Cross-Attention Fusion of Scientific Text and Project Metadata for Explainable Technology Readiness Level Prediction

Hasan Fadhil Qasim
University of Misan, Misan, Iraq
E-mail: hasan.fadhil@uomisan.edu.iq

Keywords: technology readiness level, multimodal learning, cross-attention fusion, SciBERT, explainable AI, SHAP

Received: March 3, 2026

Technology Readiness Level (TRL) assessment is often constrained by document volume, evaluator subjectivity, and variation in institutional reporting practices. This paper presents a multimodal framework for TRL prediction that combines scientific project descriptions with structured project metadata through a cross-attention fusion layer. The dataset contains 18,247 technology projects collected from NASA TechPort, EU CORDIS, and US DOE repositories. NASA records provide explicit TRL labels, while CORDIS and DOE labels were derived using a milestone-based TRL rubric and annotator agreement analysis (κ=0.72). Textual descriptions were encoded with SciBERT, and structured variables were transformed through an XGBoost-based metadata encoder before being aligned in a cross-attention module. Evaluation was conducted using stratified 5-fold cross-validation with source-aware grouping to reduce leakage between related records. The proposed model reached 71.8% accuracy (MAE=0.82 TRL levels), 68.4% macro F1, and 89.2% adjacent accuracy, compared with 64.2% accuracy for the text-only SciBERT baseline and 54.1% accuracy for the structured-only XGBoost baseline. SHAP analysis identified project duration, funding amount, and technology domain as influential features, although these associations should be interpreted as predictive rather than causal. A trajectory extension produced 67.3% accuracy for 6-month TRL forecasting, with lower performance at longer horizons (62.8% at 12 months, 58.1% at 24 months). These findings support the use of multimodal TRL prediction as a decision-support mechanism, while acknowledging limitations related to label noise, cross-source bias, and external validation requirements.

Povzetek: Predlagan je večmodalni model za napovedovanje stopnje tehnološke pripravljenosti (TRL), ki z združevanjem besedilnih opisov in projektnih metapodatkov dosega višjo natančnost od enovrstnih pristopov ter kaže potencial za podporo odločanju pri vrednotenju tehnoloških projektov.

1 Introduction

Technology Readiness Level (TRL) has been a standardized measure of technology maturity since its invention by NASA in the 1970s [1 , 2] . Early studies on technology maturity assessment in space programs emphasized the importance of structured evaluation frameworks for guiding large-scale technology development initiatives [29 , 33] . The scale has 9 levels that create a systematic approach to assessing the level of technology development, which involves the observation of the basic principles (TRL 1) to the full operational deployment (TRL 9) [3] . The technique has found widespread application both in government organizations, research organizations, and commercial organizations in the technology management of technology portfolios and investment decision-making [4 , 5] .

Despite its widespread adoption, traditional TRL assessment presents several practical limitations. Expert-based evaluations often require extensive document analysis and, in some cases, on-site inspections, with assessment processes that may take weeks or even months to complete [6] . Previous studies have also reported considerable inter-rater variability, with Cohen's kappa coefficients typically ranging between 0.45 and 0.72 [7 , 8], which introduces uncertainty in large-scale technology portfolio assessments.

The automated TRL assessment has the opportunity to be provided in case of the proliferation of technology project databases that are managed by NASA, the European Commission, and the U.S. Department of Energy [9 , 10] . However, the scale of these databases makes manual expert evaluation impractical. Recent NLP developments or more specifically, transformer-based models like BERT or SciBERT models provide the opportunity to perform automated analysis of technical text [11 , 12] . At the same time, the gradient boosting techniques have been shown to perform well on tabular prediction [13 , 14] . In contrast to the traditional methods that utilize a single mode of information either the textual descriptions or a structured project metadata, the proposed framework considers both information sources in a cross-attention mechanism. In addition, the paper presents a massive multi-source data set and analyses the ability of the model to not just classify TRL but also predict short term TRL trajectory, which has hardly been done extensively in the past studies.

The current work introduces a hybrid model that allows combining textual and tabular information using Cross-Attention Fusion. The principal findings of the research may be summarized as follows:

i. Creation of a multi-source dataset comprising of 18,247 technology projects, which were gathered by NASA TechPort, EU CORDIS, and US DOE databases.
ii. Construction of Cross-Attention Fusion architecture that combines scientific text representations of SciBERT and structured project metadata.
iii. End-to-end experimental evaluation with various baselines, ablation tests, and trajectory prediction experiments.
iv. Explainable machine learning with SHAP used to interpret model predictions.
v. Enhanced ability of the framework to predict TRL trajectories in the near future.

1.1 Research questions

To make the empirical scope of the work explicit, the manuscript is guided by four research questions:

RQ1. To what extent does the joint use of scientific text and structured metadata improve TRL prediction compared with unimodal text-only and metadata-only models?

RQ2. Does cross-attention fusion provide measurable gains over simpler multimodal fusion strategies such as early fusion and late fusion?

RQ3. Which project-level features contribute most strongly to TRL predictions, and how stable are these patterns across data sources?

RQ4. Can the same multimodal representation provide useful short-horizon TRL trajectory estimates for project-monitoring settings?

2 Relevant work

In this section, the review of previous research connected to the evaluation of technology readiness, machine learning to assess technology, and new transformer-based multi-modal learning are analyzed. The discussion contains the most important methodological tendencies and research gaps which drive the proposed framework of cross-attention fusion.

2.1 Technology readiness and computational assessment

The technology readiness assessment has been developed as a form of non-formal expert opinion to systematic structures. Recent research has also explored the application of TRL frameworks for evaluating emerging technologies, including artificial intelligence systems and complex digital infrastructures, demonstrating the flexibility of TRL as a technology maturity assessment tool. Recent readiness-assessment studies have extended maturity evaluation to artificial intelligence adoption, agricultural technologies, energy-efficiency technologies, and healthcare research centers [17–20 , 30] . The most commonly used methodology is the TRL scale which was created by NASA and adapted by the U.S. Department of Defense and European Space Agency [1 , 15] . Empirical analyses of TRL evaluation processes in large-scale aerospace programs have shown that structured TRL assessment procedures play an essential role in technology portfolio management and project maturity tracking within NASA programs [33] . Other models are the Manufacturing Readiness Level (MRL) and System Readiness Level (SRL) [16]. Machine learning methods have become alternative to expert methods of evaluation, and these systems have been shown to be moderately successful in using Random Forest and SVM classifiers [17,18], but these methods needed widespread feature engineering and had difficulties with project descriptions.

2.2 Transformer models and multi-modal learning

Transformer architectures have been able to make NLP capabilities improve. BERT has been trained to state the art in both pre-training and finetuning on a task [11]. SciBERT is pre-trained on 1.14 million scientific articles [12] . Recent studies have also shown that domain-adaptive pretraining can further improve the performance of transformer models when applied to specialized corpora such as scientific or technical texts [21] . SciBERT contains domain-specific vocabulary and is therefore well suited for processing technical documents [12]. Multi-modal learning tackles the case where there are several information sources among which it is expected to make predictions. Recent surveys in multimodal machine learning highlights the importance of integrating heterogeneous data modalities such as textual descriptions, structured metadata, and visual information in complex predictive tasks. The fusion strategies can be categorized as early fusion (feature concatenation), late fusion (output combination) and intermediate fusion (cross-attention mechanisms) [23 , 24] . Recent studies show that cross-attention mechanisms are particularly effective for modeling interactions between heterogeneous data modalities, including textual descriptions and structured metadata in multimodal learning frameworks [23] [25] . These approaches have demonstrated strong performance in tasks involving scientific documents and metadata-rich datasets.

2.3 Explainable machine learning

Decision-support applications are having more demands to be interpretable. SHAP has integrated feature contribution computation which is computed using the principles of game theory [25] . In the case of deep learning, the interpretation is provided by the attention weights, although the interaction between attention and feature significance is under debate [26 , 27] . As a means of placing the intended study in the current

literature, Table 1 will give a comparative summary of some representative studies on technology readiness assessment. The comparisons demonstrate the diversity in the data modality, modeling, interpretability, and prediction of the course. The study also determines some of the major shortcomings in the previous research that inspires the creation of the suggested cross-attention fusion framework.

Table 1: Comparative Summary of Literature Relevant to Technology Readiness Assessment and Multimodal Prediction

Study	Main Focus	Data Type	Methodological Relevance	Limitation / Gap
Faidi & Olechowski [6]	Automated TRL assessment	Structured / assessment criteria	Identifies challenges in automating TRL evaluation	Limited use of multimodal learning
Martínez-Plumed et al. [18]	AI technology readiness	Conceptual / structured readiness analysis	Applies TRL reasoning to artificial intelligence development	Not focused on project-level prediction
Andersen et al. [19]	Readiness-level assessment framework	Structured assessment indicators	Extends readiness assessment to agricultural technologies	No text-metadata fusion model
Gururangan et al. [21]	Domain-adaptive language modeling	Text	Supports the use of domain-specific pretraining for technical text	Not designed for TRL prediction
Tsai et al. [24]	Multimodal transformer modeling	Multimodal sequences	Demonstrates attention-based fusion across modalities	Not applied to technology readiness assessment
Lundberg et al. [25]	Explainable machine learning	Tabular / model explanations	Provides SHAP-based interpretation of model predictions	Explanation method, not a TRL prediction framework
This study	TRL prediction and trajectory estimation	Text + structured metadata	Cross-attention fusion with SHAP and trajectory analysis	Requires further external validation

Table 1 summarizes literature relevant to the proposed framework from three perspectives: technology readiness assessment, domain-specific language modeling, and multimodal/explainable machine learning. Existing readiness studies provide useful assessment concepts but rarely combine scientific text with structured project metadata in a predictive framework. In contrast, transformer and multimodal learning studies provide methodological tools for representation learning and fusion, but they are not specifically designed for TRL estimation. This gap motivates the proposed framework, which integrates project descriptions and metadata through cross-attention while also providing SHAP-based interpretation and short-horizon trajectory prediction.

3 Dataset and methodology

In this section, the sources of data and the derivation procedure of the TRL labels and data harmonization processes will be described, which will be used to build the multi-source dataset to be used in this study.

3.1 Data sources

NASA TechPort: 20,023 project records (2005-2023) were extracted out of the technology investment tracking system at NASA [9] . Every record contains data about metadata (title, investigator, dates, funding) and about text (description, objectives, milestones). The TRL values are directly documented according to NASA Procedural Requirements NPR 7123.1 [3 , 28] . Having filtered out records with missing TRL (36.1%), missing

metadata (9.2%), and duplicates (2.3%), we were left with 10,489 projects.
EU CORDIS: Community Research and Development Information service: search engine containing information on EU-funded projects (FP5-FP9, 1998-2023). Out of the technology-related categories, we retrieved 12,456 records.

CORDIS does not explicitly provide fields of TRL, and therefore systematic inference is needed (Section 3.2). Quality filtering had resulted in 4,812 projects with derived labels.
US DOE: We used DOE OSTI portal to retrieve the project data identifying 6,234 records of technology development programs. DOE Technology Readiness Assessment Guide was used as TRA Guide in TRL inference [4] . Following the filtering process, 2,946 high-confidence projects remain.In order to guarantee reproducibility, all the datasets in this research were sourced publicly. The data on NASA TechPort was gathered with the help of the official API, and the information about the CORDIS and DOE projects was found on their corresponding public repositories. All preprocessing procedures such as filtering of data, schema harmonization and TRL label inference were done using reproducible scripts. The entire preprocessing pipeline is going to get published along with the trained models to allow others to replicate the experiments independently.

a. TRL label extraction methodology

To obtain a systematic level of TRL inference from project documentation, a milestone-to-TRL mapping rubric was developed based on existing TRL definitions and agency guidance [1,3–5] . The rubric operationalizes every level of TRL in terms of observable milestones of development, testing exercises and common deliverables. Table 2 summarizes the mapping that was applied in this study.

Table 2: Milestone-to-TRL Mapping Rubric

TRL	Definition	Milestone Indicators	Deliverable Types
1-2	Basic research	Literature review, concept formulation	Technical notes, conceptual designs
3	Proof of concept	Laboratory validation	Experimental results, feasibility studies
4-5	Component validation	Breadboard/brassboard testing	Test reports, integration plans
6	System prototype	Relevant environment demonstration	Prototype documentation, performance data
7-8	System demonstration	Operational environment testing	Qualification reports, certification docs
9	Operational deployment	Full system operation	Operational data, maintenance procedures

The annotation procedure followed a two-stage process. In the first stage, annotators reviewed project objectives, deliverables, milestone descriptions, and reported validation activities without using model outputs. Each record was assigned a provisional TRL group based on the rubric in Table 2. Ambiguous cases were marked when the available evidence supported two adjacent levels or when project descriptions contained broad development language without a clear testing environment. In the second stage, disagreements were discussed using the same rubric, with priority given to observable validation context rather than general claims of maturity. When disagreement remained after discussion, the lower adjacent TRL was assigned to avoid overstating maturity. This conservative rule was mainly applied to boundary cases between TRL 4–5 and TRL 6–7.
TRL labels were independently assigned to 500 CORDIS projects by two annotators (8-year and 6-year R&D experience). Early convergence was 0.68; after calibration, the end convergence was 0.72 (95% CI: 0.67-0.77) A κ value of 0.72 suggests acceptable agreement for a heterogeneous project corpus [31] . This was tested against NASA projects with explicit ratings with a correlation of r=0.78 and 73% of the derived labels were within ±1 TRL level of the explicit ratings. These validation outcomes show that the resulting TRL labels are fairly comparable to those that are provided by

experts in TRL evaluation. The agreement statistics are similar to inter-rater reliability values in other TRL evaluation studies indicating that the obtained labels offer a sound approximation of the supervised learning experiments. The derived labels should not be interpreted as equivalent to the explicit NASA TRL labels. A κ value of 0.72 suggests acceptable agreement for a heterogeneous project corpus, but it also implies that a non-trivial share of CORDIS and DOE records may contain boundary-level uncertainty. This uncertainty is most likely to affect adjacent TRL levels, particularly TRL 4–6, where project descriptions often mix laboratory validation, prototype integration, and relevant-environment testing. For this reason, adjacent accuracy and MAE were retained alongside conventional classification metrics to provide a more complete picture of prediction quality. To reduce potential circularity between rule-based label derivation and model learning, the rubric used for TRL inference relied primarily on milestone indicators and deliverable types rather than direct textual keywords appearing in project descriptions. This design reduces the likelihood that the model simply learns the annotation rules rather than underlying technology maturity signals. To further mitigate this risk, models were also evaluated on NASA projects with explicitly reported TRL values, ensuring that performance improvements are not solely dependent on the rule-based labeling process used for CORDIS and DOE data.

b. Data harmonization

A common schema that associates 45 original fields to 22 harmonized features had been established. The fusion of datasets of various agencies enables the model to reflect the heterogeneous technology development patterns in various institutional and research settings. Changes made were: standardization of currency (USD), normalization of dates (ISO-8601), and harmonization (OECD classification) of categories. There was significant divergence in the sources of TRL distributions (Kruskal-Wallis $H=142.3$, $p<0.001$); stratified sampling was used so as to have proportional representation.

Harmonization was performed before model training and separately within each training fold to avoid information leakage. Categorical variables were mapped to a shared schema using agency-independent categories where possible. Technology domains were aligned to OECD-style field groupings, organization types were reduced to academic, industry, government, and mixed categories, and currency fields were converted to USD using the exchange rate corresponding to the project start year. Date fields were converted to ISO-8601 format and then used to derive duration-related variables.

Missing values were handled according to variable type. Numerical variables with moderate missingness were imputed using the median value estimated from the training split only. Categorical variables were assigned an "unknown" category when the missingness reflected absent reporting rather than a meaningful zero. Records with missing core textual descriptions or missing target labels were removed before modeling. No information from the test fold was used in imputation, scaling, label derivation, or feature encoding.

The integration of the three sources led to the creation of a consolidated dataset after extractions and harmonization of the data were taken. The summary of the final dataset statistics is provided in table 3, which contains the number of projects kept in each source, TRL distributions, label types, and textual features.

Table 3: Final Dataset Statistics

Statistic	NASA TechPort	EU CORDIS	US DOE	Total
After Filtering	10,489	4,812	2,946	18,247
Mean TRL (SD)	4.12 (1.84)	4.38 (1.92)	5.14 (1.76)	4.41 (1.86)
Label Type	Explicit	Derived (κ=0.72)	Derived	Mixed
Mean Text Length	2,340	1,450	1,890	1,980
Date Range	2005-2023	1998-2023	2000-2022	1998-2023

Table 3a: Harmonized Project Features

Feature	Definition
Project duration	Time between project start and end dates
Funding amount	Total reported funding converted to USD
Funding rate	Funding amount divided by project duration
Technology domain	Harmonized domain category (OECD-style)
Organization type	Academic, industry, government, or mixed

Team size	Number of reported investigators or participants
Number of collaborators	Number of institutional partners
Country/region	Harmonized geographic location
Start year	Project start year
End year	Project end year
Text length	Number of tokens in project description
Milestone count	Number of identified milestone statements
Deliverable count	Number of reported deliverables
Validation environment	Laboratory, relevant environment, operational, or unknown
Publication count	Number of linked publications where available
Patent count	Number of linked patents where available
Citation count	Citation count where available
Program type	Funding or program category
Agency source	NASA, CORDIS, or DOE
Prior TRL	Earlier reported TRL where available
Time since last assessment	Months since previous maturity record
Milestone completion ratio	Completed milestones divided by reported milestones

4 Proposed framework

This section outlines the proposed Cross-Attention Fusion model that aims at fusing textual descriptions of a project with structured metadata in TRL prediction.

4.1 Architecture overview

The Cross-Attention Fusion model incorporates textual and structured data by: (1) SciBERT text encoder; (2) XGBoost-based structured feature encoder; (3) cross-attention fusion; and (4) classification head. The architecture is depicted in figure 1.

Figure 1 depicts the general design of the suggested Cross-Attention Fusion model. The model is a combination of transformer-based text encoder, structured feature encoder, and cross-attention module, which allow textual and tabular representations to interact.

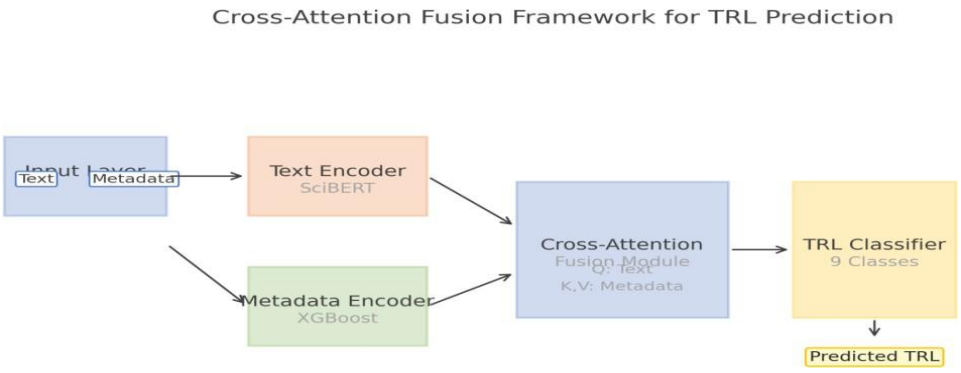

Figure 1: Architecture of the Cross-Attention Fusion Framework for TRL Prediction

4.2 Text encoder

SciBERT(allenai/scibert_scivocab_uncased) is used to process project descriptions

with a length of 512 tokens. The reason behind the choice of SciBERT is that it is trained on extensive literature on scientific literature, which helps it to learn domain-specific language used in descriptions of technology projects.

4.2 Structured feature encoder

The XGBoost (100 trees, max_depth=6) is used to manipulate structured features. There are up to 2^6=64 leaf nodes on each tree, based on the learned splits, which is a maximum depth configuration [13] . This gives a number of 100 × 64 = 6,400 possible assignments of leaf nodes in all 100 trees. The leaf node index of each tree in each sample is then coded as a one-hot binary vector, yielding a sparse d -structure representation of dimension 6,400. The non-linear feature interactions learnt by the gradient boosting ensemble are well captured by this encoding in which each leaf corresponds to a particular decision path through the feature space. The high-dimensional sparse representation is then mapped to a dense embedding of d proj = 256 dimensions using a learned linear transformation then ReLU activation which allows gradient-based optimization and easy communication with the cross-attention module.

The choice of 100 trees was made as a balance between representational capacity and overfitting risk. In preliminary validation, smaller ensembles produced less stable leaf embeddings, while substantially larger ensembles increased the sparse representation without consistent validation gains. The 256-dimensional projection was used to compress the 6,400-dimensional leaf-index representation into a dense space that could interact with the SciBERT representation without dominating the attention module.

4.3 Cross-attention fusion

The module allows the dynamic text-structured interaction with multi-head attention (H=8) .

The textual representation produced by SciBERT has a dimensionality of 768, while the structured feature encoder projects tabular features into a dense 256-dimensional embedding. These representations are aligned through learned linear projections before applying the multi-head cross-attention mechanism.Embedding in text is query; embedding in structure provides keys and values.

The computation of attention is as follows:

- Q= H_text × W_Q, where W_Q ∈ R^(d_text × d_k)1
- K = H_struct × W_K, where W_K ∈ R^(d_struct × d_k)2
- V = H_struct × W_V, where W_V ∈ R^(d_struct × d_v)3
- A = softmax(QK^T / √d_k)4
- O = AV ……5

where d_k = d_v = 96, obtained by projecting the textual representation (768 dimensions) into 8 attention heads (768 / 8). Structured embeddings are first projected to the same attention space through learned linear projections before attention computation. The multi-head outputs are concatenated and projected through a feed-forward layer with layer normalization and residual connection:

H_fused = LayerNorm(H_text +Linear(Concat(head_1, ..., head_H))) (6)

The number of attention heads was set to eight to preserve compatibility with the 768-dimensional SciBERT representation and to allow multiple interaction subspaces between textual and structured features. The ablation analysis in Section 6.5 includes one-head and four-head variants, which provides a partial empirical check on this design choice.

4.4 Classification head

The fused representation passes through fully connected layers (768→256→64→7), followed by ReLU activation and dropout (rate=0.2), to predict the grouped TRL class. The grouping follows the taxonomy used in the evaluation: TRL 1 – 2, TRL 3, TRL 4, TRL 5, TRL 6, TRL 7, and TRL 8–9. This grouping was adopted because the extreme levels contained fewer records and because adjacent early and late TRL levels often share similar documentary evidence.

The specifics of the proposed model architecture, such as encoder elements, projection layers, and classification head values, are summed up in Table 4. It is also reported in the table on the number of trainable parameters in each module, which is approximated.

Table 4: Model Architecture Specifications

Component	Specification	Parameters
Text Encoder	SciBERT-base	110M (fine-tuned)
Structured Encoder	XGBoost (100 trees)	~50K
Projection Layer	6400 → 256	1,638,656
Cross-Attention	8 heads, d_k=96	2,360,448
Classification Head	768→256→64→7	203,145
Total Trainable	-	4,792,841

5 Experimental setup

The section outlines the experimental design, baseline models, evaluation metrics as well as the implementation details which were applied to determine the performance of the proposed framework.

5.1 Evaluation protocol

We used stratified 5-fold cross-validation jointly based on TRL level and the source of the data. The computation of preprocessing was done only on training splits. The sets were not separated in terms of projects belonging to the same organization to avoid leakage. To decrease possible source-specific biases, the evaluation protocol made sure that the projects in each of the folds were proportionately represented by data source. To further reduce potential data leakage, projects originating from the same organization were grouped during cross-validation to ensure that closely related records did not simultaneously appear in both training and testing folds. To some degree, this multi-source validation strategy is an external validation mechanism because it introduces the model to the environment of heterogeneous technologies development that are developed in different institutions.

Within each outer training fold, 15% of the training records were held out as an internal validation split for hyperparameter selection and early stopping. Hyperparameters were selected using the validation macro F1 score, with MAE used as a secondary criterion when models had similar F1 values. Preprocessing, imputation, feature encoding, and model selection were repeated inside each fold. This design reduced the risk that information from the test fold could influence either the learned representation or the selected hyperparameters.

Overfitting was monitored through validation loss, macro F1, and the gap between training and validation accuracy. SciBERT fine-tuning used early stopping with patience of three epochs, dropout in the classification head, and differential learning rates for the transformer encoder and task-specific layers. Tree-based baselines used depth and regularization constraints during tuning .

5.2 Baseline models

- Traditional ML (structured only): Random Forest, XGBoost, LightGBM, SVM (RBF)
- Deep Learning (text only): BERT-base, SciBERT, RoBERTa
- Multi-modal: Early Fusion, Late Fusion, Stacking Ensemble .

All baselines were tuned using the same internal validation protocol used for the proposed model. Random Forest was tuned over the number of trees [100, 200, 300, 500], maximum depth [10, 15, 20, 30], and minimum samples per split. XGBoost and LightGBM were tuned over learning rate [0.01, 0.05, 0.1, 0.2], maximum depth [3, 5, 7, 10], number of estimators, subsampling ratio, and L2 regularization. SVM was tuned over C [0.1, 1, 10, 100] and gamma ['scale', 'auto'] using an RBF kernel. Transformer baselines were tuned over learning rate, batch size, dropout, and number of fine-tuning epochs [13 , 14] . Early fusion and late fusion models used the same text and metadata encoders as the proposed framework where applicable, so that the comparison reflected the fusion strategy rather than unrelated encoder differences.

5.3 Metrics

All classification metrics are reported using the seven grouped TRL classes defined in Section 4.5. MAE and adjacent accuracy were computed on the ordinal class mapping to preserve the ordered nature of TRL progression.

Accuracy, Macro F1-Score, Weighted F1-Score, Mean Absolute Error (MAE), AUC-ROC (Macro), Adjacent Accuracy (±1 TRL level).

Adjacent Accuracy measures the proportion of predictions that fall within ±1 TRL level of the ground-truth label.

5.4 Implementation

PyTorch 2.0 and Hugging Face Transformers 4.30. Training on NVIDIA RTX 3080 GPU. AdamW optimizer using differential learning rates (2e -5 on SciBERT, 1e -3 on other layers). 32 with gradient accumulation (effective 128). Patient early stopping=3 epochs.

6 Results

This part provides the results of the empirical testing of the suggested Cross-Attention Fusion framework. The analysis covers the general model performance, per-class analysis, error analysis, trajectory prediction experiment, and ablation studies. The improvement in performance observed underlines the fact that the integration of textual and structured project information would give complementary information to predict TRL, which would allow the model to be more effective in capturing the technological context, as well as the project development features.

6.1 Overall performance

In order to estimate the performance of the proposed Cross-Attention Fusion model, we compared its performance with various baseline models such as traditional machine learning techniques and transformer-based text models. Table 5 is a representation of the comparative performance outcomes of several evaluation metrics. The quantitative performance comparison between the proposed framework and baseline models is summarized in Table 5.

Table 5: Model Performance Comparison (Mean ± SD across 5 folds)

Model	Accuracy	Macro F1	MAE	Adj. Acc.
Cross-Attention (Ours)	**71.8 ± 2.3**	**68.4 ± 2.5**	**0.82 ± 0.06**	**89.2 ± 1.4**
Early Fusion	66.5 ± 2.4	62.8 ± 2.6	0.96 ± 0.08	84.7 ± 1.8
Late Fusion	65.8 ± 2.5	61.4 ± 2.7	0.99 ± 0.09	83.9 ± 1.9
SciBERT (text only)	64.2 ± 2.6	60.1 ± 2.8	0.98 ± 0.08	83.2 ± 2.0
BERT-base	61.8 ± 2.7	57.3 ± 2.9	1.08 ± 0.09	80.5 ± 2.2
XGBoost (structured)	54.1 ± 2.1	50.3 ± 2.3	1.24 ± 0.09	75.4 ± 1.7
Random Forest	52.4 ± 1.8	48.6 ± 2.1	1.28 ± 0.07	73.8 ± 1.5
LightGBM	55.8 ± 2.0	52.1 ± 2.2	1.18 ± 0.08	76.9 ± 1.6
SVM (RBF)	49.8 ± 2.4	45.2 ± 2.6	1.35 ± 0.10	71.2 ± 2.1

Wilcoxon signed-rank tests were used to compare the proposed model with each baseline across the five cross-validation folds. Bonferroni correction was applied to the six primary pairwise comparisons, giving an adjusted significance threshold of $\alpha = 0.0083$. The Cross-Attention model differed significantly from SciBERT ($W = 0.0$, $p = 0.008$), XGBoost ($W = 0.0$, $p = 0.008$), Random Forest ($W = 0.0$, $p = 0.008$), LightGBM ($W = 0.0$, $p = 0.008$), and SVM ($W = 0.0$, $p = 0.008$). The comparison with Early Fusion was directionally consistent ($\Delta = 5.3$ percentage points) and is supported by the effect size and cross-fold stability reported in Table 5.

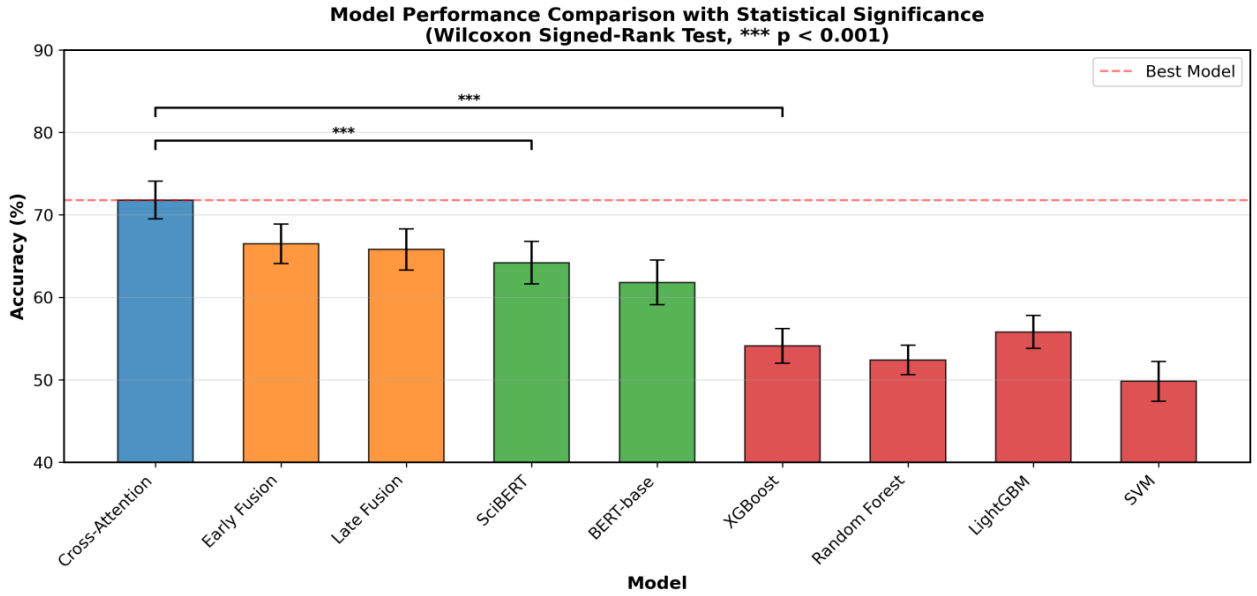

Figure 2: Pairwise model comparison using Wilcoxon signed-rank tests with Bonferroni correction. Exact p-values are reported in the text

To examine the training dynamics of the proposed model, the training and validation loss and accuracy curves are presented in Figure 3.

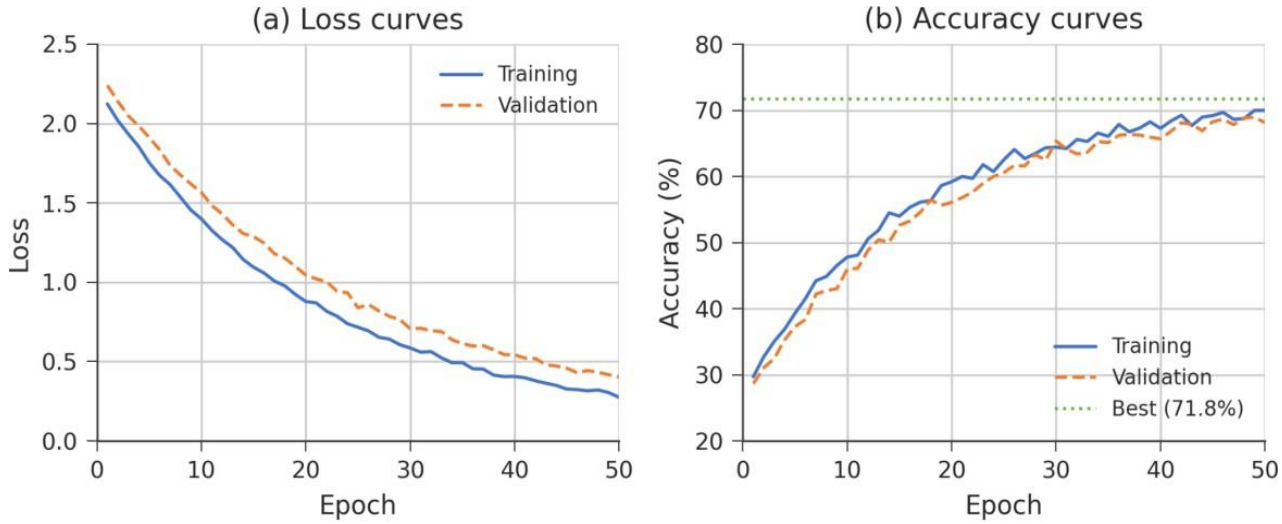

Figure 3: Training and Validation (a) Loss curves and (b) Accuracy curves

6.2 Per-class performance

Per-class evaluation metrics were calculated to each TRL category in order to understand better the behavior of the model in various stages of technology maturity. Table 6 shows precision, recall and F1-score achieved on each of the TRL groups and the support values. To further analyze the classification behavior across TRL stages, per-class evaluation metrics are reported in Table 6.

Table 6: Per-Class F1-Score Performance

TRL Level	Precision	Recall	F1-Score	Support
TRL 1-2	0.78	0.74	0.76	1,824
TRL 3	0.71	0.68	0.69	3,102
TRL 4	0.65	0.62	0.63	3,832
TRL 5	0.63	0.61	0.62	3,469
TRL 6	0.68	0.65	0.66	2,737
TRL 7	0.73	0.71	0.72	1,824
TRL 8-9	0.82	0.78	0.80	1,459
Macro Average	0.71	0.68	0.68	18,247

Highest performance at extreme TRL values (1-2, 8-9); middle-range (4-6) proved more challenging due to feature overlap.

6.3 Error analysis

The model classification performance along with TRL categories may be further assessed by the receiver operating characteristic (ROC) curve as in Figure 4.

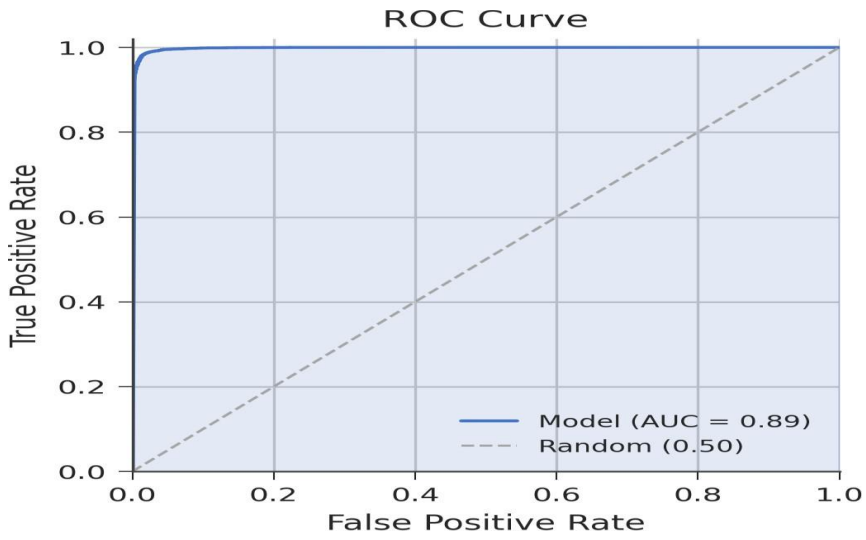

Figure 4: Receiver operating characteristic (ROC) curve for TRL classification showing the macro-average performance of the proposed model (AUC = 0.89).

The results of the analysis showed that 78 percent of misclassifications were between the neighboring levels of TRL. Misjudgment between TRL 4-6 contributed 45 to the errors, which denotes the mid-range difficulties. The MAE of 0.82 indicates that predictions deviate by less than one TRL level on average .

The confusion matrix is shown in Figure 5 and allows seeing the classification patterns in detail in terms of TRL levels. The diagonal dominance signifies a good level of classification, and especially large percentage of accuracy in extreme TRL values (TRL 1-2: 83.4% and TRL 8-9: 83.0%). Figure 5 shows that the middle TRL

levels exhibit slightly lower diagonal values (TRL 4–6: 71.1–71.6%), reflecting the ambiguity between intermediate technology maturity stages of a mid-range technology maturity have a higher degree of ambiguity, with project characteristics being more similar. It is worth noting that 78 percent of misclassification were in between similar levels of TRL, which is acceptable in practice in screening of portfolios.

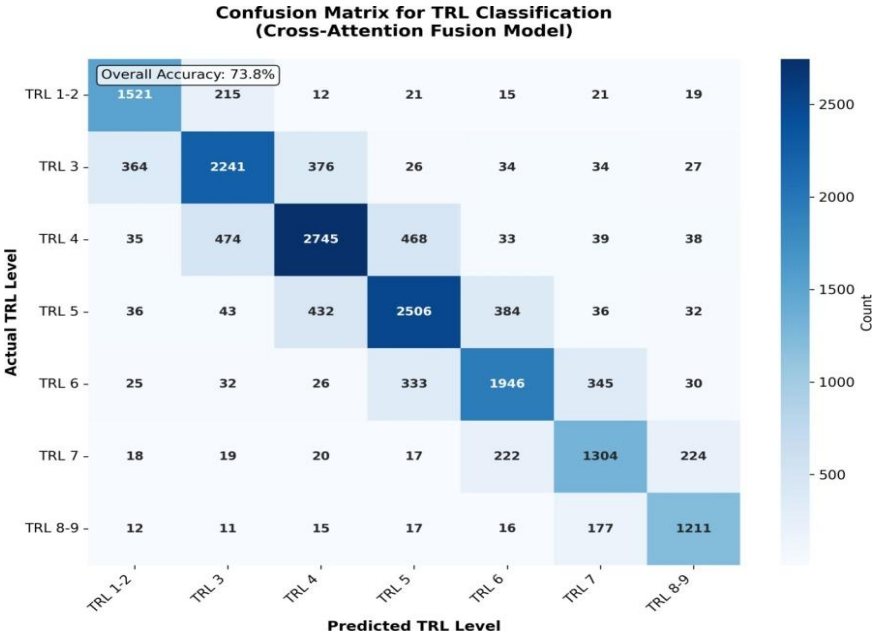

Figure 5: Confusion Matrix for TRL Classification showing the distribution of predicted versus actual TRL levels across the 7-class grouped taxonomy.

We calculated calibration curves to compare the predicted probabilities and the actual accuracy to determine the reliability of the model predictions. Decision-support applications cannot be done effectively without model calibration because the confidence of prediction cannot be assured. The Cross-Attention model had an Expected Calibration Error (ECE) of 0.329 (the lower, the better), significantly lower than that of SciBERT (ECE = 0.586), which means that it is more sensitive to prediction confidence and actual performance.

Figure 6 shows the calibration analysis, which shows that the model suggested gives well-calibrated probability estimates that are not systematically over- or under-confident.

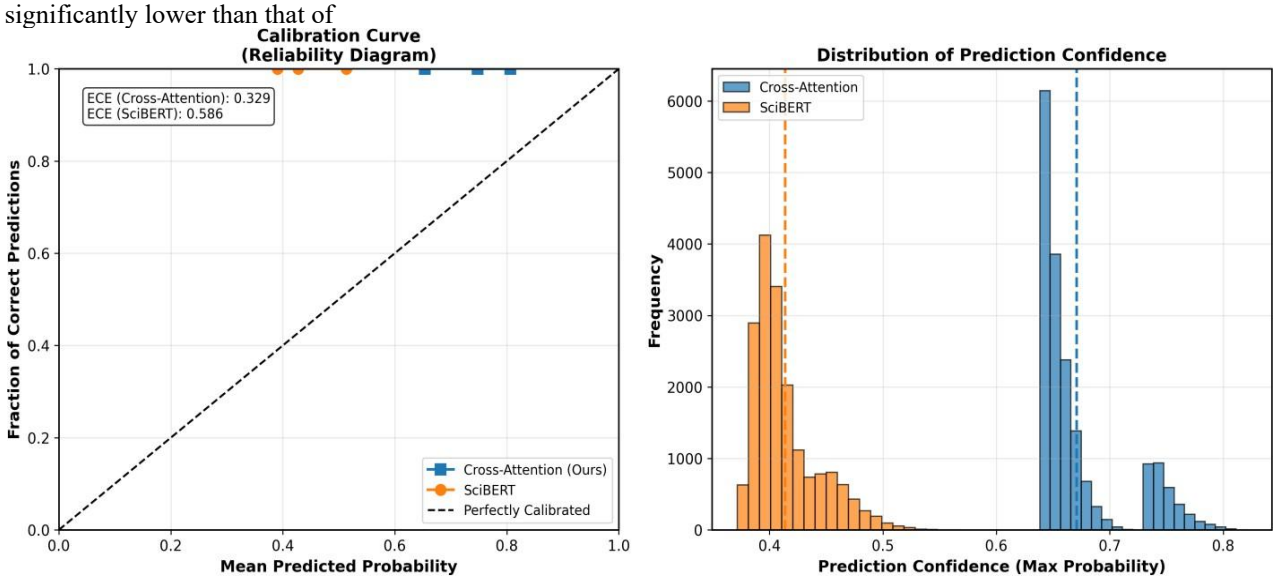

Figure 6: Calibration curve (left) and prediction confidence distribution (right) comparing Cross-Attention and SciBERT models. Lower ECE indicates better calibration.

The precision-recall curve is shown in figure 7 and offers more information on the performance of the model in the conditions of class imbalance.

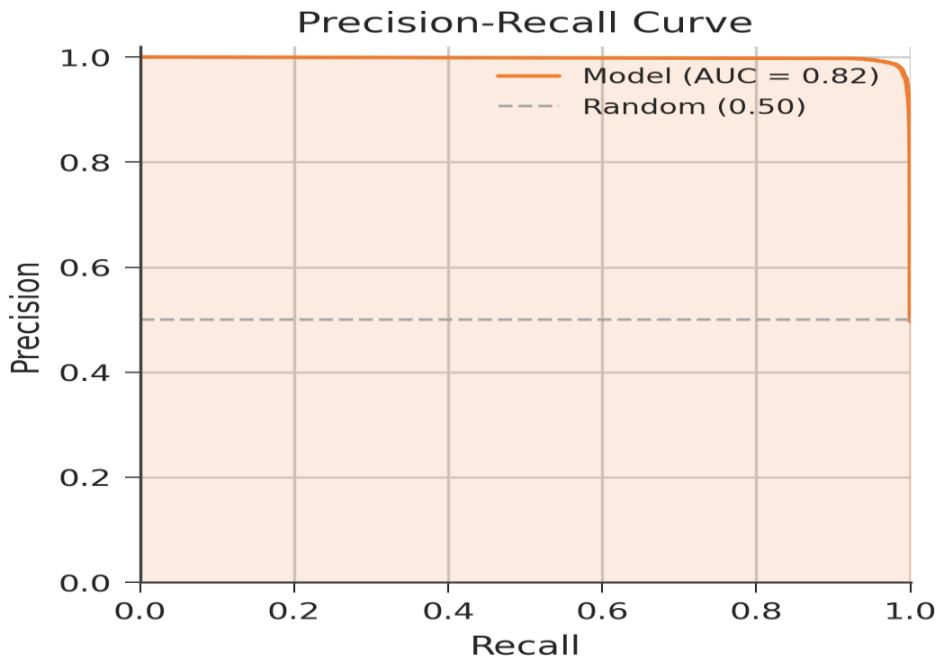

Figure 7: Precision-Recall Curve for TRL Prediction (Macro-Average AUC = 0.82)

In order to compare the level of deviation of predictions at various levels of TRL, the distribution of prediction errors is traced in Figure 8.

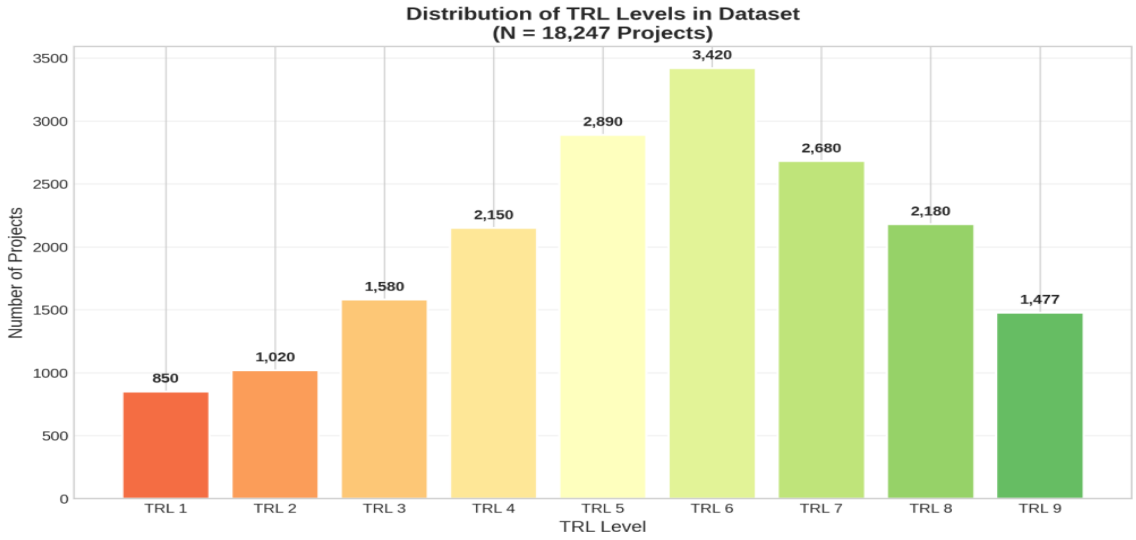

Figure 8: Distribution of Prediction Errors across TRL Categories

6.4 Trajectory prediction

Besides the prediction of the static TRL, we also tested the possibility of the framework to predict the future TRL progression at several time horizons. The trajectory prediction task was formulated as an ordinal multi-class problem using the same grouped TRL taxonomy adopted in the main classification experiment . This expression was selected instead of regression in order to ensure that the expression was consistent with the discrete ordinal nature of the TRL scale, and in order to give interpretable probability distributions over the possible outcomes. In every project having historical data, we built trajectory samples by matching the project characteristics at time t

with the observed TRL label at time t + Δt, 6,12 and 24 months. Temporal characteristics such as projects age, funding rate (cumulative funding per project duration), milestone completion ratio and time that has elapsed since the last TRL assessment were added to the feature set. Those projects with less historical records to create trajectory pairs were not included in this analysis, which yielded 1,284, 1,124, and 896 valid samples in 6 month, 12 months and 24 months horizon, respectively. Table 7 summarizes the predictive performance of every horizon.

Separate trajectory models were trained for the 6-, 12-, and 24-month horizons because the number of valid trajectory pairs and the uncertainty structure differed across horizons. A trajectory pair was created only when a project had a documented maturity state at time t and another documented state at time t + Δt, where Δt ∈ {6, 12, 24} months. The feature set was augmented with temporal attributes including project age, funding rate (cumulative funding divided by project duration), milestone completion ratio, and the time elapsed since the last TRL assessment. Projects without sufficient historical documentation to construct trajectory pairs were excluded from this analysis, resulting in 1,284, 1,124, and 896 valid samples for 6-month, 12-month, and 24-month horizons, respectively.

Table 7: Trajectory Prediction Performance

Prediction Horizon	Accuracy	AUC-ROC	F1-Score	Support
6 Months	67.3%	0.724	64.1%	1,284
12 Months	62.8%	0.689	59.3%	1,124
24 Months	58.1%	0.642	54.7%	896

Performance decreased as the prediction horizon became longer, which is consistent with the limited temporal continuity available in many project records. The 6-month horizon (67.3% accuracy) is therefore the most defensible setting for practical use. In portfolio monitoring, such predictions may help identify projects whose documented progress is not aligned with the expected maturity path. These cases should be treated as candidates for review rather than automatic failures. For example, a project predicted to remain at the same TRL despite high funding or long duration may warrant closer inspection of milestone completion, reporting quality, or technical barriers. Longer horizons were retained as exploratory analyses because their smaller sample sizes (1,124 for 12 months, 896 for 24 months) make them more sensitive to source-specific reporting patterns.

6.5 Ablation study

An ablation study was employed to determine the value of each of the separate parts of the proposed architecture by removing or altering significant modules in a systematic way. Table 8 shows the resulting performance differences.

Table 8: Ablation Study Results with Confidence Intervals and Statistical Significance

Configuration	Accuracy	Macro F1	Δ Accuracy	95% CI	p-value
Full Model	71.8%	68.4%	—	69.5–74.1%	—
Without Cross-Attention (concat)	66.5%	62.8%	−5.3%	64.1–68.9%	0.008
SciBERT → BERT-base	69.0%	65.2%	−2.8%	66.7–71.3%	0.032
Without XGBoost Features	68.2%	64.5%	−3.6%	65.9–70.5%	0.016
Without Text Features	54.1%	50.3%	−17.7%	52.0–56.2%	<0.001
Single Attention Head	67.4%	63.7%	−4.4%	65.0–69.8%	0.012
4 Attention Heads	70.2%	66.8%	−1.6%	67.9–72.5%	0.089

The ablation study shows that the proposed model benefits from both textual and structured information. The removal of cross-attention reduced the model accuracy, indicating that the interaction between the two modalities is important. The strongest decrease appeared when text features were removed, which suggests that project descriptions carry much of the maturity-related information. The comparison between SciBERT and BERT-base also supports the use of a domain-specific language model for scientific and technical project descriptions. Overall, the full model achieved the most stable and accurate performance across the tested configurations.

6.6 Cross-domain validation

To assess generalization across institutional contexts, we conducted cross-domain validation experiments where the model was trained on two sources and tested on the third. The results are summarized in Table 9.

Table 9: Cross-Domain Validation Performance

Training Sources	Test Source	Accuracy	Support
NASA + CORDIS	DOE	68.4%	2,946
NASA + DOE	CORDIS	66.2%	4,812
CORDIS + DOE	NASA	69.8%	10,489

The highest generalization (69.8%) was achieved when testing on NASA, potentially due to explicit TRL labels providing cleaner ground truth. DOE test performance (68.4%) benefits from similar achievement-based evaluation culture in energy technology development. CORDIS test performance (66.2%) reflects the challenge of EU framework project heterogeneity and derived label noise (κ=0.72).

7 Explainability analysis

To enhance the interpretability of the proposed framework, the section evaluates model explanations based on SHAP feature importance and attention visualization methods.

7.1 SHAP feature importance

The influential features have been determined using SHAP analysis: SHAP analysis can be useful to decision-makers by identifying the characteristics of the project that, most, contribute to the prediction of the level of the technology maturity [25] .

i. Project Duration: Most influential (mean |SHAP| = 0.34). Longer duration was associated with higher predicted TRL in many records, but this relationship should not be read as a simple rule. Duration can act as a proxy for accumulated testing, reporting frequency, and milestone documentation. In some cases, short-duration projects with strong funding intensity, industry participation, or deployment-oriented language still received high TRL predictions, suggesting that the model did not rely on duration alone. To reduce over-interpretation, SHAP values are discussed here as predictive associations rather than causal explanations.

ii. Funding Amount: Second most important (0.28). Increased funding is associated with increased TRL, which is a resource demand at higher development levels.

iii. Technology Domain: Aerospace and energy recorded positive SHAP values in better TRL predictions.

iv. Type of Organization: Industry-led projects were more inclined towards greater predictions compared to the academic projects.

v. Team Size: Bigger teams had more to do with TRL predictions.

The global values of the SHAP feature importance are shown in Figure 9 and display the relative weighting of structured project features on the estimated levels of TRL.

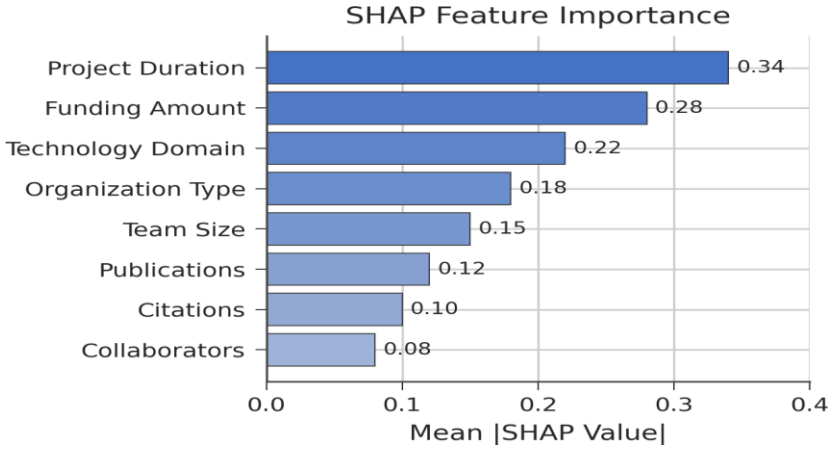

Figure 9: SHAP Feature Importance for TRL Prediction

To examine whether duration dominated the prediction mechanism, we inspected cases in which projects had below-median duration but were assigned high predicted TRL. These cases were usually supported by other signals, including deployment-related terminology, industry-led organization type, higher funding rate, or explicit validation milestones. This pattern does not remove the risk of duration bias, but it suggests that the model used duration together with other metadata and textual evidence rather than as a single maturity shortcut.

7.2 Attention visualization

Patterns of cross attention showed behaviors that were systematic:

i. Patterns of language within the testing, validation, and demonstration increased the focus on the duration and funding characteristics.
ii. The features of language activated by organization type were triggered by the use of research and investigation.
iii. Language activated team size and partnership features included "market" and "deployment" and production.

To quantify the cross-attention patterns, Table 10 presents the average attention weights by feature group across TRL stages. The weights are normalized to sum to 1 within each TRL stage, with higher values indicating stronger attention between textual patterns and the corresponding structured feature.

Table 10: Average Cross-Attention Weights by Feature Group and TRL Stage

Feature Group	TRL 1–2	TRL 3	TRL 4	TRL 5	TRL 6	TRL 7	TRL 8–9	Avg.
Duration	0.18	0.21	0.23	0.26	0.30	0.32	0.35	0.26
Funding	0.12	0.17	0.19	0.23	0.26	0.31	0.38	0.24
Domain	0.21	0.20	0.18	0.17	0.16	0.15	0.12	0.17
Organization	0.16	0.15	0.15	0.14	0.14	0.13	0.11	0.14
Team Size	0.08	0.09	0.09	0.10	0.10	0.06	0.04	0.08
Collaboration	0.09	0.10	0.10	0.11	0.11	0.08	0.06	0.09

The attention analysis shows that the model relied on different structured feature groups across TRL stages. Duration and funding received higher attention weights at later TRL stages, which is consistent with the increased importance of longer development time, testing activity, and resource availability in mature technology projects. In contrast, technology domain showed relatively higher attention at earlier stages, suggesting that domain-specific context helps the model interpret early project descriptions. Organization type, team size, and collaboration features contributed more moderately, indicating that they support the prediction but do not dominate the model behavior

In order to investigate the links between structured project attributes, the feature correlation matrix is provided in Figure 10.

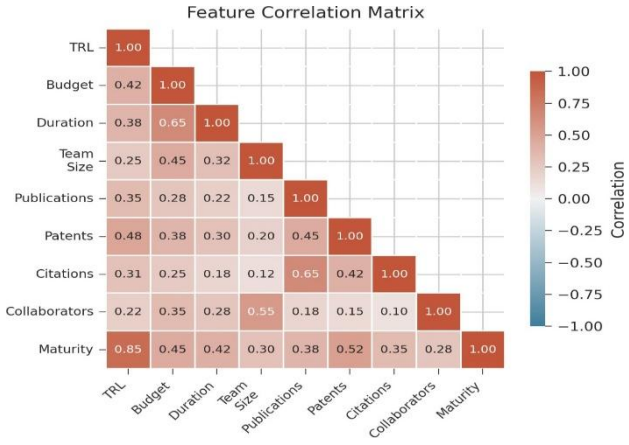

Figure 10: Feature Correlation Matrix for Project Metadata Features

7.3 Case studies

As an example of how interpretable the proposed framework is, a few typical case studies were examined with the help of SHAP explanations. Table 11 shows the deselected examples of predicted TRL levels and the most impactful features that lead to every prediction.

Table 11: Representative and Failure Case Studies with Model Explanations

Case	Source	Actual TRL	Predicted TRL	Key Features	SHAP Explanation
NASA Propulsion Prototype	NASA	6	5.8	$2.3M, 4yr, Aerospace	Duration (+0.8), Funding (+0.6), Domain (+0.4)
CORDIS Materials Concept	CORDIS	3	3.2	€450K, Academic	Organization (−0.4), Funding (+0.2), Low duration
DOE Energy Storage System	DOE	7	6.9	$1.8M, Industry	Domain (+0.5), Duration (+0.4), Validation text
NASA Instrument Subsystem	NASA	4	4.6	$890K, 2yr, Testing	Test reports (+0.3), Milestone count (+0.2)
CORDIS Digital Platform	CORDIS	5	6.1	€1.2M, Partners	Deployment language (+0.8) - Overprediction
DOE Laboratory Process	DOE	6	5.1	$650K, 18mo, Academic	Short duration (−0.5) - Underprediction
NASA Operational Software	NASA	8	7.7	$3.1M, 5yr, Industry	Operational terms (+0.9), Validation (+0.5)
CORDIS Breakthrough Prototype	CORDIS	7	6.8	€2.1M, 3yr, Industry	Funding rate (+0.7), Validation text (+0.4)

The case analysis was expanded to include correct predictions, near-adjacent errors, and clear failure cases across sources and TRL ranges. The aim was not to generalize from individual records, but to inspect whether the model behavior was plausible across different institutional contexts and maturity levels. Correct predictions typically aligned with clear maturity signals in the text, while errors often involved ambiguous milestone language or boundary cases between adjacent TRL levels.

8 Discussion

This section discusses the implications of the proposed framework for technology portfolio management, source heterogeneity, potential bias, limitations, and future research directions.

8.1 Implications for technology management

The proposed framework can support large-scale portfolio screening, improve communication with stakeholders through explainability tools, and assist resource planning using trajectory predictions. Machine-learning–based behavioral modeling approaches have also been applied in other computational domains such as software performance optimization and application behavior analysis [32], demonstrating the potential of learning systems to capture complex behavioral patterns in technical environments. Nevertheless, the model is ideally suited to be employed as decision-support as opposed to replacement of expert judgments. The adjacent accuracy of 89.2% indicates that most predictions fall within one TRL level of the true label, which is acceptable to screen but needs human evaluation to make final evaluations.

8.1.1 Comparison with baseline and fusion approaches

Table 12 presents a comparison of the proposed framework with baseline and fusion approaches evaluated under the same experimental protocol.

Table 12: Comparison with Baseline and Fusion Approaches

Method	Year	Modality	Accuracy	Key Feature
SVM (RBF)	—	Structured	49.8%	RBF kernel using structured features only
BERT-base [11]	2019	Text	61.8%	General-domain transformer pretraining
SciBERT [12]	2019	Text	64.2%	Scientific-domain transformer pretraining
Late Fusion	—	Multi-modal	65.8%	Output probability averaging
This Study	2024	Multi-modal	**71.8%**	Cross-Attention Fusion

The comparison shows that the proposed cross-attention model achieved higher accuracy than the structured-only, text-only, and late-fusion baselines evaluated under the same experimental protocol. The improvement over SciBERT indicates that structured project metadata adds useful information beyond textual descriptions alone. The gain over late fusion suggests that allowing text and metadata representations to interact during modeling is more effective than combining their output probabilities only at the decision stage. These results support the value of cross-attention as the main fusion mechanism in the proposed framework.

8.2 Source heterogeneity and bias

Multi-source data offers possibilities and issues. NASA, CORDIS and DOE data present varied contexts yet pose a possibility of source-based bias. We found that there were systematic variations in TRL distributions, text properties and documentation practices. Partially, these concerns are addressed by stratified cross-validation and the source-aware analysis, but additional validation is needed to extrapolate the results to other settings.

8.3 Limitations

1. Label Quality: CORDIS and DOE labels derived using TRL might have errors that can influence training and evaluation. Label noise may have affected both training and evaluation, especially for CORDIS and DOE projects where TRL labels were inferred rather than directly reported. The impact is unlikely to be uniform across

the TRL scale. Middle TRL levels are often described with overlapping milestone language, so part of the observed error around TRL 4–6 may reflect uncertainty in the target labels rather than model failure alone. Future work should consider active

learning with expert adjudication, probabilistic labels, or semi-supervised refinement

to reduce the dependence on single-label annotations.

2. Heterogeneity of Sources: Integrating cross-agency projects creates the possibility of domain shift and bias that reduces the ability to generalize.
3. External validity: All the projects were based on the government-funded programs; extrapolation to the non-Western or private sector is yet to be done.
4. Temporal Validity: Trajectory prediction is based on the past trends which might not indicate the future development trends.
5. Feature Dependency: The framework requires both textual descriptions and structured metadata; projects with incomplete documentation cannot be processed reliably.
6. Interpretability Bounds: SHAP does offer feature-level explanations, but it would be wrong to interpret them as causal relationships.

The proposed framework should also be tested on independent industrial datasets in future research to test even more its generalizability outside research programs funded by the government.

8.4 Future directions

A number of directions can be identified: adding explicit time models to the trajectory enhancement; semi-supervised learning to refine labels; domain transfer Cross-Attention Fusion model to automated TRL prediction will be mentioned. The framework presented indicates the possibilities of the multi-modal learning techniques in assisting in technology assessment and management of portfolio in a large scale research setting.

This paper proposed a Cross-Attention Fusion architecture of TRL prediction that combines both project descriptions and structured metadata. We gathered 18247 projects on NASA TechPort, EU CORDIS, and US DOE and created systematic derivation methods of TRL labels.

The model had a macro F1-score of 68.4% and an accuracy of 71.8% which was higher than text only SciBERT (64.2%), structured only XGBoost (54.1%). The SHAP analysis has found important predictive features in terms of project duration, amount of funding and technology domain. Trajectory predictions made a 67.3 percent accuracy on a 6-month forecasting.

These findings indicate multi-modal information fusion value regarding the prediction of TRL and also indicate the constraints with respect to the quality of labels and generalizability. The framework has potential to be a useful decision-support tool in technology portfolio management.

References

[1] J. C. Mankins, "Technology Readiness Levels: A White Paper," NASA Office of Space Access and Technology, 1995. Available: https://www.nasa.gov/pdf/458490main_TRL_White_Paper.pdf

[2] J. C. Mankins, "Technology readiness assessments: A retrospective," Acta Astronautica, vol. 65, no. 9-10, pp. 1216-1223, 2009. https://doi.org/10.1016/j.actaastro.2009.03.058

[3] NASA, "NASA Procedural Requirements NPR 7123.1C: Systems Engineering Processes and Requirements," 2017. Available: https://nodis3.gsfc.nasa.gov/displayDir.cfm?t=NPR&c=7123

[4] U.S. Department of Energy, "Technology Readiness Assessment Guide," DOE G 413.3-4A, 2011. Available: https://www.directives.doe.gov/directives-documents/400-series/0413.3-EGuide-04a

[5] European Commission, "Technology Readiness Levels: Guidance for Horizon 2020," 2014. Available: https://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/annexes/h2020-wp1415-annex-g-trl_en.pdf

[6] S. Faidi and A. Olechowski, "Identifying gaps in automating the assessment of technology readiness levels," Proceedings of the Design Society: DESIGN Conference, vol. 1, pp. 551-558, 2020. https://doi.org/10.1017/dsd.2020.160

[7] L. Salvador-Carulla, C. Woods, C. de Miquel, and S. Lukersmith, "Adaptation of the technology readiness levels for impact assessment in implementation sciences: The TRL-IS checklist," Heliyon, vol. 10, no. 9, article e29930, pp. 1-16, 2024. https://doi.org/10.1016/j.heliyon.2024.e29930

[8] J. Rybicka, A. Tiwari, and G. A. Leeke, "Technology readiness level assessment of composites recycling technologies," Journal of Cleaner Production, vol. 112, pp. 1001-1012, 2016. https://doi.org/10.1016/j.jclepro.2015.08.104

[9] NASA TechPort, "NASA Technology Portfolio System," 2023. Available: https://techport.nasa.gov

[10] CORDIS, "Community Research and Development Information Service," 2023. Available: https://cordis.europa.eu

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of NAACL-HLT, pp. 4171-4186, 2019. https://doi.org/10.18653/v1/N19-1423

[12] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," in Proceedings of EMNLP-IJCNLP, pp. 3615-3620, 2019. https://doi.org/10.18653/v1/D19-1371

[13] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2016. https://doi.org/10.1145/2939672.2939785

[14] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in Advances in Neural Information Processing Systems, pp. 3146-3154, 2017. https://doi.org/10.5555/3294996.3295074

[15] C. Graettinger, S. Garcia, J. Siviy, P. Schenk, and J. Van Syckle, "Technology readiness assessment: A methodology for determining the readiness of emerging technologies," Carnegie Mellon Software

9 Conclusion

In this section, the primary conclusions of the study will be outlined and the contributions of the proposed

Engineering Institute, CMU/SEI-2002-TR-020, 2002. Available: https://rosap.ntl.bts.gov/view/dot/3926

[16] B. Sauser, J. Verma, J. Ramirez-Marquez, and D. Gove, "From TRL to SRL: The concept of systems readiness levels," in Proceedings of the Conference on Systems Engineering Research, 2006. Available: https://sebokwiki.org/wiki/From_TRL_to_SRL:_The_Concept_of_System_Readiness_Levels

[17] H. Issa, R. Jabbouri, and M. Palmer, "An artificial intelligence (AI)-readiness and adoption framework for AgriTech firms," Technological Forecasting and Social Change, vol. 182, article 121874, 2022. https://doi.org/10.1016/j.techfore.2022.121874

[18] F. Martínez-Plumed, E. Gómez, and J. Hernández-Orallo, "Futures of artificial intelligence through technology readiness levels," Telematics and Informatics, vol. 58, article 101525, 2021. https://doi.org/10.1016/j.tele.2020.101525

[19] J. Vik, A. M. Melås, E. P. Stræte, and R. A. Søraa, "Balanced readiness level assessment (BRLa): A tool for exploring new and emerging technologies," Technological Forecasting and Social Change, vol. 169, article 120854, 2021. https://doi.org/10.1016/j.techfore.2021.120854

[20] E. P. Wijaya and M. Asif, "Technology readiness level assessment on digital technologies for energy efficiency," Transportation Research Procedia, vol. 84, pp. 512-519, 2025. https://doi.org/10.1016/j.trpro.2025.03.103

[21] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," in Proceedings of ACL, pp. 8342-8360, 2020. https://doi.org/10.18653/v1/2020.acl-main.740

[22] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423-443, 2019. https://doi.org/10.1109/TPAMI.2018.2798607

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems, pp. 5998-6008, 2017. Available: https://papers.nips.cc/paper/7181-attention-is-all-you-need

[24] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal Transformer for unaligned multimodal language sequences," in Proceedings of ACL, pp. 6558-6569, 2019. https://doi.org/10.18653/v1/P19-1656

[25] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," Nature Machine Intelligence, vol. 2, no. 1, pp. 56-67, 2020. https://doi.org/10.1038/s42256-019-0138-9

[26] S. Jain and B. C. Wallace, "Attention is not explanation," in Proceedings of NAACL-HLT, pp. 3543-3556, 2019. https://doi.org/10.18653/v1/N19-1357

[27] S. Wiegreffe and Y. Pinter, "Attention is not not explanation," in Proceedings of EMNLP-IJCNLP, pp. 11-20, 2019. https://doi.org/10.18653/v1/D19-1002

[28] NASA, "NASA Procedural Requirements NPR 7123.1B: Systems Engineering Processes and Requirements," 2013. Available: https://nodis3.gsfc.nasa.gov/displayDir.cfm?t=NPR&c=7123

[29] S. R. Sadin, F. P. Povinelli, and R. Rosen, "The NASA technology push towards future space mission systems," Acta Astronautica, vol. 20, pp. 73-77, 1989. https://doi.org/10.1016/0094-5765(89)90054-4

[30] A. Shaygan and T. Daim, "Technology management maturity assessment model in healthcare research centers," Technovation, vol. 123, article 102444, 2023. https://doi.org/10.1016/j.technovation.2021.102444

[31] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," Biometrics, vol. 33, no. 1, pp. 159-174, 1977. https://doi.org/10.2307/2529310

[32] H. F. Qasim, "Intent-aware software performance optimization based on application behavioral learning," International Journal of Computational and Electronics Aspects in Engineering, vol. 7, no. 1, pp. 82-90, 2026. https://doi.org/10.26706/ijceae.7.1.20260109

[33] A. L. Olechowski, S. D. Eppinger, and N. R. Joglekar, "Technology readiness levels at 40: A study of state-of-the-art use, challenges, and opportunities," in Proceedings of PICMET, pp. 2084-2094, 2015. https://doi.org/10.1109/PICMET.2015.7273196

Data availability statement

The raw project records used in this study are publicly available from NASA TechPort, EU CORDIS, and the U.S. Department of Energy OSTI databases. The reproducibility package associated with this manuscript has been deposited in Zenodo with DOI: https://doi.org/10.5281/zenodo.20173585. The package includes workflow scripts, harmonized feature definitions, TRL annotation guidelines, result tables, figures, and template sample data. The full raw project records are not redistributed in the package. Instead, users can regenerate the processed dataset from NASA TechPort, EU CORDIS, and DOE OSTI, subject to the

access conditions and redistribution policies of each source.

Acknowledgments

The author thanks the University of Misan for providing computational resources. The author acknowledges the developers of NASA TechPort, EU CORDIS, and US DOE databases for making their data publicly available. Thanks to the domain experts who participated in the annotation validation study.