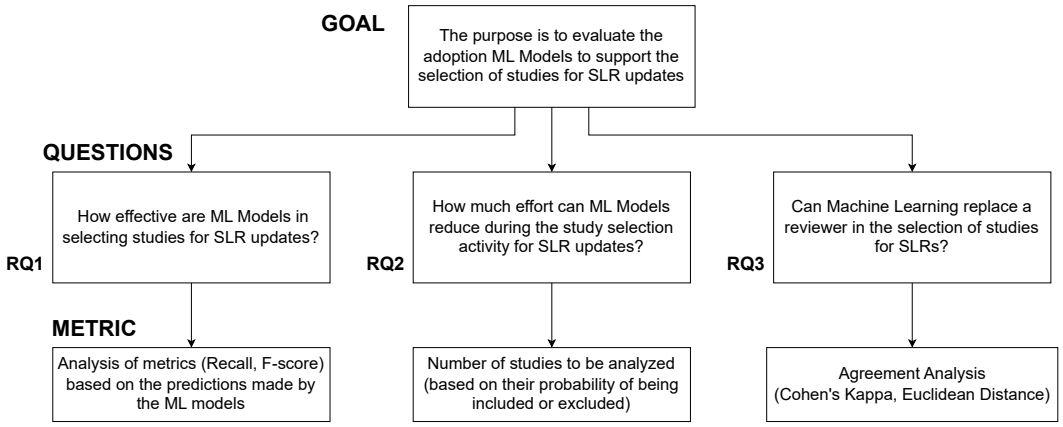

Figure 3.1: Research goal and questions diagram.

3.3 Investigation Strategy

In this study, we developed a pipeline[1] with the following steps to automate the study selection process of an SLR update by using ML and answer our research questions. Our pipeline is illustrated in Figure 3.2, describing its architecture and all the steps it performs since the moment it starts processing the list of studies until its output used to answer our research questions.

In summary, our pipeline process a set of .bib files containing the list of studies to train the ML models and the list of studies to be analyzed. After completing its execution, it returns a report file in .xlsx format informing which studies should be included and excluded, as well as metrics about the

[1]More details about our pipeline's implementation can be seen here: <https://github.com/bmnapoleao/SLR-Automated_selection_of_studies>

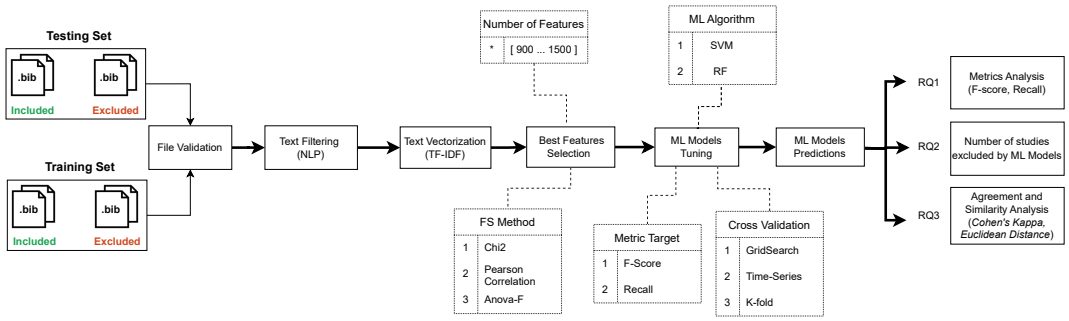

Figure 3.2: Investigation Strategy Pipeline

predictions made by the ML model and the configuration that was used to run its execution. It's important to mention that this process is not completely automated. Our pipeline expects to receive four different .bib files as input, each set (training and testing) should have one file containing the list of studies that should be excluded and one file containing the list of studies that should be included. In case there are any errors in the input files, the pipeline will stop its execution and will inform which entry was associated to each error as well as the type of error. Currently, our pipeline doesn't support automatic error resolution for invalid .bib files, they need to be manually fixed by the user.

As shown in Figure 3.2 we firstly validated the .bib files of our testing and training sets to ensure completeness of the set avoiding duplicated entries or missing keys. Each study entry must have a title, the year of publication, an abstract text and a list of authors. Secondly, we applied text filtering techniques with Natural Language Processing (NLP) (NLTK Team,), such as Lemmatization and Tokenization, to remove irrelevant characters from the texts. Thirdly, we applied Text Vectorization on the filtered texts using Term-frequency/Inverse-Document-Frequency (TF/IDF), a technique that transforms text data into a numerical matrix of features. Fourthly, we used statistical methods to compute and select the most relevant features. In the fifth step, we trained and tuned our ML Models using our training set. Finally, in the last step, we used our ML Models to predict which studies of our testing set should be included and excluded and compared the results of each one and the agreement level in comparison with the team assessment in order to answer our research questions that were previously described in Section 3.2.

Our pipeline also allows the user to pass an .env file containing the configuration to be used for each execution, where they can configure the Feature Selection (FS) method to compute the features and the number of best features that will be used in step four, as well as the ML algorithm, the metric target and the type of cross validation to be used with our ML models

during step five. If no file is passed, the pipeline will perform its execution with a default configuration. All of these parameters that can be configured are also illustrated in Figure 3.2.

Based on the promising results achieved using SVM in the work of Watanabe *et al.* (WATANABE et al., 2020) and the work of Napoleão *et al.* (NAPOLEÃO; PETRILLO; HALLÉ, 2021), which highlighted SVM as one of the most used ML classifiers for assisting the selection of studies in SLRs, we decided to evaluate SVM in our work. Additionally, considering the work of Pintas *et al.* (PINTAS; FERNANDES; GARCIA, 2021), which evaluated the most adopted ML classifiers and Feature Selection (FS) techniques for text classification and concluded that the five most used classifiers are SVM, NB, KNN, DT, and RF, we performed initial tests using these classifiers. Our first tests showed that SVM and RF were achieving better results than the others. Therefore, we decided to focus our evaluation on these two classifiers: Support Vector Machines (SVM) and Random Forest (RF).

We experimented multiple configurations of our pipeline and evaluated different configurations for FS and for training and tuning of our ML classifiers. During step four to compute the best features, we tested different statistical methods such as Chi-squared (Chi2) (Scikit-learn, b), Pearson Correlation (Scikit-learn, c) and Analysis of Variance (Anova-F) (Scikit-learn, a) as well as a different range of features. We also tested different techniques to tune our ML classifiers such as K-fold cross-validation, Time-Series cross-validations and hyperparameter tuning with GirdSearch (Scikit-learn, d).

For each evaluation, we executed the pipeline from start to finish in a clean environment using a unique configuration each time. To avoid introducing bias, the FS step was conducted solely based on the training set texts. Once the best features were identified in the training set, the same set of features was applied to the testing set to ensure consistency. To prevent overfitting, our machine learning classifiers were trained using a single type of cross-validation in each evaluation. Specifically, when utilizing GridSearch for parameter tuning, we didn't perform any other cross-validation technique, as GridSearch inherently includes cross-validation for measuring the most efficient parameter configuration. We chose this approach to maintain evaluation coherence and rigor.

3.4 Case Selection and Data Acquisition

We used as instrument of our small-scale evaluation an ongoing SLR update conducted by the same members of the team assessment. We chose

this ongoing SLR update since the inclusion and exclusion of new studies were conducted based on individual assessments and the consensus of three experienced SLR researchers by analysing title, abstract, keywords and then full-text of the studies manually, allowing us to have confidence in this data for building reliable training and testing sets. Figure 3.3 summarizes this entire process.

The team assessment performed the forward snowballing search strategy to find new evidences for the SLR update. They used the 45 studies that were included in the original SLR as their seed set of studies. After performing the search strategy, the team assessment found a total of 591 new references, of which 39 were included and 552 were excluded for the update.

The team assessment provided us all the 591 studies they analyzed during the SLR update (.bib files). We used these studies to form our testing set for our ML models, we manually filtered the studies to consider only first studies in English with a valid abstract field. At the end, we used 551 studies in our testing set, of which 38 were included by the team assessment and 513 were excluded for the update.

To train our ML models, we used a training set with 128 studies, of which 45 studies were included and 83 were excluded. The 45 studies used to train our models with what should be included were the same studies included in the original SLR. Since the team assessment did not list the studies that were excluded during the study selection phase of the original SLR, we performed a backward snowballing on the original references to obtain the 83 studies used to train our models with what should be excluded.

The forward snowballing looked for studies that were published after the original SLR, since this technique looks for the studies that cited the study being examined and use them as the seed studies for the first iteration. While the backward snowballing looked for studies that were published before the original SLR, since this technique looks for the studies cited by the study being examined and use them as the seed studies for the first iteration. This allowed us to simulate a more realistic set of excluded studies for our training set when the original set of excluded studies was not available.

It is important to point out that while our training and testing sets reflected realistic references used during the SLR update process and the original SLR process, our training set and testing set had a large difference in their sizes. Although the two sets had a similar number of included studies, the number of excluded studies in the testing set was considerably greater than in our training set. Figure 3.4 illustrates their differences, indicating the number of studies in each set as well as the percentage of studies that were

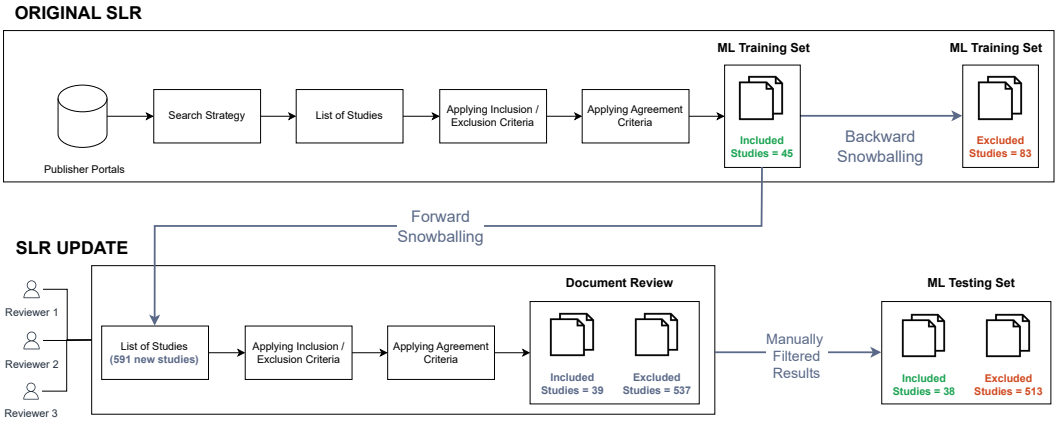

Figure 3.3: Data Acquisition Process

included and excluded in each set.

In the original table provided by the team assessment, besides what studies were included and excluded for the update, it also had the opinion of each reviewer about each study during the study selection process they performed. The reviewers could express their opinions about each study in three different level of certainty: 0 – certain that the study should be excluded, 1 – uncertain if the study should be excluded or included and 2 – certain that the study should be included. We used this information to answer RQ3 and perform the agreement and similarity analysis. Table 3.1 illustrates the format of the document provided by the assessment team, containing each reviewer's opinion about each study (decided right after the inclusion/exclusion criteria was applied) as well as the final result of each study (decided after applying the agreement criteria).

Table 3.1: Example of the document format provided by the assessment team.

Study	Final Result	R1	R2	R3
First Study	1	2	1	2
Second Study	0	2	0	0
Third Study	1	2	1	0
Fourth Study	0	0	2	1
Fifth Study	0	0	0	0

3.5 Concluding Remarks

This chapter has described the research method used to conduct our experiment. Our investigation strategy introduced a comprehensive pipeline leveraging text processing, NLP, feature selection and ML techniques to evaluate the use of ML models in each one of our research questions. The

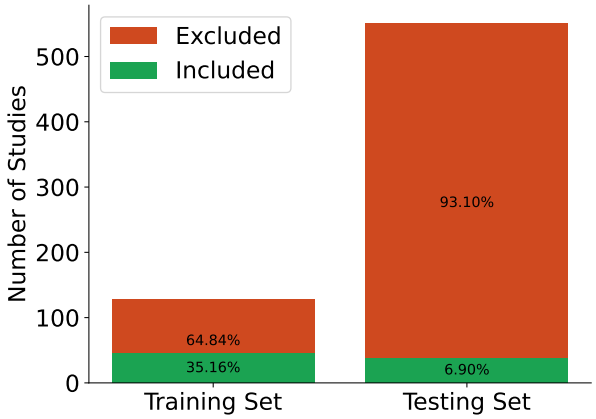

Figure 3.4: Datasets Distribution Comparison

results for each one of the three research questions is presented in the next chapter.

4
Results

4.1
Introduction

In this chapter, we present the results of our study, following the research questions presented earlier in Section 3.2. To answer each question, we analyzed the results of the ML Models and then compared them with the results obtained manually by the SLR update authors under evaluation. This chapter is divided as follows. In Section 4.2 we provide general details of how we executed and configured our pipeline in order to answer each RQ. In each of the subsequently sections, Section 4.3, Section 4.4 and Section 4.5 we describe the best configuration found for our ML models to answer each question.

4.2
Execution

To address questions RQ1 and RQ2, we replicated the same evaluation performed by SLR update authors during the selection of studies for the SLR update by using our classifiers to predict which studies should be included and excluded. For RQ1 we configured our model maximizing F-score and for RQ2 we configured our model maximizing Recall. Then we performed the agreement and similarity analysis using the predictions made by our model used in RQ1.

Precision indicates how accurate the positive predictions made by the model are, while Recall indicates how well the model captures all the actual positive instances. For Precision, values closer to 1 indicate a lower number of False Positives (FP) results and values closer to 0 indicate a higher number of FP. And for Recall, values closer to 1 indicate a lower number of False Negatives (FN) and values closer to 0 indicate a higher number of FN.

We conducted our evaluation by varying the number of best features to be considered in each execution. After applying Text Filtering and Text Vectorization techniques, presented in steps three and four of our pipeline, our training set comprised a total of 23630 features, in contrast to our testing set, which comprised 119560 features. Given that the number of features in our testing set was more than five times greater than our training set, maximizing the number of best features in our training set was crucial to the performance of our ML models.

We identified the range with the most relevant features in our training set as 900 to 1500 features, which was the range used in most of our evaluations. Notably, the best results, both in terms of F-score (RQ1) and Recall (RQ2), were consistently achieved with experiments that selected the 1200 best features.

The document provided by the team assessment contained the final result of each study (if they were included or excluded), and also had the opinion of each reviewer about each study during the study selection process they performed. The reviewers could express their opinions in three different levels of certainty: 0 – certain that the study should be excluded, 1 – uncertain if the study should be excluded or included and 2 – certain that the study should be included. We used this information to answer RQ3 and perform the agreement and similarity analysis. Table 3.1 illustrates the format of this document.

We displayed the complete information of our results for the best configurations we found for RQ1 and RQ2, as well as all the other tests executions in an *Appendix document*[1] available online.

4.3 RQ1: *How effective are ML models in selecting studies for SLR updates?*

To answer this question, during the ML models tuning step, we trained our classifiers with GridSearch focusing on maximizing the F-score. Our best result was obtained by RF with a Precision of 0.22, Recall of 0.63 and F-score of 0.33 using the Anova-F statistical method, with 1200 features. We used a default threshold of 0.5 to consider which studies should be included and excluded by our ML models. Table 7.1 shows the parameters tested and selected by GridSearch for this configuration. And the table containing all the predictions made by our ML model for each study for this question can be seen in this document[2].

Figure 4.3 illustrates the distribution of the predictions' scores made by RF with this configuration. In order to calculate each of these metrics, we compared our ML models' predictions with the final results only, obtained after the agreement criteria was applied by the team assessment, which is illustrated by the first column of Table 3.1.

[1]https://zenodo.org/records/11021614
[2]https://zenodo.org/api/records/11019279/draft/files/RQ1-RF-predictions.csv/content

4.4 RQ2: *How much effort can ML models during the study selection activity for SLR updates?*

To answer this question, we tuned the ML models with the intention of maximizing the Recall. Since the purpose of this question was to evaluate how much human effort could be reduced by the use of ML Models during the selection of studies, we wanted to mitigate the chances of a false negative (FN) result, so the reviewers could simply ignore the studies excluded by the ML model without worrying about losing a relevant study.

Our best result was obtained by using the SVM algorithm with a Precision of 0.10, Recall of 1.0 and F-score of 0.19 using the Pearson Correlation statistical method, with 1200 features. We used a default threshold of 0.5 to consider which studies should be included and excluded by our ML models. Table 7.2 shows the parameters tested and selected by GridSearch for this configuration. And the table containing all the predictions made by our ML model for each study for this question can be seen in this document[3].

According to Table 4.1, by maximizing the Recall, SVM was able to exclude a total of 187 studies, which represents 33.9% of the total amount of studies in our testing set. By increasing the threshold, we are able to see that we can reduce the human effort even more at the risk of having more FN results. For a threshold range greater than 0.5 until 0.75, only one false negative was found, while the number of TN increased by 87. Notably, this FN result was one of the few cases where the team assessment had a lot of disparity. Considering only the initial analysis of the reviewers individually (before they discussed it with each other), the reviewer R1 voted 2, R2 voted 1, and R3 voted 0. For a threshold range greater than 0.75 until 0.80, when compared to the previous threshold range, the number of FN results increased by 1, while the number of TN increased by 18. Finally, for a threshold range greater than 0.80 until 0.85, when compared to the previous threshold range, the number of FN results increased by 2, while the number of TN increased by 24.

As well as RQ1, to answer this question, we compared our ML models' predictions with the final results only, obtained after the agreement criteria was applied by the team assessment, which is illustrated by the first column of Table 3.1.

[3]https://zenodo.org/api/records/11019279/draft/files/RQ2-SVM-predictions.csv/content

Threshold(%)	RECALL (%)	TN	TP	FN	FP	Reduced (%)
0.50%	100.00%	187	38	0	326	33.9%
0.75%	97.37%	265	37	1	248	48.3%
0.80%	94.74%	283	36	2	267	51.7%
0.85%	89.49%	307	34	4	206	56.4%

Table 4.1: Tradeoff between effort reduction and number of FN.

4.5 RQ3: *Can Machine Learning Replace a Reviewer in the Selection of Studies for Systematic Literature Reviews?*

To truthfully answer this question, we conducted a two-fold analysis to evaluate both aspects of agreement and similarity, considering not only the comparison between our ML model and single reviewer results but also the final results and the average answer between multiple reviewers.

The agreement analysis, which measures the concordance of results, indicates how two or more raters make the same classifications. For this, we used the Cohen's Kappa Coefficient. The equation to calculate Kappa is:

$$Kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

where $Pr(a)$ is the relative observed agreement among raters, and $Pr(e)$ is the theoretical probability of chance agreement (CARLETTA, 1996).

The similarity analysis, which measures the resemblance between the classifications of two or more raters, indicates how close the results are, even if they are not exactly the same. For this, we used the Euclidean Distance. The equation to calculate the Euclidean Distance between two sets of data i and j is:

$$\text{EuclideanDistance}(i, j) = \sqrt{\sum_{k=1}^{n} (i_k - j_k)^2}$$

where i_k and j_k are the k-th elements of the sets i and j, respectively, and n is the number of elements in each set (SERRà; ARCOS, 2014). In this context, the elements i_k and j_k represent the scores of each study given by one of the reviewers, or the final result or our ML model, as illustrated in Table 3.1.

Firstly, we looked at the agreement and similarity levels between the reviewers among the team assessment, and analyzed the information provided by the team assessment regarding each reviewer's vote before applying the agreement criteria, which is illustrated by the last three columns of Table 3.1.

Figure 4.1 shows the distribution of votes given by each reviewer of the assessment team during the SLR update, considering only the 551 studies used in our testing set. As expected, during the study selection phase, the number of

studies selected to be excluded, after applying the inclusion/exclusion criteria, was by far the greatest, which on average the team assessment voted to exclude 87.87% of all studies. Subsequently, the second most common vote mark given by each reviewer was 2, indicating that on average the team assessment voted to include 8.3% of all studies. Last, the least common vote mark each reviewer expressed was about "uncertain", which on average the team assessment wasn't sure whether 3.87% of all studies should be excluded or included.

It's important to note that although all three reviewers had a similar vote distribution, meaning that the population standard deviation between the reviewers by each vote category was significantly low (the highest deviation being equal to 1.56% for vote 2 - "should be included"), does not directly imply the similarity neither the agreement level between the reviewers will be strong, since each reviewer can have voted for completely different studies in each category.

Considering the following rage of Cohen's Kappa coefficient values, results can be interpreted as the following agreement levels (CARLETTA, 1996):

- from 0.00 to 0.20: Poor Agreement
- from 0.21 to 0.40: Fair Agreement
- from 0.41 to 0.60: Moderate Agreement
- from 0.61 to 0.80: Substantial Agreement
- from 0.81 to 1.00: Almost Perfect Agreement

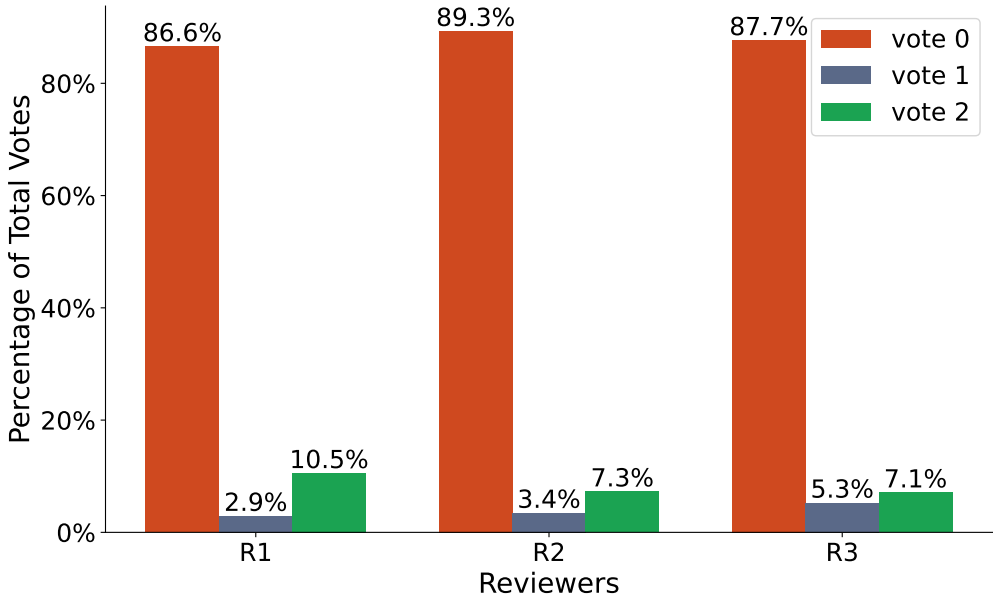

Figure 4.1: Assessment Team Votes Distribution.

	R1	**R2**	**R3**
R1	1	0.47	0.35
R2	0.47	1	0.43
R3	0.35	0.43	1

Table 4.2: Cohen's Kappa Coefficient Comparison Among the Assessment Team.

	R1	**R2**	**R3**
R1	0	13.0	14.04
R2	13.0	0	11.22
R3	14.04	11.22	0

Table 4.3: Euclidean Distance Comparison Among the Assessment Team.

Table 4.2 shows the agreement level and Table 4.3 shows the similarity level between each reviewer among the team assessment. We can see that the highest agreement was between reviewer 1 (R1) and reviewer 2 (R2) with 0.47 of agreement, which can be considered a Moderate Agreement. Whilst, the lowest agreement was between the reviewer 3 (R3) and R1 with 0.35 of agreement, which can be considered a Fair Agreement. Since the similarity is inversely proportional to the Euclidean Distance (ED), we can notice that R2 and R3 had the most similar opinions with the smallest distance of 11.22 in comparison to R1 and R3 with the highest distance of 14.04.

After that, we compared the best configuration found for RQ1 and RQ2. By looking at the distribution of RF in Fig 4.3 and SVM in Fig 4.2, we can see that the one used to answer RQ1 is closer to the behavior of the assessment team, as expected, since we focused on maximizing its F-score. In comparison, our the model used in RQ2, which was focused on maximizing the Recall,

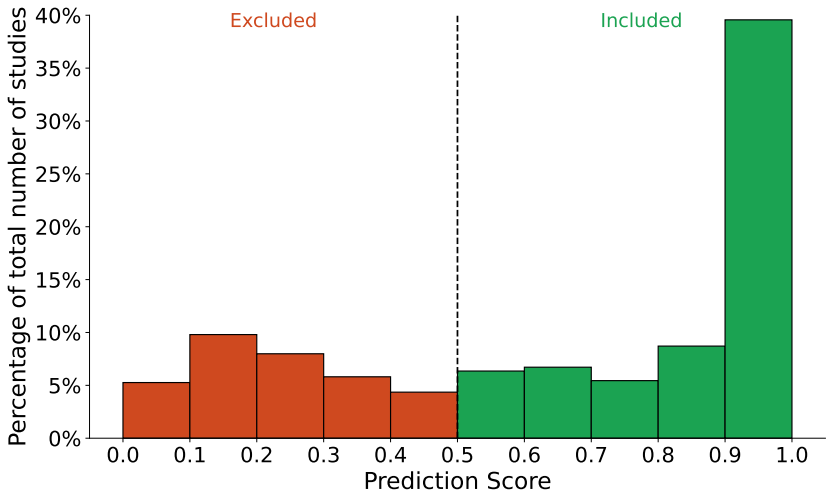

Figure 4.2: RQ2: SVM Predictions Distribution.

included most of the studies instead, which is uncommon in most cases during the process of selecting studies for SLR and SLR Updates performed by experts in the field.

In order to compare our ML models' results with the assessment team answers, we normalized our ML results into three ranges to also represent three categories. As mentioned before, by looking at 4.1 we can see that the frequency of occurrence of a 0 vote is way greater than the others, because of this we decided it would be more fair if the threshold for the exclusion of studies would be greater than the rest, as well as having the smallest range for the votes of type 1. We decided to use the following predictions' score range for each category.

- from 0.00 to 0.50: should be excluded
- from 0.51 to 0.60: uncertain
- from 0.61 to 1.00: should be included

Figure 4.4 illustrates the normalized distribution of our ML model used in RQ1 according to the range previously mentioned.

After we converted the probabilities given by our RF model used in RQ1 into the three categories, we compared the normalized results of the RF model with the results of each reviewer and calculated the *Cohen's Kappa Coefficient* and the Euclidean Distance between them. The agreement level between the RF model and each reviewer was: RF | R1 = 0.27, RF | R2 = 0.37, RF | R3 = 0.30. The Euclidean Distance between the RF model and each reviewer was: RF | R1 = 17.89, RF | R2 = 16.76, RF | R3 = 17.52. Table 4.4 illustrates both results.

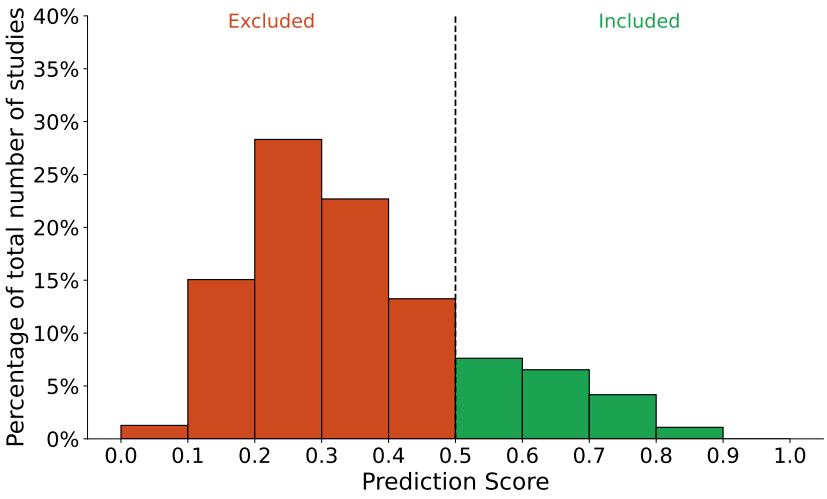

Figure 4.3: RQ1: RF Predictions Distribution.

Comparison	Cohen's Kappa Coefficient	Euclidean Distance
R1 vs RF	0.27	17.89
R2 vs RF	0.37	16.76
R3 vs RF	0.30	17.52

Table 4.4: Agreement and Similarity between ML and Reviewers

We used the Euclidean Distance to evaluate the similarity between the answers of our RF model and reviewers considering the Final Results (FR), when compared individually and collectively. In order to evaluate the Euclidean Distance between our model and reviewers against the FR accurately, we multiplied the FR answers by two, to normalize the binary answers from FR with the three-category answers from reviewers and ML, so if a reviewer voted to include a study with certainty and the study was indeed included, the distance between these two points is zero.

Finally, we evaluated the Euclidean Distance in three different ways, as follows:

– Similarity between single answers and FR:

$$\textit{EuclideanDistance}(i, FR) \text{ where } i \in \{\text{R1}, \text{R2}, \text{R3}, \text{RF}\}$$

– Similarity between pairs and FR:

$$\textit{EuclideanDistance}\left(\frac{i+j}{2}, \text{FR}\right) \text{ where } i \neq j \text{ and } i, j \in \{\text{R1}, \text{R2}, \text{R3}, \text{RF}\}$$

– Similarity between groups and FR:

$$\textit{EuclideanDistance}\left(\frac{i+j+k}{3}, \text{FR}\right) \text{ where } i \neq j \neq k \text{ and } i, j, k \in \{\text{R1}, \text{R2}, \text{R3}, \text{RF}\}$$

Table 4.5 shows the Euclidean Distance (ED) measured in each case. As we can see, the smallest distance in comparison to the FR in each case was given by: R2 with ED = 9.95, pair(R2,R3) with ED = 8.86 and team assessment with ED = 8.17.

4.6 Concluding Remarks

In this chapter, we presented the results of our experiments using our ML models to replicate the study selection performed by the team assessment. The results highlighted how specific configurations significantly influence model

Comparison	**Euclidean Distance**
R1 vs FR	12.00
R2 vs FR	**9.95**
R3 vs FR	11.00
RF vs FR	17.94
avg(R1,R2) vs FR	8.90
avg(R1,R3) vs FR	9.12
avg(R2,R3) vs FR	**8.86**
avg(R1,RF) vs FR	12.37
avg(R2,RF) vs FR	11.84
avg(R3,RF) vs FR	12.03
avg(R1,R2,R3) vs FR	**8.17**
avg(RF,R2,R3) vs FR	10.06
avg(R1,RF,R3) vs FR	10.20
avg(R1,R2,RF) vs FR	10.15

Table 4.5: Euclidean Distance Analysis

performance, particularly in terms of precision and recall. Our analyses provide a foundation for deeper discussions that will be presented in the next chapter.

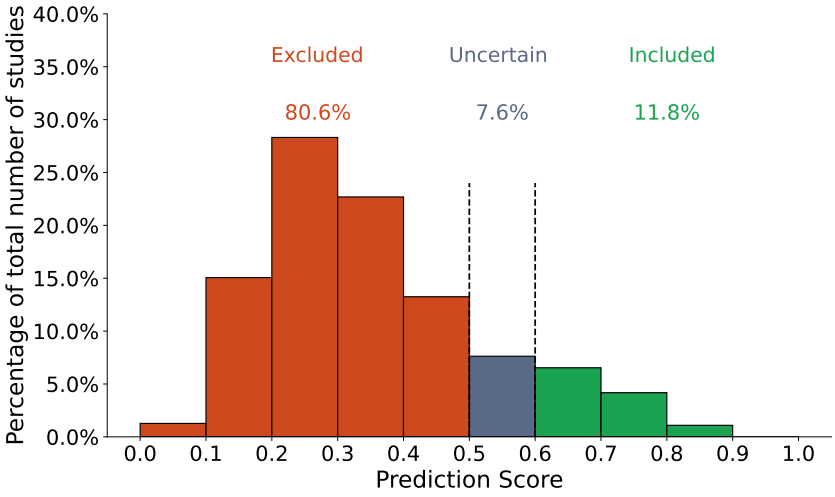

Figure 4.4: RQ3: RF Predictions Distribution considering "uncertain" range.

5
Discussion and Threats to Validity

5.1
Introduction

In this chapter, we present discussions about the outcomes and the inherent challenges associated with validating the findings from our investigation into the application of our ML models for SLR updates in SE. By analyzing the effectiveness and reliability of our ML models as outlined in the previous Chapter 4, in the first three sections, Section 5.2, Section 5.3 and Section 5.4 we discuss the meaning of the results found for each RQ and their implications. Then, in Section 5.5 we address some validity threats that might influence the interpretation and generalizability of our results.

5.2
Research Question 1

Regarding RQ1 - "How effective are ML models in selecting studies for SLR updates?", based on our results, we concluded that the best configuration to maximize the F-score of our ML models was using RF with Anova-F. Not only that, but in almost all of our tests configurations, RF outperformed SVM in terms of F-score with exception for one test, where we selected only 900 best features and used the Pearson Correlation method to compute the best features. We also noticed that for most cases, the Anova-F improved the F-score rating at the cost of lowering its Recall. However, even noticing that the RF was more successful in this task than SVM and considering its best result, its F-score was still not good enough to be considered to automate the process of selecting studies for the SLR update.

5.3
Research Question 2

Regarding RQ2 - "How much effort can ML models reduce during the study selection activity for SLR updates?", our results showed that the best configuration to maximize the Recall of our ML models was using SVM with Pearson Correlation. On one hand, it's possible to see that in all of our tests, SVM had higher Recall marks than RF. Particularly, when used with Pearson Correlation to select the best features, it had the highest marks for Recall in most cases (although even in executions using Anova-F it still reached a high

Recall of 0.97 for some tests). On the other hand, its F-score and Precision marks were very low. However, even if a big number of FP studies remained after that, we believe that our SVM model showed potential for reducing human effort during the study selection task of SLR updates, by automatically excluding part of the irrelevant studies.

5.4 *Research Question 3*

Regarding RQ3 - "Can Machine Learning Replace a Reviewer in the Selection of Studies for Systematic Literature Reviews?" our evaluation showed that the best result was achieved by the RF classifier with the same configuration used for RQ1. As we can see in Table 4.2, the strongest level of agreement was between R1 and R2 with a score of 0.47 for Cohen's Kappa Coefficient, followed by an agreement of 0.43 between R2 and R3, and followed by the weakest agreement between R1 and R3 with a score of 0.35. In this case, two pairs of reviewers had a moderate level of agreement and one pair had only a fair level of agreement.

On the other hand, when compared with our RF classifier, it had its strongest agreement of 0.37 with R2, followed by an agreement of 0.30 with R3 and an agreement of 0.27 with R1, all considered as fair level of agreement. Even though the pair RF classifier and R2 had a stronger agreement than the pair R1 and R3, both still had the same agreement level (fair agreement), also the R2 had a Moderate Agreement with both reviewers, while it had only a fair agreement with the RF classifier. Also, when comparing the agreement of RF with R3 and its respective pairs (R3|R1 and R3|R2), it had a weaker agreement than both, the same was true when comparing RF with R2 and its respective pairs.

Looking at Table 4.3, we can see that R2 and R3 had the most similar answers, with an Euclidean Distance (ED) of 11.22 between them. As showed in Table 4.5 R2 also had the smallest distance from FR when compared individually (with an ED of 9.95), as expected, the pair R2 and R3 had the strongest similarity with FR (with an ED of 8.86) in contrast to the other pairs. Finally, the closest distance was given by the team assessment with an ED of 8.17, which is expected since the FR was generated from their answers. It's possible to see that our RF model had the highest distances in all comparisons and even though looking at the distance to FR obtained when working together with R1 didn't cause much negative impact (ED(R1,RF) = 12.37 vs ED(R1) = 12.00) as the rest, it shows that our ML model didn't help any of the reviewers to get closer to the FR.

Therefore, we concluded that our supervised ML Models are not ready to replace a reviewer during the selection of studies for SLR updates.

It is worth mentioning that the similarity and agreement levels between our classifier and the team assessment could increase or decrease depending on how we configure the thresholds to normalize the probabilities given by our ML classifier into the three categories used by the team assessment. We noticed that reducing the range to consider a vote as uncertain would increase the level of agreement with all members of the assessment team. But in most cases, our classifier still had only a fair level of agreement, so our conclusion was the same.

5.5 Threats to Validity

In this section, we discuss the main threats to our experiment based on the categorization presented in Wohlin *et al.* (WOHLIN et al., 2012) and the adopted mitigation strategies.

Construct Validity. Our evaluation results might have been affected by the choice of ML algorithms. Other algorithms could have been explored in our study and can be considered as part of future work.

Internal Validity. Our training dataset comprised only studies including in the SLR replication (WOHLIN et al., 2022) (training included) and those obtained through backward snowballing (training excluded). We deliberately excluded studies not in English or those categorized as Ph.D. dissertations or book chapters from our testing set, the same approach adopted by the SLR replication. This focused approach aimed to provide our models with only relevant and essential data. Another potential threat is that during the manual process performed by the team assessment, the authors could end up reading other sections of the studies besides the title and abstracts when they are not completely sure if the study should be included or excluded just by reading its abstract. This is a possible advantage manual process could have over our models that consider only content from title and abstract.

External Validity. The dataset used in our analysis might not represent the diversity of SLR Updates in SE. Similar analyses could have been conducted based on other SLRs to improve the generalizability of our results. However, replicating our results on other SLRs to strengthen external validity would require significant effort. Moreover, it is challenging to acquire a reliable and detailed SLR dataset for SLRs updates that could be considered in our evaluation.

Reliability. One limitation of our study is associated with the dataset

used in our evaluation and the possibility of sample bias. The data used in our evaluation was acquired from the same authors who performed the SLR replication, also through a rigorous analysis process. In addition, to improve the reliability of our results, our ML models and the small-scale evaluation dataset are openly available.

5.6 Concluding Remarks

In this chapter, we underscored the potential and limitations of using our ML models to support the process of performing SLRs. While the models demonstrate a promising reduction in manual effort required for study selection, their capacity to replace a reviewer remains below expectations. The next chapter will build upon these findings, exploring potential enhancements and proposing alternative strategies for future work, aiming to highlight opportunities to explore even further the topic of supervised ML models within the context of SLR updates in SE.

6
Contributions and Future Work

6.1
Contributions

This study advances the application of supervised ML models as a supporting tool for researchers during SLRs updates, by developing and testing a comprehensive supervised ML-based pipeline to automate the study selection process. We have demonstrated through our tests, using realistic data to build our datasets, that while our ML models, have shown promise in reducing human effort to perform the study selection activity by pre-filtering irrelevant studies, they currently lack the precision required to completely automate the selection of studies for SLR updates. We also concluded that they did not achieve a level of similarity and agreement strong enough to be considered sufficient to replace a human reviewer during an SLR update in the study selection phase. Our work also highlights different configurations used for our ML models that correlates to their recall and F-score, providing results that can be useful for further exploration in this area.

6.2
Limitations

The study's scope was limited by several factors that could be addressed in future research. Primarily, we only tested a few ML algorithms in our experiments, as well as the number of strategies for NLP and FS tested. Additionally, we did not explore the potential of Large Language Models (LLMs) of automating the selection of studies activity in comparison to our ML models, which showed a lot of potential in our research topic (BOLANOS et al., 2024). Also, we did not leverage from the work of Pintas *et al.* (PINTAS; FERNANDES; GARCIA, 2021) to perform a deeper analysis of our dataset structure and decide a specific FS strategy related to it. Last, we did not find more primary studies to increase the number of excluded studies in our training set and decrease the contrast in the size of our training and testing sets, which could have improved our results.

6.3 Future Work

In summary, this dissertation lays the groundwork for significant advancements in the use of supervised ML to automate and enhance the efficiency of SLR updates in SE. While our ML models currently supplement rather than replace human expertise, they mark a substantial step forward in the integration of supervised ML models into the process of performing SLR updates. The exploration of more sophisticated models and methodologies, as outlined in our future work, shows potential for further investigation.

7
Appendix

This chapter contains tables with the configuration used to tune our ML models for RQ1 and RQ2, describing the parameters that were tested, and the best configuration selected after performing GridSearch.

7.1
GridSearch Configurations

Grid Search Parameters Configuration

Table 7.1: RQ1: Grid Search - Best configuration for RF

RF/GridSearch	**class_weight**	**criterion**	**max_depth**	**min_samples_split**	**n_estimators**
Best Params	balanced	gini	10	10	100
Tested Params	[None, 'balanced']	['gini', 'entropy']	[10, 20, 30, None]	[2, 10, 25, 50, 75, 100]	[2, 5, 100]

Table 7.2: RQ2: Grid Search - Best configuration for SVM

SVM/GridSearch	**class_weight**	**cv**	**scoring**	**C**	**gamma**
Best Params	balanced			0.65	1
Tested Params	['balanced', None]	5	recall_macro	[0.5, 0.65, ..., 3.35]	[1, 0.1, ..., 0.0001]

8
Bibliography

AMPATZOGLOU, A. et al. Identifying, categorizing and mitigating threats to validity in software engineering secondary studies. **Information and Software Technology**, v. 106, p. 201 – 230, 2019. ISSN 0950-5849.

aO, B. M. N. et al. Establishing a search string to detect secondary studies in software engineering. In: **2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)**. [S.l.: s.n.], 2021. p. 9–16.

BOLANOS, F. et al. Artificial intelligence for literature reviews: Opportunities and challenges. **arXiv preprint arXiv:2402.08565**, 2024.

BRERETON, P. et al. Lessons from applying the systematic literature review process within the software engineering domain. **Journal of Systems and Software**, Elsevier Science, New York, NY, USA, v. 80, n. 4, p. 571–583, 2007.

CARLETTA, J. Assessing agreement on classification tasks: the kappa statistic. **arXiv preprint cmp-lg/9602004**, 1996.

COHEN, K. et al. The structural and content aspects of abstracts versus bodies of full text journal articles are different. **BMC Bioinformatics**, v. 11, p. 492, 2010.

DIESTE, O.; LÓPEZ, M.; RAMOS, F. Formalizing a systematic review updating process. In: **6^{th} Int. Conference on Software Engineering Research, Management and Applications (SERA'08)**. [S.l.: s.n.], 2008. p. 143–150.

FABBRI, S. et al. Externalising tacit knowledge of the systematic review process. **IET Software**, v. 7, n. 6, p. 298–307, 2013.

FELIZARDO, K. et al. Using forward snowballing to update systematic reviews in software engineering. In: **International Symposium on Empirical Software Engineering and Measurement (ESEM)**. [S.l.: s.n.], 2016.

FELIZARDO, K. et al. A visual analysis approach to update systematic reviews. In: **International Conference on Evaluation and Assessment in Software Engineering (EASE)**. [S.l.]: ACM, 2014. p. 1–10.

FELIZARDO, K. R. et al. Knowledge management for promoting update of systematic literature reviews: An experience report. In: **2020 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)**. [S.l.: s.n.], 2020. p. 471–478.

GARCéS, L. et al. An experience report on update of systematic literature reviews. In: . [S.l.: s.n.], 2017. p. 91–96.

KITCHENHAM, B. **Procedures for Performing Systematic Reviews**. [S.l.], 2004.

KITCHENHAM, B.; BUDGEN, D.; BRERETON, P. **Evidence-Based Software Engineering and Systematic Reviews**. [S.l.]: Chapman & Hall/CRC, 2015. (Chapman & Hall/CRC Innovations in Software Engineering and Software Development Series).

KITCHENHAM, B.; CHARTERS, S. **Guidelines for performing Systematic Literature Reviews in Software Engineering**. [S.l.], 2007.

MENDES, E. et al. When to update systematic literature reviews in software engineering. **Journal of Systems and Software**, v. 167, p. 110–167, 2020.

NAPOLEÃO, B. M.; PETRILLO, F.; HALLÉ, S. Automated support for searching and selecting evidence in software engineering: A cross-domain systematic mapping. In: IEEE. **2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)**. [S.l.], 2021. p. 45–53.

NAPOLEÃO, B. M. et al. Towards continuous systematic literature review in software engineering. In: IEEE. **2022 48th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)**. [S.l.], 2022. p. 467–474.

NLTK Team. **Natural Language Toolkit**. <https://pypi.org/project/nltk>. Online; accessed 21 April 2024.

OCTAVIANO, F. et al. Scas-ai: a strategy to semi-automate the initial selection task in systematic literature reviews. In: IEEE. **2022 48th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)**. [S.l.], 2022. p. 483–490.

PINTAS, J. T.; FERNANDES, L. A.; GARCIA, A. C. B. Feature selection methods for text classification: a systematic literature review. **Artificial Intelligence Review**, Springer, v. 54, n. 8, p. 6149–6200, 2021.

Scikit-learn. **Sklearn Feature Selection ANOVA F**. <https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_classif.html>. Online; accessed 21 April 2024.

Scikit-learn. **Sklearn Feature Selection Chi2**. <https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.chi2.html>. Online; accessed 21 April 2024.

Scikit-learn. **Sklearn Feature Selection Pearson's r**. <https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.r_regression.html>. Online; accessed 21 April 2024.

Scikit-learn. **Sklearn Model Selection GridSearchCV**. <https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html>. Online; accessed 21 April 2024.

SERRà, J.; ARCOS, J. L. An empirical evaluation of similarity measures for time series classification. **Knowledge-Based Systems**, v. 67, p. 305–314, 2014. ISSN 0950-7051.

STOL, K.-J.; FITZGERALD, B. A holistic overview of software engineering research strategies. In: **CESI**. [S.l.]: IEEE Press, 2015. p. 47–54.

WATANABE, W. M. et al. Reducing efforts of software engineering systematic literature reviews updates using text classification. **Information and Software Technology**, v. 128, 2020. ISSN 0950-5849.

WOHLIN, C. et al. Successful combination of database search and snowballing for identification of primary studies in systematic literature studies. **Information and Software Technology**, Elsevier, v. 147, p. 106908, 2022.

WOHLIN, C. et al. Guidelines for the search strategy to update systematic literature reviews in software engineering. **Information and Software Technology**, v. 127, p. 106366, 2020. ISSN 0950-5849.

WOHLIN, C.; RAINER, A. Is it a case study? a critical analysis and guidance. **Journal of Systems and Software**, v. 192, p. 111395, 2022. ISSN 0164-1212.

WOHLIN, C. et al. **Experimentation in software engineering**. [S.l.]: Springer Science & Business Media, 2012.

ZHANG, L. et al. Empirical research in software engineering — a literature survey. **Journal of Computer Science and Technology**, v. 33, p. 876–899, 2018.

ZHOU, Y. et al. Quality assessment of systematic reviews in software engineering: A tertiary study. In: **International Conference on Evaluation and Assessment in Software Engineering (EASE)**. New York, NY, USA: Association for Computing Machinery, 2015. p. 1–14. ISBN 9781450333504.