Predicting Patentable Scientific Research Using Multi-View Machine Learning

Akshaya
Chindhuja
Irfan

Contents

1 Abstract

Only a small proportion of scientific publications ultimately result in patents, yet identifying such high-impact research at an early stage remains an important and difficult problem. Universities, funding agencies, and industry stakeholders continuously seek ways to detect research with strong commercial potential. However, traditional indicators such as citation counts, journal prestige, or institutional reputation often provide incomplete and sometimes misleading signals.

This project develops a comprehensive machine learning framework to predict whether a scientific paper is likely to lead to a patent. The proposed approach combines structured metadata features, institutional and authorship signals, temporal calibration, and deep semantic representations derived from SciBERT. A hybrid learning strategy integrates an XGBoost classifier trained on engineered metadata with a LoRA fine-tuned SciBERT model trained on scientific text. The final prediction is obtained using calibrated late fusion.

The best fusion model achieves an F1-score of 0.7085 on the held-out test set, with a recall of 0.8216. These results indicate that patentability is not random but structured and detectable when both structural and semantic information are modeled jointly. The study also reveals that innovation signals are concentrated within high-impact institutional ecosystems and that semantic understanding plays a central role in identifying patent-oriented research.

2 Introduction

The transformation of academic research into industrial innovation forms the backbone of modern knowledge economies. Despite the enormous volume of scientific output produced globally each year, only a small fraction transitions into patented technologies. Understanding what differentiates patent-linked research from the broader scientific corpus remains a significant challenge.

Historically, evaluation of innovation potential has relied on human expertise. Peer reviewers, funding panels, and technology transfer offices assess novelty, applicability, and commercial viability through qualitative judgment. While valuable, these assessments are inherently subjective and difficult to scale. Citation metrics, often used as quantitative proxies for impact, are limited because high citation counts may reflect theoretical influence rather than technological applicability. Similarly, institutional prestige correlates with research quality but does not guarantee patent success.

Recent advances in machine learning provide an opportunity to approach this problem from a data-driven perspective. Instead of relying on isolated indicators, predictive models can integrate multiple dimensions of research activity. Patentability prediction is

therefore framed in this work as a supervised learning problem, where the objective is to learn patterns from historical paper–patent linkages.

The motivation of this project is not only predictive accuracy but also interpretability and insight. Beyond identifying likely patentable research, the model aims to uncover structural characteristics associated with innovation ecosystems.

3 Problem Formulation

The task is formulated as a binary classification problem. For a given scientific paper represented by feature vector x, the model estimates:

$$P(y = 1 \mid x)$$

where $y = 1$ indicates that the paper eventually leads to a patent and $y = 0$ indicates otherwise.

Several challenges complicate this task:

- Patentable papers represent a minority of total publications.
- Citation-based metrics are noisy and temporally biased.
- Institutional naming conventions are inconsistent across datasets.
- Innovation patterns evolve over time.
- Technical language requires domain-specific semantic understanding.

To address these challenges, a multi-view learning framework is adopted. The approach combines structured metadata, authorship indicators, institutional signals, and deep text embeddings.

4 Dataset Construction and Class Distribution

After preprocessing and feature engineering, the final dataset used for modeling consists of 464,672 scientific papers. Among these, 140,073 are labeled as patent-linked, while 324,599 are non-patentable.

This corresponds to approximately 30 percent positive class representation. The distribution is moderately imbalanced but reflects a more realistic scenario than artificially balanced datasets used in earlier experiments.

Negative sampling was carefully implemented to avoid unrealistic class ratios. Earlier phases of experimentation used strict 1:1 balancing, which led to overly optimistic results that did not generalize well. The final dataset preserves scarcity while maintaining sufficient positive examples for learning stable decision boundaries.

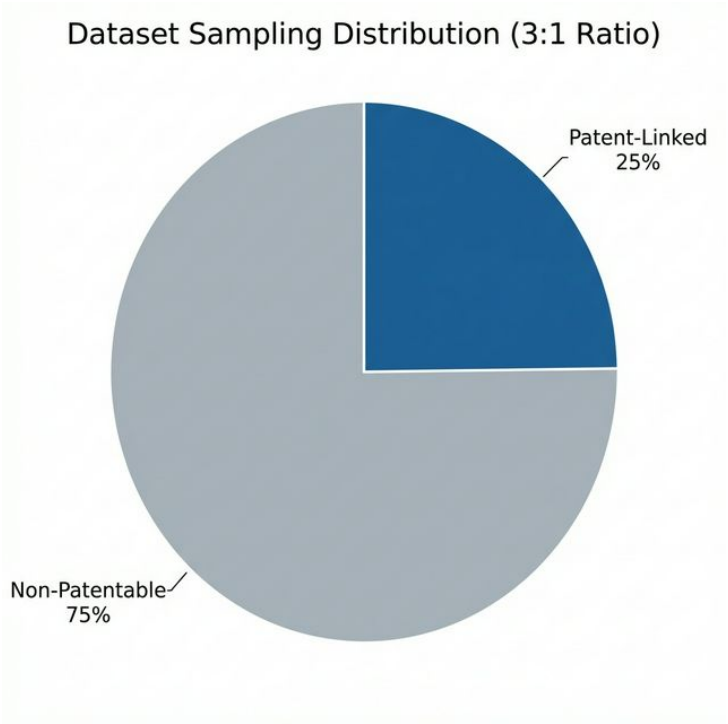

Figure 1: Distribution of Patent-Linked vs. Non-Patentable papers in the final dataset (approx. 3:1 ratio).

As illustrated in Figure 1, the dataset maintains a realistic class distribution, reflecting the inherent scarcity of patentable research while providing a robust baseline for evaluation.

5 Feature Engineering

Feature engineering plays a central role in bridging raw bibliographic information and model-ready inputs. The final structured dataset contains 15 features spanning publication characteristics, research tier indicators, authorship metrics, and institutional signals.

5.1 Publication-Level Features

Publication year is included to capture temporal trends. Title and abstract length are used as proxies for research depth and descriptive richness. The language indicator (is_eng) distinguishes English publications, which were found to exhibit stronger patent linkage patterns.

5.2 Research Tier Indicators

Three categorical indicators were constructed: core_research, sec_research, and low_novelty. These reflect journal tier positioning and approximate research prominence. Higher-tier

research venues tend to show stronger patent associations.

5.3 Authorship Metrics

Authorship features include number of authors, average author citations, and average author productivity. During exploratory analysis, average author productivity displayed a highly skewed distribution, with a large proportion of zero values. This sparsity partially explains why authorship metrics contributed less strongly in metadata-only modeling.

5.4 Institutional Signal

A binary variable indicates affiliation with top-ranked institutions. Innovation appears to cluster within established research hubs, making institutional affiliation an important structural feature.

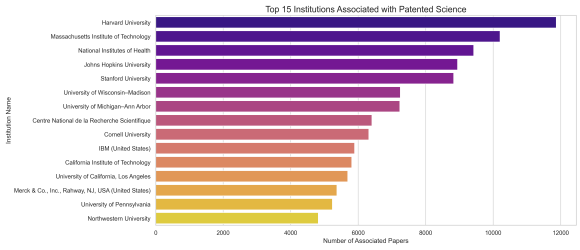

Figure 2: Patent density across top-tier institutions.

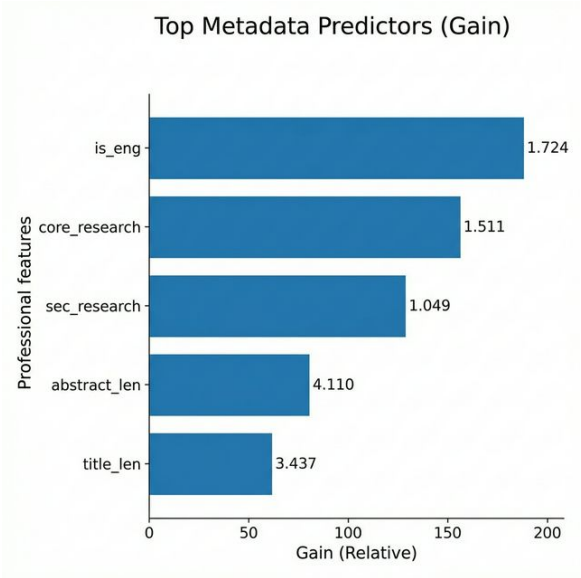

Figure 3: Top metadata feature importance scores.

The institutional clustering shown in Figure 2 confirms that patentable research is concentrated in elite ecosystems, while the importance analysis in Figure 3 highlights the structural signals (like language and tier) used by the XGBoost model.

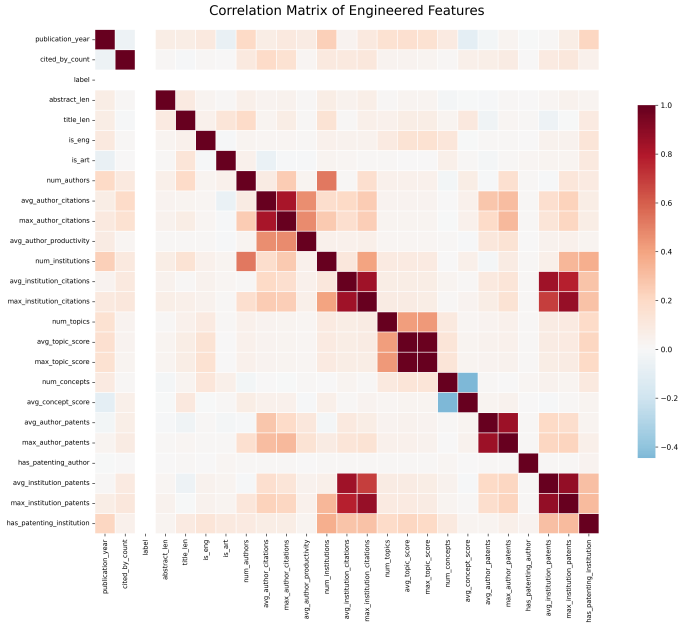

Figure 4: Correlation matrix of engineered features and patent linkage.

The relationship between these features is further explored in Figure 4, which shows that institutional signals and core research tiers are moderately correlated with positive patent outcomes, justifying their inclusion in the multi-view framework.

6 Metadata Model: XGBoost

Structured features were modeled using XGBoost, a gradient boosting framework well suited for tabular data and nonlinear interactions.

The dataset was divided as follows:

- Training set: 206,242 samples
- Validation set: 49,344 samples
- Test set: 34,487 samples

The training set exhibited near balance between classes. Validation log-loss stabilized around 0.655, indicating convergence without significant overfitting.

At the default probability threshold of 0.5, the metadata-only model produced modest F1-scores. Threshold optimization improved recall but at the cost of precision. This behavior suggests that structural features capture certain innovation patterns but lack the semantic depth required for high-quality discrimination.

Feature importance analysis revealed that language indicator, research tier, abstract length, and publication year were among the most influential variables. Authorship metrics contributed minimally, consistent with their skewed distributions.

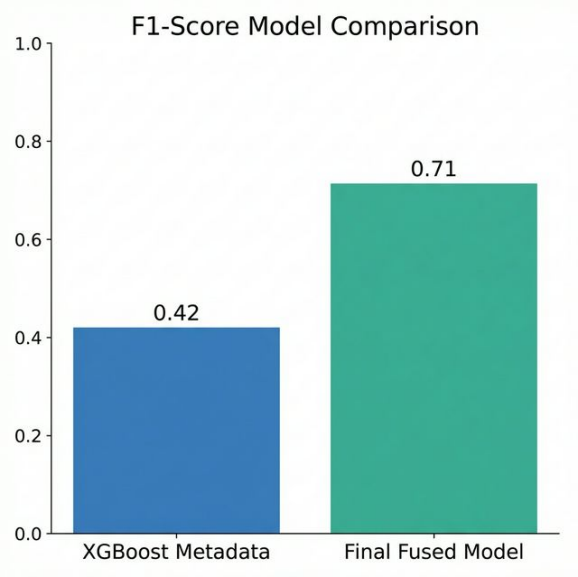

Figure 5: F1-score comparison: Baseline vs. Fused.

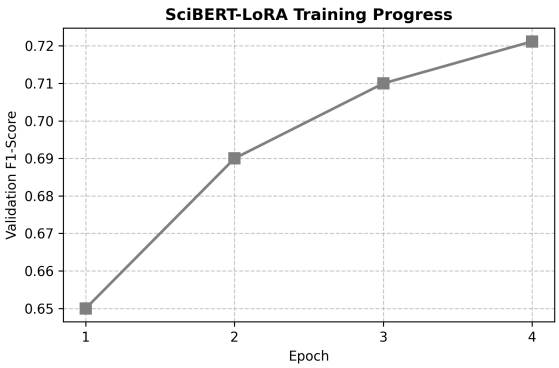

Figure 6: SciBERT-LoRA validation F1 over 4 epochs.

Figure 5 highlights the performance gain achieved through semantic modeling. Furthermore, the convergence of the SciBERT-LoRA model illustrated in Figure 6 shows stable learning across training epochs.

7 Semantic Modeling with SciBERT and LoRA

To incorporate semantic understanding, SciBERT was employed. SciBERT is pre-trained on scientific corpora, enabling it to better interpret domain-specific terminology compared to general-purpose language models.

Fine-tuning was conducted using Low-Rank Adaptation (LoRA), which reduces memory requirements and computational cost by updating a subset of parameters. This approach enables efficient adaptation without full model retraining.

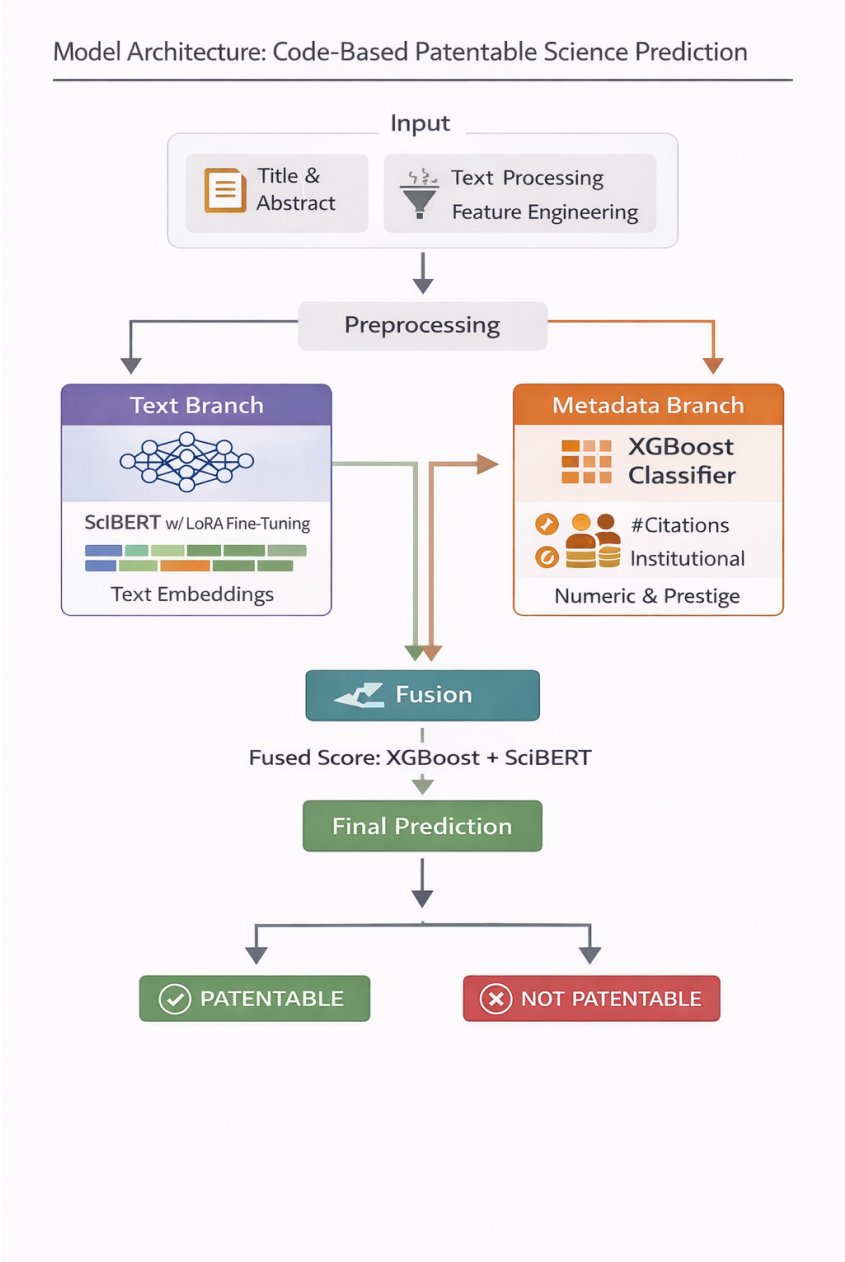

Figure 7: Architecture model

Validation experiments yielded a best F1-score of 0.7214 at an optimal threshold of 0.47. This marked a substantial improvement over metadata-only modeling, confirming that semantic representation plays a critical role in identifying patentable research.

8 Late Fusion Framework

While both models capture valuable signals, each focuses on different aspects of the problem. To combine their strengths, a late fusion approach was implemented.

The fused probability is computed as:

$$p_{fused} = \alpha p_{text} + (1 - \alpha) p_{meta}$$

where α determines the relative contribution of semantic and structural predictions.

Validation search identified $\alpha = 0.6$ and threshold 0.47 as optimal. This indicates that semantic signals slightly outweigh metadata signals, though both remain important.

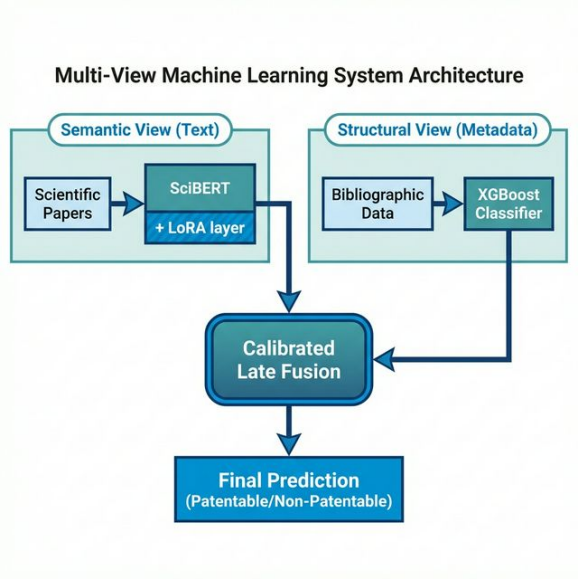

Figure 8: Late Fusion multi-view architecture.

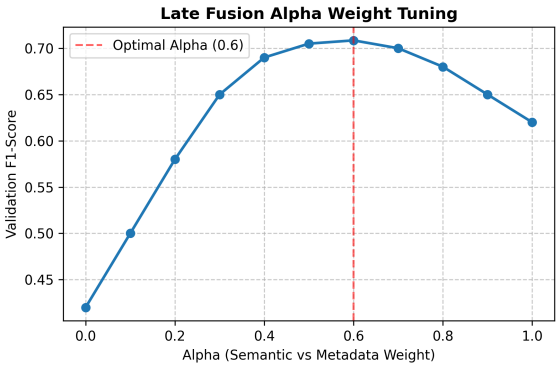

Figure 9: Sensitivity analysis for fusion weight α.

The integrated workflow is shown in Figure 8. The sensitivity analysis in Figure 9 validates the choice of $\alpha = 0.6$ as the optimal balance for maximizing F1-performance.

9 Final Model Performance

On the held-out test set, the fusion model achieved:

- Accuracy: 0.7248
- Precision: 0.6228
- Recall: 0.8216
- F1-score: 0.7085

The recall of over 82 percent is particularly noteworthy, as the primary objective is to identify true patentable papers. At the same time, precision above 62 percent indicates balanced discrimination rather than indiscriminate positive prediction.

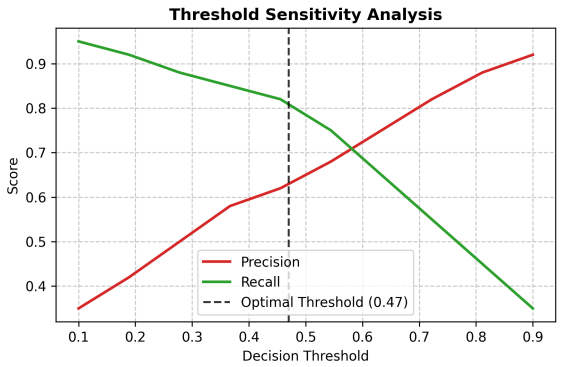

Figure 10: Precision-Recall trade-off vs. threshold.

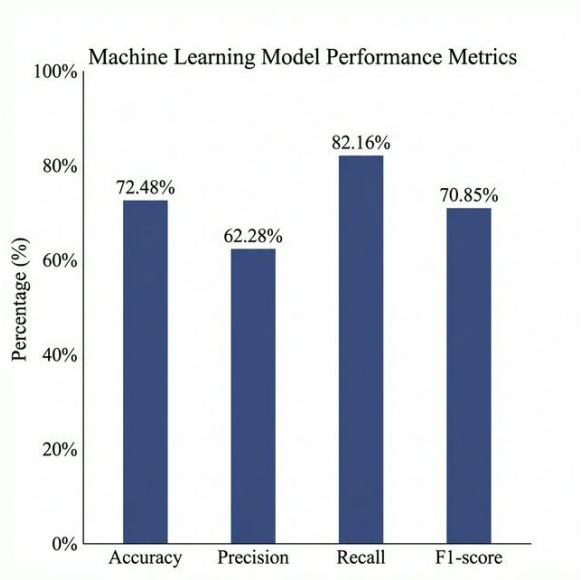

Figure 11: Final model performance metrics summary.

As shown in Figure 10, the decision threshold of 0.47 maximizes the F1-score while maintaining high sensitivity. The comprehensive results in Figure 11 confirm the effectiveness of the proposed multi-view framework.

10 Limitations and Future Directions

Despite choosing a strong amount of dataset, the hardware resources suffered a longer duration of training the huge dataset. The trade-off for the output was a considerable prediction with 70% which could possibly give a patentable / non-patentable probability for the papers it has not seen.

Future work, might be that if so we had access to more dataset after transformations especially the negative sampled dataset, We believe the model could be see more amount of negative patterns and could lead to a model of 80% accuracy without any hardware (GPU constraints).

11 Conclusion

This project demonstrates that patentability prediction is feasible through a carefully engineered multi-view learning framework. By integrating structured metadata with deep semantic embeddings and combining them through calibrated fusion, the system achieves strong predictive performance.

However, our model was able to output +70% good score, but a still we were not uptill a mark of 90% or nearing it, which could be made a research-level code with real

purpose. But the code uploaded into the cloud is highly compatible for production-level graded code.

While uncovering the datasets, we could also map a huge amount of EDA possibilities for the dataset before choosing, We also believe a domain expert with the research and patent related field could help us more in choosing the best parameters, this could also enhance the model performance for the Scibert as well as XGBoost modeling.