Artificial Intelligence-Assisted Living Systematic Review: A Case Study in the Monte Carlo Electron Transport Research

Federica Frabetti[1] and Elisabetta Ronchieri[2(✉)]

[1] Department of Statistical Sciences, University of Bologna, Bologna, Italy
federica.frabetti6@studio.unibo.it
[2] INFN CNAF, Bologna, Italy
elisabetta.ronchieri@cnaf.infn.it

Abstract. Continuously monitoring the literature in Living Systematic Reviews (LSRs) is computationally demanding, particularly during the reference screening stage, where newly retrieved studies must be classified as relevant or irrelevant. In highly specialized fields like Monte Carlo Electron Transport, manual screening is not only tedious but also mentally taxing due to the domain-specific lexicon. Existing Artificial Intelligence (AI) tools often are not easily adaptable to specific contexts. In this research, we introduce and assess a modular AI-assisted screening framework characterized by four main components: a feature extractor, a classifier, a querier, and a balancer. We define and evaluate nine different configurations, ranging from traditional TF-IDF representations paired with classical classifiers to more advanced transformer-based embeddings, such as Sentence-BERT and Multilingual-e5-large. We take into consideration the title and abstract for each reference.

In screening datasets, there is the class imbalance issue. Therefore, to tackle this problem, we incorporate a resampling strategy along with Active Learning querying methods, including uncertainty sampling and certainty-based prioritization. Our experimental findings on a Monte Carlo Electron Transport dataset reveal that the configuration based on Multilingual-e5-large, Support Vector Machine, and certainty-based prioritization achieves the best balance between recall and screening efficiency with the highest WSS@95% value of 0.89838.

The proposed pipeline provides a reproducible way to automate document classification in specialized domains and reduces one of the main bottlenecks in maintaining LSRs with high-quality.

Keywords: Living Systematic Reviews · Active Learning · Automated Reference Screening · Monte Carlo Electron Transport · Transformer Embeddings · Imbalanced Learning · Supervised Classification

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16758, pp. 84–99, 2027.
https://doi.org/10.1007/978-3-032-30497-1_6

1 Introduction

The proliferation of scientific publications across diverse disciplines has made the aggregation of extant knowledge substantially difficult. Systematic Reviews (SRs) constitute a methodologically rigorous and transparent approach to synthesizing evidence [7]; however, in rapidly advancing domains, such syntheses tend to lose importance soon after dissemination and publication. Living Systematic Reviews (LSRs) mitigate this constraint through the implementation of ongoing surveillance mechanisms and iterative revision cycles [15,22].

The reference screening phase represents a critical limitation of LSRs during their routine maintenance [2,5]. Newly discovered publications must undergo some level of manual evaluation to determine their relevance or irrelevance to the review scope. This classification burden is especially demanding in narrowly focused research areas such as Monte Carlo Electron Transport, where accurate assessment requires a sophisticated grasp of specialized concepts and a deep understanding of domain-specific terminology. Furthermore, the problem is exacerbated by severe class imbalance—typically, fewer than 5% of retrieved records are relevant to the research question [12]. Such conditions render the screening process both inefficient and demanding for human reviewers.

Computational approaches available in Artificial Intelligence and Natural Language Processing offer substantial promise for automating reference screening [?] [1,3,8,28]. Nevertheless, contemporary systems exhibit three fundamental constraints that limit their practical application. First, most existing methods are suitable for general-domain tasks and therefore inadequately capture the nuanced technical lexicon of specialized fields. Second, many such systems function as unknown decision-making architectures, thereby constraining interpretability and restricting the capacity for independent verification, limiting transparency and reproducibility. Third, these tools show insufficient performance when they are confronted with severe class imbalance, producing unacceptably elevated false-negative rates when high-recall screening becomes essential.

To address these challenges, we propose a modular, AI-assisted framework for automated reference screening in the specialized domain of Monte Carlo Electron Transport. This framework decomposes the screening pipeline into four components—feature extractor, classifier, querier, and balancer—allowing systematic evaluation and customization. We define nine configurations combining traditional sparse representations (i.e., TF-IDF) [18] and transformer-based dense embeddings (i.e., Sentence-BERT, and Multilingual-e5-large) [24,29], together with multiple classifiers [11] (that include Support Vector Machines, Logistic Regression, Naive Bayes, and Random Forests) and Active Learning querying strategies [10] (specifically Uncertainty Sampling and Certainty-based Prioritization). Moreover, we integrate a dedicated resampling layer to mitigate class imbalance and ensure high recall for rare includable documents.

Our querying strategies fall within the Active Learning paradigm characterized by two querying approaches [17]: uncertainty sampling, which prioritizes documents near the decision boundary to improve model learning; and certainty-

based prioritization, which selects documents with the highest predicted relevance to accelerate the identification of includable studies. This distinction allows us to evaluate the trade-off between exploration (uncertainty sampling) and exploitation (certainty-based prioritization) within the same Active Learning framework. Moreover, while uncertainty sampling is widely adopted, we investigate whether certainty-based prioritization is more effective in extreme class imbalance settings where rapid identification of relevant studies is critical.

Through a simulation on a Monte Carlo Electron Transport dataset, we show that configurations combining dense embeddings with certainty-based prioritization significantly improve screening efficiency. Our results demonstrate that, in contrast to common practice, exploitation-driven querying can outperform exploration strategies in high-imbalance, high-recall contexts.

While the individual components of the proposed framework are based on established techniques, the novelty of this work lies in their systematic integration and controlled comparative evaluation within a highly specialized and extremely imbalanced domain. The preliminary literature review conducted during the early stages of the research yielded a substantial number of documents characterized by highly domain-specific vocabulary, which poses significant challenges for traditional feature extraction methods. In particular, we provide: (i) a modular architecture enabling controlled comparison across feature extraction, classification, querying, and balancing strategies; (ii) a comprehensive empirical analysis of certainty-based versus uncertainty-based prioritization under extreme class imbalance; and (iii) reproducible evidence demonstrating that exploitation-driven querying strategies outperform traditional approaches in this setting.

2 Related Works

Active Learning (AL) is widely used to accelerate the screening process in Systematic Reviews. It is already a well-established research area. Tools, such as ASReview [27] and LitStudy [9], use the human in the loop system: in this case, a machine learning model keeps suggesting iteratively the most relevant documents based on what the reviewer has labeled previously.

Most frameworks [27] use uncertainty sampling, where the model picks instances close to the decision boundary [19,23]. Other frameworks go with certainty-based prioritization [13], which means focusing on documents with the highest chance of being included. However, the effectiveness of these strategies really depends on the initial seed set and how well the classifier deals with the sparse, includable documents. Our work builds on this approach by adding a dedicated Querier module. It benchmarks those two sampling approaches across nine different model setups.

The performance of screening AI is fundamentally limited by its Feature Extractor [4,20]. Early methods relied on Bag-Of-Words and TF-IDF representations. They were efficient and straightforward, but they missed semantic synonyms and the nuances in technical language.

Now transformer-based language models, such as BERT or RoBERTa, use dense embeddings [16,25]. Models such as Sentence-BERT and Multilingual-e5-large perform better in cross-lingual tasks and domain-specific ones. They map text to a high-dimensional vector space where similar concepts are clustered together. However, their computational cost remains a significant limitation. In our study, we compare traditional TF-IDF to newer ones like mxbai-embed-large and Multilingual-e5-large in a physics context, specifically because of specialized areas that require more work.

Class imbalance is a significant challenge in LSRs [21,22]. Relevant papers usually are less than 5% of the results. Standard classifiers typically overfit to the irrelevant class, leading to a dangerous drop in recall.

There are various Balancer strategies like Synthetic Minority Over-sampling Technique and random resampling [26]. However, few solutions look at how they work with physics-based lexicons. In fields like Monte Carlo Electron Transport research, the vocabulary is sufficiently specialized that general tools (often trained on PubMed or legal data) might get it wrong. Our framework addresses this gap by making a Balancer module central to the pipeline. It keeps recall high even when the imbalance is extreme.

Tools like Rayyan or Covidence are user-friendly, but they act like black boxes with limited ways to customize the internal architecture [14]. Researchers cannot swap classifiers easily or adjust querying for a physics review. We propose a modular architecture to set up a white box that allows us to test and validate properly for computational science.

3 Experimental Setup

3.1 Dataset and Gold Standard

The dataset used in this study consists of 3,330 records (titles and abstracts) retrieved from major scientific databases, including arXiv and Google Scholar, using a Boolean search string tailored to Monte Carlo Electron Transport (see Fig. 1). To evaluate the models, we established a "gold standard" labeled dataset where each relevant record was manually annotated and classified by domain experts. This process resulted in 133 includable (relevant) documents and 3,197 irrelevant documents.

This distribution corresponds to an inclusion rate—defined as the proportion of relevant papers over the total number of retrieved (relevant + irrelevant) papers, as shown in Eq. 1—of 3.99%, reflecting the pronounced class imbalance typical of specialized systematic reviews.

$$\text{Inclusion Rate } (\%) = \left(\frac{\text{Relevant Papers}}{\text{Relevant Papers} + \text{Irrelevant Papers}} \right) \times 100 \qquad (1)$$

Irrelevant papers are defined as publications that do not strictly address the specific topic of interest, despite sharing similar terminology, keywords, or belonging to closely related research areas. These records are retrieved from

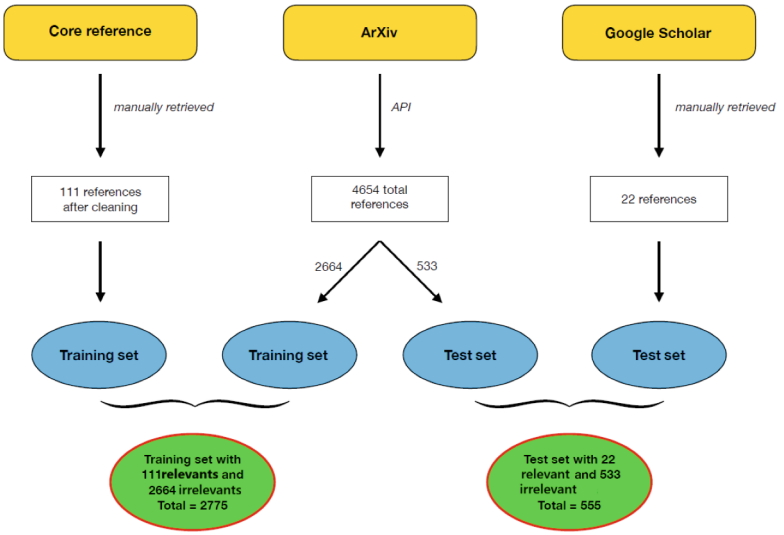


Fig. 1. Training and test sets.

databases due to their lexical or thematic similarity to relevant studies, but they do not provide meaningful contributions to the core objectives of the Living Systematic Reviews. In contrast, relevant papers are those that align closely with the central focus of the LSR and contribute with substantive insights, such as additional characteristics, methodological details, or domain-specific findings, improving the understanding of the topic under investigation.

3.2 Training and Test Set Definition

To simulate the screening process, we employed a stratified hold-out validation strategy. Unlike a standard random split, this approach ensures that the model's ability to identify rare "includable" papers is tested rigorously.

To initialize the Active Learning process, we defined a seed set composed of a small number of labeled relevant and irrelevant documents. Specifically, the seed set includes 20 documents, with a balanced representation of both classes where possible. This initial set simulates the early stage of a systematic review, where a researcher has already identified a limited number of key studies.

In addition to the seed set (the initial library), the training data includes a document pool, the remainder of the training set from which the querier selects documents for labeling to present to the researcher. The seed set (the initial library) and the document pool are subsets of the same training resource.

During the simulation, the querier selected documents from the training set based on its logic (e.g., Uncertainty sampling). Once a document was "queried," its label was revealed to the model (simulating a human's manual decision), and the model was retrained iteratively.

The test set includes the unseen library and is used exclusively for final performance evaluation. We reserved 16.67% of the total labeled dataset as a dedicated test set. This set was never seen by the feature extractor or the classifier during the iterative querying process. This ensures that the reported metrics accurately reflect the model's generalization ability in a realistic "needle-in-a-haystack" screening scenario.

The balancing (e.g., oversampling or undersampling) was only applied to the training set. To maintain the validity of the simulation, the test set was kept in its original, imbalanced state. This ensures that the reported performance metrics reflect the true difficulty of the "needle-in-a-haystack" screening task.

At each iteration of the Active Learning loop, a batch of 2 documents is selected by the querier and presented for labeling. Once labeled, these documents are added to the training set, and the model is retrained. The process continues iteratively until a predefined stopping criterion is reached. In our experiments, the stopping condition corresponds to achieving a target recall threshold (e.g., 95%) or exhausting the document pool. Unless otherwise specified, models are updated after each iteration using the augmented labeled set.

We wish to detail that the training set is characterized by relevant and irrelevant references as shown in Fig. 2: the former is obtained by experts and existing bibliography like "Unravelling Tungsten atom ionization at high energies: A relativistic electron impact study" (https://doi.org/10.1016/j.elspec.2025.147552), the latter derives from the ArXiv database like "Modification of the uniform electron gas polarizational stopping power due to the interaction of the projectile with new collective modes at moderate and strong coupling" (https://doi.org/10.48550/arXiv.2505.17229). The test set also contains both relevant and irrelevant references, maintaining the same proportion between the two classes as the training set: the relevant entries are obtained via Google Scholar, while the irrelevant entries are obtained via ArXiv.

3.3 Evaluation Metrics

Standard accuracy is misleading in imbalanced screening tasks. Therefore, we defined success using the following metrics [6]:

1. Recall (Sensitivity): The percentage of all relevant papers found. In LSRs, a recall of <95% is generally unacceptable.
2. Workload Savings (WSS@95): The percentage of the library a researcher did not have to read because the AI correctly predicted those papers as irrelevant before the 95% recall threshold was reached.
3. Extra Relevant Records found (ERF): It is the proportion of relevant records found after correcting for the number of relevant records found via random screening.
4. Average Time to Discovery (ATD): The average number of documents a reviewer must screen before encountering a relevant (includable) one.
5. Loss: It is the proportion of "includable" (relevant) documents that the model incorrectly classified as irrelevant and were therefore never shown to the researcher.

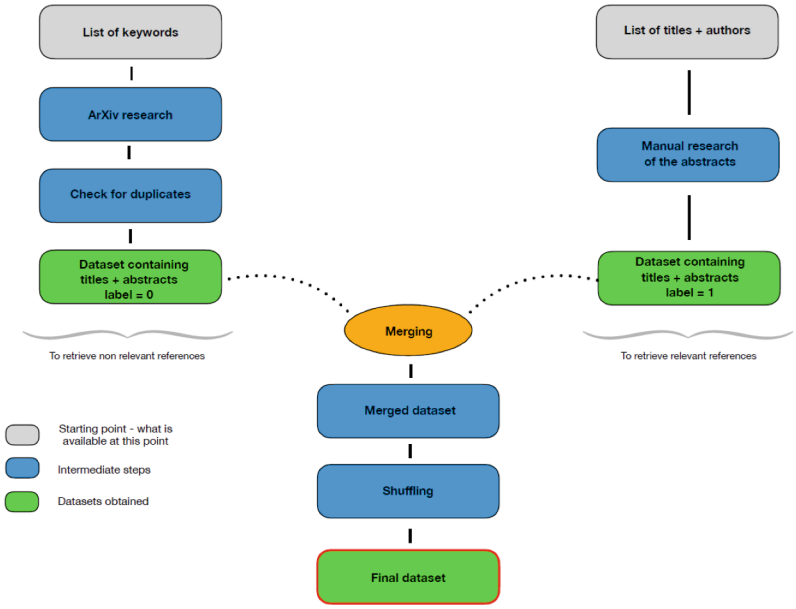


Fig. 2. Training set definition.

The terms ERF (Extra Relevant Records found), Loss, and ATD (Average Time to Discovery) are standard metrics in the broader field of Active Learning for Systematic Reviews. ERF is closely related to WSS, which quantifies the percentage of the library the researcher did not have to read.

WSS quantifies how much human effort was saved automating the screening process, given a recall level. It measures how many documents the reviewer did not have to screen compared to random screening, while achieving a target recall. It is computed as Eq. 2:

$$WSS@R = R - \frac{n}{N} \tag{2}$$

where R defines the target recall, n represents the number of documents screened to reach that recall, and N is the total number of documents in the dataset. WSS at 95% provides insight into the global efficiency of the model under a safety constraint, given by the high recall level required.

ERF at 10% is the proportion of relevant records found after correcting at screening 10% of the total number of records available in the dataset.

ATD determines how long, on average, the model takes to discover a new relevant reference, where the time is interpreted as the number of records needed to screen. The ATD measures the average number of labeling decisions (documents screened) required to find a relevant record. It represents the speed of discovery. It has a low value if relevant documents are ranked at the top. This metric provides the number of screening decisions that are necessary to find each relevant record.

Loss typically refers to the Recall Loss—the fraction of relevant documents that were missed because the model incorrectly classified them as irrelevant at the point where screening was stopped.

ATD and Loss cover aspects of the screening process not captured by WSS and ERF, such as discovery speed and learning stability, respectively.

To ensure robustness, experiments can be repeated across multiple random initializations of the seed set. Due to computational constraints, results are reported for a single representative run. While this limits statistical generalization, the consistency of trends across configurations suggests robustness of the observed behavior. Future work will include repeated runs and statistical validation.

3.4 Configurations Summary

Table 1 summarizes the nine distinct, defined configurations of the modular AI-assisted screening framework according to the four components: feature extractor, classifier, querier and balancer.

The feature extraction layer evaluates both traditional sparse representations and dense embeddings to capture the specialized lexicon of Monte Carlo Electron Transport. TF-IDF (Sparse) represents documents based on the importance of terms relative to the corpus, calculated through the product of term frequency and inverse-document frequency. It produces high-dimensional sparse vectors where each dimension corresponds to a unique term in the vocabulary. Transformer-based embeddings (Dense) models (SBERT-mpnet, SBERT-mxbai, and Multilingual-e5-large) map text into a dense, high-dimensional vector space. Unlike TF-IDF, these capture semantic synonyms and contextual nuances critical for interpreting technical physics terms.

The classification layer consists of four distinct discriminative models tailored to handle high-dimensional data. Support Vector Machine (SVM) is used in Models A, F, H, and I, finding the optimal hyperplane to separate relevant and irrelevant classes. It is particularly effective in sparse spaces like those generated by TF-IDF. Logistic Regression is employed in Models B, E, and G. This probabilistic classifier produces a confidence score representing the likelihood of a document being relevant. Model C utilizes Naive Bayes to estimate class probability based on feature frequency, while Model D uses Random Forest to manage complex non-linear relationships in the data.

Models A through D and also H rely on TF-IDF to figure out how important certain words are in the documents based on term frequency and inverse-document frequency. These are paired with different classifiers, like SVM for A and H, then Logistic Regression in B, Naive Bayes for C, and Random Forest in D. Model A specifically takes the SVM and uses uncertainty sampling to pick documents right around the decision boundary in those sparse spaces, which helps find the best separation. Model H is highly similar to A but it switches over to certainty-based prioritization instead, just to isolate the querying strategy.

The transformer-based models, from E to G and I, use dense embeddings with Sentence-BERT (SBERT) or Multilingual-e5-large to capture deep semantic

Table 1. Configurations Summary

Model	Feature Extractor	Classifier	Querier	Balancer
A	TF-IDF (Sparse)	SVM (Linear)	Uncertainty Sampling	Balanced
B	TF-IDF	Logistic Reg.	Uncertainty Sampling	Balanced
C	TF-IDF	Naive Bayes	Certainty-based prioritization	Balanced
D	TF-IDF	Random Forest	Certainty-based prioritization	Balanced
E	SBERT-mpnet (Dense)	Logistic Reg.	Certainty-based prioritization	Balanced
F	Multilingual-e5-large	SVM	Certainty-based prioritization	Balanced
G	SBERT-mxbai (Dense)	Logistic Reg.	Uncertainty Sampling	Balanced
H	TF-IDF	SVM	Certainty-based prioritization	Balanced
I	SBERT-mxbai	SVM	Certainty-based prioritization	Balanced

meaning. Model F uses Multilingual-e5-large along with SVM, and the certainty-based approach performed well in terms of recall and precision.

Models A, B and G employ uncertainty sampling, prioritizing documents near the decision boundary to make the model learn better over time. Then the others models– from C to F, H and I, use certainty-based prioritization, picking the documents with the highest predicted relevance to accelerate the identification of includable studies. The TF-IDF models seem simpler overall, but the transformers capture more meaning.

Table 2 summarizes the hyperparameter settings used to initialize the nine model configurations (Models A–I). These settings ensure that the comparative analysis isolates the impact of the feature extractor and querying logic rather than variance in tuning.

The Balancer sets a strategy to Balanced. This standardized setup allows for a rigorous white-box evaluation of how different AI architectures respond to the extreme class imbalance (3.99% inclusion rate) found in specialized research screening. The Balancer component across all configurations was set to a synthetic oversampling ratio of 0.5 to mitigate the 3.99% inclusion rate.

3.5 Architecture

Figure 3 shows three integrated stages: input and component definition, the core iterative Active Learning workflow, and the 9-model comparative matrix mapping models A–I.

Stage 1 defines the raw input (the Monte Carlo Electron Transport dataset) and explicitly lists the sub-options for each main pillar: feature extractors shows the transition from sparse vectors (TF-IDF) to dense contextual embeddings (mpnet, e5-large, mxbai); classifiers lists the four heads models, i.e., SVM, Logistic Regression, Naive Bayes, Random Forests; queriers visually contrasts Active Learning with uncertainty sampling versus certainty-based prioritization; balancer shows the resampling layer to manage class imbalance.

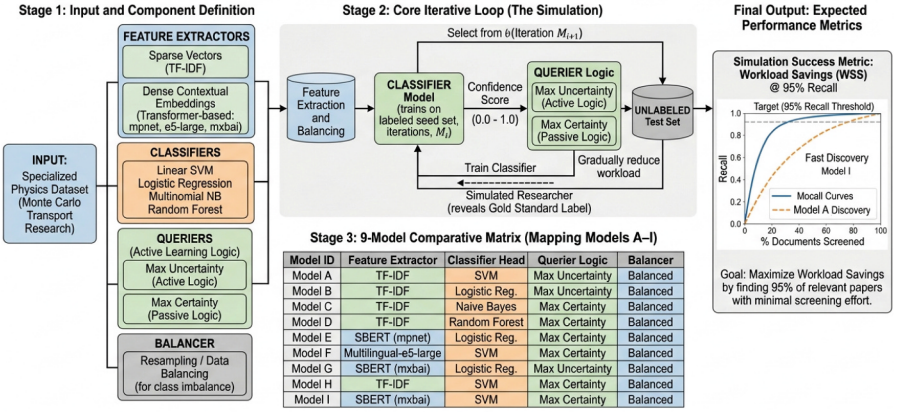
Table 2. Hyperparameter Settings

Component	Parameter	Configuration/Value	Model
TF-IDF	N-gram Range	(1, 2) (Unigrams and Bigrams)	H
	N-gram Range	(1, 1) (Unigrams)	A
	Max Features	5,000	A-D, H
	Vector Type	Sparse high-dimensional	A-D, H
	Max Df	1.0	A
	Max Df	0.95	H
	Sublinear TF	False	A
	Sublinear TF	True	H
SBERT/	Max Seq. Length	512 tokens	E-G, I
Multilingual-e5-large	Pooling	Mean Pooling	E-G, I
	Normalize	L2	E-G
	Normalize	True	I
SVM	Kernel	Linear	A, F, H-I
	C (Regularization)	1.0	A, F
	C (Regularization)	0.11	H
	C (Regularization)	0.067	I
	Penalty	L2	A, F, H-I
	Max Iter	1000	A, F, H
	Max Iter	5000	I
Logistic Regression	Type	Probabilistic Classifier	B, E, G
	Penalty	L2	B, E, G
	Solver	lbfgs	B, E, G
	Max Iter	100	B, E, G
	C (Regularization)	1.0	B, E, G
Random Forest	Num. Estimators	100	D
	Max Depth	None	D
	Max Features	sqrt	D
	Criterion	gini	D
Naive Bayes	Type	Multinomial	C
	Alpha	1.0 (Laplace smoothing)	C
	Fit prior	True	C

Stage 2 is the central part of the figure. It shows how the simulation set is fed into the configured feature extractor and balancer. The data then flows into the classifier, which outputs a confidence score (0 to 1). The querier then decides which documents from the test set are presented to the simulated researcher (or gold standard set). This creates a feedback loop that continuously retrains the model as more labels are revealed.

Stage 3 is a detailed, matrix-style table that defines the specific configuration for each of your nine test subjects. The matrix explicitly differentiates between Model A (TF-IDF + SVM + Uncertainty Sampling) and Model H (TF-IDF + SVM + Certainty-based prioritization). This isolates the effect of the querier logic. Similarly, it contrasts Model G (mxbai + Logistic Regression + Uncertainty Sampling) and Model E (mpnet + Logistic Regression + Certainty-based prioritization) to isolate the combined effect of a new feature extractor and a new querier.

The final panel in Fig. 3 shows the simulation success metric. The entire Figure points toward a graph of Workload Savings (WSS) @ 95% Recall. It illustrates competing Recall Curves for Model A (slower, uncertainty-based discovery) versus Model I (faster, certainty-based discovery), visually proving why one configuration is superior for an LSR. Figure 3 contrasts Model A (uncertainty-based, slower discovery) with Model I (certainty-based, faster discovery). A Fast Discovery model aims to minimize the ATD by prioritizing documents with the highest predicted relevance.



Stage 3: 9-Model Comparative Matrix (Mapping Models A–I)

Model ID	Feature Extractor	Classifier Head	Querier Logic	Balancer
Model A	TF-IDF	SVM	Max Uncertainty	Balanced
Model B	TF-IDF	Logistic Reg.	Max Uncertainty	Balanced
Model C	TF-IDF	Naive Bayes	Max Certainty	Balanced
Model D	TF-IDF	Random Forest	Max Certainty	Balanced
Model E	SBERT (mpnet)	Logistic Reg.	Max Certainty	Balanced
Model F	Multilingual-e5-large	SVM	Max Certainty	Balanced
Model G	SBERT (mxbai)	Logistic Reg.	Max Uncertainty	Balanced
Model H	TF-IDF	SVM	Max Certainty	Balanced
Model I	SBERT (mxbai)	SVM	Max Certainty	Balanced

Fig. 3. Modular AI-Assisted Screening Framework (Models A–I Analysis).

The image illustrates two competing logics for the querier module: uncertainty sampling versus certainty-based prioritization. In the first case, the model (a TF-IDF vectorizer + an SVM or Logistic Regression head) predicts which documents it is most uncertain about (those closest to the decision boundary). These documents are queried first to improve the model's accuracy. This logic corresponds to Models A, B, and G. In the second case, the model (a deep SBERT or Multilingual-e5-large embedding + a classifier head) prioritizes documents it is most certain are relevant. This logic corresponds to Models C, D, E, F, H, and I.

4 Results

Table 3 shows the performance metrics across all simulated models. Models F (Multilingual-e5-large + SVM + certainty-based prioritization) and B (TF-IDF + Logistic Regression + uncertainty sampling) achieve the best and worst performance, respectively.

Table 3. Performance Metrics on Training Set

Model	WSS@95%	ERF@10%	ATD [sec]	Loss
Model F	0.89838	0.90090	68.89	0.00484
Model I	0.89802	0.90090	69.32	0.00500
Model C	0.89802	0.90090	69.40	0.00503
Model H	0.89730	0.90090	69.56	0.00509
Model E	0.89730	0.90090	69.93	0.00523
Model G	0.89694	0.90090	70.27	0.00536
Model D	0.89586	0.90090	70.33	0.00538
Model A	0.89514	0.90090	71.51	0.00582
Model B	−0.05622	0.48649	883.19	0.31051

Although most configurations achieve high WSS@95%, substantial differences emerge when considering practical efficiency metrics, such as ATD and Loss. These differences highlight an important aspect: similar recall values may mask substantial variation in efficiency and reliability.

A key finding is the advantage of certainty-based prioritization in extreme class imbalance settings (≈4% inclusion rate), where rapid identification of rare relevant instances is critical. Certainty-based querying prioritizes high-confidence relevant predictions, effectively favoring exploitation over exploration, which leads to faster discovery and reduced screening effort.

The Multilingual-e5-large models achieved the highest WSS@95%. Interestingly, model C (TF-IDF + Naive Bayes + certainty-based prioritization) achieved a WSS@95% similar to model I (SBERT-mxbai + SVM + certainty-based prioritization). Models G and E, although they use more complex feature extractors, achieve a lower WSS@95% than Model C. In this case, it is crucial to consider interactions with all components: model G employs SBERT-mxbai, as does Model I.

The poor performance of Model B can be attributed to the combination of TF-IDF features and Logistic Regression under uncertainty sampling: the negative WSS@95% indicates that the screening effort exceeds that of random sampling, an extremely low ERF@10% indicates that the model increases workload instead of reducing it, while a large ATD indicates that the model takes a significant time to discover new relevant references (it requires nearly thirteen times

more screening effort compared to Model F). In sparse feature spaces, the decision boundary learned by Logistic Regression may be insufficiently expressive, leading to unreliable uncertainty estimates. As a result, the querying strategy prioritizes uninformative samples, significantly degrading performance.

In general, based on the ranking obtained, it can be noticed that no clear pattern emerges in terms of feature extractors: TF-IDF is placed both in the top part of the ranking as well as the bottom. The same can be said for embedding-based feature extractors such as SBERT, even if they never occupy the very last positions, appearing both at the top and in intermediate positions of the ranking. TF-IDF can still compete in some configurations, such as Model C; however, the best-performing model uses embeddings. Analyzing the ranking of the models in terms of the classifier employed, one can notice that often setups with Support Vector Machine show a better result compared to others, while Logistic Regression and Random Forest tend to show lower performance. In terms of querier, the division is more pronounced: the common factor among the best-performing models is the use of certainty-based querying, followed by uncertainty sampling. Uncertainty sampling is the querying strategy implemented in model B.

Then, we evaluated the best model with the test set, as summarized in Table 4. Performance remains strong on the test set, although some metrics show a moderate decline: WSS@95% goes from 0.8984 to 0.8018; ERF@10% goes from 0.9009 to 0.8636; ATD goes from 68.89 to 35.18; Loss goes from 0.00484 to 0.04443.

Our best model achieved an ERF@10% of 0.8636 on the test set, indicating that by screening only 10% of the data, the model effectively captured the vast majority of relevant papers, significantly reducing the effort required. ATD dropped from 68.89 to 35.18. A lower ATD is better, as it indicates the researcher finds "needle-in-a-haystack" much faster, which is the primary goal of the Fast Discovery logic mentioned in our framework. Loss increased from 0.00484 to 0.04443 on the test set. This is an increase, it remains below the 5% threshold (0.05), which aligns with your stated goal of maintaining a 95% Recall (meaning a maximum allowable loss of 5%).

Table 4. Model F Performance Metrics on Test Set

Metric	Purpose	Test Result	Interpretation
WSS@95%	Workload Savings	80.18%	≈80% of the documents did not need to be read
ERF@10%	Early Efficiency	0.8636	High success in early discovery
ATD	Discovery Speed	35.18	Average documents screened per seconds
Loss	Evidence Safety	4.44%	≈4.4% of relevant papers were missed

In highly imbalanced screening scenarios typical of Living Systematic Reviews, researchers should prioritize certainty-based querying strategies combined with semantically rich embeddings. Uncertainty sampling may be suboptimal when the primary goal is rapid identification of rare relevant studies. In this study, the proposed modular framework may enable researchers to tailor screening pipelines to domain-specific requirements, balancing recall, workload, and discovery speed.

5 Conclusions

This study provides a systematic and reproducible evaluation of AI-assisted screening strategies in a highly specialized domain of Monte Carlo Electron Transport, highlighting how the interaction between model components—rather than individual methods—drives performance under extreme class imbalance. We provided a white-box alternative to existing proprietary platforms and tools via an architecture composed of four distinct components—feature extractor, classifier, querier, and balancer.

We defined and assessed nine configurations, revealing that Model F, characterized by Multilingual-e5-large, SVM, and certainty-based prioritization, is the optimal configuration for this domain. It achieved a Workload Savings (WSS@95%) of 80.18% on an unseen test set, showing that researchers can significantly reduce their screening effort while maintaining the strict 95% recall safety threshold required for systematic evidence synthesis.

Key findings include transformer superiority with respect to traditional TF-IDF to get high-stakes recall scenarios. In fact, dense embeddings like Multilingual-e5-large are crucial for capturing the semantic nuances of technical physics terminology. Certainty-based prioritization is a querier logic that consistently led to fast discovery, minimizing the Average Time to Discovery (ATD) and allowing researchers to identify relevant evidence much earlier in the screening process. Furthermore, despite the extreme class imbalance (3.99% inclusion rate), the integrated Balancer component ensured that the model's Loss remained below 5%, confirming its reliability for practical LSR updates.

This study is subject to some limitations. There is the absence of repeated runs and statistical testing, which will be addressed in future work to further validate robustness. Results are based on a single dataset from a highly specialized domain, which may limit generalizability. Experiments were conducted using a single initialization of the seed set. Hyperparameter tuning was not exhaustively explored, as the focus was on comparative analysis across configurations. Future work will address these limitations through broader datasets and repeated experimental runs.

Future work will focus on integrating these modular components into a screening activity to customize researcher screening pipelines according to their specific recall and workload requirements.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abramovich, F., Grinshtein, V., Levy, T.: Multiclass classification by sparse multinomial logistic regression. IEEE Trans. Inf. Theory **67**(7), 4637–4646 (2021). https://doi.org/10.1109/tit.2021.3075137
2. Akl, E.A., et al.: The life and death of living systematic reviews: a methodological survey. J. Clin. Epidemiol. **156**, 11–21 (2023). https://doi.org/10.1016/j.jclinepi.2023.02.005
3. Anisuzzaman, D., Malins, J.G., Friedman, P.A., Attia, Z.I.: Fine-tuning large language models for specialized use cases. Mayo Clin. Proc.: Digit. Health **3**(1), 100184 (2025). https://doi.org/10.1016/j.mcpdig.2024.11.005
4. Ayash, L., Algarni, A., Alqahtani, O.: Advancements in feature selection and extraction methods for text mining: a review. Discov. Appl. Sci. **7**(8) (2025). https://doi.org/10.1007/s42452-025-07587-w
5. Burgard, T., Bittermann, A.: Reducing literature screening workload with machine learning: a systematic review of tools and their performance. Zeitschrift für Psychologie **231**(1), 3–15 (2023). https://doi.org/10.1027/2151-2604/a000509
6. Chen, M., et al.: Impact of human and artificial intelligence collaboration on workload reduction in medical image interpretation. NPJ Digit. Med. **7**(1) (2024). https://doi.org/10.1038/s41746-024-01328-w
7. Finocchi, I., Ribichini, A., Schaerf, M.: A large-scale temporal analysis of scientific production across disciplines and countries. J. Informet. **19**(2), 101665 (2025). https://doi.org/10.1016/j.joi.2025.101665
8. He, Y., Zhu, X., Li, D., Wang, H.: Enhancing large language models for specialized domains: a two-stage framework with parameter-sensitive LoRA fine-tuning and chain-of-thought rag. Electronics **14**(10), 1961 (2025). https://doi.org/10.3390/electronics14101961
9. Heldens, S., et al.: litstudy: A Python package for literature reviews. SoftwareX **20**, 101207 (2022). https://doi.org/10.1016/j.softx.2022.101207
10. Hoarau, A., Lemaire, V., Le Gall, Y., Dubois, J.C., Martin, A.: Evidential uncertainty sampling strategies for active learning. Mach. Learn. **113**(9), 6453–6474 (2024). https://doi.org/10.1007/s10994-024-06567-2
11. Idrissi Khaldi, M., Erraissi, A., Hain, M., Banane, M.: Comparative analysis of supervised machine learning classification models. In: Farhaoui, Y., Herawan, T., Imoize, A.L., Allaoui, A.E. (eds.) ICAISE 2024. LNNS, vol. 1353, pp. 321–326. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-88304-0_44
12. Iversen, I.A., Quintana, D.S.: The value of living systematic reviews. Nat. Hum. Behav. (2026). https://doi.org/10.1038/s41562-026-02428-1
13. Jafarzadeh, H., Heidary-Dahooie, J., Akbari, P., Qorbani, A.: A project prioritization approach considering uncertainty, reliability, criteria prioritization, and robustness. Decis. Support Syst. **156**, 113731 (2022). https://doi.org/10.1016/j.dss.2022.113731
14. Kellermeyer, L., Harnke, B., Knight, S.: Covidence and Rayyan. J. Med. Libr. Assoc. **106**(4) (2018). https://doi.org/10.5195/jmla.2018.513
15. Khamis, A.M., Kahale, L.A., Pardo-Hernandez, H., Schünemann, H.J., Akl, E.A.: Methods of conduct and reporting of living systematic reviews: a protocol for a living methodological survey. F1000Research **8**, 221 (2019). https://doi.org/10.12688/f1000research.18005.2
16. Kotei, E., Thirunavukarasu, R.: A systematic review of transformer-based pre-trained language models through self-supervised learning. Information **14**(3), 187 (2023). https://doi.org/10.3390/info14030187

17. Kumar, P., Gupta, A.: Active learning query strategies for classification, regression, and clustering: a survey. J. Comput. Sci. Technol. **35**(4), 913–945 (2020). https://doi.org/10.1007/s11390-020-9487-4
18. Liang, M., Niu, T.: Research on text classification techniques based on improved tf-idf algorithm and LSTM inputs. Procedia Comput. Sci. **208**, 460–470 (2022). https://doi.org/10.1016/j.procs.2022.10.064
19. Liu, S., Li, X.: Understanding uncertainty sampling (2023). https://doi.org/10.48550/ARXIV.2307.02719
20. Malekipirbazari, M., Aksakalli, V., Shafqat, W., Eberhard, A.: Performance comparison of feature selection and extraction methods with random instance selection. Expert Syst. Appl. **179**, 115072 (2021). https://doi.org/10.1016/j.eswa.2021.115072
21. Mujahid, M., et al.: Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering. J. Big Data **11**(1) (2024). https://doi.org/10.1186/s40537-024-00943-4
22. Nevill, C.R., et al.: Capturing the influx of living systematic reviews: a systematic methodological survey. J. Clin. Epidemiol. **186**, 111904 (2025). https://doi.org/10.1016/j.jclinepi.2025.111904
23. Nguyen, V.L., Shaker, M.H., Hüllermeier, E.: How to measure uncertainty in uncertainty sampling for active learning. Mach. Learn. **111**(1), 89–122 (2021). https://doi.org/10.1007/s10994-021-06003-9
24. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using Siamese BERT-networks (2019). https://doi.org/10.48550/ARXIV.1908.10084
25. Salem, M., Mohamed, A., Shaalan, K.: Transformer models in natural language processing: a comprehensive review and prospects for future development. In: Hassanien, A.E., Rizk, R.Y., Darwish, A., Alshurideh, M.T.R., Snášel, V., Tolba, M.F. (eds.) AISI 2025. LNDECT, vol. 238, pp. 463–472. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-81308-5_42
26. Sasirekha, R., Kanisha, B.: Adaptive ensemble framework with synthetic sampling for tackling class imbalance problem. Eng. Rep. **7**(4) (2025). https://doi.org/10.1002/eng2.70109
27. van de Schoot, R., et al.: An open source machine learning framework for efficient and transparent systematic reviews. Nat. Mach. Intell. **3**(2), 125–133 (2021). https://doi.org/10.1038/s42256-020-00287-7
28. Supriyono, Wibawa, A.P., Suyono, Kurniawan, F.: Advancements in natural language processing: implications, challenges, and future directions. Telemat. Inform. Rep. **16**, 100173 (2024). https://doi.org/10.1016/j.teler.2024.100173
29. Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., Wei, F.: Multilingual e5 text embeddings: a technical report (2024). https://doi.org/10.48550/ARXIV.2402.05672