國立陽明交通大學

管理學院科技管理學程

碩士論文

Degree Program of Management of Technology

National Yang Ming Chiao Tung University

Master Thesis

融合結構化與語意特徵之機器學習模型於半導體專利存續預測

Integrating Structured and Semantic Features for Semiconductor Patent Survival Prediction Based on Machine Learning Modeling

研究生：陳冠中（Chen, Kuan-Chung）

指導教授：李昕潔（Lee, Shin-Jye）

中華民國一一四年七月

July 2025

融合結構化與語意特徵之機器學習模型於半導體專利存續預測

Integrating Structured and Semantic Features for Semiconductor Patent Survival Prediction Based on Machine Learning Modeling

研 究 生：陳冠中　　Student：Kuan-Chung Chen
指導教授：李昕潔 博士　　Advisor：Dr. Shin-Jye Lee

國立陽明交通大學
管理學院科技管理學程

碩士論文

A Thesis
Submitted to Graduate Institute of Management of Technology
College of Management
National Yang Ming Chiao Tung University
in Partial Fulfillment of the Requirements
for the Degree Program of
Master of Business Administration
in
Management of Technology

July 2025

Taiwan, Republic of China

中華民國　一一四年七月

誌謝

本論文之完成，承蒙多方協助與鼓勵，謹在此致上誠摯謝意。

首先，衷心感謝指導教授李昕潔老師於本研究過程中給予我莫大支持與啟發。李老師在學術上的嚴謹態度與研究上的前瞻視野，不僅引導我建立清晰的研究架構，更讓我在探索與反思中持續精進。老師耐心細緻的指導與實務導向的建議，是本論文得以順利完成的關鍵。

同時，感謝國立陽明交通大學科技管理研究所的師長與同儕，課堂上的討論與課後的交流，豐富了我對資料分析與科技管理議題的理解，也激發了本研究的發想與實踐動力。

最後，感謝我的家人和朋友們長期以來的支持與陪伴，讓我能無後顧之憂地全心投入研究。若本研究有絲毫成就，皆因有你們在背後默默守護與鼓勵。

謹以此文，向所有曾經給予我啟發與協助的人致上最深的謝意。

陳冠中 謹誌

中華民國一一四年七月

於新竹陽明交通大學

融合結構化與語意特徵之機器學習模型於半導體專利存續預測

研究生：陳冠中　　指導教授：李昕潔 博士

國立陽明交通大學管理學院科技管理學程

摘要

隨著專利申請量與技術複雜度日益提升，如何在授權初期即預測專利之長期價值，已成為技術管理與政策規劃的重要課題。本研究聚焦半導體領域（IPC H01L），以「是否維持至最大法定存續期」作為可觀察之價值代理指標，建構一套結合結構化變數與語意嵌入的機器學習預測框架。資料涵蓋 2000 至 2022 年間共 159,945 件美國發明專利，並根據專利維護費繳納紀錄進行標記。模型訓練與預測僅使用授權當下可取得之特徵，包含 34 項結構化變數與由預訓練語言模型產生之專利摘要語意嵌入。實驗採用多種監督式分類器進行比較，並採用堆疊式集成模型以探索進一步的效能提升。結果顯示，結構與語意特徵融合可明顯提升預測表現，異質資訊整合對於捕捉潛在價值訊號具顯著貢獻。此外，語意模型之選擇亦會影響預測結果，與預測任務語境契合度較高者表現更為優異。同時，與最佳單一模型相比，堆疊式集成模型在原始訓練集下未見明顯優勢，但在擴增訓練資料後則呈現一致的效能提升。本研究驗證以授權當下資料進行早期專利價值預測的可行性，提供一套兼具解釋性與應用性的分析架構，對專利管理、投資決策與技術密集產業的資源分配具有實務價值。

關鍵詞：半導體專利存續預測、異質特徵融合、文本分析、機器學習

Integrating Structured and Semantic Features for Semiconductor Patent Survival Prediction Based on Machine Learning Modeling

Student: Chen, Kuan-Chung Advisor: Dr. Lee, Shin-Jye

Degree Program of Management of Technology
College of Management
National Yang Ming Chiao Tung University

Abstract

As patent filings grow in volume and complexity, predicting their long-term value at the grant stage has become a critical issue in technology management and policy planning. This study focuses on the semiconductor domain (IPC H01L) and examines whether a patent is maintained through its full statutory term as a proxy for long-term value. A machine learning framework integrating structured variables and semantic embeddings is developed for this task. The dataset covers 159,945 U.S. utility patents granted between 2000 and 2022, labeled from maintenance fee payment records. Training and prediction rely solely on grant-time features, including 34 structured variables and semantic embeddings of patent abstracts generated by pretrained language models. Multiple supervised classifiers are compared, and a stacking ensemble is implemented to explore performance gains. Results show that combining structured and semantic features substantially improves predictive performance, underscoring the contribution of heterogeneous information integration to capturing latent value signals. The choice of semantic embedding model also affects performance, with those better aligned to the task context achieving superior results. Compared with the best single model, the stacking ensemble showed no clear advantage on the original training set but delivered consistent improvements when trained on the expanded dataset. This study demonstrates the feasibility of early-stage patent value prediction using grant-time data and provides an interpretable, practically applicable framework to support patent management, investment decisions, and resource allocation in technology-intensive sectors.

Keywords: Semiconductor patent survival prediction, Heterogeneous feature fusion, Text mining, Machine learning

Contents

List of Figures

List of Tables

1. Introduction

This chapter presents the background, motivation, and objectives of the study. It first outlines the challenges of early-stage patent value prediction, then explains the rationale for adopting machine learning and semantic features, and introduces the domain focus and aims of the proposed framework.

1.1 Research Background

Patents provide legal and strategic mechanisms for protecting technological innovation, particularly in sectors characterized by rapid development and intensive R&D activity (Ernst, 2003). They serve not only to secure legal rights but also to deter competitors and facilitate licensing (Hall et al., 2014). In fields such as semiconductors, patents are used not only to safeguard intellectual property but also to guide investment and manage competitive dynamics, reflecting their strategic function in technology management (Ernst, 2003). As patent filings continue to increase in both volume and complexity, there is a growing need for systematic approaches to assess the potential value of individual patents at an early stage, when key strategic decisions are made (Reitzig, 2004).

Traditional indicators of patent value—such as forward citations, litigation outcomes, and licensing activity—have long served as proxies for a patent's technological or commercial significance (Jaffe & Trajtenberg, 2002). However, their applicability to early-stage prediction is limited due to their delayed availability and domain-specific variability (Hall et al., 2005). To address these limitations, recent studies have explored patent survival as a more stable and assignee-driven alternative (Choi et al., 2020; Liu et al., 2024). Although also realized over time, survival reflects internal valuation decisions through structured maintenance fee payments, offering a more consistent and operationalizable signal of perceived utility (Bessen, 2008).

In recent years, advances in machine learning have significantly enhanced the feasibility

of data-driven approaches to patent value prediction. Researchers have increasingly leveraged grant-stage data—such as bibliographic metadata and other early indicators—to support earlier and more scalable evaluations (Choi et al., 2020; Deng & Zhang, 2025). As a result, this shift reflects a broader movement from rule-based systems to AI-driven models across various patent analysis tasks, including valuation, classification, and technology forecasting (Shomee et al., 2024). As methodological approaches continue to evolve, many research gaps remain, as further elaborated in Section 1.2. Collectively, these issues underscore the need for more targeted research, thereby motivating the present study.

1.2 Research Motivation

With the accelerating pace and growing intricacy of patent activity, organizations face mounting challenges in identifying high-potential inventions and allocating resources effectively (Reitzig, 2004). In this context, the ability to predict a patent's long-term value at an early stage—ideally around the time of grant—has become increasingly important for firms managing intellectual property portfolios, investors seeking to evaluate intangible assets, and policymakers aiming to support innovation (Bessen, 2008). Early-stage prediction can help prioritize patents for commercialization, inform licensing strategies, and mitigate the risks of overinvestment in low-value inventions.

With the rise of machine learning, researchers have increasingly explored its potential for patent value prediction (Shomee et al., 2024). However, several methodological and contextual gaps remain. First, most studies adopt a cross-domain perspective, with limited focus on specific high-value technology sectors (Chung & Sohn, 2020). In particular, few have examined the semiconductor domain (IPC H01L), which is not only strategically important but also rapidly expanding, with intensifying innovation and patenting activity (Hall & Ziedonis, 2001; Henry, 2025).

Second, although some researchers have begun to use patent survival as a proxy for commercial value (Hwang et al., 2021; van Zeebroeck, 2011), such applications remain relatively rare compared to the widespread use of forward citations. Third, while structured metadata and semantic representations are both commonly used, studies that systematically integrate and evaluate these feature types remain uncommon, limiting the potential to capture complementary value signals (Deng & Zhang, 2025; Lin et al., 2018).

Finally, while prior studies have employed a range of models—including deep learning architectures such as multi-layer perceptrons (MLPs) and advanced tree-based methods like XGBoost—most adopt single-model approaches. Stacking, a meta-learning technique that combines heterogeneous base learners, remains rarely applied in this context. These observations collectively point to the need for further exploration of more flexible and integrated modeling strategies.

1.3 Research Objectives and Questions

This study aims to develop a machine learning-based framework for early-stage patent value prediction by leveraging both structured metadata and semantic representations to assess whether a granted patent is likely to be maintained until its full statutory expiration. Accordingly, the specific research objectives are as follows:

1. To construct a domain-specific prediction model focused on semiconductor patents (IPC H01L), a strategically important and rapidly evolving technology sector. By isolating this domain, the model aims to capture value-related patterns that may be obscured in cross-domain analyses.
2. To adopt patent survival—measured by whether a patent is maintained through its full statutory term—as a broadly applicable proxy for commercial value, offering a stable outcome variable suitable for early-stage prediction in practical settings.

3. To integrate grant-time-accessible structured metadata with semantic embeddings derived from patent abstracts, enabling a richer and more complementary representation of patent value signals.
4. To evaluate the predictive performance of various machine learning models—including a stacking ensemble of heterogeneous base learners—and to quantify the added value of semantic features through comparative experiments.

To further clarify the research focus and guide the experimental design, the following research questions are formulated:

1. Can the long-term maintenance status of semiconductor patents be accurately predicted using only grant-time-accessible features?
2. Do semantic embeddings derived from patent abstracts significantly enhance predictive performance when combined with structured metadata?
3. How do different semantic embedding models (e.g., BERT for Patents, SPECTER, MiniLM) compare in the context of early-stage patent value prediction?
4. Can a stacking ensemble model deliver additional performance gains over individual classifiers in predicting patent survival?

This study is anticipated to yield several contributions. Methodologically, it advances early-stage patent value prediction by integrating heterogeneous features—structured metadata and semantic embeddings—within machine learning frameworks, and by systematically comparing their predictive roles. Practically, it provides an interpretable and grant-time-applicable framework that can assist firms, investors, and policymakers in making informed intellectual property management and resource allocation decisions in technology-intensive sectors, particularly in the semiconductor domain, where innovation dynamics are most pronounced.

2. Related Work

This chapter reviews prior research relevant to patent value prediction, with a focus on methodological developments that inform the design of this study. As machine learning becomes increasingly integrated into patent analytics, understanding how value indicators, domain characteristics, feature types, and modeling strategies affect predictive performance is essential (Deng & Zhang, 2025; Shomee et al., 2024).

Section 2.1 provides an overview of the evolution of patent value prediction research. It outlines three foundational dimensions: the use of proxy indicators such as forward citations and patent survival, the importance of domain-specific modeling, and the design of predictive strategies that integrate structured and semantic features.

Section 2.2 introduces core machine learning concepts, emphasizing tree-based models and ensemble learning methods—particularly stacking—as robust solutions for handling high-dimensional and heterogeneous patent data. Section 2.3 traces the progression of text mining approaches, from early statistical models to modern contextual embeddings, and highlights the growing role of semantic representations in capturing patent content.

Section 2.4 discusses heterogeneous feature fusion techniques, classifying them into early, late, and intermediate strategies, and demonstrating how they enable models to jointly leverage structured indicators and textual cues. Finally, Section 2.5 summarizes these developments and identifies three key directions—survival-based proxies, ensemble modeling, and heterogeneous feature integration—that shape the methodological foundation of this study.

Together, these strands of literature motivate the predictive framework presented in Chapter 3, which combines bibliographic metadata and abstract embeddings to support early-stage value prediction for patents in the semiconductor domain.

2.1 Patent Value Prediction

This section reviews key strategies for predicting patent value, with a focus on proxy indicators, domain-specific modeling, and predictive design. Given the absence of a direct measure, prior studies have adopted proxies such as forward citations and patent survival to approximate long-term value (Trajtenberg, 1990; van Zeebroeck, 2011). However, citation-based metrics often emerge unpredictably and vary across domains (Alcácer & Gittelman, 2006; Criscuolo & Verspagen, 2008; Squicciarini et al., 2013), making them less suitable for consistent early-stage labeling. Additionally, patent value is highly domain-sensitive, underscoring the need for tailored models that reflect industry-specific dynamics (Belderbos & Mohnen, 2020; Harhoff et al., 2003; Lanjouw & Schankerman, 2004). With advancements in machine learning, recent studies have increasingly adopted tree-based and ensemble models, as well as hybrid features that integrate structured data with semantic embeddings (Deng & Zhang, 2025; Eom et al., 2021; Liu et al., 2024). These developments highlight the growing sophistication and effectiveness of predictive approaches in the patent domain.

2.1.1 Proxy Indicators of Patent Value

Assessing patent value is inherently difficult due to the lack of a universally accepted metric (Hsieh, 2013; van Zeebroeck, 2011). To approximate a patent's technological or commercial significance, researchers have relied on observable proxy indicators such as forward citations, patent renewals or survival duration, licensing or sale activities, and litigation occurrences (Gambardella et al., 2007; Lanjouw & Schankerman, 2001; Schankerman & Pakes, 1986; Trajtenberg, 1990). Each of these proxies captures a different dimension of value and operates on a distinct temporal scale within the patent lifecycle.

Among them, forward citations have received the most attention, often interpreted as signals of technological relevance and knowledge diffusion (Jaffe & Trajtenberg, 2002).

However, their utility for early-stage prediction is limited. In emerging or niche technology domains, citations tend to accumulate slowly or remain sparse due to narrow applicability or delayed market uptake (Squicciarini et al., 2013). Furthermore, citation patterns may be shaped by assignee strategies, examiner behavior, and institutional factors, introducing noise unrelated to intrinsic patent value (Alcácer & Gittelman, 2006; Criscuolo & Verspagen, 2008).

To address these limitations, recent studies have increasingly adopted patent renewal or survival data as an alternative proxy for long-term value, as it reflects an assignee's internal judgment about whether a patent merits continued investment (Danish et al., 2022; Hwang et al., 2021). These decisions, typically made at fixed maintenance intervals, offer a structured and predictable signal of perceived value. As such, survival outcomes—though realized ex post—can serve as long-term but well-defined prediction targets based on grant-time-accessible features (van Zeebroeck, 2011). Compared to proxies that rely on external validation, survival provides a decision-based, institutionally grounded measure that aligns more closely with real-world value assessment processes (Hwang et al., 2021).

2.1.2 Domain-Specific Perspectives

Patent value prediction is intrinsically domain-sensitive, as innovation cycles, protection mechanisms, and value indicators differ markedly across technological fields (Harhoff et al., 2003; Kim & Magee, 2017). Prior empirical research has demonstrated that patent quality and research productivity vary substantially between industries, reflecting fundamental structural differences in innovation processes and value appropriation strategies (Lanjouw & Schankerman, 2004). Furthermore, the heterogeneity of R&D spillover effects across sectors and national contexts further reinforces the need to account for domain-specific dynamics in predictive modeling (Belderbos & Mohnen, 2020).

A compelling example of domain-specific modeling is provided by He et al. (2025), who

developed the BioTexVal system for predicting the value of biomedical textile patents. By aligning the training data and modeling approach with domain-specific characteristics, the system achieved an accuracy of 88.38%, clearly outperforming baseline methods (He et al., 2025). While the study did not benchmark against cross-domain models, its performance nonetheless underscores the benefits of aligning modeling strategies with the structural patterns and technological context of a given domain. Despite these advantages, few predictive modeling studies have explicitly focused on the semiconductor domain (IPC subclass H01L), which remains underexplored relative to its strategic importance (Chung & Sohn, 2020; Hall & Ziedonis, 2001).

2.1.3 Predictive Strategy Design

As machine learning algorithms continue to evolve, the strategies for predicting patent value have likewise grown more sophisticated (Shomee et al., 2024). Recent studies highlight key methodological choices—particularly in model selection and feature design—that can influence predictive performance. As an early contribution, Choi et al. (2020) employed a feedforward neural network (FFNN) to predict patent lifespan based solely on structured features available at the time of grant. A simplified schematic of a feedforward neural network is presented in Figure 2.1 to illustrate the model architecture employed in the study. While the study demonstrated the feasibility of early-stage prediction, it also suggested that more advanced machine learning approaches—such as tree-based ensemble methods—could be explored to further enhance predictive outcomes. Building on this work, Liu et al. (2024) extended the same prediction task by replacing the FFNN with a gradient boosting model—LightGBM—augmented by a customized focal loss function. Using the same dataset design as Choi et al. (2020), their model achieved a test accuracy of 0.7447 and an AUC-ROC of 0.8185, outperforming all baseline methods. This empirical contrast supports the view that advanced

machine learning approaches, particularly tree-based ensemble models, are better suited for capturing the complex relationships inherent in structured patent data.

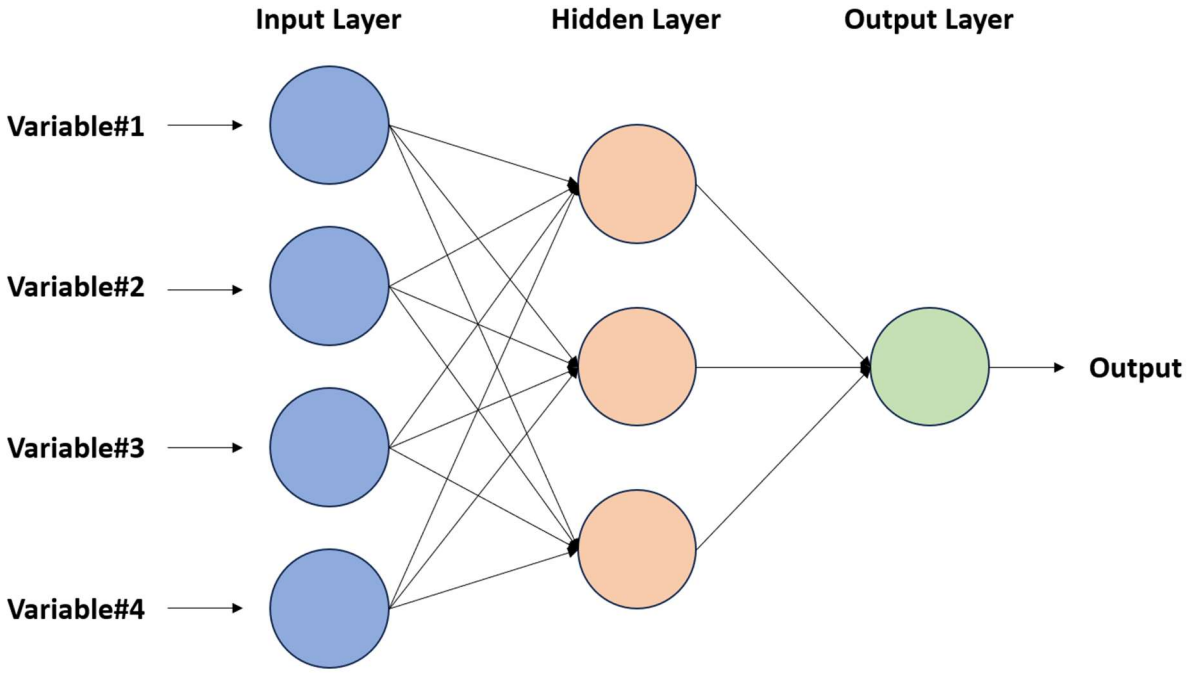

Figure 2.1. Architecture of a Simple Feedforward Neural Network

Beyond individual model selection, ensemble learning strategies such as stacking have been adopted to further enhance predictive performance (Zhou, 2012). Stacking integrates multiple base learners through a meta-learner, enabling the model to capture complementary decision patterns. In the context of patent value prediction, Eom et al. (2021) applied a stacking ensemble approach to predict the marketable value of U.S. patents in the electricity sector. Through a comprehensive comparison with linear regression, random forest, MLP, and CNN models, their stacking ensemble consistently achieved the lowest prediction errors across MAE, MSE, and RMSE. These results highlight the robustness and predictive advantage of ensemble learning strategies in heterogeneous patent valuation scenarios.

In addition to model architecture, the composition of input features exerts a critical

influence on predictive performance (Domingos, 2012). Advances in natural language processing have facilitated the integration of semantic representations into patent valuation tasks, enabling the capture of information beyond what is available in structured metadata (Cohan et al., 2020). A notable example is provided by Deng and Zhang (2025), who developed the High-value Patent Multi-Feature Fusion Method (HMFM) for identifying high-value patents using data from the Chinese patent system. Their approach integrated structured indicators with semantic embeddings extracted from patent abstracts via a BERT-DPCNN framework. When implemented with an XGBoost classifier, the hybrid model yielded a performance improvement, with accuracy increasing from 84.1% to 87.2%. These results highlight the efficacy of hybrid feature design in improving predictive performance in patent valuation tasks.

2.2 Machine Learning

Given the growing reliance on predictive modeling in patent value assessment (Shomee et al., 2024), machine learning (ML) offers an effective approach for capturing complex, nonlinear relationships among high-dimensional patent features. Building on the strategic and methodological considerations introduced in Section 2.1, this section outlines the foundational concepts of ML most relevant to this task. It focuses on supervised learning and the specific advantages of decision tree-based models, which naturally accommodate mixed data types, handle missing values effectively, and require minimal preprocessing when applied to structured patent metadata. In particular, the discussion emphasizes ensemble learning strategies—including bagging, boosting, and stacking—which enhance model robustness and accuracy by integrating diverse learners and mitigating the limitations of individual models. These techniques have demonstrated strong performance in predictive analytics and are especially effective in heterogeneous patent data environments.

2.2.1 Basic Concepts of Machine Learning

Machine learning (ML) refers to a class of computational methods that enable systems to automatically learn patterns and make predictions from data, without being explicitly programmed with task-specific rules (Mitchell, 1997). In contrast to traditional rule-based systems, which rely on handcrafted heuristics, machine learning algorithms can generalize from observed examples and adapt to new, unseen data (Mitchell, 1997). This data-driven paradigm has proven particularly useful in complex domains where feature relationships are nonlinear, high-dimensional, or poorly understood a priori.

Supervised learning, the most relevant paradigm for patent value prediction, involves training a model on labeled data to predict an outcome variable (Domingos, 2012). Formally, given a training set of input-output pairs (x_i, y_i), the model seeks a function f(x) that minimizes prediction error on new data. The quality of this approximation depends on several factors, including the representativeness of training data, the expressiveness of the model class, and the balance between underfitting and overfitting (Domingos, 2012).

Building on these principles, a wide range of supervised learning algorithms can be employed, each offering different trade-offs between expressiveness, interpretability, and data requirements (Goodfellow et al., 2016). Common model types include linear models (e.g., logistic regression), which are efficient and transparent but limited in capturing nonlinearity (Goodfellow et al., 2016); decision trees, which segment the feature space into interpretable rule-based regions (Hastie et al., 2009); and neural networks, which are highly expressive but require large amounts of data and tend to sacrifice interpretability (Goodfellow et al., 2016). These distinctions help inform model selection in predictive frameworks that must balance accuracy, efficiency, and real-world applicability.

As patent value prediction increasingly relies on data-driven models, particularly in early-stage settings, understanding these foundational principles is essential for evaluating design

choices and ensuring robust, interpretable outcomes (Shomee et al., 2024).

2.2.2 Tree-Based Models

Tree-based models constitute a class of non-parametric supervised learning algorithms that use a hierarchical structure of decision rules to partition the input space (Breiman et al., 1984). At each node of a decision tree, the algorithm selects a feature and a threshold to split the data in a way that maximizes a particular objective—such as information gain or reduction in impurity—ultimately producing a tree-like structure of conditions that lead to different output predictions (Breiman et al., 1984).

One of the main advantages of decision trees is their interpretability: the resulting model can be easily visualized and understood in terms of if–then rules (Hastie et al., 2009). In addition, decision trees can naturally handle both numerical and categorical variables, are robust to feature scaling, and are capable of capturing nonlinear relationships between features and the target variable. However, they are also prone to overfitting, especially when the tree grows too deep, resulting in high variance and poor generalization on unseen data (Murphy, 2012).

To overcome the limitations of individual decision trees, ensemble learning techniques have emerged as powerful alternatives. Among them, bagging (bootstrap aggregating) is designed to reduce variance by training multiple models independently on different bootstrapped subsets of the data. A well-known implementation of bagging is the Random Forest algorithm, which builds an ensemble of decision trees and averages their predictions to improve generalization (Breiman, 2001).

In contrast, boosting methods—such as XGBoost and LightGBM—build decision trees sequentially, with each new model trained to correct the residual errors of the previous ones (Chen & Guestrin, 2016; Ke et al., 2017). These methods are particularly effective for structured tabular data and have demonstrated strong predictive performance across a wide range of real-

world tasks. Figure 2.2 provides a visual summary of the contrasting workflows between bagging and boosting, highlighting how they differ in data sampling, training sequence, and model integration strategy.

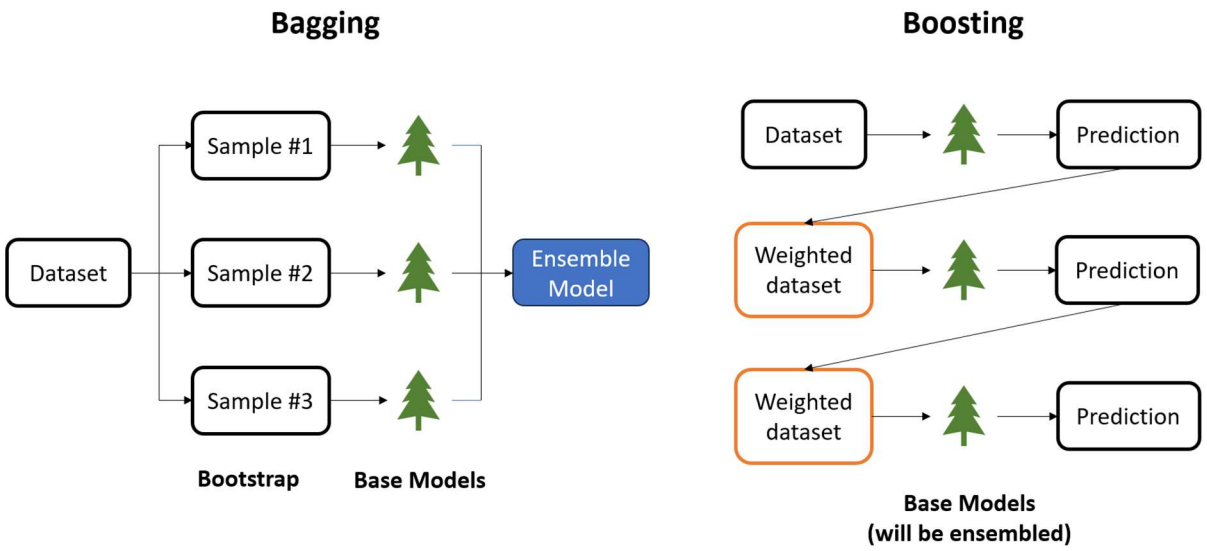

Figure 2.2. Conceptual Illustration of Bagging and Boosting Workflows

Numerous empirical studies have validated the effectiveness of tree-based models in structured prediction tasks. For instance, gradient boosting decision trees have demonstrated superior performance over neural networks and support vector machines in applications such as credit scoring (Lessmann et al., 2015), fraud detection (Correa Bahnsen et al., 2016), and healthcare analytics (Rajkomar et al., 2018). These results underscore the practical strength of tree-based methods in capturing complex decision boundaries across structured input spaces. In particular, their strong performance, interpretability, and computational efficiency make them a foundational component for more advanced ensemble techniques such as stacking (Zhou, 2012).

2.2.3 Ensemble Learning: Stacking

Stacking, or stacked generalization, is a widely used ensemble learning technique that combines multiple base models—potentially of different algorithmic types—by training a meta-model to aggregate their predictions (Wolpert, 1992). Unlike bagging and boosting, which typically rely on homogeneous ensembles of weak learners such as decision trees (Breiman, 1996; Dietterich, 2000), stacking allows for the integration of heterogeneous models that capture different patterns in the data (Ting & Witten, 1999). This flexibility makes stacking especially suitable for tasks where no single model consistently performs well across the entire input space (Zhou, 2012).

In a typical stacking framework, illustrated in Figure 2.3, several base learners are first trained independently. Their outputs—either predicted class probabilities or raw predictions—are then used as inputs to a meta-model, which learns an optimal combination strategy. Cross-validation is often employed during the training process to prevent overfitting and ensure that the meta-learner does not simply memorize the outputs of the base models (Ting & Witten, 1999; Zhou, 2012).

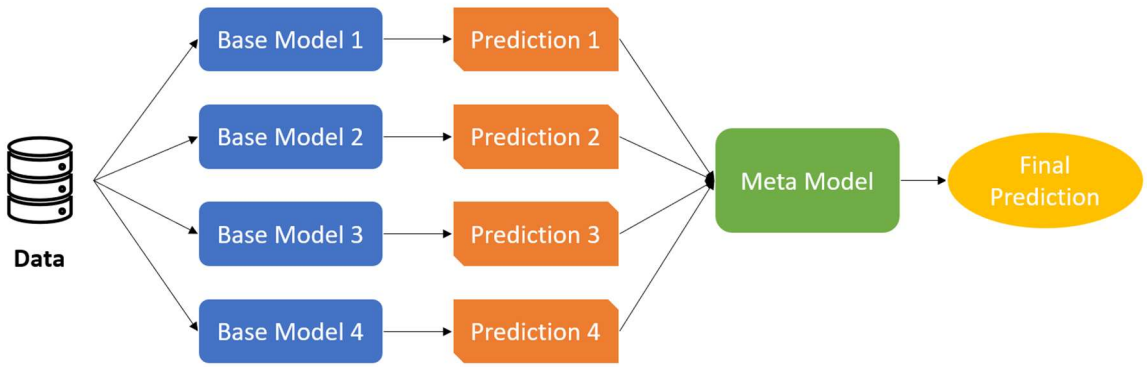

Figure 2.3. Architecture of a Stacking Ensemble

Stacking has demonstrated strong applicability in handling heterogeneous feature types, noisy data distributions, and complex decision boundaries (Zhou, 2012). It has been successfully applied across various domains, including recommendation systems, sequence-based classification with NLP techniques, and biomedical imaging prediction (Ren et al., 2022; Sill et al., 2009; Zhang et al., 2024). Its strength lies not only in enhancing predictive performance but also in integrating complementary learning mechanisms, such as combining tree-based models with linear classifiers or neural networks (Ting & Witten, 1999; Wolpert, 1992). These characteristics make stacking particularly well-suited for tasks involving mixed data types or multi-modal feature representations.

2.3 Text Mining

Text mining refers to a collection of computational techniques designed to extract meaningful information from unstructured textual data (Aggarwal & Zhai, 2012). Core tasks include text classification, clustering, topic modeling, and information retrieval. These methods support downstream machine learning tasks by transforming raw text into structured representations (Feldman & Sanger, 2006).

Early approaches to text representation primarily relied on statistical techniques such as the Bag-of-Words (BoW) and Term Frequency–Inverse Document Frequency (TF-IDF). These methods represent documents as high-dimensional vectors based on word occurrences, but they disregard word order and semantic context (Manning et al., 2008; Salton et al., 1975). To capture latent thematic structures, topic models such as Latent Dirichlet Allocation (LDA) were introduced, enabling each document to be viewed as a mixture of topics (Blei et al., 2003). However, topic models still struggle to capture contextual nuances or the dynamic meanings of words across different usage scenarios (Camacho-Collados & Pilehvar, 2018).

A significant breakthrough in text mining came with the development of word embeddings

(Mikolov et al., 2013; Pennington et al., 2014). Models like Word2Vec and GloVe encode words as dense vectors in a continuous space, such that semantically similar words are placed closer together (Mikolov et al., 2013; Pennington et al., 2014). Derived from local context windows, these embeddings capture distributional semantics via word co-occurrence patterns (Mikolov et al., 2013; Pennington et al., 2014). Nevertheless, these fixed representations fail to account for polysemy—where a single word may carry multiple meanings depending on context—and lack sentence-level understanding (Devlin et al., 2019; Peters et al., 2018).

The introduction of contextualized language models, particularly Bidirectional Encoder Representations from Transformers (BERT), represented a major advancement. BERT generates dynamic embeddings for each word token based on its surrounding context, enabling a more nuanced understanding of sentence structure and meaning. Its transformer-based architecture allows for bidirectional attention, capturing both preceding and following information in a text sequence. This has led to substantial improvements across a wide range of natural language understanding tasks (Devlin et al., 2019).

Beyond word-level embeddings, sentence and document-level representations have gained prominence. Models such as Sentence-BERT and SPECTER aggregate contextual embeddings into fixed-length vectors suitable for downstream applications like classification, similarity measurement, and regression. These semantic embeddings allow models to harness linguistic signals that are often inaccessible to structured features, providing richer representations of complex textual input (Cohan et al., 2020; Reimers & Gurevych, 2019). This process is illustrated in Figure 2.4, which shows how textual input is transformed into a semantic vector via a pretrained language model. Importantly, they can be integrated with machine learning pipelines in the form of feature vectors, facilitating hybrid modeling strategies that combine textual and non-textual information (Cohan et al., 2020; Reimers & Gurevych, 2019).

Figure 2.4. Conceptual Illustration of Semantic Embedding Generation

In summary, the evolution from statistical methods to semantic embedding models has fundamentally transformed text mining. Context-aware representations now offer more expressive and flexible inputs for machine learning, which has substantially enhanced the generalizability and performance of models dealing with textual data (Cambria & White, 2014).

2.4 Heterogeneous Feature Fusion

In modern predictive tasks, it is increasingly common to encounter inputs from heterogeneous sources, such as structured metadata, textual descriptions, or image embeddings, which differ in both format and semantic characteristics (Baltrušaitis et al., 2019). In response, heterogeneous feature fusion has emerged as a critical modeling strategy to leverage the complementary strengths of diverse data modalities. Within the context of machine learning, this typically involves integrating structured attributes (e.g., categorical or numerical variables) with unstructured representations (e.g., semantic embeddings derived from natural language) to construct more expressive and informative feature sets (Baltrušaitis et al., 2019; Ngiam et al., 2011).

According to Baltrušaitis et al. (2019), fusion strategies are typically classified into three main categories: early fusion, late fusion, and intermediate fusion. *Early fusion* refers to the concatenation or joint encoding of all feature types prior to model input, allowing a single learner to optimize over the combined representation (Baltrušaitis et al., 2019). This approach

is conceptually simple and preserves cross-modal interactions at the feature level, but may be sensitive to dimensionality mismatches and scale heterogeneity. *Late fusion*, by contrast, involves training separate models on each feature type and combining their predictions at the decision level, typically through averaging, voting, or a meta-learner (Baltrušaitis et al., 2019). This strategy is robust to heterogeneous feature spaces but may miss synergistic patterns available only at the input level. *Intermediate fusion* lies between these two extremes and seeks to jointly learn representations while maintaining some modality-specific processing (Baltrušaitis et al., 2019). This is often implemented using neural architectures with shared and modality-specific branches or attention mechanisms. A visual comparison of these three strategies is presented in Figure 2.5.

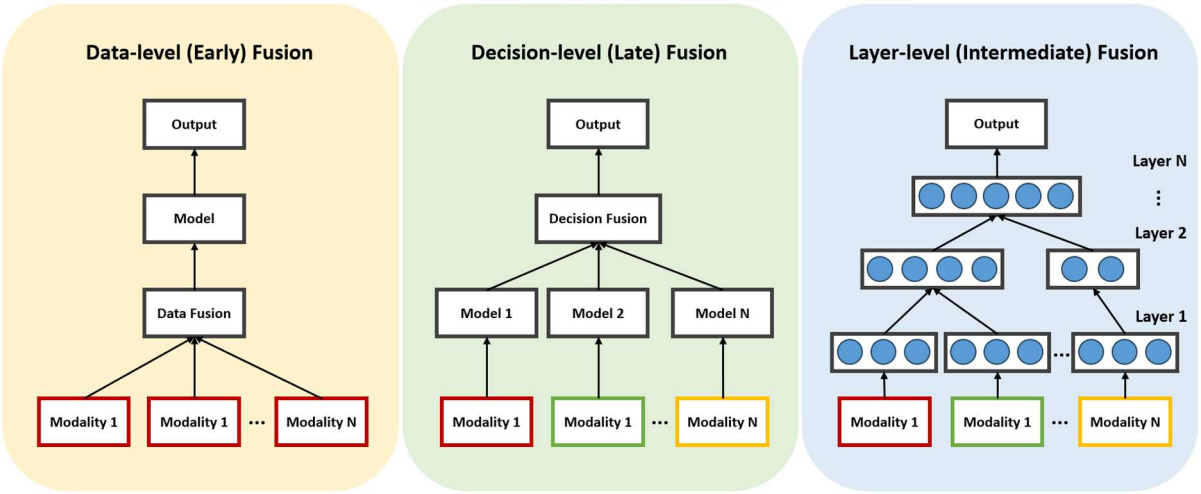

Figure 2.5. Schematic Illustration of Heterogeneous Feature Fusion Strategies

The fusion of structured and semantic features is particularly valuable in domains where structured indicators provide stable, high-precision signals, while semantic embeddings capture contextual richness and abstract correlations (Yin et al., 2020; Zhang et al., 2023). For instance, in biomedical risk prediction, structured patient records are often fused with clinical notes embeddings to improve diagnostic accuracy (Rajkomar et al., 2018). Similarly, financial default

models have shown gains when combining tabular transaction features with text-based risk assessments (Tavakoli et al., 2025). These examples underscore the generalizability of fusion strategies across diverse application domains.

Despite its benefits, heterogeneous fusion introduces modeling challenges. These include alignment mismatches between modalities, imbalance in representational capacity, and difficulties in interpreting fused models (Bao et al., 2023; Li & Tang, 2024). Moreover, when embeddings are derived from large pretrained models, issues of domain transferability and training stability may arise (Li et al., 2025). Careful feature engineering, normalization, and model calibration are often necessary to achieve optimal integration (Baltrušaitis et al., 2019).

In summary, heterogeneous feature fusion provides a conceptually grounded and practically relevant approach for capturing the multifaceted nature of patent data. By integrating structured indicators with semantic representations, it becomes possible to model both the formal and contextual dimensions of value-relevant information (He et al., 2025). Prior studies in fields such as biomedicine (Hemker et al., 2024) and finance (Wei et al., 2024) have demonstrated the benefits of this strategy, suggesting its potential applicability to patent valuation tasks. These considerations highlight the theoretical basis for incorporating feature fusion in predictive frameworks for early-stage value assessment.

2.5 Summary

This chapter reviewed the evolving methodologies and design considerations in patent value prediction, with a focus on integrating machine learning and textual features. Section 2.1 examined the foundations of patent valuation research, outlining three key dimensions: the choice of proxy indicators, the importance of domain-specific modeling, and the design of predictive strategies. While forward citations have traditionally served as value proxies, their limitations in early-stage prediction have prompted a shift toward alternatives such as patent

survival. Moreover, the heterogeneity of innovation dynamics across technology fields underscores the need for domain-aware models, particularly in underexplored areas like semiconductors. Studies discussed in Section 2.1.3 have also highlighted the importance of model and feature design, suggesting that hybrid approaches combining structured and semantic features can improve predictive performance.

Section 2.2 introduced core machine learning techniques, including tree-based models and ensemble strategies. Methods such as bagging, boosting, and stacking have demonstrated strong performance in handling high-dimensional and heterogeneous data, with stacking in particular offering a flexible framework for integrating diverse base learners. Section 2.3 traced the evolution of text mining—from statistical models to contextualized embeddings—emphasizing the growing utility of semantic representations in capturing nuanced textual signals.

Finally, Section 2.4 reviewed heterogeneous feature fusion strategies, categorizing them into early, late, and intermediate designs. These approaches allow models to jointly leverage the precision of structured data and the contextual richness of semantic embeddings.

Taken together, these insights from prior studies establish a conceptual foundation for developing machine learning frameworks that combine survival-based proxy labeling, ensemble modeling, and hybrid feature representations. Such strategies may be particularly valuable in domains like semiconductors (IPC H01L), where early-stage patent value prediction involves complex, heterogeneous signals. These methodological directions inform the experimental design introduced in the next chapter.

3. Research Methodology

This chapter presents the methodological framework designed to operationalize early-stage patent value prediction—a task motivated by the need for timely and informed decision-making at the point of grant. Building upon the literature reviewed in Chapter 2, the research design emphasizes practical applicability, reproducibility, and the integration of heterogeneous data sources available at the time of grant.

To achieve this goal, the study combines two types of features: structured metadata extracted from bibliographic and legal records, and semantic embeddings derived from patent abstracts using pretrained language models. These features are used to train and evaluate a series of supervised classification models, thereby enabling a comparative analysis of predictive strategies under varying feature configurations and model architectures.

This chapter is organized into five sections. Section 3.1 details the study's methodological approach and defines the early-stage patent value prediction task. Section 3.2 introduces the dataset and labeling strategy. Section 3.3 describes the feature construction process, including the engineering of structured variables and extraction of semantic embeddings. Section 3.4 outlines the predictive modeling approach, covering baseline classifiers, tree-based models, and a stacking ensemble. Section 3.5 defines the evaluation metrics used to assess classification accuracy and class-specific performance.

Together, these components form the empirical foundation for the experiments presented in Chapter 4.

3.1 Study Approach

This study tackles the challenge of early-stage patent value prediction by estimating whether a granted semiconductor patent will be maintained through its full statutory term. Patent survival is adopted as a proxy for long-term commercial value, offering an observable

and objective outcome variable. The focus on grant-time-accessible features ensures that predictions can be made under real-world constraints—where future citation, litigation, or licensing data are not yet available. This setting reflects practical decision-making scenarios in R&D management, portfolio evaluation, and technology transfer.

To operationalize this task, the study approach is structured into four key stages:

(1) Data collection and labeling: integrating structured patent metadata with maintenance fee event records (e.g., PatentsView and USPTO datasets) to build a unified dataset and assign CBP labels.

(2) Feature construction: constructing 34 structured features from bibliographic/legal metadata and generating semantic embeddings from patent abstracts using pretrained language models.

(3) Model training: applying supervised classifiers, including both single models and a stacking ensemble, under consistent experimental settings.

(4) Performance evaluation: assessing binary classification performance using Area Under the Receiver Operating Characteristic Curve (AUC-ROC), Accuracy, Precision, Recall, and F1-Score.

The framework evaluates the predictive performance of individual features and models, along with the incremental benefit of heterogeneous integration across modalities and architectures.

All modeling procedures are grounded in three core principles: early accessibility of features, consistency of experimental settings, and reproducibility of results. Only information available at the time of patent grant is used, and all models are trained on a fixed train-test split with uniform configurations to ensure fair comparison. Semantic features are derived from pretrained language models without task-specific fine-tuning, thereby preserving the linguistic priors inherent in publicly available patent corpora. This process is illustrated in Figure 3.1, which details the methodological pipeline.

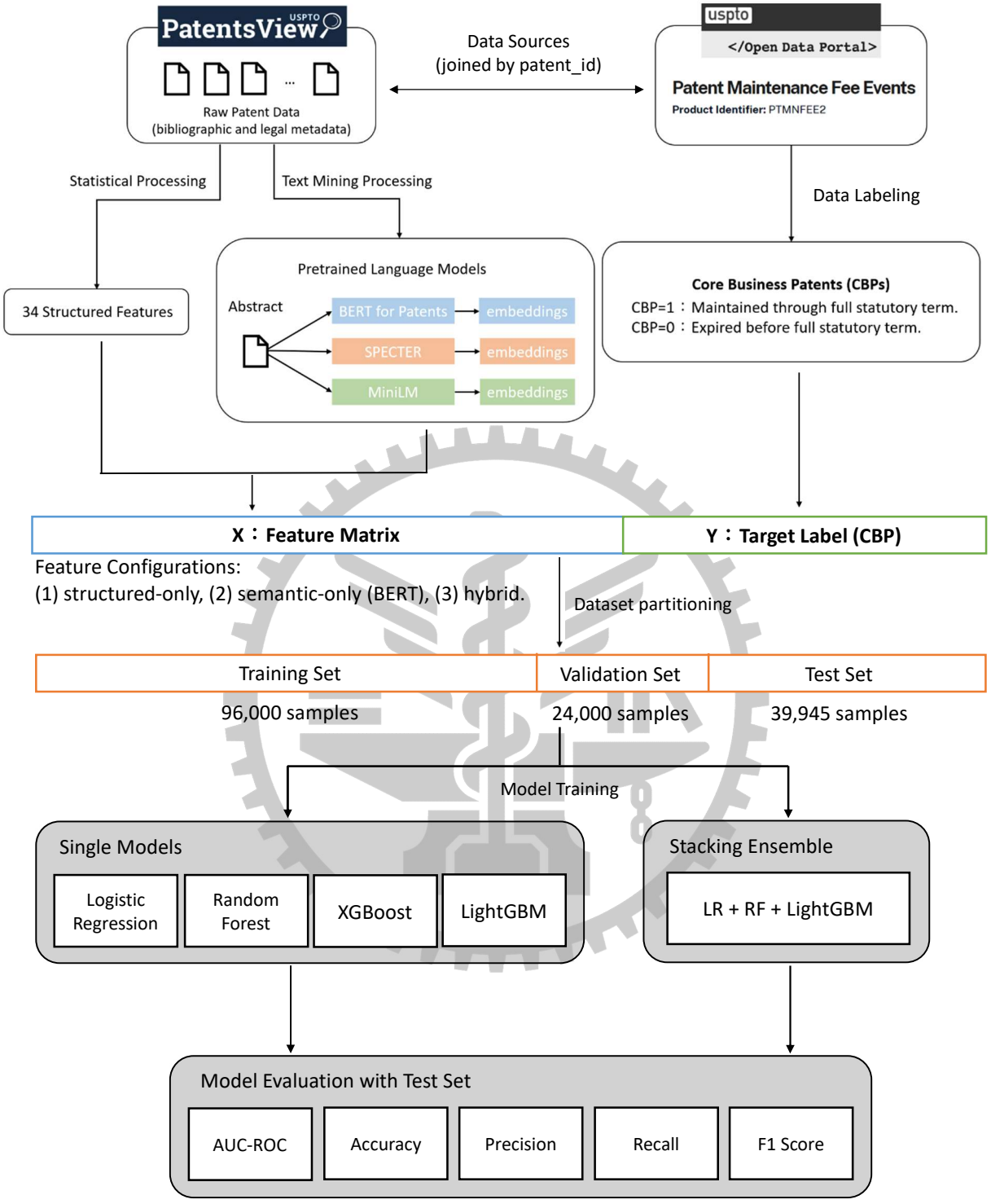

Figure 3.1. Illustration of the methodological pipeline

3.2 Data Sources and Labeling

This section introduces the sources and structure of the data used in this study, outlining the types of information collected, the integration of multiple datasets, and the filtering criteria applied to define the sample scope. It also describes the procedure for labeling the prediction target based on post-grant maintenance behavior, while ensuring that all feature variables remain observable at the time of patent grant.

3.2.1 Data Sources

PatentsView is an open-access data platform supported by the Office of the Chief Economist (OCE) of the United States Patent and Trademark Office (USPTO), in collaboration with academic and technical partners. It provides bulk downloads of structured metadata on granted U.S. patents, enabling large-scale empirical research on innovation and intellectual property. All data files used in this study were manually retrieved from the official download site (https://patentsview.org/download/data-download-tables). The platform offers a wide range of tab-delimited files (e.g., TSV) containing information on patent bibliographic details, inventors, assignees, classification codes (e.g., IPC), and citation relationships. For this study, the raw data tables were parsed and integrated using patent_id as the primary key to construct a unified dataset. All structured features were derived exclusively from information available at the time of patent grant. The sample was limited to utility patents granted between 2000 and 2022, and further restricted to patents classified under IPC subclass H01L, which pertains to semiconductor-related technologies.

To supplement the structured metadata, this study also incorporated patent maintenance data from the USPTO Maintenance Fee Events dataset, accessible via the official USPTO open data portal (https://data.uspto.gov/bulkdata/datasets/ptmnfee2). This dataset records historical fee payment events for utility patents granted since 1981 and is updated weekly. Each event

record includes basic administrative fields such as the patent number, event code, and event entry date. The maintenance fee data were linked to PatentsView records based on patent number to support the construction of outcome labels, as further described in Section 3.2.2.

3.2.2 Data Labeling

This study defines the target variable as a binary indicator of whether a granted patent has been maintained through its full statutory term, serving as a proxy for long-term commercial value. In the U.S. patent system, utility patents are subject to three scheduled maintenance fee payments, typically due at 3.5, 7.5, and 11.5 years after the grant date. Only patents that complete the third-stage (12th-year) payment and are not subsequently subject to a refund or expiration due to non-payment are considered to have been maintained until expiration.

Patent maintenance status was determined using the USPTO Maintenance Fee Events dataset. A patent was labeled as a Core Business Patent (CBP = 1) if its event history included at least one valid third-stage payment event (e.g., M1553, M2553, M3553, etc.), and no subsequent refund or expiration event occurred following that payment (e.g., R1553, R2553, R3553, EXP, etc.). Conversely, a patent was labeled as a non-CBP (CBP = 0) if its latest recorded event indicated either a refund of the third-stage payment or expiration due to non-payment, suggesting that the patent was not maintained through its full term. To ensure outcome validity, patents that did not meet either condition—such as those that had not yet reached the 12th-year payment window or lacked sufficient event history—were excluded from the dataset. A summary of key event codes used in this labeling process is provided in Table 3.1.

Table 3.1. Summary of Event Codes for USPTO Maintenance Fee Events

Event Code(s)	Description
M1553, M2553, M3553, F172, F175, F275, M172, M175, M185, M275, M285	Payment of the 12th-year maintenance fee (for large, small, and micro entities)
R1553, R2553, R3553, R172, R175, R185, R275, R285	Refund of a previously paid 12th-year maintenance fee
EXP.	Patent expired due to failure to pay maintenance fees
EXPX	Patent reinstated following confirmed maintenance fee payment
REM.	Maintenance fee reminder issued by USPTO

3.3 Feature Construction

This section introduces the construction of two types of input features: structured metadata features and semantic embeddings. The structured features consist of 34 variables categorized into six conceptually grounded groups based on prior research. The semantic features are derived from patent abstracts using three pretrained language models, each representing a distinct embedding strategy.

3.3.1 Structured Features

To support early-stage prediction of patent value, this study incorporates 34 structured features derived from metadata available at the time of grant. These features were grouped into six conceptual categories based on prior research that links specific patent characteristics to long-term commercial value. Each group represents a distinct dimension of patent-level information, encompassing legal, organizational, and technical perspectives.

(1) R&D Scale

This group includes the number of inventors and applicants associated with each patent. Larger inventor teams often indicate broader knowledge integration and greater organizational investment in the inventive process. Prior studies have shown that patents involving more inventors tend to exhibit higher technological impact and commercial potential, likely due to the enhanced capacity for combining diverse expertise and solving complex problems (Breitzman & Thomas, 2015; Ernst, 2003; Singh & Fleming, 2010).

(2) Ownership Structure

Features in this category capture the presence and characteristics of patent assignees, including whether a patent is assigned, the number of distinct organizational entities involved, and their national affiliations. Patents held by organizational entities—particularly those associated with multinational firms—are more likely to exhibit higher commercial value, broader licensing potential, and longer maintenance periods. This is because institutional ownership often reflects stronger resource commitment, strategic deployment of intellectual property, and clearer market orientation (Criscuolo, 2009; Guellec & van Pottelsberghe, 2000; Lanjouw & Schankerman, 2004).

(3) Legal and Filing Strategy

This category includes indicators such as the presence of priority claims, the number and geographical scope of priority filings, participation in the Patent Cooperation Treaty, time from application to grant, and applicant entity type (e.g., micro, small, or large). These variables reflect strategic behaviors that signal an applicant's resource level, international commercialization intent, and engagement with the patent system. Prior research has linked such filing characteristics with higher renewal rates and increased patent value (Graham et al., 2009; Harhoff & Wagner, 2009; Putnam, 1996).

(4) Textual and Claims Structure

This category covers structural features of the patent document's written content, including claim count and length, as well as the word volume of abstract and descriptive sections. These indicators reflect how applicants define protection scope and document their inventions, and have shown varying degrees of association with patent value. While longer text or more numerous claims may suggest broader coverage or higher drafting effort, empirical findings indicate that such relationships are context-dependent and not necessarily linear. Nonetheless, textual and claim-level structures remain relevant for understanding how patents are constructed and how they relate to value proxies (Jansen, 2009; Marco et al., 2019; Okada et al., 2016).

(5) Citation and Disclosure

This category captures backward citation behaviors and disclosure breadth, as reflected in the number and type of documents referenced in a patent. Features include the count of cited U.S. patents, U.S. applications, foreign patents, and non-patent literature, as well as the number of unique countries from which cited patents originate. These indicators serve as proxies for the knowledge base upon which the invention builds, and may reflect the applicant's effort in prior art disclosure and the technological scope of the invention. Prior studies have associated higher citation volume and broader citation diversity with greater patent value and technological significance, though such relationships may vary depending on the technological domain or the origin of the citations (Criscuolo, 2006; Harhoff et al., 2003; Michel & Bettels, 2001).

(6) IPC Diversity

This category captures the breadth of technological domains associated with each patent, as reflected in the number of distinct IPC codes assigned at the time of grant. A greater number of IPC codes may suggest that the invention spans multiple technical areas, indicating broader applicability, cross-domain relevance, or greater potential for recombination. Prior

research has associated such technological diversity with increased innovation potential and patent value (Squicciarini et al., 2013).

By organizing the structured features into six theoretically grounded categories, this study achieves comprehensive coverage of key patent attributes linked to long-term value while ensuring temporal consistency at the time of grant. This feature schema not only improves interpretability but also establishes a modular basis for integrating structured and semantic inputs within the subsequent modeling framework. A complete overview of all structured features is presented in Table 3.2.

All structured features described above were retained in the final experiments. Although feature selection methods were preliminarily evaluated, they did not lead to improvements in predictive performance; detailed results are provided in Appendix A. Given that the primary classifiers—particularly tree-based models—can inherently handle irrelevant or redundant inputs (see Section 2.2.2), retaining the complete feature set ensures both methodological consistency and the inclusion of all available information in the predictive modeling process.

Table 3.2. Summary of Structured Features by Category

Feature Category	Feature Name	Description
R&D Scale	team_size	Number of inventors.
	number_applicant	Number of applicants.
Ownership Structure	d_assignee	Whether the patent has at least one assignee.
	number_assignee	Number of assignees.
	number_assignee_org	Number of assignees that are organizations.
	number_assignee_nation	Number of countries represented by the assignees.
	assignee_has_org	Whether any assignee is an organization.

Feature Category	Feature Name	Description
Legal and Filing Strategy	has_priority	Whether the patent claims any priority.
	number_priority_claim	Number of priority claims.
	number_priority_nation	Number of countries in which priority is claimed.
	has_pct	Whether the patent is linked to a Patent Cooperation Treaty filing.
	delivery_time	Years between application and grant.
	entity_micro	Whether the applicant is a micro entity.
	entity_small	Whether the applicant is a small entity.
	entity_large	Whether the applicant is a large entity.
Textual and Claim Structure	num_claims	Total number of claims.
	number_claim_ind	Number of independent claims.
	number_claim_dep	Number of dependent claims.
	ratio_claim_ind	Ratio of independent claims to total claims.
	ratio_claim_dep	Ratio of dependent claims to total claims.
	number_avgword_indep	Average number of words in independent claims.
	number_word_claim	Total word count in all claims.
	number_avgword_claim	Average number of words across all claims.
	number_word_abst	Number of words in the abstract.
	number_word_desc	Number of words in the description.
	number_word_ft	Total word count in abstract, claims, and description.

Feature Category	Feature Name	Description
Citation and Disclosure	number_us_patent_citation	Number of cited U.S. granted patents.
	number_us_application_citation	Number of cited U.S. patent applications.
	number_foreign_patent_citation	Number of cited foreign patents.
	num_total_us_citation	Total number of U.S. patent citations (granted + applications).
	num_total_citation	Total number of patent citations (U.S. and foreign).
	number_otherref_citation	Number of cited non-patent literature references.
	number_citation_nation	Number of countries from which cited patents originate.
IPC Diversity	number_ipc_at_issue	Number of IPC codes assigned at the time of grant.

3.3.2 Semantic Embedding Features

To complement the structured features introduced in the previous section, this study incorporates semantic representations derived from patent abstracts. As concise and standardized descriptions of technical content, abstracts are available at the time of grant and provide a compact textual source for modeling. Prior research has highlighted their utility in semantic tasks due to their accessibility and information density (Cohan et al., 2020), making them well-suited for embedding-based prediction frameworks.

This study employs three pretrained sentence embedding models to generate semantic features: BERT for Patents, SPECTER, and MiniLM. These models were selected based on their architectural diversity, domain specificity, and practical relevance. A summary of their key characteristics is presented in Table 3.3.

- **BERT for Patents** is a BERT-Large model (24 layers, 1024 hidden units) pretrained from scratch on the full corpus of United States patent documents, as described by Google's Patents team (Google, 2020). Although the original pretrained weights have not been released, this study adopts a publicly available variant on Hugging Face (*anferico/bert-for-patents*), which follows the same architecture and produces 1024-dimensional embeddings. The model is designed to represent the semantic structure of patent language and support tasks such as classification, retrieval, and semantic similarity detection.
- **SPECTER** is a 768-dimensional model based on the SciBERT architecture and trained using a citation-informed triplet loss function to capture semantic similarity among scientific publications (Cohan et al., 2020). This study employs the pretrained version available on Hugging Face (*allenai/specter*), which generates document-level embeddings that encode citation-based relatedness and topical coherence. While it was not trained on patent texts, SPECTER is effective in modeling technical writing and has been widely used for tasks such as document classification, retrieval, and clustering in research domains.
- **MiniLM** is a lightweight Transformer model developed by Microsoft Research to reduce computational cost through a self-attention distillation framework, which compresses larger models while retaining core semantic capabilities (Wang et al., 2020). This study adopts the 384-dimensional variant **all-MiniLM-L6-v2**, fine-tuned for sentence embeddings using the Sentence-Transformers framework (Reimers & Gurevych, 2019) and publicly available on Hugging Face (*sentence-transformers/all-MiniLM-L6-v2*). It is optimized for semantic search, clustering, and other low-latency natural language processing applications, providing a practical balance between representational accuracy and efficiency.

Table 3.3. Overview of Pretrained Semantic Embedding Models

Attribute	BERT for Patents	SPECTER	MiniLM
Embedding Dimension	1024	768	384
Training Corpus	Full text of U.S. patents	Scientific papers (title + abstract) with citation links	General sentence-level corpus
Model Type	BERT-large	SciBERT + Triplet Loss	Distilled Transformer (MiniLM-L6)
Domain Specificity	Patent-specific	Technical/scientific	General-purpose

To integrate these semantic representations into the predictive framework, each patent abstract was tokenized using the default tokenizer associated with each pretrained model and then encoded into a fixed-length embedding vector. These embedding vectors were then concatenated with the 34 structured features described in Section 3.3.1 to construct hybrid input representations, which served as the input for subsequent predictive modeling.

Importantly, the inclusion of all three models also enables a preliminary comparison of semantic embedding strategies. These models differ in corpus domain, architecture, and embedding dimensionality. This diversity allows for a preliminary comparative assessment of how different semantic representation strategies may influence early-stage patent value prediction in a domain-specific setting.

3.4 Predictive Modeling

This study evaluates each classifier under three feature configurations—structured-only, semantic-only, and hybrid—to address the heterogeneous nature of the input features, which

include both structured metadata and semantic embeddings. The modeling strategy emphasizes flexibility, robustness, and the ability to capture complex, potentially nonlinear relationships between input variables and the target outcome.

Four widely used classification algorithms are employed to establish baseline and tree-based performance comparisons: logistic regression, random forest, XGBoost, and LightGBM. Logistic regression serves as a simple and interpretable baseline, while the three tree-based models are selected for their ability to model nonlinear interactions and effectively handle mixed-type features. All models are trained under consistent experimental settings using a fixed train-test split to ensure comparability across feature configurations.

To further enhance predictive performance, this study implements a stacking ensemble strategy that integrates the outputs of multiple base classifiers. The base learners—logistic regression, random forest, and LightGBM—are deliberately chosen from the same set of classifiers used in the single-model experiments. This design ensures that any observed improvement can be attributed to the ensemble mechanism itself, rather than to differences in model types. Logistic regression is employed as the meta-learner to combine the base learners' predictions into a final output, as it offers interpretability, stability against overfitting when aggregating a small set of model outputs, and a natural compatibility with probability-based inputs produced by the base classifiers.

The stacking ensemble is implemented as a two-stage process with cross-validation, as illustrated in Figure 3.2, to prevent information leakage and ensure robustness. The base learners are trained on stratified folds of the training set, and their out-of-fold predictions are then used to train the meta-learner on held-out data. This approach enables the ensemble to capture complementary decision patterns while maintaining generalizability and interpretability. All model hyperparameters were tuned under the hybrid feature configuration (structured + BERT for Patents) and then fixed across all experiments to ensure consistency and comparability.

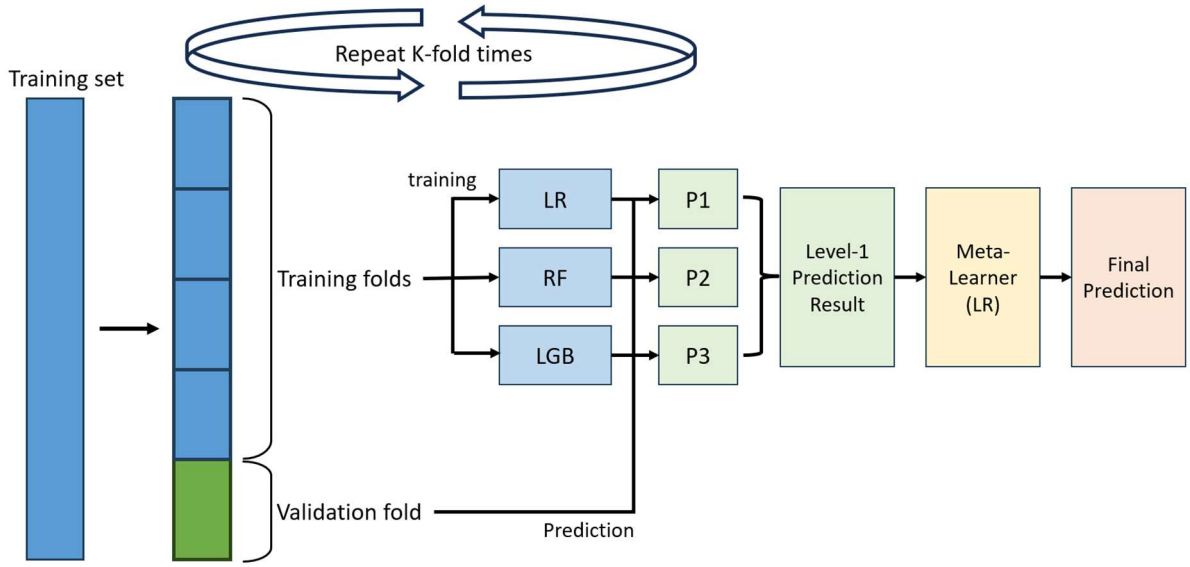

Figure 3.2. Architecture of the Two-Stage Stacking Ensemble with Cross-Validation

3.5 Evaluation Metrics

To evaluate model performance in predicting whether a granted patent will be maintained through its full legal term, this study employs several standard metrics for binary classification. Given the near-balanced nature of the dataset, two primary metrics are emphasized: Area Under the Receiver Operating Characteristic Curve (AUC-ROC) and Accuracy.

- AUC-ROC measures the model's ability to distinguish between the two classes across all possible classification thresholds. It is threshold-independent and particularly useful for evaluating probabilistic outputs, providing a robust assessment of overall discriminative performance.
- Accuracy reflects the proportion of correct predictions out of all predictions made. As the dataset is approximately balanced, accuracy remains a reliable and interpretable metric in this context.

To supplement these main indicators, the study also reports Precision, Recall, and F1-Score. These metrics offer additional insights into the model's class-specific behavior. Precision

indicates the proportion of predicted positive samples that are correct, while recall measures the proportion of actual positives that are correctly identified. The F1-Score, as the harmonic mean of precision and recall, summarizes the trade-off between these two aspects. All evaluation metrics are computed on the test set using a fixed threshold of 0.5, and the results are analyzed in Chapter 4 to compare model and feature group performance.

4. Experiments and Results

Building on the predictive framework introduced in Chapter 3, this chapter presents a series of experiments designed to evaluate the effectiveness of integrating structured metadata and semantic embeddings for early-stage patent value prediction. The evaluation is conducted on U.S. utility patents in the semiconductor domain, using only information available at the time of grant to ensure applicability in real-world decision contexts.

This chapter comprises four experimental sections followed by a summary. Section 4.1 describes the experimental setup, including dataset composition, feature configurations, model specifications, and evaluation metrics. Section 4.2 evaluates model performance across three input settings—structured-only, semantic-only, and hybrid—to assess the individual and combined contributions of heterogeneous features. Section 4.3 evaluates the effect of semantic model choice by comparing three pretrained embedding models (BERT for Patents, SPECTER, and MiniLM) under the hybrid configuration. Section 4.4 evaluates a stacking ensemble that integrates the predictions of multiple base classifiers to enhance accuracy and robustness. Section 4.5 summarizes the empirical results and highlights the key findings.

Overall, the results show that hybrid features consistently outperform single-source inputs, confirming the complementary value of combining bibliographic, legal, and textual signals. Among the semantic models tested, BERT for Patents achieves the highest predictive performance, likely due to its domain-specific pretraining and closer alignment with the linguistic characteristics of the task. In contrast to expectations, the stacking ensemble shows no notable advantage over the best single model under the unified data split, though results from a larger training set suggest potential for further gains with expanded data. These findings provide empirical support for the proposed framework and underscore the importance of heterogeneous feature fusion and ensemble design in predicting long-term patent maintenance.

4.1 Experimental Settings

To ensure consistency and reproducibility in model evaluation, all experiments in this study adopt a unified set of experimental settings. The dataset comprises 159,945 U.S. utility patents granted between 2000 and 2022 in the semiconductor domain (IPC subclass H01L). After filtering for complete metadata and valid maintenance event records, the final sample includes 77,763 patents labeled as Core Business Patents (CBP = 1) and 82,182 as non-CBP (CBP = 0). The dataset was randomly split into three subsets: 96,000 patents for training, 24,000 for validation, and 39,945 for testing, corresponding to approximately 60%, 15%, and 25% of the full dataset, respectively. All subsets maintain a near-balanced distribution of CBP and non-CBP classes, as summarized in Table 4.1. A fixed random seed (42) was used for all stochastic operations to ensure reproducibility. Given the near-balanced class distribution, stratified sampling was not considered necessary.

Table 4.1. Dataset Composition and CBP Distribution

Subset	Total Samples	CBP (n, %)	Non-CBP (n, %)
Training Set	96,000	46,689 (48.6%)	49,311 (51.4%)
Validation Set	24,000	11,723 (48.8%)	12,277 (51.2%)
Test Set	39,945	19,351 (48.4%)	20,594 (51.6%)
Total	159,945	77,763 (48.6%)	82,182 (51.4%)

Each classification model is trained under three feature configurations: (1) structured-only features, comprising 34 metadata variables described in Section 3.3.1; (2) semantic-only features, derived from sentence embeddings of patent abstracts generated by BERT for Patents;

and (3) hybrid features, which integrate structured metadata with semantic representations. For the hybrid configuration, three pretrained embedding models—BERT for Patents, SPECTER, and MiniLM—are evaluated, as introduced in Section 3.3.2. These models produce embeddings of 1024, 768, and 384 dimensions, respectively, which are concatenated with the 34 structured variables to form composite input representations.

The classifiers used in this study include logistic regression (LR), random forest (RF), XGBoost, and LightGBM. These models were chosen to represent a combination of linear and tree-based approaches, offering a range of complexity and interpretability. Given the iterative nature and higher variance of boosting algorithms, early stopping is enabled for XGBoost and LightGBM in the single-model experiments using a 24,000-sample validation set to prevent overfitting. In contrast, logistic regression and random forest are trained without early stopping due to their lower variance and relative robustness to overfitting. For the stacking ensemble experiments, logistic regression, random forest, and LightGBM are used as base learners, trained via cross-validation without early stopping to fully utilize the available training data while avoiding information leakage. All model hyperparameters are tuned based on performance under the hybrid feature setting using BERT for Patents, and the resulting configurations are applied consistently across all experiments. All embedding models are used as fixed encoders without task-specific fine-tuning.

Five evaluation metrics are reported for all experiments: area under the ROC curve (AUC), accuracy, precision, recall, and F1-Score. Among these, AUC and accuracy are emphasized as the primary indicators, consistent with the evaluation strategy described in Section 3.5. This consistent evaluation strategy allows for reliable performance comparison across models and feature configurations.

4.2 Performance Comparison Across Feature Configurations

This section presents a comparative analysis of model performance under three feature configurations: structured-only, semantic-only, and hybrid. The goal is to examine whether semantic features extracted from patent abstracts can complement structured metadata, and whether their integration enhances predictive performance beyond what either type of input can achieve individually.

As shown in Table 4.2, the hybrid configuration—combining 34 structured features with BERT-based semantic embeddings (BERT for Patents)—consistently yields the best performance. The XGBoost classifier achieves the highest AUC (0.7882) and accuracy (0.7106) under the hybrid setting, followed closely by LightGBM (AUC = 0.7880; accuracy = 0.7097). The hybrid XGBoost model outperforms its semantic-only counterpart by 4.6 percentage points in AUC (0.7882 compared with 0.7425) and 3.2 percentage points in accuracy (0.7106 compared with 0.6791), confirming the synergistic value of combining structured and semantic information, a key premise of this study.

When using only structured features, overall performance is noticeably lower across all evaluation metrics, though the results still reflect a baseline level of predictive capability. For instance, the AUC of XGBoost drops to 0.6932 and accuracy to 0.6349, representing a decline of 9.5 percentage points in AUC compared to the hybrid input. This pattern is consistent across all classifiers, illustrating the limitations of relying solely on bibliographic and legal metadata for early-stage value prediction.

Interestingly, semantic-only models—based on BERT embeddings—perform slightly better than structured-only models in most cases. For example, the semantic-only version of XGBoost achieves an AUC of 0.7425 and accuracy of 0.6791. This indicates that semantic signals encoded in patent abstracts do capture value-relevant patterns, aligning with recent findings in the literature that highlight the predictive utility of textual representations.

Table 4.2. Model Performance Across Feature Configurations

Feature Configuration	Model	AUC	Accuracy	Precision	Recall	F1-Score
Structured-only	Logistic Regression	0.6638	0.617	0.5911	0.6798	0.6323
	Random Forest	0.6972	0.6357	0.6166	0.6556	0.6355
	XGBoost	0.6932	0.6349	0.6096	0.6848	0.645
	LightGBM	0.6951	0.6372	0.6112	0.6898	0.6482
Semantic-only (BERT for Patents)	Logistic Regression	0.6735	0.6252	0.616	0.601	0.6084
	Random Forest	0.7341	0.6729	0.6802	0.6131	0.6449
	XGBoost	0.7425	0.6791	0.6785	0.6416	0.6595
	LightGBM	0.7443	0.6795	0.678	0.6444	0.6608
Hybrid (BERT for Patents)	Logistic Regression	0.7196	0.6556	0.6373	0.6707	0.6536
	Random Forest	0.7602	0.6918	**0.6993**	0.6382	0.6673
	XGBoost	**0.7882**	**0.7106**	0.6979	**0.7097**	**0.7038**
	LightGBM	0.788	0.7097	0.6971	0.7088	0.7029

Beyond aggregate metrics, semantic-only models also exhibit a distinct prediction behavior: higher precision but lower recall compared to structured-only models. This pattern suggests that semantic embeddings tend to adopt a more conservative decision strategy, identifying high-value patents primarily when strong linguistic signals are present in the abstract. As a result, they are less likely to misclassify non-valuable patents as valuable (higher precision), but also more prone to missing truly valuable patents with less overt textual cues

(lower recall). This trade-off reinforces the notion that semantic information alone, while informative, is insufficient to fully capture the multifaceted nature of patent value—particularly in cases where textual signals are subtle or underrepresented.

Taken together, these results underscore the complementary strengths of structured and semantic features. While semantic embeddings contribute contextual richness, structured metadata provides stability and coverage, particularly for patents whose value is not easily inferred from text alone. The consistent superiority of hybrid models across all classifiers demonstrates that combining these modalities offers a more comprehensive foundation for early-stage patent value prediction.

4.3 Comparative Performance of Semantic Embedding Models

To examine how different types of semantic embeddings affect predictive performance, this section compares three pretrained models—BERT for Patents, SPECTER, and MiniLM—under the hybrid feature configuration. In all cases, embeddings are derived from patent abstracts and concatenated with the 34 structured features described in Section 3.3.1. The classifiers, training setup, and hyperparameter configurations remain consistent across models, ensuring that any observed differences can be attributed solely to the semantic representation.

Table 4.3 summarizes the performance of each embedding model across four classifiers. Among the three models, BERT for Patents consistently delivers the highest performance across all metrics and classifiers, achieving the best overall results with XGBoost (AUC = 0.7882, Accuracy = 0.7106, F1-Score = 0.7038). This superior performance may be attributed to its domain-specific pretraining on the full corpus of U.S. patents, enabling it to better capture the linguistic patterns and technical terminology prevalent in patent documents. Its higher embedding dimensionality (1024) may also contribute to the richer semantic representation.

SPECTER, which is trained on scientific abstracts with a citation-informed objective,

achieves its best performance with XGBoost (AUC = 0.7680, Accuracy = 0.6932, F1-Score = 0.6897). Although it lacks exposure to patent-specific language, its alignment with technical discourse still allows it to encode meaningful semantic content. However, its training objective—centered on modeling inter-document similarity rather than supervised classification—may limit its ability to capture task-specific discriminative signals, potentially contributing to its slightly lower performance compared to BERT for Patents.

MiniLM, the most lightweight model in terms of architecture and dimensionality (384), records its best results with LightGBM (AUC = 0.7672, Accuracy = 0.6925, F1-Score = 0.6896). While it lacks domain specificity, its efficiency and general-purpose sentence representations offer a practical trade-off between accuracy and computational cost. Nonetheless, like SPECTER, it falls short of the performance attained by BERT for Patents.

Table 4.3. Performance Comparison of Semantic Embedding Models

Embedding Model	Classifier	AUC	Accuracy	Precision	Recall	F1-Score
BERT for Patents (1024d)	Logistic Regression	0.7196	0.6556	0.6373	0.6707	0.6536
	Random Forest	0.7602	0.6918	0.6993	0.6382	0.6673
	XGBoost	0.7882	0.7106	0.6979	0.7097	0.7038
	LightGBM	0.788	0.7097	0.6971	0.7088	0.7029
SPECTER (768d)	Logistic Regression	0.689	0.6307	0.6104	0.6571	0.6329
	Random Forest	0.7495	0.6804	0.6825	0.6364	0.6586
	XGBoost	0.768	0.6932	0.6761	0.7038	0.6897
	LightGBM	0.7665	0.6918	0.6742	0.7041	0.6888

Embedding Model	Classifier	AUC	Accuracy	Precision	Recall	F1-Score
MiniLM (384d)	Logistic Regression	0.6913	0.6325	0.6108	0.6651	0.6368
	Random Forest	0.7565	0.6881	0.6846	0.6605	0.6723
	XGBoost	0.7632	0.6887	0.672	0.6983	0.6849
	LightGBM	0.7672	0.6925	0.6746	0.7053	0.6896

Overall, these findings suggest that semantic embeddings can substantially influence model performance, even with consistent structured inputs and modeling settings. BERT for Patents outperformed the other models across all classifiers, likely benefiting from domain-specific pretraining. However, differences in embedding size and training objectives may also have contributed, as the models vary in corpus, architecture, and representation strategy. These results highlight the value of semantically aligned embeddings in heterogeneous feature fusion for early-stage patent value prediction.

4.4 Evaluation of Stacking Ensemble Performance

In addition to individual classifiers, this study implements a stacking ensemble to further enhance predictive performance by leveraging the complementary strengths of multiple base learners. As described in Section 3.4, the ensemble integrates three base classifiers—logistic regression, random forest, and LightGBM—using logistic regression as a meta-learner. All models are trained using the hybrid feature set constructed from 34 structured variables and 1024-dimensional BERT for Patents embeddings, which demonstrated the strongest performance among all configurations in previous experiments.

The stacking procedure follows a two-stage process to avoid information leakage. Base learners are first trained using stratified folds of the training data, and their out-of-fold

predictions are then used to train the meta-learner on held-out validation data. This design ensures that the meta-learner receives reliable input while preserving generalization. The evaluation results of the stacking ensemble model are presented in Table 4.4.

Using the 96k training set, the stacking ensemble achieves an AUC of 0.7868 and an accuracy of 0.7092, with the highest precision (0.7014) among the three models compared in this setting. However, these values are slightly lower than those of the overall best single model, XGBoost (AUC = 0.7882, Accuracy = 0.7106), and also below its strongest base learner, LightGBM (AUC = 0.7880, Accuracy = 0.7097). Although stacking produces competitive results, under the 96k configuration it does not outperform either benchmark in terms of overall predictive performance.

Table 4.4. Performance of Stacking Ensemble and Best Single Models

Training Set	Model	AUC	Accuracy	Precision	Recall	F1-Score
96k	XGBoost (BERT+Structured)	0.7882	0.7106	0.6979	**0.7097**	0.7038
96k	LightGBM (BERT+Structured)	0.7880	0.7097	0.6971	0.7088	0.7029
96k	Stacking Ensemble (LR + RF + LGBM)	0.7868	0.7092	0.7014	0.6960	0.6987
120k	Stacking Ensemble (LR + RF + LGBM)	**0.7945**	**0.7168**	**0.7102**	0.7017	**0.7059**

To examine the effect of training data size, an additional experiment was conducted using a 120k training set, obtained by combining the original 96k training set with the 24k validation set from earlier single-model experiments. Because the base learners in the stacking ensemble do not employ early stopping and cross-validation is used during training, the full 120k set could be utilized without risk of information leakage. Compared with the 96k version, the 120k

stacking ensemble shows consistent gains across all metrics, with AUC increasing from 0.7868 to 0.7945, accuracy from 0.7092 to 0.7168, and F1-Score from 0.6987 to 0.7059. As the model architecture and feature configuration remain identical, these improvements can be attributed to the larger training set.

In summary, while stacking under the 96k setting performs comparably to the strongest single classifiers, enlarging the training set to 120k produces clear and consistent improvements, suggesting that the current data scale may still be insufficient to fully capture the patterns underlying this complex predictive task.

4.5 Summary

This chapter evaluated the proposed predictive framework for early-stage patent value prediction in the semiconductor domain across three experimental dimensions: feature configuration, semantic embedding choice, and ensemble integration. All experiments strictly used grant-time-accessible inputs to reflect real-world decision settings.

First, comparing feature configurations showed that hybrid models—concatenating 34 structured metadata variables with abstract-based semantic embeddings—consistently outperformed single-source inputs across all classifiers. The best result was achieved by XGBoost under the hybrid setting (AUC = 0.7882, Accuracy = 0.7106), closely followed by LightGBM (AUC = 0.7880, Accuracy = 0.7097). Semantic-only models generally surpassed structured-only counterparts, but exhibited higher precision and lower recall, indicating a more conservative decision behavior. These results confirm that bibliographic/legal indicators and linguistic signals capture complementary aspects of patent value.

Second, under the hybrid setting, the choice of semantic embedding materially affected performance. Among BERT for Patents, SPECTER, and MiniLM, the domain-specific BERT for Patents delivered the highest scores across all classifiers—peaking with XGBoost—while SPECTER and MiniLM were competitive but consistently lower. This outcome aligns with the

closer alignment of BERT for Patents' pretraining corpus and linguistic representations to patent-specific terminology and task-relevant semantics, resulting in stronger capture of value-relevant textual patterns.

Third, in the stacking ensemble experiments (logistic regression, random forest, and LightGBM as base learners with logistic regression as the meta-learner), results under the 96k training set were comparable to those of the strongest single model, showing no significant advantage. However, when the training set was expanded to 120k, performance improved consistently across all evaluation metrics. This finding suggests that the current data scale may still be insufficient to fully capture the patterns underlying this complex predictive task, and that further expanding the training data could enhance the predictive performance of all model types.

Overall, the experiments validate the effectiveness of heterogeneous feature fusion and underscore the importance of semantically aligned embeddings in this task. While stacking ensembles can match strong single models under the baseline configuration, their incremental benefits appear contingent on larger training data. These results set up Chapter 5, which summarizes the study's key findings, highlights its contributions, and discusses methodological limitations and practical implications.

5. Conclusion

This chapter synthesizes the study's key empirical findings and discusses their methodological and practical implications within the context of early-stage patent value prediction. It then acknowledges the design limitations, outlines promising avenues for future research, and positions the study within the broader landscape of predictive analytics in technology-intensive domains. By addressing these elements in a coherent sequence, the chapter provides a structured closure to the investigation and emphasizes its relevance to both academic inquiry and real-world application.

5.1 Key Findings

This study proposed a machine learning framework that integrates structured metadata and semantic embeddings to predict whether a granted semiconductor patent will be maintained through its full statutory term—a proxy for long-term commercial value. Based on extensive experiments using grant-time features only, the following key findings were obtained:

First, models that combine structured and semantic features consistently outperformed those using either type alone across all classifiers. This confirms the complementary nature of heterogeneous inputs: while structured features offer legal, organizational, and bibliographic context, semantic embeddings derived from patent abstracts capture value-relevant textual cues that are otherwise inaccessible.

Second, although semantic-only models performed worse than hybrid configurations, they still achieved reasonable accuracy and AUC scores—often surpassing structured-only models. Notably, these models exhibited higher precision but lower recall, indicating a conservative prediction pattern that favors strong linguistic signals. This suggests that even without structural context, abstract-based semantic embeddings can capture latent value signals in cases where the textual expressions are sufficiently salient, but may fail to identify valuable patents whose

technical content is less explicitly conveyed.

Third, the choice of semantic embedding model significantly affects predictive performance. Among the three models evaluated, BERT for Patents consistently outperformed SPECTER and MiniLM across all classifiers. This advantage likely reflects a combination of factors, including domain-specific pretraining, higher embedding dimensionality, and training objectives better aligned with classification tasks. These differences underscore the importance of selecting semantically appropriate models when applying embeddings in specialized domains such as patents.

Fourth, with the 96k training set, the stacking ensemble achieved performance comparable to the strongest single classifiers, showing no significant advantage. However, when the training data was increased to 120k, performance improved consistently across all evaluation metrics, suggesting that the current dataset may still be insufficient to fully capture the latent patterns of this complex predictive task and that further data expansion could enhance the performance of all model types.

These empirical findings provide the foundation for the study's academic and practical contributions, which are elaborated in the next section.

5.2 Research Contributions

This study advances the field of early-stage patent value prediction by proposing a domain-specific, grant-time-constrained machine learning framework that integrates heterogeneous features and evaluates model performance across multiple experimental dimensions. The contributions span both methodological innovations and practical applications, as detailed below:

(1) Framework grounded in real-world decision constraints.

Unlike many prior studies that rely on post-grant indicators such as citations or litigation records, this study constructs a prediction framework using only features available at the time of patent grant. This design mitigates information leakage and enhances the practical utility of the model in real-time decision-making scenarios, such as portfolio evaluation, licensing prioritization, and R&D resource allocation.

(2) Domain-specific modeling in a strategically critical yet underexplored sector.

By focusing exclusively on IPC subclass H01L (semiconductors), the study captures sector-specific innovation dynamics often diluted in cross-domain analyses. This targeted approach addresses a gap in the literature, where domain-specific applications remain limited despite the heterogeneity of patent value determinants across technological fields.

(3) A reproducible taxonomy of early-accessible structured features.

The study systematically categorizes 34 grant-time metadata variables into six conceptually grounded groups: R&D scale, ownership structure, legal and filing strategy, textual and claims structure, citation and disclosure, and IPC diversity. This taxonomy not only enhances feature interpretability but also provides a modular structure that can be adapted in future research or extended to other domains.

(4) Empirical validation of heterogeneous feature fusion.

The experiments demonstrate that hybrid models—combining structured metadata with semantic embeddings derived from patent abstracts—consistently outperform models using only one type of input. This finding empirically confirms the complementary nature of bibliographic/legal signals and contextual linguistic representations, supporting the broader adoption of feature fusion strategies in intellectual property analytics.

(5) Comparative analysis of semantic embedding models.

Through a controlled comparison of three pretrained language models—BERT for Patents, SPECTER, and MiniLM—the study provides evidence that domain-specific pretraining

significantly improves predictive performance in patent evaluation tasks. This insight offers practical guidance for selecting embedding models in future applications and underscores the importance of semantic alignment between model corpus and prediction context.

(6) Clarification of stacking ensemble utility.

The study evaluates a stacking ensemble integrating logistic regression, random forest, and LightGBM as base learners. Contrary to expectations that such architectures would outperform individual models by leveraging diverse predictive patterns, the results show that the stacking ensemble offered no clear advantage over the best single classifier—XGBoost—when trained on the same dataset. This finding suggests that, despite its conceptual appeal, stacking may not provide additional benefits in this setting, and that well-optimized single models can already achieve competitive performance with lower implementation complexity and resource requirements.

Together, these contributions establish a rigorous and interpretable framework for early-stage patent value prediction, with methodological elements that are both scalable and adaptable to broader applications in technology management and innovation policy.

5.3 Research Limitations

Like all empirical studies, this research entails several design choices that introduce inherent limitations. Although these choices were made in alignment with the study's objectives and practical constraints, they may affect the generalizability, interpretability, or extensibility of the findings. This section identifies and explains the most salient limitations to provide a transparent foundation for future research.

(1) This study adopts patent survival as a practical proxy for long-term patent value; however, it may not fully reflect the multifaceted nature of economic or strategic value. Survival status captures the assignee's willingness to maintain a patent through fee payments, but it does

not account for other value dimensions such as licensing revenue, litigation potential, or strategic importance at the portfolio level. Therefore, while appropriate for early-stage prediction, survival remains an imperfect surrogate for patent value in broader contexts.

(2) All semantic features in this study were generated using frozen pretrained embedding models without task-specific fine-tuning. This design choice facilitates scalability and reduces computational complexity but may limit the ability of domain-specific models to adapt to nuanced value signals. Although three embedding models were compared, the current experiment was not designed to systematically isolate the relative impact of model architecture, corpus specificity, or embedding dimensionality—factors that may interact in complex ways and influence semantic representation quality.

(3) The textual representation is derived solely from patent abstracts, which, although consistent and grant-time-accessible, may not capture the full technical or legal substance of a patent. The exclusion of claims, descriptions, or full-text data may constrain the model's ability to identify deeper value signals embedded in detailed disclosures or the scope of legal protection.

(4) The dataset is restricted to U.S. utility patents classified under IPC subclass H01L (semiconductors). While domain alignment may reduce inter-domain variability and improve internal consistency, this choice limits the ability to evaluate the framework's generalizability across technological fields. The observed performance gains from narrowing were moderate, suggesting that the effectiveness of domain-specific modeling may depend on additional factors such as feature composition, model architecture, or the nature of the prediction target. A systematic cross-domain comparison would be required to assess the broader applicability of the proposed framework.

In sum, these limitations do not undermine the study's core contributions, but they highlight several avenues for methodological refinement and broader empirical validation. Future research may address these issues by (i) evaluating alternative or composite proxy indicators to better capture patent value, (ii) exploring fine-tuning strategies for embedding

models or task-specific representation learning, (iii) incorporating full-text information such as claims and descriptions, and (iv) testing the framework across diverse IPC domains or cross-sectoral datasets to assess its generalizability and robustness in broader innovation contexts.

5.4 Future Work

While the preceding section focused on limitations that invite direct methodological refinements, this section shifts attention to broader innovations that have gained traction in recent research. These directions are conceptually compatible with the present framework and have potential to serve as natural extensions to enhance its predictive capacity, scalability, and representational depth.

First, graph neural networks (GNNs) offer a promising avenue for capturing the relational structure inherent in patent metadata. Many bibliographic attributes—such as citations, assignee affiliations, inventor collaborations, and IPC co-classifications—form natural graph structures that are difficult to model with traditional tabular approaches. By representing patents as nodes and their relationships as edges, GNN-based models can learn from structural dependencies across the patent ecosystem, potentially improving prediction accuracy and interpretability (Han et al., 2024).

Second, multimodal learning presents an opportunity to further enrich feature representation. While this study integrates structured and semantic inputs, future work could incorporate additional modalities such as patent drawings, classification hierarchies, or examiner communications. Recent advances in cross-modal attention and transformer-based fusion architectures enable more effective alignment across heterogeneous inputs, allowing models to harness complementary signals from diverse data sources (Song et al., 2025).

Third, the temporal dynamics of innovation merit closer integration into predictive frameworks. While patent value prediction often relies on longitudinal datasets, technological

relevance can shift rapidly—especially in fast-moving domains like semiconductors, where once-prominent inventions may quickly lose significance. Although this study focuses on grant-time features to preserve early-stage applicability, future research could explore time-aware models, such as temporal GNNs, that integrate historical trends, organizational trajectories, or domain evolution without violating the ex-ante prediction principle. These approaches may help capture latent temporal patterns and enhance model robustness and generalizability over time (Ding et al., 2024; Ji et al., 2019).

Taken together, these directions point to a new generation of predictive frameworks capable of modeling the structural, semantic, and temporal complexities of patent data—offering richer, more context-aware insights while preserving the ex-ante constraints essential for responsible and timely decision-making.

5.5 Concluding Remarks

This study has proposed a machine learning-based framework for early-stage patent value prediction, integrating structured metadata and semantic embeddings to assess whether a granted patent is likely to be maintained through its full statutory term. By focusing on grant-time-accessible features and avoiding information leakage, the framework aligns with practical constraints faced by firms and policymakers in real-world decision-making. The empirical results confirm the effectiveness of heterogeneous feature fusion and ensemble learning strategies, based on experiments conducted in the semiconductor domain.

Building on this foundation, several forward-looking directions—such as graph-based modeling, multimodal integration, and time-aware architectures—have been identified as promising avenues to further advance this line of research and expand its applicability in dynamic, real-world environments. Taken together, these findings underscore the importance of combining methodological rigor with domain-specific awareness, and demonstrate that

early-stage predictions, while inherently challenging, can yield valuable and actionable insights for managing innovation, guiding R&D investment, and evaluating the long-term technological potential of emerging inventions under conditions of uncertainty and complexity.

In doing so, this study contributes not only to practical applications in patent evaluation, but also to the broader research agenda in predictive analytics, by illustrating how heterogeneous data sources and machine learning techniques can be effectively integrated and extended under early-stage constraints in specialized domains.

References

Aggarwal, C. C., & Zhai, C. (Eds.). (2012). *Mining text data*. Springer. https://doi.org/10.1007/978-1-4614-3223-4

Alcácer, J., & Gittelman, M. (2006). Patent citations as a measure of knowledge flows: The influence of examiner citations. *The Review of Economics and Statistics, 88*(4), 774–779. https://doi.org/10.1162/rest.88.4.774

Correa Bahnsen, A., Aouada, D., Stojanovic, A., & Ottersten, B. (2016). Feature engineering strategies for credit card fraud detection. *Expert Systems with Applications, 51*, 134–142. https://doi.org/10.1016/j.eswa.2015.12.030

Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(2), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607

Bao, R., Sun, Y., Gao, Y., Wang, J., Yang, Q., Mao, Z.-H., & Ye, Y. (2023). A recent survey of heterogeneous transfer learning [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2310.08459

Belderbos, R., & Mohnen, P. (2020). *Inter-sectoral and international R&D spillovers* (UNU-MERIT Working Paper Series No. 047). United Nations University – Maastricht Economic and Social Research Institute on Innovation and Technology (UNU-MERIT). https://ideas.repec.org/p/unm/unumer/2020047.html

Bessen, J. E. (2008). The value of U.S. patents by owner and patent characteristics. *Research Policy, 37*(5), 932–945. https://doi.org/10.1016/j.respol.2008.02.005

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*, 123–140. https://doi.org/10.1007/BF00058655

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees.* Wadsworth International Group.

Breitzman, A., & Thomas, P. (2015). Inventor team size as a predictor of the future citation impact of patents. *Scientometrics, 103*(2), 631–647. https://doi.org/10.1007/s11192-015-1550-5

Camacho-Collados, J., & Pilehvar, M. T. (2018). From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research, 63*, 743–788. https://doi.org/10.1613/jair.1.11259

Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine, 9*(2), 48–57. https://doi.org/10.1109/MCI.2014.2307227

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Choi, J., Jeong, B., Yoon, J., Coh, B.-Y., & Lee, J.-M. (2020). A novel approach to evaluating the business potential of intellectual properties: A machine learning-based predictive analysis of patent lifetime. *Computers & Industrial Engineering, 145*, Article 106544. https://doi.org/10.1016/j.cie.2020.106544

Chung, P., & Sohn, S. Y. (2020). Early detection of valuable patents using a deep learning model: Case of semiconductor industry. *Technological Forecasting and Social Change, 158*, Article 120146. https://doi.org/10.1016/j.techfore.2020.120146

Cohan, A., Feldman, S., Beltagy, I., Downey, D., & Weld, D. S. (2020). SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 2270–2282). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.207

Criscuolo, P. (2006). The 'home advantage' effect and patent families: A comparison of OECD triadic patents, the USPTO and the EPO. *Scientometrics, 66*(1), 23–41. https://doi.org/10.1007/s11192-006-0003-6

Criscuolo, P. (2009). Inter-firm reverse technology transfer: The home country effect of R&D internationalization. *Industrial and Corporate Change, 18*(5), 869–899. https://doi.org/10.1093/icc/dtp028

Criscuolo, P., & Verspagen, B. (2008). Does it matter where patent citations come from? Inventor vs. examiner citations in European patents. *Research Policy, 37*(10), 1892–1908. https://doi.org/10.1016/j.respol.2008.07.011

Danish, M. S., Ranjan, P., & Sharma, R. (2022). Valuation of patents in emerging economies: A renewal model-based study of Indian patents [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2208.06157

Deng, N., & Zhang, J. (2025). HMFM: A method for identifying high-value patents by fusing multiple features. *Computers, Materials & Continua, 82*(2), 2235–2254. https://doi.org/10.32604/cmc.2024.058103

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Dietterich, T. G. (2000). Ensemble methods in machine learning. In J. Kittler & F. Roli (Eds.), *Multiple classifier systems* (Lecture Notes in Computer Science, Vol. 1857, pp. 1–15). Springer. https://doi.org/10.1007/3-540-45014-9_1

Ding, M., Yu, W., Zeng, T., & Wang, S. (2024). PTNS: Patent citation trajectory prediction based on temporal network snapshots. *Scientific Reports, 14*, Article 24034. https://doi.org/10.1038/s41598-024-75913-0

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55*(10), 78–87. https://doi.org/10.1145/2347736.2347755

Eom, H., Choi, S., & Choi, S. O. (2021). Marketable value estimation of patents using ensemble learning methodology: Focusing on U.S. patents for the electricity sector. *PLOS ONE, 16*(9), e0257086. https://doi.org/10.1371/journal.pone.0257086

Ernst, H. (2003). Patent information for strategic technology management. *World Patent Information, 25*(3), 233–242. https://doi.org/10.1016/S0172-2190(03)00077-2

Feldman, R., & Sanger, J. (2006). *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge University Press. https://doi.org/10.1017/CBO9780511546914

Gambardella, A., Giuri, P., & Luzzi, A. (2007). The market for patents in Europe. *Research Policy, 36*(8), 1163–1183. https://doi.org/10.1016/j.respol.2007.07.006

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press. http://www.deeplearningbook.org

Google. (2020). *BERT for patents* [Computer software]. GitHub. https://github.com/google/patents-public-data/blob/master/models/BERT%20for%20Patents.md

Graham, S. J. H., Merges, R. P., Samuelson, P., & Sichelman, T. M. (2009). High technology entrepreneurs and the patent system: Results of the 2008 Berkeley Patent Survey. *Berkeley Technology Law Journal, 24*(4), 255–327.

Guellec, D., & van Pottelsberghe de la Potterie, B. (2000). Applications, grants and the value of patent. *Economics Letters, 69*(1), 109–114. https://doi.org/10.1016/S0165-1765(00)00265-2

Hall, B. H., Helmers, C., Rogers, M., & Sena, V. (2014). The choice between formal and informal intellectual property: A review. *Journal of Economic Literature, 52*(2), 375–423. https://doi.org/10.1257/jel.52.2.375

Hall, B. H., Jaffe, A. B., & Trajtenberg, M. (2005). Market value and patent citations. *The RAND Journal of Economics, 36*(1), 16–38. https://doi.org/10.2307/1593752

Hall, B. H., & Ziedonis, R. H. (2001). The patent paradox revisited: An empirical study of patenting in the U.S. semiconductor industry, 1979–1995. *The RAND Journal of Economics, 32*(1), 101–128. https://doi.org/10.2307/2696400

Han, S., Huang, H., Huang, X., Li, Y., Yu, R., & Zhang, J. (2024). Core patent forecasting based on graph neural networks with an application in stock markets. *Technology Analysis & Strategic Management, 36*(8), 1680–1694. https://doi.org/10.1080/09537325.2022.2108781

Harhoff, D., Scherer, F. M., & Vopel, K. (2003). Citations, family size, opposition and the value of patent rights. *Research Policy, 32*(8), 1343–1363. https://doi.org/10.1016/S0048-7333(02)00124-5

Harhoff, D., & Wagner, S. (2009). The duration of patent examination at the European Patent Office. *Management Science, 55*(12), 1969–1984. https://doi.org/10.1287/mnsc.1090.1069

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

He, Y., Deng, K., & Han, J. (2025). Patent value prediction in biomedical textiles: A method based on a fusion of machine learning models. *PLOS ONE, 20*(4), e0322182. https://doi.org/10.1371/journal.pone.0322182

Hemker, K., Simidjievski, N., & Jamnik, M. (2024). HEALNet: Multimodal fusion for heterogeneous biomedical data. In *Advances in Neural Information Processing Systems* (Vol. 37). Neural Information Processing Systems Foundation, Inc. https://proceedings.neurips.cc/paper_files/paper/2024/hash/765871e77d2ca65126d3d64d31aa6908-Abstract-Conference.html

Henry, M. K. (2025). Semiconductor technology: An overview of the global patent landscape. *Henry Patent Law Firm*. https://henry.law/blog/semiconductor-technology-an-overview-of-the-global-patent-landscape/

Hsieh, C.-H. (2013). Patent value assessment and commercialization strategy. *Technological Forecasting and Social Change, 80*(2), 307–319. https://doi.org/10.1016/j.techfore.2012.09.014

Hwang, J.-T., Kim, B.-K., & Jeong, E.-S. (2021). Patent value and survival of patents. *Journal of Open Innovation: Technology, Market, and Complexity, 7*(2), 119. https://doi.org/10.3390/joitmc7020119

Jaffe, A. B., & Trajtenberg, M. (2002). *Patents, citations, and innovations: A window on the knowledge economy*. The MIT Press. https://doi.org/10.7551/mitpress/5263.001.0001

Jansen, W. (2009). *Examining the relation between patent value and patent claims* [Master's thesis, Eindhoven University of Technology]. Eindhoven University Repository. https://research.tue.nl/files/46937839/642246-1.pdf

Ji, T., Chen, Z., Self, N., Fu, K., Lu, C.-T., & Ramakrishnan, N. (2019). Patent citation dynamics modeling via multi attention recurrent networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)* (pp. 2621–2627). International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2019/364

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30, pp. 3146–3154). Curran Associates, Inc. https://dl.acm.org/doi/10.5555/3294996.3295074

Kim, J., & Magee, C. L. (2017). Dynamic patterns of knowledge flows across technological domains: Empirical results and link prediction [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1706.07140

Lanjouw, J. O., & Schankerman, M. (2001). Characteristics of patent litigation: A window on competition. *The RAND Journal of Economics, 32*(1), 129–151. https://doi.org/10.2307/2696401

Lanjouw, J. O., & Schankerman, M. (2004). Patent quality and research productivity: Measuring innovation with multiple indicators. *The Economic Journal, 114*(495), 441–465. https://doi.org/10.1111/j.1468-0297.2004.00216.x

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research, 247*(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Li, M., Zhang, P., Xing, W., Zheng, Y., Zaporojets, K., Chen, J., Zhang, R., Zhang, Y., Gong, S., Hu, J., Ma, X., Liu, Z., Groth, P., & Worring, M. (2025). Using large language models to tackle fundamental challenges in graph learning: A comprehensive survey [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2505.18475

Li, S., & Tang, H. (2024). Multimodal alignment and fusion: A survey [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2411.17040

Lin, H., Wang, H., Du, D., Wu, H., Chang, B., & Chen, E. (2018). Patent quality valuation with deep learning models. In J. Pei, Y. Sato, & X. Lin (Eds.), *Database systems for advanced applications* (Lecture Notes in Computer Science, Vol. 10828, pp. 474–490). Springer. https://doi.org/10.1007/978-3-319-91458-9_29

Liu, J., Li, P., & Liu, X. (2024). Patent lifetime prediction using LightGBM with a customized loss. *PeerJ Computer Science, 10*, e2044. https://doi.org/10.7717/peerj-cs.2044

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval.* Cambridge University Press. https://doi.org/10.1017/CBO9780511809071

Marco, A. C., Sarnoff, J. D., & deGrazia, C. A. W. (2019). Patent claims and patent scope. *Research Policy, 48*(9), Article 103790. https://doi.org/10.1016/j.respol.2019.04.014

Michel, J., & Bettels, B. (2001). Patent citation analysis: A closer look at the basic input data from patent search reports. *Scientometrics, 51*(1), 185–201. https://doi.org/10.1023/A:1010577030871

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1301.3781

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. The MIT Press.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* (pp. 689–696). PMLR. https://proceedings.mlr.press/v15/ngiam11a/ngiam11a.pdf

Okada, Y., Naito, Y., & Nagaoka, S. (2016). *Claim length as a value predictor of a patent* (IIR Working Paper No. WP#16-04). Institute of Innovation Research, Hitotsubashi University. https://hermes-ir.lib.hit-u.ac.jp/hermes/ir/re/28165/070iirWP16-04.pdf

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 2227–2237). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1202

Putnam, J. D. (1996). *The value of international patent rights* [Unpublished doctoral dissertation]. Yale University.

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Sun, M., Sundberg, P., Yee, H., Zhang, K., Duggan, G. E., Flores, G., Irvine, J., Le, Q., Litsch, K., Marcus, J., Mossin, A., ... Dean, J. (2018). Scalable and accurate deep learning with electronic health records. NPJ Digital Medicine, 1, 18. https://doi.org/10.1038/s41746-018-0029-1

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Reitzig, M. (2004). Improving patent valuations for management purposes—validating new indicators by analyzing application rationales. *Research Policy, 33*(6–7), 939–957. https://doi.org/10.1016/j.respol.2004.02.004

Ren, Z.-H., Yu, C.-Q., Li, L.-P., You, Z.-H., Guan, Y.-J., Li, Y.-C., & Pan, J. (2022). SAWRPI: A stacking ensemble framework with adaptive weight for predicting ncRNA–protein interactions using sequence information. *Frontiers in Genetics, 13*, 839540. https://doi.org/10.3389/fgene.2022.839540

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613–620. https://doi.org/10.1145/361219.361220

Schankerman, M., & Pakes, A. (1986). Estimates of the value of patent rights in European countries during the post-1950 period. *The Economic Journal, 96*(384), 1052–1076. https://doi.org/10.2307/2233173

Shomee, H. H., Wang, Z., Ravi, S. N., & Medya, S. (2024). A comprehensive survey on AI-based methods for patents [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2404.08668

Sill, J., Takács, G., Mackey, L., & Lin, D. (2009). Feature-weighted linear stacking [Preprint]. arXiv. https://doi.org/10.48550/arXiv.0911.0460

Singh, J., & Fleming, L. (2010). Lone inventors as sources of breakthroughs: Myth or reality? *Management Science, 56*(1), 41–56. https://doi.org/10.1287/mnsc.1090.1072

Song, Z., Liu, Z., & Li, H. (2025). Research on feature fusion and multimodal patent text based on graph attention network [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2505.20188

Squicciarini, M., Dernis, H., & Criscuolo, C. (2013). Measuring patent quality: Indicators of technological and economic value. *OECD Science, Technology and Industry Working Papers,* No. 2013/03. OECD Publishing. https://doi.org/10.1787/5k4522wkw1r8-en

Tavakoli, M., Chandra, R., Tian, F., & Bravo, C. (2025). Multi-modal deep learning for credit rating prediction using text and numerical data streams. *Applied Soft Computing, 171*, 112771. https://doi.org/10.1016/j.asoc.2025.112771

Ting, K. M., & Witten, I. H. (1999). Issues in stacked generalization. *Journal of Artificial Intelligence Research, 10*, 271–289. https://doi.org/10.1613/jair.594

Trajtenberg, M. (1990). A penny for your quotes: Patent citations and the value of innovations. *The RAND Journal of Economics, 21*(1), 172–187. https://doi.org/10.2307/2555502

van Zeebroeck, N. (2011). The puzzle of patent value indicators. *Economics of Innovation and New Technology, 20*(1), 33–62. https://doi.org/10.1080/10438590903038256

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2002.10957

Wei, Y., Xu, K., Yao, J., Sun, M., & Sun, Y. (2024). Financial risk analysis using integrated data and transformer-based deep learning. *Journal of Computer Science and Software Applications, 4*(7), 1–8. https://doi.org/10.5281/zenodo.15714892

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks, 5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Yin, P., Neubig, G., Yih, W.-t., & Riedel, S. (2020). TaBERT: Pretraining for joint understanding of textual and tabular data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8413–8426). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.745

Zhang, Y., Gong, K., Zhang, K., Li, H., Qiao, Y., Ouyang, W., & Yue, X. (2023). Meta-Transformer: A unified framework for multimodal learning [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2307.10802

Zhang, H., Wang, Y., Hu, B., Song, B., Wen, Z., Su, L., Chen, X., Wang, X., Zhou, P., Zhong, X., & Pang, H. (2024). Using machine learning to develop a stacking ensemble learning model for the CT radiomics classification of brain metastases. *Scientific Reports, 14,* Article 28575. https://doi.org/10.1038/s41598-024-80210-x

Zhou, Z.-H. (2012). *Ensemble methods: Foundations and algorithms*. Chapman & Hall/CRC.

Appendix A. Feature Selection Results

This appendix presents the results of a preliminary feature selection analysis based on feature importance scores derived from Random Forest (RF) and XGBoost (XGB) classifiers. Importance scores were obtained using each model's built-in feature_importances_ attribute, which measures the normalized total contribution of each feature to reducing the model's loss function—Gini impurity for RF and gradient boosting gain for XGB. Models used identical hyperparameters and the same train/validation/test split as the main experiments; only the input feature set differed (all 34 structured features vs. the top-11). Feature selection was applied exclusively to the 34 structured variables; no dimensionality reduction was performed on semantic embeddings to preserve the integrity of pretrained representations.

For each model, features were ranked in descending order, and the top one-third (11 features) are reported here. These results highlight notable differences in feature ranking between the two algorithms: RF prioritized features related to textual and claim structure, while XGB emphasized ownership structure, entity type, and certain legal or procedural attributes.

To evaluate whether feature reduction could improve predictive performance, models with identical hyperparameters were trained using (i) all 34 structured features and (ii) the top 11 features from each model's ranking. As shown in Tables A.3 and A.4, performance declined for both RF and XGB after feature reduction, most notably in AUC-ROC and accuracy. Consequently, all 34 structured features were retained for the final experiments (see Section 3.3.1).

Table A.1. Top 11 Features Ranked by Random Forest Importance

Rank	Feature Name	Importance Score
1	number_word_desc	0.0703
2	number_word_ft	0.0685

Rank	Feature Name	Importance Score
3	number_avgword_indep	0.0682
4	number_word_abst	0.0678
5	number_avgword_claim	0.0668
6	number_word_claim	0.0651
7	num_total_citation	0.0462
8	number_us_patent_citation	0.0444
9	num_total_us_citation	0.0432
10	ratio_claim_dep	0.0409
11	ratio_claim_ind	0.0408

Table A.2. Top 11 Features Ranked by XGBoost Importance

Rank	Feature Name	Importance Score
1	entity_large	0.1545
2	number_ipc_at_issue	0.0988
3	entity_small	0.0714
4	number_applicant	0.0644
5	has_priority	0.0517
6	number_assignee_org	0.0416
7	team_size	0.0359
8	number_assignee	0.0303
9	number_assignee_nation	0.0279
10	has_pct	0.0246

Rank	Feature Name	Importance Score
11	number_us_application_citation	0.0242

Table A.3. Random Forest Performance: Full vs. Selected Features

Feature Set	AUC	Accuracy	Precision	Recall	F1-Score
All features (34)	0.6972	0.6357	0.6166	0.6556	0.6355
Top 11 features	0.6142	0.5805	0.5713	0.5375	0.5539

Table A.4. XGBoost Performance: Full vs. Selected Features

Feature Set	AUC	Accuracy	Precision	Recall	F1-Score
All features (34)	0.6932	0.6349	0.6096	0.6848	0.6450
Top 11 features	0.6662	0.6135	0.5870	0.6820	0.6309