Journal of the Korean Institute of Industrial Engineers
Vol. 52, No. 4, pp. 320-334, August 2026.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2026.52.4.320

설명가능한 머신러닝 기반 특허 소송 위험 평가 방법

설영진[1] · 홍준희[2] · 이승현[3] · 최재웅[4] · 강민수[5] · 윤장혁[1†]

[1]건국대학교 산업공학과 / [2]한화에어로스페이스 IPS센터 / [3]삼일회계법인 AX Node
[4]단국대학교 경영경제대학 경영학부/ [5]광개토연구소

An Explainable Machine Learning-Based Approach to Patent Litigation Risk Assessment

Youngjin Seol[1] · Junhee Hong[2] · Seunghyun Lee[3] · Jaewoong Choi[4] · Minsoo Kang[5] · Janghyeok Yoon[1]

[1]Department of Industrial Engineering, Konkuk University
[2]Integrated Product Support Center, Hanwha Aerospace
[3]AX Node, Samil PricewaterhouseCoopers
[4]School of Business Administration, Dankook University
[5]Kwanggaeto Lab

Machine learning-based patent litigation prediction has gained attention as a means of assessing patent litigation risk proactively. However, prior studies have faced limitations due to insufficient model interpretability and inadequate incorporation of the multifaceted technological and organizational contexts underlying patent litigation. This study proposes an explainable machine learning-based framework for patent litigation risk assessment that integrates technological attributes and applicant characteristics while ensuring interpretability. The framework comprises four stages: (1) patent and litigation data collection from the USPTO(United States Patent and Trademark Office) and PACER(Public Access to Court Electronic Records), (2) extraction of 16 technology-level and 5 organization-level indicators, (3) a two-stage prediction model－litigation occurrence classification followed by severity assessment－with comparative evaluation of machine learning algorithms including Random Forest, XGBoost, and LightGBM, and (4) stage-specific key determinant identification through SHAP(Shapley additive explanations) analysis. Empirical analysis using 33,603 patents and 81,335 litigation records demonstrates effective prediction of technology infringement risk, with key determinants consistent with existing literature. A risk classification scheme is further introduced to support strategic decision-making in intellectual property management.

Keywords: Patent Litigation Risk, Patent Litigation, Machine Learning, SHAP Analysis, Patent and Patentee Information

1. 서 론

최근 기술의 복잡성이 증가하고 특허 장벽이 밀집된 환경이 형성됨에 따라, 특허 소송 위험을 체계적으로 관리해야 할 필요성이 더욱 커지고 있다(Shapiro, 2000; Lanjouw and Schankerman, 2001; Ming *et al.*, 2024). 유효한 특허로 보호된 기술을 특허권자가 아닌 타인이 무단으로 사용하는 특허 침해가 발생할 경우, 특허권자는 특허 소송을 통해 권리를 행사할 수 있으나, 소송 과정에서 발생하는 법률 비용, 분쟁 대응을 위한 경영 자원의 투입, 혁신 활동의 지연 등은 기업에 상당한 부담을 초래할 수 있다(Bessen and

† 연락저자 : 윤장혁 교수, 05029 서울특별시 광진구 능동로 120 건국대학교 산업공학과, Tel : 02-450-0453, Fax : 02-450-3525, E-mail : janghyoon@konkuk.ac.kr
2026년 2월 13일 접수; 2026년 3월 11일 게재 확정.

Meurer, 2012; Lee *et al.*, 2018). 따라서 특허 소송의 발생 가능성을 예측하고 관련 위험을 효과적으로 관리할 수 있는 방안이 필요하다.

초기 특허 소송 연구들은 소송 결정요인을 규명하는데 초점을 맞추었다. 특허의 기술적 특성(피인용, 청구항 수 등)과 조직적 특성(특허권자 유형, 국적 등)이 소송 발생과 유의미한 관계가 있음을 밝혔다(Lanjouw and Schankerman, 2001; Cremers, 2004). 특허 소송이 주로 기술 간 기능적 중복이나 권리 범위의 충돌에서 비롯된다는 점에 주목하여, 일부 연구들은 특허의 기술적 요소가 담긴 청구항 간의 의미적 유사성을 분석하기도 했다(Lee *et al.*, 2013; Park *et al.*, 2012). 특허 소송을 사후적으로 설명하는 것을 넘어, 발생 가능성을 사전에 예측하는 연구가 수행되었으며(Kim *et al.*, 2021; Wongchaisuwat *et al.*, 2017), 이들 연구는 특허의 소송 위험을 개별 특허 수준에서 정량화함으로써, 기업이 특허 포트폴리오를 선제적으로 관리할 수 있는 기반을 마련하였다.

선행 연구들의 기여에도 불구하고, 특허 소송 예측을 활용하여 개별 특허의 위험도를 분석하기에는 여전히 한계점이 존재한다. 첫째, 기존 연구들은 기술의 내재적 특성을 반영하는 특허 지표들을 주로 활용해 왔으나(Kim *et al.*, 2021), 특허권의 보유 주체인 출원인에 대한 고려는 상대적으로 부족했다. 특허권자의 특성뿐 아니라 이들의 과거 소송 이력과 침해 대응 방식이 특허 소송 가능성에 유의미한 영향을 가지고 있음에도 (Lanjouw and Schankerman, 2003; Surdeanu *et al.*, 2011), 특허권자에 대한 조직적 정보가 특허 소송 예측에는 활용되지 못했다. 둘째, 머신러닝 모델의 내부 작동 원리와 예측 과정이 외부에 명확히 드러나지 않는 불투명한 특성은, 실제 의사결정 과정에서 모델의 예측 결과를 활용하는 데 어려움을 초래한다. 특히, 예측 결과가 전문가의 판단과 상충할 경우 오히려 혼란을 유발하여, 전문가 간 효과적인 의사소통이나 합의 형성을 저해할 수 있다(Kim *et al.*, 2022).

이러한 한계를 보완하기 위해 본 연구는 특허 및 출원인 정보를 활용한 설명가능한 머신러닝 기반의 특허 소송 위험 예측 접근법을 제안한다. 본 연구에서는 특허 소송과 관련된 지표를 두 가지 유형으로 정의하고 이를 추출한다; (1) 기술의 내재적 특성을 반영하는 16개의 정량적 특허지표(권리 범위, 독창성, 과학 집약도 등)와 (2) 조직적 특성을 반영하는 5개의 특허권자 중심 지표(기술 역량, 소송 집약도 등)이다. 다음으로, 본 연구는 두 단계로 구성된 머신러닝 기반 프레임워크를 설계한다; (1) 특허의 소송 발생 가능성을 분류하는 소송 위험 선별 모델을 개발하며, (2) 소송 발생이 예상되는 특허 중 반복적 소송에 연루될 가능성이 높은 고위험 특허를 식별하여 소송 위험의 심각도를 평가하는 모델을 개발한다. 특허 소송 위험은 단일 차원으로 평가하기 어려우며, 소송 발생 가능성과 심각도는 서로 다른 전략적 대응을 요구한다(Somaya, 2012). 소송 발생 가능성이 높은 특허에 대해서는 잠재적 침해에 대한 모니터링 강화, 사전 라이센싱 협상 등 예방적 조치가 우선된다(Crampes and Langinier, 2002). 반면, 반복적 소송에 노출된 고위험 특허에 대해서는 해당 기술 영역의 특허 장벽 강화나 교차 라이센싱을 통한 분쟁 구조의 재편 등 보다 적극적인 위험 관리가 요구된다(Shapiro, 2000; Lanjouw and Schankerman, 2004). 나아가 본 연구는 소송 발생 가능성과 심각도를 분리하여 각각 독립적인 예측 모델을 구축하고, Shapley additive explanations(SHAP) 분석을 적용함으로써, 소송 발생을 촉진하는 요인과 소송 위험의 심화를 유발하는 요인을 구별하여 식별하고, 예측 결과에 대한 전문가의 해석과 활용 가능성을 높이고자 한다.

본 연구는 제안한 접근법의 유효성을 검증하기 위해, 총 81,335건의 소송 사례와 33,603개의 특허를 대상으로 실증 분석을 수행하였다. 특허와 소송 사례에 대한 고품질 정보를 확보하기 위해 Public Access to Court Electronic Records(PACER)와 United States Patent and Trademark Office(USPTO) 데이터베이스를 활용하였다. 이를 통해 사건번호, 소송일자, 원·피고 정보 등 소송 정보와 특허의 서지사항 및 출원인 정보를 수집하였다. 사례분석 결과, 제안된 접근법은 특허 침해 소송 위험의 발생 가능성과 심각도를 효과적으로 예측할 수 있음을 보여주었으며, 기술의 독창성과 권리 범위는 소송 위험의 발생 가능성에 긍정적인 영향을 미치는 반면, 특허권자의 소송 집약도는 소송 위험의 심각도에 긍정적인 영향을, 기술 역량은 부정적인 영향을 미친다는 사실을 밝혀냈다. 제안된 접근법이 특허 소송 위험 평가 시 전문가의 의사결정을 보완하는 도구로 활용될 수 있을 뿐 아니라, 기술적 특성 및 조직적 특성이 특허 소송 위험에 어떻게 연관되는지를 심층적으로 이해하는 기반을 제공할 수 있을 것으로 기대한다.

본 논문의 구성은 다음과 같다. 제2장에서는 본 연구의 이론적 배경을 제시하고, 제3장에서는 본 연구에서 제안된 접근법을 단계별로 상세하게 설명한다. 제4장에서는 연구 결과가 구체적으로 설명되며, 제5장에서는 제안된 접근법을 정량적으로 검증하고, 이의 학문적 및 실무적 시사점을 논의한다. 마지막으로, 제6장에서는 본 연구에 대한 요약 및 시사점과 한계점에 따른 향후 연구 방향을 제시한다.

2. 배경연구

2.1 특허 소송

특허 소송은 특허권과 관련된 법적 분쟁을 해결하기 위한 절차로, 침해 여부뿐만 아니라 특허권의 유효성, 귀속 문제, 실시허가계약 위반 등 다양한 쟁점을 포함한다(Krestel *et al.*, 2021). 기술 혁신이 가속화되고 지식재산의 경제적 가치가 높아짐에 따라, 특허를 둘러싼 분쟁은 지속적으로 증가하고 있다(Shapiro, 2000; Juranek *et al.*, 2020). 특허 소송은 침해 소송, 무효확인 소송, 실시계약 관련 분쟁 등으로 구분되며, 이 중에서도 가장 일반적으로 발생하는 유형은 특허 침해 소송이다.

미국 특허법에 따르면, 특허 침해는 특허권자의 허락 없이 특허받은 발명을 사용, 제조, 판매하는 모든 행위를 포함한다. 일반적으로 특허권자(원고)는 자신의 특허가 무단으로 사용되었다고 판단될 경우, 해당 기술을 사용하거나 제조 및 판매한 침해 혐의자(피고)를 상대로 법적 절차를 개시한다. 즉, '누가 소송을 제기하는가?'에 대한 질문에는 특허권자가 해당되며, '누구를 상대로 소송을 제기하는가?'에 대한 답은 무단으로 해당 기술을 사용하는 자(제조사, 판매사 등)가 된다. 이처럼 침해 소송은 특허권자와 침해 혐의자 간의 분쟁 구조를 중심으로 구성된다. 미국에서의 침해 소송은 일반적으로 원고인 특허권자가 연방 지방법원에 소장을 제출하면서 시작된다. 소장에서 원고는 피고의 특허 침해 사실을 주장하며, 이에 따른 손해배상과 침해 금지 명령을 함께 청구한다. 이후 과정은 상당한 시간과 비용을 수반하기 때문에, 양측이 합의하여 소송을 종결하거나(Cremers, 2009), 재판 절차로 이어지며, 법원이 실제 침해 여부를 판단하게 된다.

특허 소송은 단순한 법적 분쟁을 넘어, 기업에게 상당한 경제적 부담을 수반한다(Rajwani, 2010). 다국적 회계법인인 PricewaterhouseCoopers에 따르면, 미국 내 특허 소송 한 건당 손해배상액은 약 590만 달러에 이른다(PricewaterhouseCoopers, 2014). 또한 미국은 'American Rule'을 채택하고 있어, 승소하더라도 각 당사자가 자신의 변호사 비용을 부담해야 하며(Chen, 2013), 이는 자금 여력이 부족한 중소기업에게 치명적인 부담이 될 수 있다. 특허 소송은 법률적 비용 뿐만 아니라 기업의 기술 전략, 경영 의사결정, 산업 경쟁 구조 등 광범위한 분야에도 영향을 미친다(Juranek *et al.*, 2020). 예를 들어, 핵심 기술을 둘러싼 침해 소송이 진행되는 동안 해당 제품의 상용화가 지연되거나 중단되어 시장 점유율과 수익성이 타격을 받을 수 있다(Lee *et al.*, 2018). 일부 기업은 이러한 소송을 경쟁사의 기술을 견제하거나, 자사 기술 포트폴리오 가치를 부각시키는 협상 수단으로 활용하기도 한다(Yang, 2019).

본 연구는 다양한 특허 소송 유형 중에서도 특히 특허 침해를 원인으로 발생하는 침해 소송에 주목하며, 특허 침해 소송 위험이 어떠한 기술적 또는 조직적 요인에 의해 사전에 설명되고 예측 가능한지를 분석하는 것을 핵심 목표로 삼는다.

2.2 특허 소송 위험 분석

기술 복잡성과 특허 장벽의 밀도가 높아지면서, 기술 혁신 전략 수립에 있어 특허 소송 위험을 사전에 평가하고 관리하는 것이 점점 더 중요해지고 있다(Reitzig *et al.*, 2007). 특허는 그 자체로 완벽한 보호 수단이 아니며, 침해자를 식별하고 소송, 합의, 또는 묵인 중 적절한 대응을 선택하는 과정에서 상당한 비용과 전략적 판단이 수반된다(Crampes and Langinier, 2002). 따라서 기업은 잠재적 침해 소송 위험을 사전에 예측함으로써, 자사의 지식재산을 선제적으로 보호하고 소송 관련 비용과 경영 자원의 불필요한 투입을 최소화할 필요가 있다.

특허 소송의 주요 원인이나 영향 요인을 분석하려는 다양한 연구들이 수행되어 왔다. Lanjouw and Schankerman(2001)은 특허청 및 법원 소송 데이터를 활용하여 인용 수, 청구항 수, 기술 유사성 등이 특허 소송 발생 가능성에 유의미한 영향을 미친다는 사실을 실증적으로 확인했다. 또한 특허 소송의 확률은 특허 보유자의 유형(기업, 개인) 및 국적(국내, 해외)에 따라 상이하게 나타남을 밝혔다. Cremers(2004)는 특허의 인용 수와 청구항 수, 그리고 특허 패밀리 규모를 중심으로 통계분석을 실시해, 이들이 소송에 취약한 특허의 특징과 밀접하게 관련되어 있음을 밝혀냈다. Lim(2014)은 미국의 소송 특허 데이터를 기반으로 원고-피고 간 특허 인용 네트워크를 구축하여 다층적 인용 관계를 분석했다. 직접 인용보다 간접 인용 및 잠재 인용 관계가 특허 소송 발생에 더 큰 영향을 미친다는 점을 실증적으로 밝혔다. Marco and Miller(2019)는 특허청과 RPX(Rational Patent eXchange) 데이터를 기반으로, 특허 심사 과정에서의 변수들이 특허 소송 발생에 어떤 영향을 미치는지를 분석하였다. RPX는 회원사들을 특허 소송으로부터 보호하기 위해 특허를 매입하는 방어적 특허 집적 기업으로, 부가 서비스로 미국 지방법원의 특허 소송 데이터를 수집·제공한다. 그 결과, 청구항 수, 재심사 요청, 정보공개서 제출, 출원인 유형 등 일부 심사 특성이 소송 위험 예측에 유의미한 신호가 될 수 있음을 보였다. 이러한 연구들은 특허 소송의 주요 원인과 영향 요인에 대한 이해를 한층 심화시켰으며, 기업이 특허 소송 위험을 관리하고 기술 전략을 수립하는 데 중요한 기초 자료를 제공했다.

특허는 수십 개의 분석 항목을 포함하고 있으며, 일반적으로 정형 데이터(인용, 발명자 정보 등)와 비정형 데이터(초록, 설명 등)로 구분된다. 정형 데이터는 일관된 의미와 형식을 갖는 반면, 비정형 데이터는 텍스트 기반으로 다양한 구조와 표현을 가진다. 이러한 특허 내 텍스트를 기반으로, 기술적 유사성이 높은 잠재적 침해 특허를 식별하거나 특허 청구항을 분석하여 침해 여부를 판단하는 접근법이 제안되었다. Shin and Park(2005)은 USPTO의 청구항 데이터를 기반으로 텍스트 마이닝과 네트워크 분석을 결합하여 청구항 간 중첩 관계를 분석하고, 이를 시각화한 특허 청구항 맵을 제안하였다. Park *et al.*(2012)은 USPTO의 특허 명세서 데이터로부터 Subject-Action-Object 구조를 추출해 기술 구성요소 간의 관계를 정량화하고, 이를 기반으로 특허 간 기술 유사도를 계산하여 침해 가능성이 높은 특허를 자동으로 탐지하는 방법을 제안하였다. 또한 제품 - 특허 침해 맵을 생성함으로써 잠재적 침해 관계를 시각적으로 표현하였다. Lee *et al.*(2013)은 USPTO의 청구항 데이터에서 키워드를 추출하여 요소 간 종속 관계를 반영한 트리 구조를 구성한 뒤, 트리 매칭 알고리즘을 통해 특허 간 의미 유사도를 계산하였다. 해당 연구는 기술 구성의 논리적 구조를 고려한 침해 가능성이 높은 특허의 자동 식별 방법을 제안하였다. 텍스트 유사도 기반 접근법은 일반적으로 두 특허 간의 기술적 유사성을 정량화하는 데 초점을 맞추며, 침해 가능성이 높은 특허들을 군집화하거나 유사 특허를 탐지하는 데 효

과적이다. 그러나 특허 텍스트 간 유사도에 의존한 접근법은 특허의 침해 소송 위험도를 독립적으로 예측하지는 못하며, 비교 대상이 반드시 존재해야 한다는 구조적 제약이 있다. 또한 개별 특허의 위험 수준을 수치화하거나, 우선순위를 설정하기 위한 기준으로 활용하기 어렵다는 한계도 존재한다.

특허 침해 소송의 발생 여부는 기술 유사성뿐만 아니라 특허 내재적 요인, 출원인 특성, 산업 분야 등 다양한 요인들에 의해 결정된다. 이러한 맥락 속에서, 여러 연구자들은 복잡한 요인 간의 비선형적 관계를 반영하여 특허 침해 소송의 발생 가능성과 위험 수준을 정량적으로 예측하기 위해 머신러닝 모델을 활용했다. 예를 들어, Chien(2011)은 특허 소송 발생여부를 예측하기 위해 로지스틱 회귀모델을 구축하였다. 입력 변수로는 특허의 내재적 특성인 인용 수, 청구항 수, 특허 범위 외에도, 출원 후 발생하는 획득 특성으로서 양도 여부, 재심사 여부, 유지기간, 인용 변화량 등을 포함하였다. Wongchaisuwat *et al.*(2017)은 특허의 정량정보와 텍스트 데이터를 결합하고, 출원인의 재무 정보를 포함한 다양한 입력 변수를 활용하여 소송 발생 가능성과 소송 시점을 예측하는 머신러닝 기반 모델을 제안하였다. Kim *et al.*(2021)은 정량적 특허지표와 특허 텍스트 임베딩 정보를 결합하여, 랜덤 서바이벌 포레스트를 활용한 생존 분석 모델을 구축하고, 특허 소송 발생 시점을 예측하는 기법을 제안하였다.

선행 연구들은 특허 소송 위험을 이해하고 소송 가능성을 예측하는 데 중요한 진전을 이루었으나, 여전히 해결해야 할 과제가 남아 있다. 첫째, 대부분 소송 발생 여부를 이분법적으로 분류하는 데 초점을 맞췄다. 이러한 접근법은 소송 발생 가능성을 대략적으로 파악하는 데 유용하나, 고위험 특허와 저위험 특허를 세분화하여 특허 소송 위험을 정량적으로 평가하는 데에는 한계가 있다. 둘째, 특허 텍스트 정보와 내부 지표에 주로 의존하며, 특허권자의 소송 이력이나 보유 특허 포트폴리오와 같은 정보를 충분히 반영하지 못했다. 소송을 제기하는 주체의 특성을 고려하는 것은 특허 소송 위험의 정밀한 예측과 대응 전략 수립에 중요한 요소이다(Reitzig *et al.*, 2007). 마지막으로, 기존 연구들은 소송 발생 가능성에 대한 예측 모델 개발에 중점을 두었으나, 소송 위험에 영향을 미치는 주요 요인을 구체적으로 규명하기에는 제한적이었다. <Table 1>에는 선행 연구와 제안된 접근방식의 차이점이 요약되어 있다.

3. 분석절차

본 연구는 설명가능한 머신러닝 기반의 특허 소송 위험도 예측모형을 제안한다. 먼저, 특허의 서지 정보와 침해 소송 이력을 수집하고, 이를 연계하여 통합 데이터베이스를 구축한다. 다음으로, 특허의 기술적 속성과 특허권자의 조직적 특성을 반영할 수 있는 정량적 특허지표들을 정의하고 추출한다. 이후, 다양한 머신러닝 알고리즘을 활용하여 특허 소송의 발생 여부와 위험 수준을 각각 예측하는 2단계 모델을 학습한다. 마지막으로, SHAP 기반의 해석 기법을 적용하여 특허 소송 위험에 영향을 미치는 주요 요인을 식별한다. 연구의 전반적인 절차는 <Figure 1>과 같다.

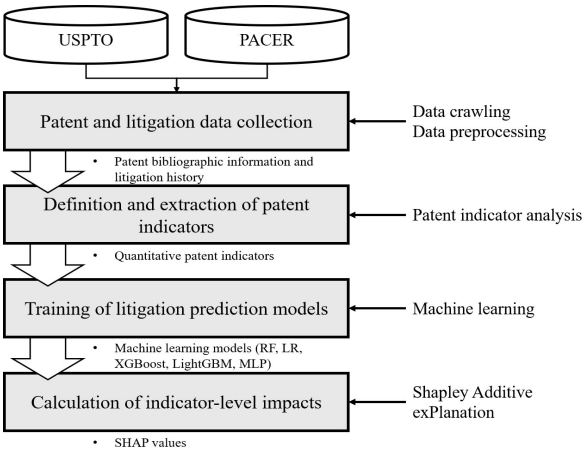

Figure 1. Overall Procedure of the Proposed Approach

3.1 특허 및 소송 데이터 수집

본 접근법은 특허 데이터를 수집하는 것으로 시작된다. 미국, 유럽, 한국 등 주요 국가들은 각국의 특허청을 통해 특허

Table 1. Comparison between Prior Studies and Current Study

Approach	Input	Method	Use of litigation data	Interpretability
Statistics-based approaches	Quantitative patent indicators	Statistical analysis, regression models	Used	Interpretable through regression coefficients
Text-based approaches	Patent texts (claims, specifications, etc.)	Text mining, similarity analysis	Not used	Similarity scores imply potential technological overlap
Machine learning-based approaches	Quantitative patent indicators	Machine learning models	Used	Limited interpretability through feature importance
This study	Quantitative patent indicators, patentee-level indicators	Machine learning, SHAP analysis	Used	Identification of litigation risk factors through SHAP analysis

데이터를 일반 대중들에게 공개하고 있다. 특히 미국의 USPTO는 발명자 및 출원인 정보를 명확히 식별할 수 있도록 체계적으로 정리된 데이터를 제공한다. 또한, USPTO는 미국 특허 데이터 분석 플랫폼인 PatentsView를 통해 방대한 미국 특허 데이터를 정기적으로 공개하고 있다. 이에 따라, 본 연구는 PatentsView를 통해 미국 특허 데이터를 수집하며, 그 대상은 1976년부터 등록된 미국 특허 전체이다. 수집된 데이터에는 특허 제목, 초록, 발명자, 출원인, 인용, CPC(Cooperative Patent Classification) 코드 등 다양한 서지적 정보가 포함되며, 분석 목적에 맞추어 항목별로 정제 및 구조화하고, 일관된 형식의 메타데이터로 재구성한다.

한편, 특허 소송 데이터는 미국 연방 법원의 전자소송정보시스템인 PACER를 기반으로 수집한다. PACER는 미국 연방법원에서 제기된 다양한 소송 사건에 대한 기록을 포함하고 있으며, 본 연구는 이 중 특허와 관련된 사건들에 주목한다. 각 소송 사례는 사건 번호, 제소일, 원고 및 피고 정보, 소송에 포함된 특허 번호, 사건 개요 등의 정보를 포함하며, 사건 단위로 제공되는 데이터를 분석 가능하도록 특허 단위로 변환한다. 하나의 사건에 복수의 특허가 포함된 경우, 각 특허를 개별 항목으로 분리하여 일대일 대응이 가능한 구조로 재구성한다. 이를 통해 특정 특허가 실제로 소송에 연루된 적이 있는지를 식별할 수 있으며, 이후 예측모형의 학습 데이터로 활용한다. 전체 특허 소송에는 다양한 사건 유형이 포함되어 있으며, 이 중에는 무효 판단, 권리 이전 분쟁, 행정적 청구, 외국 특허 관련 소송 등 기술 침해와 직접적인 관련이 없는 유형도 포함되어 있다. 본 연구는 특허 소송 중 특허 침해 사례에 초점을 두고 있으므로, 사건 유형이 '1--Patent Infringement suit non-DJ'로 분류된 사례만을 분석 대상으로 한정한다. 이는 특허권자가 원고로서 특허 침해를 직접 주장한 사건에 해당하며, PACER 소송 데이터 내 명시된 'case_type' 항목을 통해 정확하게 선별할 수 있다.

이와 같이 수집 및 정제된 특허 및 소송 데이터는 본 연구의 전반적인 분석 체계의 기반이 된다. 특허 데이터는 개별 기술의 내재적 특성을 정량화하는 데 활용되며, 소송 데이터는 특허권자 단위의 소송 이력을 축적하여 조직적 특성을 파악하는 데 사용된다. 두 데이터는 특허 등록번호를 기준으로 연계되며, 이를 통해 각 특허의 기술적 특성과 해당 출원인의 과거 소송 이력을 통합적으로 분석할 수 있는 구조를 갖춘다. 특히, 침해 소송에 연루된 특허만을 선별하여, 각 특허에 실제 소송 발생 여부에 따른 이진 라벨을 부여하고, 이를 기반으로 특허 소송 위험 분석을 위한 분류 문제를 정의한다. 결과적으로 본 연구는 이러한 통합 데이터셋을 바탕으로, 특허 소송 위험의 다양한 원인을 식별하고, 침해 발생 가능성에 대한 정밀한 예측을 가능하게 하는 분석 기반을 구축한다.

3.2 특허지표 정의 및 추출

본 단계에서는 특허 소송 예측모형 학습을 위한 특허지표들을 정의하는 과정으로, 특허의 기술적 특성과 특허를 보유한 특허권자의 조직적 특성을 반영하는 정량적 지표들을 추출한 뒤 이를 예측모형에 활용한다. 선행 연구를 검토하여, <Table 2>에 요약된 바와 같이 21개의 특허지표를 정의했다. 구체적으로, 기술적 특성에 해당하는 지표는 (1) 기술 범위 (2) 참신성 (3) 과학 집약도 (4) 우선권 범위 (5) 개발 노력의 다섯 가지 차원으로 구분된다. 한편, 조직적 특성을 반영하는 지표는 (1) 기술 보유 역량 (2) 소송 강도 차원으로 구분되며, 특허권자의 특허 포트폴리오 규모 및 과거 소송 이력을 기반으로 구성된다.

첫 번째 범주인 내재적 기술 특성의 기술 범위는 특정 특허 기술이 얼마나 다양한 기술 분야에 적용되거나 확장될 수 있는지를 나타낸다. Main Class Diversity와 Subclass Diversity는 각각 해당 특허가 포함된 IPC 메인 클래스 및 서브클래스의 개수를 나타내며, 기술 응용 범위의 넓이를 파악할 수 있는 지표이다(Khachatryan and Muehlmann, 2019). Technological Domain은 특허가 IPC의 섹션 A부터 H까지 섹션에 속한 여부를 이진 변수로 표시하여 기술의 다분야 적용성을 평가한다(Park and Yoon, 2017). Commercial Scope는 동일 발명이 등록된 국가 수를 통해 기술의 상업적 활용 범위를 나타낸다. Abstractive Technological Scope는 기술 설명의 추상성 및 서술 밀도를 간접적으로 나타내며, Core Technological Scope와 Peripheral Technological Scope는 기술 보호의 범위와 확장 가능성을 각각 보여준다(Fischer and Leidinger, 2014). Core Technological Complexity는 청구항의 해석 난이도와 기술적 복잡성을 정량화하는 지표로 활용된다(Lanjouw and Schankerman, 2004). 다음으로, 참신성은 특허가 참조한 선행 특허의 수인 Prior Knowledge를 통해 기술 발전의 축적 정도와 기존 기술 대비 차별성을 나타낸다(Reitzig, 2003). 이는 기술의 독창성과 발명의 새로운 기여도를 평가하는 데 중요한 역할을 한다. 과학 집약도는 비특허 문헌의 인용수인 Scientific Knowledge를 바탕으로 측정되며, 기술의 과학적 기반과 학술적 신뢰도를 간접적으로 드러낸다(Allison and Lemley, 1998). 우선권 범위는 특허의 우선권이 설정된 국가 수인 Patent Priority Scope를 통해 기술이 확보한 국제적 권리 보호의 범위를 나타낸다. 이는 다국적 시장에서의 기술 활용도와 함께, 국가 간 법적 해석 차이로 인한 분쟁 가능성까지 고려할 수 있게 한다(Dechezleprêtre *et al.*, 2017). 개발 노력은 기술 개발 과정의 협업 구조와 기술 문서의 완성도를 반영한다. Collaboration Degree는 해당 특허의 출원인 수를 통해 협력의 규모를, International Collaboration은 출원인이 속한 국가 수를 통해 협업의 국제화 정도를 보여준다(Guellec and de la Potterie, 2000). Inventor Efforts는 발명자 수를 통해 투입된 인력 규모와 전문성 수준을 나타내며, Foreign Inventor Efforts는 발명자의 국적 수를 통해 다국적 공동개발 여부를 드러낸다. Examination Time은 공개일과 출원일 사이의 기간을 측정하여 기술 설명의 명확성과 복잡성에 따른 행정 처리 소요를 반영한다(Allison and Tiller, 2003).

두 번째 범주인 조직적 특성의 기술 보유 역량은 소송 시점을 기준으로 출원인이 보유한 전체 특허 수인 Technological

Table 2. Summary of Quantitative Patent Indicators Used as Input Variables

Category	Dimension	Indicator	Description
Intrinsic technological characteristics	Technological scope	Main class diversity (MCD)	Number of IPC codes at the main-class level
		Subclass diversity (SCD)	Number of IPC codes at the subclass level
		Technological domain (TD)	1 if the patent belongs to each IPC section from A to H, and 0 otherwise
		Commercial scope (CS)	Number of patents registered in multiple countries for the same invention
		Abstractive technological scope (ABS)	Number of words in the abstract
		Core technological scope (CTS)	Number of independent claims
		Core technological complexity (CTC)	Average number of words in the independent claims
		Peripheral technological scope (PTS)	Number of dependent claims
	Novelty	Prior knowledge (PK)	Number of backward citations
	Scientific intensity	Scientific knowledge (SK)	Number of non-patent literature citations
	Priority scope	Patent priority scope (PPS)	Number of countries in which the patent holds priority
	Development efforts	Collaboration degree (CD)	Number of applicants
		International collaboration (IC)	Number of countries to which the applicants belong
		Inventor efforts (IE)	Number of inventors
		Foreign inventor efforts (FIE)	Number of countries to which the inventors belong
		Examination time (ET)	Time lag between the application and publication dates of the patent
Organizational characteristics	Technological capability	Technological capability of patentee (TCP)	Average number of patents held by the patentee at the time of litigation
	Immediate litigation intensity	1-year litigation intensity (LI_1)	Average number of the patentee's litigation cases during the one year preceding each lawsuit involving the patent
	Short-term litigation intensity	3-year litigation intensity (LI_3)	Average number of the patentee's litigation cases during the three years preceding each lawsuit involving the patent
	Mid-term litigation intensity	5-year litigation intensity (LI_5)	Average number of the patentee's litigation cases during the five years preceding each lawsuit involving the patent
	Long-term litigation intensity	7-year litigation intensity (LI_7)	Average number of the patentee's litigation cases during the seven years preceding each lawsuit involving the patent

Capability of Patentee를 측정하여, 해당 기업 또는 기관의 기술 자산 축적 수준과 기술 기반의 전략적 활용 가능성을 나타낸다(Coombs and Bierly III, 2006). 마지막으로, Litigation Intensity는 소송 시점에서 1년, 3년, 5년, 7년 동안 해당 출원인이 연루된 특허 소송 횟수(LI_1~LI_7)를 각각 측정함으로써, 조직의 소송 빈도와 대응 성향을 정밀하게 파악할 수 있도록 한다(Wongchaisuwat *et al.*, 2017). 본 연구에서는 기술적·조직적 특성을 정량화한 지표들을 특허의 소송 위험도를 다차원적으로 평가하기 위한 기초 자료로 활용하며, 이를 통해 소송 위험의 구조적 요인을 규명하는 데 활용하고자 한다.

3.3 특허 소송 예측모형 학습

본 단계에서는 특허 소송 위험을 정량적으로 평가하기 위한 예측모형 학습 과정을 설명한다. 본 연구는 Multi-Layer Perceptron(MLP), Random Forest(RF), XGBoost, Logistic

Regression(LR), Light Gradient-Boosting Machine(GBM) 등 총 5가지 머신러닝 알고리즘을 활용하여 특허의 잠재적 소송 발생 가능성과 위험 수준을 예측한다. 이 모형들은 각기 상이한 학습 구조와 데이터 처리 방식에 기반하고 있어, 데이터의 특성과 예측 목적에 따라 다양한 성능 차이를 보일 수 있다. 따라서 모형 간 성능을 비교·분석함으로써 최적의 접근 방식을 도출하는 것을 목표로 한다.

본 연구는 두 단계 방식으로 특허 소송 위험도를 평가한다(<Figure 2>). 첫 번째 단계에서는 전체 특허를 대상으로 침해 소송 발생 여부를 판단하는 이진 분류를 수행하며, 소송이 발생할 확률에 따라 특허를 두 가지 등급(L, NL)으로 분류한다. 첫 번째 단계에서 개발하는 모형은 개별 특허의 소송 발생 가능성을 확인할 수 있다. 두 번째 단계에서는 소송 발생군(L)로 분류된 특허를 대상으로 침해 소송 빈도가 낮은 특허는 저위험군(L2), 반복적으로 소송에 연루되는 특허는 고위험군(L1)으로 정의하여 특허를 분류한다. 이는 동일한 특허가 반복적으로 침해 소송에 연루된다는 사실이, 해당 기술이 타인에 의해 권리 범위가 침해 논란을 일으키기 쉬운 구조를 가졌다고 해석할 수 있기 때문이다. 두 번째 단계에서 개발하는 모형은 개별 특허의 소송 위험의 심각도를 확인할 수 있다.

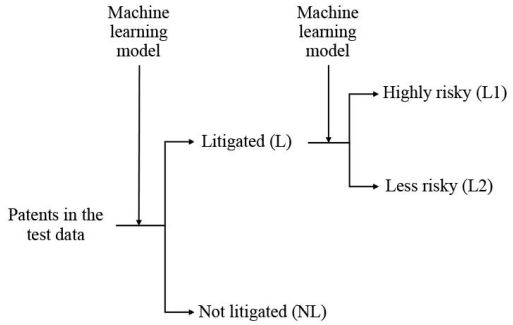

Figure 2. Description of the Data and Machine Learning Models for Two Primary Analyses

특허 소송 위험도 평가를 위한 5가지 머신러닝 모형의 성능은 Confusion Matrix가 생성된 후, k-Fold Cross-validation 기법에 기반한 정량적 성능 평가 지표를 사용하여 평가한다. k-Fold Cross-validation은 데이터를 동일한 크기의 k개로 분할하고 훈련과 검증을 k번 반복한다(Hastie, 2009). Confusion Matrix는 모형이 각 클래스의 인스턴스를 얼마나 정확하게 또는 잘못 예측했는지 보여주는 표이며, 행렬의 행은 실제 값을 나타내고 열은 예측된 값을 나타낸다(<Figure 3>). 이때, True Positive(TP), True Negative(TN)은 올바른 예측 횟수를 나타내며, TP는 실제 소송이 발생했거나 고위험으로 분류되는 특허를 정확히 예측한 경우, TN은 실제 비소송 혹은 저위험 특허를 정확히 예측한 경우를 의미한다. 반면, False Positive(FP), False Negative(FN)은 오류 횟수를 나타내며, FP는 실제로는 위험도가 낮거나 소송이 발생하지 않았음에도 고위험 혹은 소송특허로 잘못 예측한 경우이며, FN은 실제 고위험 혹은 소송특허를 반대로 오분류한 경우를 의미한다.

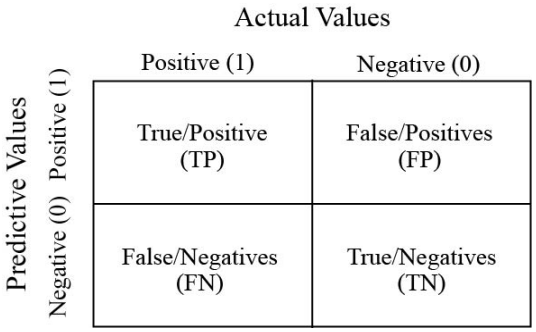

Figure 3. Confusion Matrix

본 연구에서는 Confusion matrix를 통해 Accuracy, Precision, Recall, F1-score, F2-score를 계산하여 모형의 성능을 평가한다(수식 (1)~(5)). 이때, accuracy는 전체 예측 중 올바른 예측의 비율을 의미한다. Precision은 전체 양성 예측 중 정확한 양성 예측의 비율이며, recall은 전체 양성 레이블 중 정확한 양성 예측의 비율을 의미한다. F1-score는 Precision과 Recall의 조화 평균을 의미하며, Fβ-score는 β 파라미터를 활용하여 Precision과 Recall의 가중치를 조절한 F1-score의 변형 상태를 의미한다. 예를 들어, F2-score는 Recall을 Precision보다 두 배 더 중요하게 여기는 경우로, False-negative Rate을 최소화할 때 분류 성능을 적절히 평가할 수 있는 지표다. 본 연구에서는 특허 소송 위험 관리에서 오류 유형 간 비용의 비대칭성을 고려하여 F2-score를 주요 평가 지표로 설정한다. 특허 소송이 실제 발생할 경우, 기업은 건당 수백만 달러에 달하는 손해배상 비용은 물론, 법률 비용, 제품 상용화 지연, 경영 자원의 긴급 투입 등 상당한 규모의 직간접적 피해를 입을 수 있다(Bessen and Meurer, 2012; PricewaterhouseCoopers, 2014). 반면, 안전한 특허를 위험하다고 오판할 경우(False Positive), 추가적인 모니터링이나 법률 검토 비용이 발생하지만 이는 상대적으로 관리 가능한 수준의 비용이다. 따라서 기업의 의사결정 관점에서 False Negative의 최소화가 우선되며, 이를 반영하여 Recall에 더 높은 가중치를 부여하는 F2-score를 주요 평가 지표로 채택한다.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$\text{F1-score} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (4)$$

$$\text{F2-score} = (1+2^2) \cdot \frac{Precision \cdot Recall}{(2^2 \cdot Precision) + Recall} \quad (5)$$

그러나 전통적 성능 지표들은 클래스 분포가 한쪽으로 치우친 경우, 불균형 샘플에 대한 성능을 반영하는 데 한계가 존재할 수도 있다. 다수 샘플을 한 클래스로 분류할 경우, 과대평가의 함정에 빠질 수 있기 때문이다. 이를 보완하기 위해, 본 연구에서는 불균형 데이터셋의 모형 평가에 주로 사용되는 지표인 Balanced Accuracy, Geometric Mean(G-mean), Matthews Correlation Coefficient(MCC)를 추가로 활용한다. Balanced Accuracy는 Sensitivity(TPR; True Positive Rate)와 Specificity (TNR; True Negative Rate)의 평균을 통해 두 클래스를 균형적으로 평가하며, 클래스 분포가 불균형할 때 단순 Accuracy가 주는 편향을 완화하는 효과가 있다(Brodersen *et al.*, 2010).

$$Sensitivity\,(TPR) = \frac{TP}{TP+FN} \tag{6}$$

$$Specificity\,(TNR) = \frac{TN}{TN+FP} \tag{7}$$

$$Balanced\,Accuracy = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{8}$$

G-mean은 TPR과 TNR 기하평균으로, 한쪽 클래스에 치우친 예측을 수행하는 모형이 높은 점수를 받지 못하도록 설계되어 있다. 즉, 양성과 음성 모두에서 일정 수준 이상의 분류 능력을 확보해야 G-mean이 높아지도록 한다. 또한, MCC는 모형 예측의 전반적인 상관관계를 단일 지표로 나타내는 방법으로, 1에 가까울수록 완벽한 분류, 0이면 무작위 분류, -1이면 정반대 분류를 의미한다. 즉, MCC가 높다는 것은 TP, TN, FP, FN의 모든 요소에서 예측 성능이 고르게 우수하다는 것을 나타내며, 이는 극도로 희소한 소송 사례를 다루는 과정에서도 예측의 정확성과 편향 정도를 종합적으로 파악할 수 있다.

$$G-mean = \sqrt{TPR \times TNR} \tag{9}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{10}$$

3.4 특허지표별 영향도 산출

본 절에서는 설명가능한 머신러닝 기법인 SHAP을 학습된 예측모형에 적용하여 각 특허지표별 영향도를 정량적으로 산출한다. 이전 절에서 도출된 최적의 예측 모형에 대해 테스트 데이터셋을 기반으로 SHAP 값을 계산한다. SHAP 값은 개별 데이터에 대해 각 변수들이 모형 예측에 얼마나 영향을 미쳤는지를 나타내는 값으로, 모든 테스트 데이터셋에 대해 산출한다. 이를 통해 예측 결과에 대한 입력 변수의 기여도를 직관적으로 해석할 수 있으며, 소송 예측에 각 특허의 어떤 특성이 결정적인 역할을 했는지를 설명할 수 있다. <Figure 4>는 개별 데이터의 SHAP 값을 시각화한 것이다. 평균 예측값(2.215)을 기준으로, 해당 특허의 예측값(2.846)에 도달하기까지 각 입력 변수가 어떻게 기여했는지를 보여준다. 그림에서는 예측값을 높이는 방향으로 기여한 변수는 빨간색, 낮추는 방향으로 기여한 변수는 파란색으로 시각화된다. 예를 들어, MedInc 변수의 값이 6.232인 경우, 해당 데이터에 대한 예측 결과에 가장 큰 영향을 미쳤으며, 예측값을 높이는 방향으로 기여한 것을 확인할 수 있다.

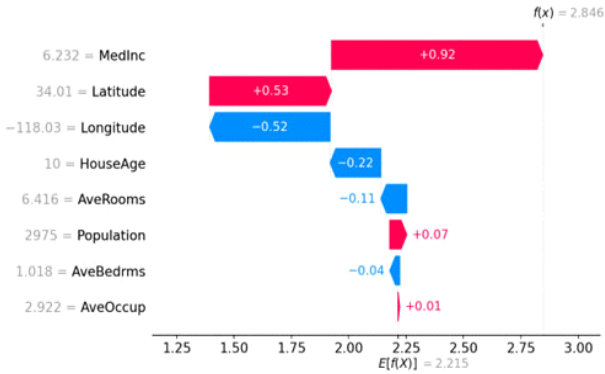

Figure 4. Example of deriving SHAP Values

SHAP 값을 기반으로 산출된 특허지표별 영향도는 각 지표가 예측모형의 의사결정에 기여한 방향성과 크기를 모두 포함한다. 이는 개별 특허의 특성이 특허 소송 위험 예측에 미친 영향을 정량적으로 해석할 수 있게 하며, 기술 특성과 소송 위험 간의 복잡한 연관관계를 설명 가능하게 만든다. 결과적으로, 본 접근법은 전문가가 소송 위험도를 평가하거나 대응 전략을 수립할 때, 의사결정을 보다 정량적이고 체계적으로 보완할 수 있는 해석 도구로 활용될 수 있다.

4. 분석 결과

4.1 데이터

본 연구는 특허의 소송 여부를 구분하기 위해 특허 데이터베이스와 특허 소송 사례 데이터를 활용하였다. 구체적으로, 1976년부터 2016년까지 등록된 6,150,348개의 특허 모집단을 포함하는 USPTO의 데이터셋을 수집했다. 이 데이터셋에는 특허의 청구항, 특허 패밀리, 출원인 정보, 인용 정보 등 특허 등록 시점에서의 다양한 서지 정보가 포함되어 있다. 다음으로, 각 특허의 소송 발생 여부를 식별하기 위해 1976년부터 2016년까지의 특허 소송 사례 총 81,335건이 수록된 PACER의 소송 사례 데이터셋을 사용했다. 데이터의 각 항목은 <Table 3>에 요약된 대로 사건 번호 및 날짜, 원고/피고, 관련 특허번호와 같은 세부 정보가 포함되어 있다. 전처리 과정에서는 먼저 소송 사례에 포함된 특허 등록번호를 기준으로 USPTO 데이터와 매칭하여 소송 여부를 라벨링 하였다. 최종적으로, 과거에 소송이 한 번이라도 발생한 사례가 있는 소송 특허 총 33,603개를 추출했다.

Table 3. Example of Patent Litigation Data

Case_ number	Date_filed	Patent	…	Party_type	Name
0:03-cv-00868	2003-01-16	6384088	…	Defendant	Dr Martin Hinz
0:03-cv-00968	2003-01-30	5600912	…	Plaintiff	Magnum Research, Inc
0:03-cv-00993	2003-02-03	5108321	…	Plaintiff	Crestliner, Inc.
…	…	…	…	…	…
2:16-cv-08559	2016-11-16	9049283	…	Defendant	Ispeaker Co Ltd
2:16-cv-01430	2016-12-20	6825780	…	Plaintiff	Codec Technologies LLC

이후, 특허 침해 위험을 정량적으로 예측하기 위한 두 가지 단계의 모형을 구성하였다. 본 연구의 첫 번째 모형에서는 전체 특허 데이터를 대상으로 소송 발생 여부를 예측하는 이진 분류 모형을 학습하였다. 이를 위해, 전체 특허 데이터를 먼저 8:2 비율로 무작위 분할하여 학습용과 테스트용 데이터셋을 구성하였다. 이때, 테스트 셋에는 원본 분포(소송 특허 대 비소송 특허의 실제 비율 1:168)를 유지하여 향후 평가 시 현실적인 소송 발생 비율을 반영하도록 하였다. 한편, 학습 데이터셋에서는 극단적인 클래스 불균형을 완화하기 위해 소송 특허와 비소송 특허를 1:1 비율로 맞추는 언더샘플링을 수행하였다. 언더샘플링은 다수 클래스의 샘플 수를 소수 클래스 수준으로 축소하여 클래스 간 균형을 맞추는 기법으로, SMOTE 등 오버샘플링 기법에 비해 합성 데이터 생성에 따른 과적합 위험이 낮고, 학습 효율성 측면에서도 이점이 있다(He and Garcia, 2009). 이처럼 학습 시에는 균형 비율을 적용하여 모델이 소수 클래스의 패턴을 충분히 학습할 수 있도록 하되, 테스트 시에는 실제 비율을 유지함으로써 현실 환경에서의 예측 성능을 검증하고자 하였다.

본 연구의 두 번째 모형은 특허 소송의 발생 빈도를 통해 소송 위험도를 추가로 구분하기 위해 설계되었다. 구체적으로, 이미 소송이 발생한 특허 중에서 양수인 정보 추출이 가능한 특허를 식별하고, 소송 발생 횟수가 1~2회인 경우를 저위험, 3회 이상인 경우를 고위험으로 설정하였다 (<Table 4>).

Table 4. Summary of the Datasets

Analysis level	Dataset	Patent category	Number of patents
L vs. NL	Training set	Litigated	26,882
		Not litigated	26,882
	Test set	Litigated	6,721
		Not litigated	1,127,616
L1 vs. L2	Training set	Highly risky	7,860
		Less risky	5,764
	Test set	Highly risky	1,965
		Less risky	1,441

4.2. 특허 소송 위험 추정

<Table 5>와 같이, 본 연구에서 추출한 특허 지표들은 예측 모형의 입력 변수로 사용되었다. 이때, 특허권자 관련 지표는 실제 소송이 발생한 시점에서의 조직적 특성을 파악하기 위한 것이므로, 소송 이력이 전무한 특허에는 해당 정보가 존재하지 않아 표 내에 “-”로 표시되었다. 본 연구에서는 이러한 정량적 특허지표를 바탕으로, 특허 소송 위험을 더욱 정교하게 평가할 수 있는 모형을 구축하였다.

머신러닝 기반의 특허 소송 예측모형은 Python의 ‘TensorFlow’, ‘Scikit-learn’, ‘XGBoost’, ‘LightGBM’ 라이브러리를 활용하여

Table 5. Sample of the Input Data

Reg no	Patent indicators									Category	
	MCD	SCD	…	CTC	PTS	ABS	…	LI_5	LI_7		
3930280	2	2	…	167	11	1	…	-	-	NL	-
4092518	1	1	…	92	2	1	…	0	0	L	L2
4383272	1	1	…	80	18	1	…	2	2	L	L1
…	…	…	…	…	…	…	…	…	…	…	…
6978597	1	2	…	135	5	1	…	-	-	NL	-
7030152	1	1	…	86	5	1	…	36	54	L	L1
…	…	…	…	…	…	…	…	…	…	…	…
8945621	1	1	…	216	12	2	…	52	62	L	L2

Table 6. Hyperparameters for Grid Search

Algorithm	Hyperparameter candidates
RF	Number of estimators (100, 200, and 300); Maximum depth (None, 10, and 20); Minimum samples split (2 and 5); Maximum feature (sqrt and log2)
LR	Penalty (L1 and L2); Regularization strength (0.01, 0.1, 1, 10, 100, 1000); Solver (liblinear and saga); Maximum iteration (100, 300 and 500)
XGBoost	Number of estimators (50, 100, and 200); Learning rate (0.05, 0.1, and 0.2); Maximum depth (3, 5, and 7); Minimum child node weight (1 and 3)
LightGBM	Number of estimators (100, 300, and 500); Learning rate (0.01, 0.05, and 0.1); Number of leaves in a tree (31, 63, and 127); Maximum depth (-1, 10, and 20)
MLP	Number of hidden layers (1, 2, and 3); Number of nodes in a hidden layer (64, 128, and 256); Dropout rate (0, 0.2, and 0.5); Learning rate (0.001, 0.005, and 0.01); Batch size (128 and 256)

구현되었다. 각 알고리즘은 다양한 하이퍼파라미터 구성을 지원하기 때문에, 각 알고리즘에 대해 Grid search와 5 Fold Cross-validation을 수행하여 다양한 하이퍼파라미터 조합에 대한 모형 성능을 평가했다. Grid search는 미리 정의된 모든 하이퍼파라미터 값을 체계적으로 테스트하여 모형 성능을 극대화하는 최적의 구성을 식별한다(Bergstra and Bengio, 2012). 각 알고리즘에 대한 하이퍼파라미터 후보 값은 <Table 6>에 제시되어 있다.

Grid Search를 거쳐 선정된 하이퍼파라미터로 학습된 머신러닝 모형들의 성능은 <Table 7>에 제시된 종합적인 평가 지표를 통해 확인하였다. 전반적으로 L과 NL을 분류하는 1단계에서는 모든 모형이 낮은 Precision이라는 공통된 경향을 보였으며, 이는 극도로 불균형한 데이터 환경에서 소수 클래스(소송특허)를 최대한 탐지하기 위해 모형의 예측 범위를 넓힌 결과, 실제 소송 이력이 없는 특허까지 소송특허로 오분류하는 경우가 증가한 결과로 해석할 수 있다. 이 중 LightGBM과 XGBoost가 상대적으로 양호한 결과를 보여주었으며, 소송이 드물게 발생하는 소송 발생을 일정 수준 이상 식별해 내는 데 GBM의 접근 방식이 효과적이었음을 시사한다.

한편, 2단계 예측에서는 이미 소송이 발생한 특허 중에서도 소송 빈도가 높은 사례(L1)를 분류해야 하므로, 반복 소송 가능성을 좀 더 정교하게 포착하는 역량이 요구된다. 이때 LightGBM이 가장 우수한 성능을 보였으며, 이는 리프 중심의 세분화된 분기 구조를 통해 고위험 특허의 미묘한 패턴을 효과적으로 구분한 것으로 해석할 수 있다. 결과적으로, 소송 발생 여부 판별과 소송 빈도 기반 위험도 예측이라는 두 과제에서, 데이터 구성과 분류 목적의 차이에 따라 모형별로 상이한 학습 효율을 보였으며, 이를 통해 각 모형이 서로 다른 환경에서 가지는 상대적 이점을 확인할 수 있다.

4.3 주요 동인 분석

SHAP 값의 정확한 계산은 가능한 모든 지표의 하위 집합을 고려해야 하므로 기하급수적인 계산 복잡성을 수반한다. 이러한 이유로 본 연구에서는 트리 기반 모형의 구조를 활용해 SHAP 값을 효율적으로 계산할 수 있는 Tree SHAP을 사용했다. Tree SHAP은 어떤 지표의 하위 집합이 트리의 특정 리프 노드에 도달하는 경로를 재귀적으로 추적함으로써, 계산 복잡도를 지수 수준에서 다항식 수준으로 줄이는 방식이다(Lundberg and Lee, 2017). 이를 통해 본 연구는 모든 특허에 대해 다양한 지표별 SHAP 값을 산출할 수 있었으며, 글로벌 수준에서 모형의 작동 방식과 예측 결과를 효과적으로 해석할

Table 7. Performance Evaluation of Machine Learning Models for Patent Litigation Prediction and Risk Classification

Algorithm	Analysis level	Balanced Accuracy	Precision	Recall	F1 score	F2 score	G-mean	MCC
RF	L vs. NL	0.695	0.015	0.627	0.030	0.070	0.691	0.070
	L1 vs. L2	0.815	0.705	0.740	0.722	0.733	0.812	0.620
LR	L vs. NL	0.692	0.015	0.622	0.030	0.070	0.688	0.069
	L1 vs. L2	0.725	0.620	0.574	0.596	0.583	0.709	0.461
XGBoost	L vs. NL	0.696	0.016	0.618	0.031	0.073	0.692	0.072
	L1 vs. L2	0.809	0.687	0.738	0.711	0.727	0.806	0.605
LightGBM	L vs. NL	0.700	0.016	0.629	0.031	0.073	0.697	0.073
	L1 vs. L2	0.823	0.679	0.775	0.724	0.754	0.821	0.620
MLP	L vs. NL	0.682	0.017	0.563	0.032	0.075	0.690	0.070
	L1 vs. L2	0.717	0.502	0.670	0.574	0.628	0.715	0.400

수 있었다. SHAP 값을 시각화하기 위해 입력 변수가 예측 결과에 어떻게 기여하는 지를 보여주는 두 가지 유형의 플롯을 사용했다. 구체적으로 요약 플롯과 막대 플롯을 사용하면 지표들의 기여도를 파악할 수 있다. 특히 요약 플롯은 모든 데이터셋에 대한 각 지표의 SHAP 값 분포를 시각화하며, 막대 플롯은 각 지표의 절대SHAP 값의 평균을 표시하여 예측에서 어떤 지표가 상대적으로 더 중요한 역할을 하는지 직관적으로 나타낸다. <Figure 5>는 이러한 요약 플롯과 막대 플롯을 통해, L 대 NL 분류와 L1 대 L2 분류 각각에서 어떠한 지표들이 모형 예측에 큰 영향을 미쳤는지 시각적으로 보여준다.

먼저 1단계(L vs. NL)에서 PK, CS, PTS, CTS는 특허 소송이 발생할 가능성에 영향을 미치는 주요 지표로 식별되었다. <Figure 5>의 (a)와 (b)에서 볼 수 있듯이, CS(Commercial Scope)를 제외한 나머지 주요 지표들은 SHAP 값과 양의 관계를 갖고 있으며, 해당 지표들의 값이 클수록 소송 발생 확률을 높이는 방향으로 기여하고 있음을 나타낸다. 예를 들어, PK(Prior Knowledge)가 높다는 것은 특허가 기술 축적으로 인해 기존 기술들과의 관계가 밀접함을 의미하며, 이에 따라 권리 충돌 가능성이 커져 침해 소송으로 이어질 확률이 높아진다고 예상할 수 있다. 또한 예측에 대한 기여도는 상대적으로 작지만, MCD(Main Class Diversity) 및 SCD(Subclass Diversity)와 SHAP 값 사이에는 음의 관계를 보였으며, 이는 특허의 클래스 수가 많을수록 소송 발생 가능성이 일부 낮아지는 경향이 있음을 시사한다. 이는 기술 분야가 넓게 분산되어 있을수록 해당 특허의 핵심을 직접 침범하지 않을 수 있으며, 오히려 기술분야가 한정적일수록 해당 기술군 내에서 침해 가능성이 높아짐을 짐작할 수 있다. 한편, ABS(Abstractive tech-

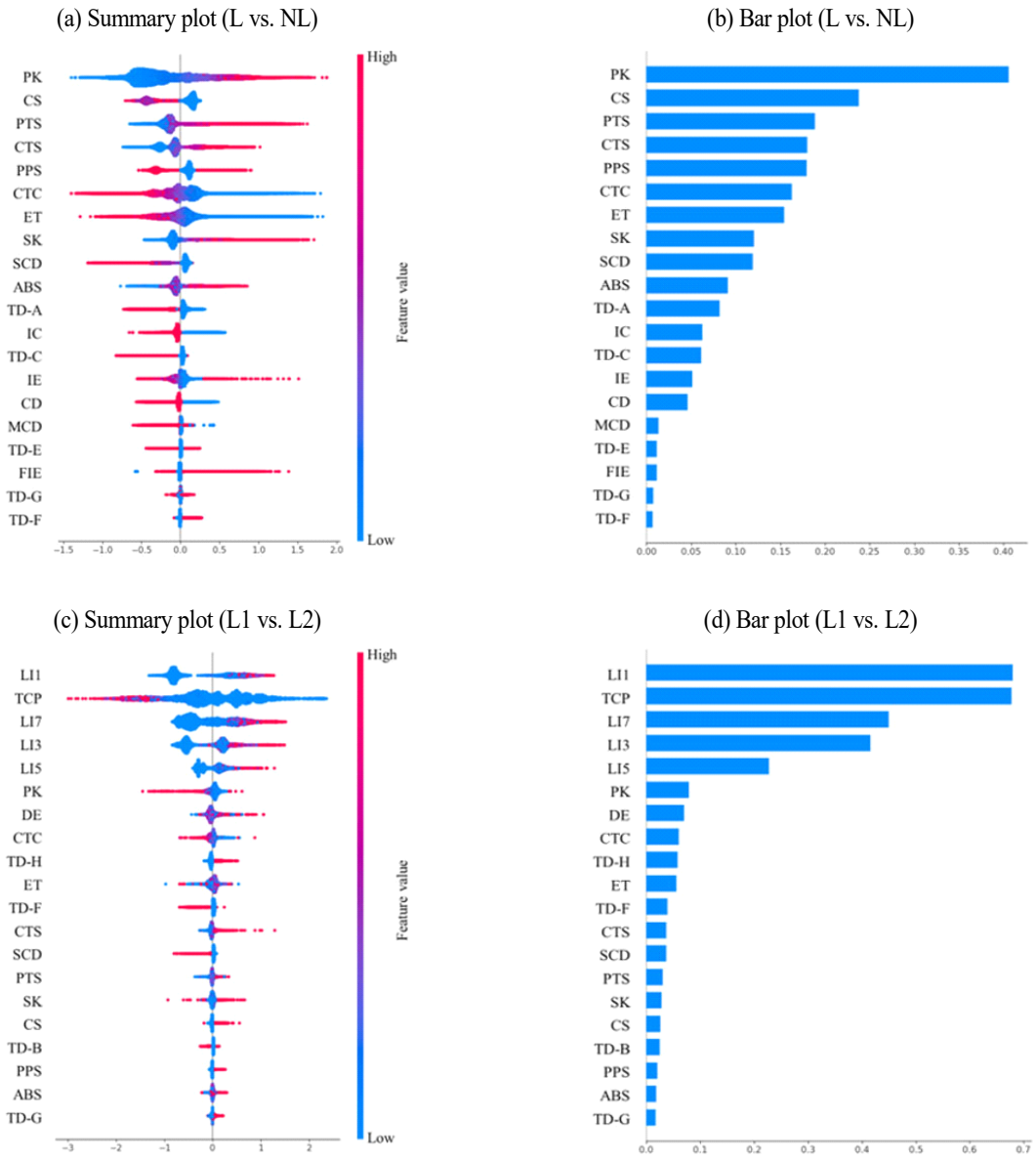

Figure 5. SHAP Summary and Bar Plots of the Proposed Models

nological scope)가 클수록 특허 초록의 길이가 길다는 것을 의미하며, 이는 기술 내용을 보다 상세하고 포괄적으로 설명했을 가능성을 시사한다. 이러한 기술 설명의 포괄성은 권리 범위에 대한 해석 여지를 넓히고, 그 결과 제 3자의 기술과 중첩될 가능성을 높여 특허 침해로 이어질 위험이 증가한 것으로 해석할 수 있다.

L1 대 L2 분류와 관련해서는 LI_1, TCP, LI_3, LI_7과 같은 특허권자 관련 지표들이 고위험 소송 특허를 식별하는 데 중요한 역할을 하는 것으로 나타났다. <Figure 5>의 (c)와 (d)에서 볼 수 있듯이, 단기 소송 이력을 나타내는 LI_1 값이 높을수록 해당 특허가 고위험군으로 분류될 가능성이 크게 증가함을 확인할 수 있다. 반면, TCP(Technological capability of patentee)는 SHAP 값과 음의 상관관계를 가지며 고위험 특허로 분류될 가능성을 오히려 낮추는 요인으로 작용하였다. 특히, LI_1이 LI_3, LI_5, LI_7보다 상대적으로 높은 중요도를 갖는다는 점은, 예측모형이 고위험 여부 판단에 있어 장기적 이력보다 단기 소송 경험을 더 강한 신호로 간주하고 있음을 시사한다.

이러한 결과는 기존의 선행연구에서도 유사하게 관찰된 바 있다. Surdeanu *et al.*(2011)은 특허권자의 과거 소송이력이 향후 침해 소송의 발생가능성에 유의미한 영향을 미친다고 보고했으며, 실제로 반복적으로 소송에 관여한 이력이 있는 특허권자가 향후에도 소송에 연루될 확률이 높다고 분석하였다. 또한, Lanjouw and Schankerman(2004)은 특허권자의 전체 포트폴리오 규모가 클수록 개별 특허의 소송 위험도는 낮아지는 경향이 있다는 점을 실증적으로 제시한 바 있으며, 이는 본 연구에서 소송 당시 특허권자가 보유한 평균 특허 수로 나타나는 특허권자의 기술 보유 역량이 고위험 분류에 부정적인 영향을 미친다는 분석 결과와 일맥상통한다.

표면적으로는 소송 이력이 많은 기업의 특허가 향후에도 고위험군으로 분류되는 결과는 당연하게 보일 수 있다. 이처럼 특허권자 관련 지표만으로도 일정 수준의 위험 예측이 가능하다고 판단될 경우, 복잡한 예측 모형의 필요성에 의문이 제기될 수 있다. 그러나 본 연구의 핵심은 단일 지표의 단편적인 해석을 넘어, 기술적 지표와 조직적 지표 간의 상호작용을 함께 반영하여 소송 위험의 맥락을 정량적으로 예측하고 해석했다는 데에 있다. 특히, 기존 선행연구에서 제시된 이론적 가설들을 실증적으로 재확인했다는 점에서, 본 연구는 예측모형의 설명력을 확보함과 동시에, 실무에서 활용 가능한 수준의 적용 가능성을 보여주었다는 점에서 의의가 있다.

5. 토의

2단계에서 개발한 분류 모형과 테스트 셋을 활용해, 개별 특허의 소송 위험의 심각도를 등급화하였다. 본 연구에서는 소송 위험 수준을 등급화하기 위해 Stanine 등급 시스템을 적용하였다(<Table 8>). Stanine 등급 시스템은 평균 5등급, 표준편차 2등급을 기준으로 데이터를 9등급으로 나누는 방식이다(Salkind, 2006). 실제로 높은 등급에 위치한 특허들은 고위험군과 밀접한 관련성을 보였으며, 반대로 낮은 등급에 위치한 특허들은 침해 소송이 적게 발생하는 경향을 보였다. 이를 통해 다양한 특허들을 대상으로 상대적인 소송 위험도를 더욱 직관적으로 표현할 수 있다. 이와 같은 결과는 소송 위험 등급이 실제 소송 건수와 연관을 갖고 있다는 점을 시사하며, 이는 제시한 연구가 단순 예측 성능 이상의 이론적 확장 가능성을 보여준다.

이와 같은 특허 소송 위험도의 명확한 등급화는 기업이 자사의 특허 포트폴리오를 전략적으로 관리하는 데 유용하게 활용될 수 있다. 예를 들어, 1~2등급에 해당하는 고위험 특허는 장벽 강화나 교차 라이센싱을 통한 분쟁 구조의 재편을 사전에 검토할 수 있다. 나아가, 본 연구에서 제안한 분석 방식은 경쟁사의 특허 소송 위험 프로필을 정량적으로 파악하는 데에도 응용 가능성이 있다. 기업은 이를 바탕으로 경쟁사의 고위험 특허나 핵심 기술을 선제적으로 파악하고, 시장 진입 전략, R&D 방향성, 또는 분쟁 대응 전략 등을 보다 구체적으로 수립할 수 있다. 결과적으로, 본 연구의 접근 방식은 특허 포트폴리오 관리 능력을 강화하여 기업의 특허 및 기술 전략 수립에 활용할 수 있는 의사결정 근거를 제공할 수 있을 것이다(Choi *et al.*, 2020; Choi *et al.*, 2025).

Table 8. Assessment Results of Patent Litigation Risk Levels

Risk grade	Number of patents	Threshold	Actual level	
			Less risky	Highly risky
1	222	0.910301	4	218
2	385	0.726693	56	329
3	660	0.433162	192	468
4	935	0.201640	655	280
5	1,101	0.073132	1,066	35
6	935	0.055531	927	8
7	660	0.045468	658	2
8	385	0.034951	385	0
9	219	0.014468	219	0
Total	5,502		4,162	1,340

6. 결론 및 추후 연구

본 연구는 설명가능한 머신러닝 기반의 특허 소송 예측 프레임워크를 제안함으로써, 특허 소송 위험을 정량적으로 평가하고 그 수준을 체계적으로 등급화하는 방법론을 제시하였다. 기존 연구들이 주로 소송 발생 여부의 이분법적 예측에 집중했던 것과 달리, 본 연구는 특허 소송 위험의 다차원적 맥락에 주목하여 소송 여부(L vs. NL)뿐 아니라, 소송 빈도(L1 vs. L2)를 함께 고려한 이단계 예측 구조를 설계하였다. 이를 위해 특

허의 기술적 속성(청구항 수, 인용 수, 과학 인용 등)과 출원인 중심의 조직적 특성(포트폴리오 규모, 과거 소송 이력 등)을 정량화한 21개의 지표를 도출하고, 이를 다양한 머신러닝 알고리즘에 적용하였다. 모형 성능 검증 결과, LightGBM과 XGBoost가 불균형 데이터 환경에서도 비교적 안정적이고 우수한 예측 성능을 보였으며, 특히 2단계에서는 LightGBM이 F2-score 0.754, G-mean 0.821 수준의 성능을 달성하였다. 이는 실제로 소수에 해당하는 소송 특허를 효과적으로 탐지함으로써, 특허 포트폴리오의 고위험 영역을 선별해 내는 데 유용함을 실증적으로 확인시켜 준다. 또한, 블랙박스 모형의 해석 한계를 극복하기 위해SHAP 기반 설명 기법을 도입함으로써, 각 예측 결과에 대한 입력 특성의 기여도를 정량적으로 분석하였다. 특히 PK, CTS, PTS 등의 기술 지표와 함께, 특허권자의 단기 소송 이력(LI_1)과 기술 보유 역량(TCP)이 예측 결과에 큰 영향을 미친다는 사실이 확인되었으며, 이는 기존 선행연구의 실증 결과들과도 일치함을 보여주었다.

제안한 접근은 특허 소송 위험을 정량적으로 예측하고 그 수준을 체계적으로 등급화함으로써, 기업의 특허 포트폴리오를 더욱 전략적으로 관리할 수 있는 실질적인 분석 기반을 제공한다. 특허의 기술적 속성과 특허권자의 조직적 특성을 통합적으로 고려하여 소송 위험도를 평가한 본 연구는, 기업이 자원과 법적 대응 전략을 더욱 효율적으로 배분하고, 침해 가능성이 높은 특허에 대해 사전 대응책을 수립하는 데 실용적인 인사이트를 제공한다. 아울러, SHAP 기반 분석을 통해 도출된 주요 특허 요인은 기업의 기술 개발 방향성 조정, 경쟁사의 핵심 특허 모니터링, 사전 협상 전략 수립 등 기술 경영 전반에 걸친 전략적 의사결정을 정교화하는 데 기여할 수 있다. 또한, 단기 소송 이력이나 포트폴리오 규모와 같은 특허권자 중심의 변수들이 소송 위험 예측에 중요한 역할을 한다는 결과는 기존의 이론적 논의를 경험적으로 뒷받침하는 동시에, 향후 산업 내 경쟁구조 분석이나 출원인 행태 연구에도 응용 가능성을 보여준다. 나아가, 설명가능한 머신러닝 기법과 공리 기반 설명 일관성 평가를 결합한 본 연구의 분석 체계는 지식재산 분야에서 설명가능한 인공지능의 실무 적용 가능성을 실증적으로 입증했다는 점에서 학문적으로도 의의가 크다. 이러한 분석 구조는 단순히 예측 성능에 그치지 않고, 예측 결과에 대한 투명성과 신뢰성을 확보함으로써 기술 리스크 분석 및 특허 기반 전략 수립을 위한 실질적 의사결정 지원 도구로 활용될 수 있을 것이다.

제안한 연구는 다음 한계점을 가지고 있다. 첫째, 미국의 특허 및 소송 데이터를 기반으로 수행되었기 때문에, 국가별 특허 제도, 소송 절차, 기술 분류 체계 등의 차이로 인해 다른 국가나 글로벌 시장에 일반화하는 데 한계가 있다. 특히, 한국과 같은 국가의 경우, 특허권자의 분쟁 대응 방식이나 소송의 빈도, 공개 데이터의 범위 등이 미국과 다를 수 있기 때문에 동일한 예측 모형을 직접 적용하기에는 제약이 존재한다. 따라서 추후연구에서는 한국을 포함한 다양한 국가 및 지역을 대상으로 예측 모형을 재검증하거나, 해당 국가의 제도적 특성을 반영할 수 있는 맞춤형 모형을 설계함으로써, 글로벌 차원에서 적용 가능한 특허 소송 위험 예측 프레임워크로 발전시킬 필요가 있다. 둘째, 본 연구는 과거 특정 기간에 수집된 소송 데이터를 바탕으로 하고 있어, 최근의 기술 트렌드 변화나 특허 제도, 법적 환경의 최신 동향을 충분히 반영하지 못했을 가능성이 있다. 또한, 최신 소송 데이터를 확보하는 데에는 상당한 비용적 제약이 있어 본 연구에서는 포함하지 못하였으나, 추후연구에서는 보다 최신의 소송 데이터를 통합하여 모형을 주기적으로 업데이트하고, 그 적합성을 재검증하는 과정이 필요하다. 이를 통해 급변하는 기술시장 환경을 더욱 현실적으로 반영하는 예측 프레임워크를 구축할 수 있을 것이다. 셋째, 특허 원문에 포함된 기술적 설명이나 표현의 의미, 시대적 기술 흐름과 같은 내재적 컨텍스트를 충분히 반영하지 못했다는 한계가 있다. 실제로 특허는 시간이 지남에 따라 기술의 초점이 변화하고(Seol *et al.*, 2025), 각 시점별로 기술의 최신성도 달라질 수 있기 때문에, 이러한 변화를 반영하지 않으면 소송 위험도 예측에 오차가 발생할 수 있다. 따라서 추후연구에서는 특허 원문 텍스트를 활용하여 기존 지표와 통합적으로 분석함으로써, 특허의 기술적 의미와 시간적 변화를 반영한 예측 프레임워크로 발전시킬 필요가 있다. 또한, 본 연구는 특허 쌍 간의 기술적 유사도를 직접적으로 산출하여 입력하지는 않았다. 기술적 중첩성이나 청구항 간 유사도는 침해의 본질적 요인으로서 소송 위험 예측의 정밀도를 높일 수 있는 잠재적 변수이나, 수백만 건 규모의 특허 데이터셋에 대해 쌍별(pairwise) 유사도를 산출하는 것은 계산 비용 측면에서 현실적 제약이 존재한다. 따라서 추후 연구에서는 효율적인 유사도 산출 기법을 활용하여 기술적 중첩성 지표를 통합함으로써, 소송 징후 포착과 기술적 침해 가능성 평가를 결합한 예측 프레임워크로 확장할 필요가 있다. 마지막으로, 본 연구는 주로 정량적이고 정적인 특허 지표를 중심으로 분석을 수행하였으나, 특허 소송 위험은 기술 분야의 특성과 시장 경쟁 구조와 같은 맥락적 요인과도 밀접한 관련이 있다. 향후 연구에서는 특허가 속한 기술 분야의 특성, 경쟁 기업 간의 전략적 관계, 시장 내 특허 분쟁 현황과 같은 외부 환경 요소를 추가적으로 고려하여, 예측 모형의 설명력과 실효성을 높이는 방향으로 연구를 확장할 필요가 있다.

참고문헌

Allison, J. R. and Lemley, M. A. (1998), Empirical evidence on the validity of litigated patents, *AIPLA QJ*, **26**, 185.

Allison, J. R. and Tiller, E. H. (2003), The business method patent myth, *Berkeley Tech. LJ*, **18**, 987.

Bergstra, J. and Bengio, Y. (2012), Random search for hyper-parameter optimization, *The Journal of Machine Learning Research*, **13**(1), 281-305.

Bessen, J. and Meurer, M. J. (2012), The private costs of patent litigation,

JL Econ. & Pol'y, **9**, 59.

Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. (2010), The balanced accuracy and its posterior distribution, *2010 20th International Conference on Pattern Recognition.*

Chen, E. H. (2013), Making abusers pay: Deterring patent litigation by shifting attorneys' fees, *Berkeley Technology Law Journal*, **28**, 351-381.

Chien, C. V. (2011), Predicting patent litigation, *Tex. L. Rev.*, **90**, 283.

Choi, J., Jeong, B., Yoon, J., Coh, B.-Y., and 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**, 106544.

Choi, J., Yoon, J., and Lee, C. (2025), Early screening of potential breakthrough technologies with enhanced interpretability: A patent-specific hierarchical attention network model, *Computers & Industrial Engineering*, **203**, 111034.

Coombs, J. E. and Bierly III, P. E. (2006), Measuring technological capability and performance, *R&D Management*, **36**(4), 421-438.

Crampes, C. and Langinier, C. (2002), Litigation and settlement in patent infringement cases, *RAND Journal of Economics*, 258-274.

Cremers, K. (2004). Determinants of patent litigation in Germany.

Cremers, K. (2009), Settlement during patent litigation trials. An empirical analysis for Germany, *The Journal of Technology Transfer*, **34**(2), 182-195.

Dechezleprêtre, A., Ménière, Y., and Mohnen, M. (2017), International patent families: from application strategies to statistical indicators, *Scientometrics*, **111**(2), 793-828.

Fischer, T. and Leidinger, J. (2014), Testing patent value indicators on directly observed patent value: An empirical analysis of Ocean Tomo patent auctions, *Research Policy*, **43**(3), 519-529.

Guellec, D. and de la Potterie, B. v. P. (2000), Applications, grants and the value of patent, *Economics Letters*, **69**(1), 109-114.

Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. (2005), The elements of statistical learning: Data mining, inference and prediction, *The Mathematical Intelligencer*, **27**(2), 83-85.

He, H. and Garcia, E. A. (2009), Learning from imbalanced data, *IEEE Transactions on Knowledge and Data Engineering*, **21**(9), 1263-1284.

Juranek, S., Quan, T., and Turner, J. L. (2020), Patents, litigation strategy and antitrust in innovative industries, *Review of Industrial Organization*, **56**(4), 667-696.

Khachatryan, D. and Muehlmann, B. (2019), Measuring technological breadth and depth of patent documents using Rao's Quadratic Entropy, *Journal of Applied Statistics*, **46**(15), 2819-2844.

Kim, J., Lee, G., Lee, S., and Lee, C. (2022), Towards expert–machine collaborations for technology valuation: An interpretable machine learning approach, *Technological Forecasting and Social Change*, **183**, 121940.

Kim, Y., Park, S., Lee, J., Jang, D., and Kang, J. (2021), Integrated survival model for predicting patent litigation hazard, *Sustainability*, **13**(4), 1763.

Krestel, R., Chikkamath, R., Hewel, C., and Risch, J. (2021), A survey on deep learning for patent analysis, *World Patent Information*, **65**, 102035.

Lanjouw, J. O. and Schankerman, M. (2001), Characteristics of patent litigation: a window on competition, *RAND Journal of Economics*, **32**(1), 129-151.

Lanjouw, J. O. and Schankerman, M. (2003), Enforcing patent rights: an empirical study, LSE Research.

Lanjouw, J. O. and Schankerman, M. (2004), Protecting intellectual property rights: Are small firms handicapped? *The Journal of Law and Economics*, **47**(1), 45-74.

Lee, C., Song, B., and Park, Y. (2013), How to assess patent infringement risks: A semantic patent claim analysis using dependency relationships, *Technology Analysis & Strategic Management*, **25**(1), 23-38.

Lee, J., Oh, S., and Suh, P. (2018), Inter-firm patent litigation and innovation competition, Available at SSRN 3258087.

Lim, J. (2014), Analysis of the relationship between patent litigation and citation: Subdivision of citations, *Applied Mathematics & Information Sciences*, **8**(5), 2515.

Lundberg, S. M. and Lee, S.-I. (2017), A unified approach to interpreting model predictions, *Advances in Neural Information Processing Systems*, 30.

Marco, A. C. and Miller, R. D. (2019), Patent examination quality and litigation: Is there a link? *International Journal of the Economics of Business*, **26**(1), 65-91.

Ming, C., Yu, X., and Zhang, B. (2024), Assessing the infringement risk of patent portfolios using network analysis and IF-TOPSIS: A case of standard-essential patent portfolios in the ICT industry, *Technology in Society*, **78**, 102663.

Park, H., Yoon, J., and Kim, K. (2012), Identifying patent infringement using SAO based semantic technological similarities, *Scientometrics*, **90**(2), 515-529.

Park, Y. and Yoon, J. (2017), Application technology opportunity discovery from technology portfolios: Use of patent classification and collaborative filtering, *Technological Forecasting and Social Change*, **118**, 170-183.

PricewaterhouseCoopers (2014), *2014 patent litigation study: As case volume leaps, damages continue downward trend*, PwC.

Rajwani, C. (2010), Controlling costs in patent litigation, *Journal of Commercial Biotechnology*, **16**(3), 266-271.

Reitzig, M. (2003), What determines patent value?: Insights from the semiconductor industry, *Research Policy*, **32**(1), 13-26.

Reitzig, M., Henkel, J., and Heath, C. (2007), On sharks, trolls, and their patent prey—Unrealistic damage awards and firms' strategies of "being infringed", *Research Policy*, **36**(1), 134-154.

Salkind, N. J. (2006), *Encyclopedia of measurement and statistics*, SAGE publications.

Seol, Y., Choi, J., Lee, S., Son, C., and Yoon, J. (2025), Dynamic technology impact analysis: A multi-task learning approach to patent citation prediction, *Computers & Industrial Engineering*, 111757.

Shapiro, C. (2000), Navigating the patent thicket: Cross licenses, patent pools, and standard setting, *Innovation Policy and the Economy*, **1**, 119-150.

Shin, J. and Park, Y. (2005), Generation and application of patent claim map: Text mining and network analysis, *Journal of Intellectual Property Rights*, **10**(3), 198-205.

Somaya, D. (2012), Patent strategy and management: An integrative review and research agenda, *Journal of Management*, **38**(4), 1084-1114.

Surdeanu, M., Nallapati, R., Gregory, G., Walker, J., and Manning, C. D. (2011), Risk analysis for intellectual property litigation, *Proceedings of the 13th International Conference on Artificial Intelligence and Law.*

Wongchaisuwat, P., Klabjan, D., and McGinnis, J. O. (2017), Predicting litigation likelihood and time to litigation for patents, *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law.*

Yang, D. (2019), Patent litigation strategy and its effects on the firm, *International Journal of Management Reviews*, **21**(4), 427-446.

저자소개

설영진: 건국대학교 산업공학과에서 2021년 학사 학위를 취득하고, 건국대학교 산업공학과 석박사통합과정에 재학 중이다. 주요 연구 분야는 Computational patent analytics for technology opportunities, Technology recommendation method and system, Machine & Deep learning-based system for industrial processes이다.

홍준희: 건국대학교 산업공학과에서 2025년 석사 학위를 취득하고, 한화에어로스페이스 IPS센터에서 재직 중이다. 주요 연구 분야는 Computational patent analytics for technology opportunities, Technology recommendation method and system, Machine & Deep learning-based system for industrial processes이다.

이승현: 건국대학교 산업공학과에서 2026년 박사 학위를 취득하였다. 현재 삼일회계법인 AX Node에서 재직 중이다. 주요 연구 분야는 Web & Patent data analytics for social problem-solving R&D, Natural language processing for business intelligence, Data-driven prognostics and health management이다.

최재웅: 건국대학교 산업공학과에서 2022년 박사 학위를 취득하였다. 고려대학교 행정학과에서 박사후 연구원을 거쳐, 현재 단국대학교 경영경제대학 경영학부에서 재직 중이다. 주요 연구분야는 AI transformation, Applied deep learning, Technology management 이다.

강민수: (주)광개토연구소의 대표로, 글로벌 특허 데이터에 인공지능(AI)과 빅데이터 기술을 결합한 차세대 특허 검색·분석 플랫폼 PatentPia를 개발·운영하고 있다. 전 세계 특허 문헌을 대상으로 기업, 기술 인력, 기술 분야 간의 연결 관계를 계량적으로 분석하고, 기술 동향 탐색, 유망 기술 예측, 혁신 자산 발굴 및 추천을 위한 데이터 기반 의사결정 지원 시스템을 제공하고 있다.

윤장혁: POSTECH 산업공학과에서 학사, 석사 학위를 취득한 후, LG CNS에서 4년간 재직하였으며, POSTECH 산업경영공학과에서 박사 학위를 취득하였다. 한국지식재산연구원을 거쳐 현재는 건국대학교 산업공학과 정교수로 재직 중이다. 주요 연구분야는 대량 데이터 분석 기반의 Business intelligence, Patent analytics, Social media analytics, Industrial artificial intelligence, Healthcare data analytics이다.