Journal of the Korean Institute of Industrial Engineers
Vol. 49, No. 3, pp. 223-236, June 2023.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2023.49.3.223

해석가능 인공지능을 활용한 바이오화학 기술의 비즈니스 잠재성 평가

이지호[1] · 이승현[1] · 손은수[2] · 윤장혁[1†] · 이재민[2†]

[1]건국대학교 산업공학과 / [2]한국과학기술정보연구원 데이터분석본부

Evaluating the Business Potential of Bio-based Chemical Technologies Using Explainable AI

Jiho Lee[1] · Seunghyun Lee[1] · Eunsoo Sohn[2] · Janghyeok Yoon[1] · Jae-Min Lee[2]

[1]Department of Industrial Engineering, Konkuk University
[2]Division of Data Analysis, Korea Institute of Science and Technology Information

Since the fossil fuel-based industry significantly contributes to air pollution and climate change, better living through fossil fuel has come at a cost. In this connection, Bio-based chemical technologies based on reusable biomass such as cells or other living things are receiving great attraction. But at the same time, they are considered as high-risk investments that require a long-term effort to be adopted by businesses. Therefore, building on a common academic consensus that there is a strong correlation between patent lifetime and business potential, this study proposes a machine learning model to predict the lifetime of bio-based chemical technologies. To this end, CAS (Chemical Abstract Service) patent database and PATSTAT (Worldwide Patent Statistical Database) are used to identify global bio-based chemical technology patents. The proposed model identifies bio-based chemical technologies that have high business potential with an accuracy of 81%. Further, the application of an explainable AI algorithm to the model found that the geographical scope of technologies and the size of stakeholders of a business significantly influence the business potential of bio-based chemical technologies. Our research results can be used for the investment and management process for bio-based chemical technologies with high business potential.

Keywords: Business potential, Patent lifetime, Technology management, Machine learning, Explainable AI

1. 서 론

인류의 풍요는 석유화학산업의 발전에 기반한다 하여도 과언이 아닌데, 땅속에 존재하는 원유를 인류에게 필요한 에너지원으로 활용함으로써 인류는 최대의 풍요와 평안을 누리고 있다(Nam, 2015). 하지만, 개도국 중심의 인구증가 및 소득증대는 지속적인 에너지 수요 증가로 이어지며, 이로 인해 석유자원 고갈 논쟁이 심화되고 있다. 또한, 풍요와 평안 이면에 숨겨져 있던 환경 문제, 천연 자원 고갈 문제 등이 수면위로 부상하면서, '굴뚝산업' 탈피에 대한 관심이 그 어느 때보다 높아진 상황이다(Alemu, 2020). 인류의 풍요와 평안을 유지하면서 이면에 숨겨진 다양한 문제를 해결할 수 있는 지속 가능성의 위

본 연구는 2022년도 한국과학기술정보연구원(KISTI) 주요사업, '수요 기반의 과학기술혁신을 위한 미래기술 분석체제 고도화'와 산업통상자원부 수탁과제 '빅데이터기반 차세대 첨단 바이오화학소재 발굴 및 공정설계 최적화 플랫폼 기술개발'의 지원을 받아 수행한 것입니다.
† 연락저자 : 윤장혁 교수, 05029 서울특별시 광진구 능동로 120 건국대학교 산업공학과, Tel: 02-450-0453, Fax: 02-450-3525, E-mail: janghyoon@konkuk.ac.kr
이재민 박사, 02456 서울특별시 동대문구 회기로 66 한국과학기술정보연구원 데이터분석본부 미래기술분석센터, Tel: 02-3299-6292, Fax: 02-3299-6117, E-mail: jmlee@kisti.re.kr
2022년 10월 31일 접수; 2022년 12월 15일 수정본 접수; 2023년 01월 13일 게재 확정.

기에 대응하기 위한 새로운 기술이 필요하다는 의미이다.

이러한 상황에서, 재생가능한 바이오매스 자원을 에너지원으로 하는 바이오화학 기술이 주목받고 있다(Lee, 2011). 바이오화학 기술의 주요 특징은 모든 생산과정에 투입되는 원료가 재생가능한 원료라는 것이다(Seo and Young, 2010). 기존의 기술들과 달리, 살아있는 생명체를 개발의 자원 또는 활용 대상으로 하기 때문에 기술개발을 위한 R&D는 연구실 수준에서 수행되며, 개발된 기술이 제품의 형태로 시장에 나올 때까지 긴 시간과 많은 투자가 요구된다(Choi, 2007). 이는 기술사업화에 대한 불확실성이 높음을 의미하는데, 투자자의 입장에서는 기술의 비즈니스 잠재성을 정량적으로 평가할 수 있는 방법이 필요하며 기술권리자의 입장에서는 개발한 기술의 비즈니스 잠재성에 대한 객관화된 평가 결과를 제공할 방법이 필요한 상황이다.

기술의 비즈니스 잠재성은 타겟 시장의 규모, 기술의 완성도 등 다양한 외부요인에 영향을 받지만, 기술 수명이 기술의 비즈니스 잠재성을 대변할 수 있는 것으로 알려져 있다(Lee *et al.*, 2012). 실제로, 속지주의 원칙에 의거하여 유지료를 지불한 국가에서 기술에 대한 배타적 권리를 최장 20년 보장해주는 특허를 활용한 기술의 비즈니스 잠재성 평가 연구가 수행되었다(Cho, 2019). 특허의 유지료는 시간이 지나면서 점점 높아지기 때문에, 기술이 창출할 수 있는 비즈니스 가치가 높지 않다면 특허에 대한 권리를 포기하는 것이 일반적이다(Lee and Yoon, 2006). 반대로, 기술을 통해 창출가능한 이익이 특허의 유지료보다 높다면, 특허의 권리를 유지하게 된다. 이처럼, 특허의 수명은 기술의 비즈니스 가치와 높은 상관관계가 있으며 (Guellec *et al.*, 2000), 권리가 오래 유지되는 특허일수록 사업화에 활용되는 경우가 많고 발명의 질적수준이 우수한 것으로 알려져 있다(Choi *et al.*, 2020).

따라서 본 연구는 바이오화학 기술을 다루는 특허, 즉, 바이오화학 특허의 수명예측을 통해 비즈니스 잠재성이 높은 특허를 선별하고, 해석가능 인공지능 알고리즘을 통해 비즈니스 잠재성에 영향을 미치는 주요 지표들을 분석하여 기술의 비즈니스 잠재성 평가에 대한 근거를 제시할 수 있는 방법을 제시하고자 한다. 특허수명을 활용하여 기술의 비즈니스 잠재성을 평가하기 위한 연구들이 존재하였으나, 선행연구들은 기술분야의 수명을 주로 분석하였다(Yoo, 2004). 즉, 기술분야에 속하는 개별기술의 비즈니스 잠재성 보다는 기술분야 자체의 수명주기 분석에 초점을 맞춘 것으로, 본 연구가 제시하고자 하는 바이오화학 기술의 비즈니스 잠재성 평가에 활용되기는 적합하지 않다. 또한 선행연구는 특정 국가에 등록된 특허의 수명만을 활용하여 비즈니스 잠재성을 분석하였다(Choi *et al.*, 2020). 이는 특정 국가에서의 기술 비즈니스 잠재성을 평가하는데 적합하지만, 바이오화학 기술과 같이 국가별 기술수준과 시장규모의 차이가 큰 하이테크 산업에 적용하기에는 적합하지 않다. 동일한 기술이라도, 타겟 시장의 규모와 시장의 기술수준에 따라 비즈니스 잠재성은 달라질 수 있기 때문에, 국가별 차이를 모두 고려할 수 있는 분석방법이 필요하다.

본 연구는 1) 바이오화학 특허의 수명을 정의하고, 2) 바이오화학 특허의 서지정보를 통해 예측모델 학습에 활용될 특허지표를 추출한 뒤, 3) 바이오화학 특허의 수명예측 모델을 구축하여 바이오화학 기술의 비즈니스 잠재성을 평가하고, 4) 해석가능 인공지능 알고리즘을 통해 기술의 비즈니스 잠재성 평가결과에 대한 근거를 제시하는 네 단계를 따른다. 본 연구는 세 가지 기여를 갖는다. 첫째, 본 연구가 제시한 바이오화학 특허의 수명예측 모델은 비즈니스 잠재성이 높은 바이오화학 기술을 식별할 수 있다. 따라서, 국가와 기업이 투자 대상 바이오화학 기술을 선별하는 과정에 활용될 수 있다. 둘째, 본 연구는 하나의 국가가 아닌, 유럽, 한국, 미국, 등 다양한 국가에서 출원된 특허를 활용한다. 따라서, 제시된 바이오화학 특허의 수명예측 모델은 범국가적으로 활용될 수 있다. 셋째, 본 연구는 바이오화학 특허의 수명, 즉, 비즈니스 잠재성에 영향을 미치는 지표들을 분석 및 제시한다. 따라서, 분석 결과는 기술에 대한 비즈니스 잠재성 평가의 근거자료로 활용될 수 있다.

본 논문의 남은 구성은 다음과 같다. 먼저 제2장에서는 특허를 통해 기술의 비즈니스 잠재성을 평가한 선행연구와 해석가능 인공지능 알고리즘에 대해 설명한다. 다음으로 제3장에서는 본 연구가 사용한 데이터와 연구절차에 대해 설명한다. 다음으로 제4장에서는 연구 결과를 제시하고, 제5장을 통해 본 연구의 기여 및 한계점, 추후 연구에 대한 방향을 제시한다.

2. 배경연구

2.1 특허기반의 기술 비즈니스 잠재성 평가 연구

기술의 비즈니스 잠재성은 기술을 통한 이익창출 가능성을 의미하며, 이는 기술 판매, 기술기반의 비즈니스 가능성 등을 포괄하는 개념이다. 기술의 비즈니스 잠재성을 직접적으로 평가하는 것은 어렵지만, 기술에 대한 권리를 보장하는 제도인 특허를 통해 기술의 비즈니스 잠재성을 평가할 수 있음이 다양한 선행연구를 통해 제시되었다. 실제로, 특허의 인용정보가 기술의 비즈니스 잠재성에 영향을 미치며, 인용 횟수당 기술의 시장가치가 3% 증가하는 것으로 연구되었다(Hall *et al.*, 2005).

특허는 거래가 가능하기 때문에, 특허의 거래가능성을 통해 기술의 비즈니스 가치를 분석한연구들이 수행되었다. 먼저, 한국특허의 거래이력정보를 활용하여 거래가능성이 높은 특허를 식별하는 예측모델이 제시되었다(Ko *et al.*, 2019). 특허를 기술의 우수성을 의미하는 내재적 지표와 기술생태계를 의미하는 외재적 지표를 통해 표현하였으며, 신경망 모델을 활용하여 거래가능성이 높은 특허를 식별하였다는 점에서 의의가 있다. 다음으로, 미국특허의 거래정보를 활용하여 전자상거래 기술의 거래가능성을 분석한 연구가 수행되었다(Kim

and Geum, 2020). 빠르게 변화하는 산업을 대상으로 기술거래를 예측했다는 측면에서 의의가 존재한다.

또한, 특허의 수명을 활용한 비즈니스 잠재성 평가 연구가 수행되었다. 특허의 수명을 활용해 비즈니스 잠재성을 평가하는 이유는, 잠재가치가 있는 기술을 개발하고, 기술의 수명기간 내에 기술기반의 비즈니스를 수행, 기술가치를 회수하는 것이 기술 기업의 생태이기 때문이다. 즉, 기술의 비즈니스 잠재성은 기술수명에 필연적으로 영향을 받게 되는데, 수명이 긴 기술일수록 기술가치를 회수할 수 있는 기간이 길어지는 것으로 이해할 수 있다(Cheng and Lee, 2008). 먼저, 특허의 인용 정보분석을 통해 기술의 수명을 분석한 연구들이 선행되었다(Yoon, 2004; Yoo *et al.*, 2006; Jun *et al.*, 2012). 선행연구들은 특허가 더 이상 사용되지 않거나 타 특허에 의해 대체되는 기간을 특허가 처음 인용된 시점부터 마지막으로 특허가 인용된 시점으로 정의하여, 특허가 다루고 있는 기술수명의 대리지표로 활용하였다. 하지만, 특허의 인용정보를 기반으로 하는 수명예측 연구들은, 특정 기술군의 수명주기를 대표하는 값을 예측하는 것에 초점을 맞추었고, 개별 기술의 수명에 영향을 미칠 수 있는 기술적 속성, 외부 요인 등을 예측에 반영하지는 못했다.

이후, 특허의 인용정보에 기반한 기술수명 예측연구의 한계점을 해결하고자 DNN 모델을 활용한 특허수명 예측방법이 제시되었다(Choi *et al.*, 2020). 해당 연구는, 특허의 최대 수명이 20년이라는 점에서, 최대 수명까지 유지될 특허를 예측하는 모델을 제시하였다. 이를 위해 미국 특허 200,000건을 수집하였으며, 기술보호 범위, 기술개발에 투입된 자원, 기술 완성도 등을 의미하는 특허지표를 정의하여 DNN 모델의 입력 값으로 활용하였다. 제시된 모델은 F2 score 0.84를 보였으며, 기술 수명예측을 통해 기술의 비즈니스 가치를 평가하였다는 점에서 큰 의의가 있다.

이처럼 특허의 거래가능성과 수명정보를 활용해 기술의 비즈니스 잠재성을 평가할 수 있지만, 선행연구에서 제시된 방법들은 바이오화학 기술의 비즈니스 잠재성을 평가하는데 활용하기에는 적합하지 않다. 먼저, 특허의 거래는 각 국가의 특허청을 통해 이루어지며, 특허 거래정보를 제공하지 않는 특허청이 다수 존재하기 때문에 다국가의 특허거래내역을 수집하는 것은 사실상 불가능하다. 또한, 특허의 거래가 기업 합병, 라이센싱, 내부자 거래 등 다양한 목적에 의해 수행되는 점을 고려하였을 때, 기술의 비즈니스 잠재성으로 인한 특허거래정보를 식별하는 데 어려움이 존재한다. 다음으로, 단일국가 대상의 수명분석은 다른 국가의 특허에는 적용하기 어렵다는 한계가 존재한다. 바이오화학 기술과 같이 국가별 기술수준의 격차가 크고, 시장규모의 차이가 큰 기술의 비즈니스 잠재성 평가를 위해서는, 범국가적 수명 예측모델이 필요하다. 마지막으로, 선행연구가 제시한 수명예측 모델은 기술분야에 대한 구분 없이 모든 특허를 대상으로 학습된 수명예측 모델이기 때문에, 특정 기술분야에 적용시킬 경우, 성능이 낮아질 가능성이 존재한다.

따라서 본 연구는 유럽 특허, 한국 특허, 미국 특허 등 주요국에 출원된 바이오화학 특허의 수명정보와 특허의 권리가 보호되는 지역의 정보, 특허 자체의 우수성을 의미하는 특허지표를 활용하여 바이오화학 기술의 비즈니스 잠재성 평가모형을 제시하고자 한다. 또한, 해석 가능한 인공지능을 활용하여 기술의 비즈니스 가치에 영향을 주는 주요 요인들을 식별하며, 다음 장에서는 해석 가능한 인공지능에 대하여 자세히 살펴본다.

2.2 해석 가능한 인공지능

인공지능은 기술의 최첨단(State of the art)을 넘어서 다양한 산업분야의 핵심요소로 자리잡았다(Russell, 2010). 특히, 복잡한 데이터를 학습하기 위한 목적으로 대량의 은닉 계층을 쌓아서 만든 인공지능 기술인 딥러닝 모델이 활발하게 연구되고 있으며, 대부분의 연구에서 기존의 방법론들보다 높은 성능을 보여주고 있다(Shrestha and Mahmood, 2019). 딥러닝 모델의 높은 성능 이면에는 대규모 파라미터의 조합, 복잡한 학습 알고리즘이 숨겨져 있기에, 그 과정은 대표적인 블랙박스(Black-box)로 여겨진다(Castelvecchi, 2016). 하지만, 의학, 보안, 자율주행 등 의사결정에 대한 이유가 중요한 문제에 딥러닝 모델이 적극 활용됨에 따라, 모델의 블랙박스 해석에 대한 요구가 증가하고 있는 현실이다(Preece *et al.*, 2018).

일반적으로, 예측모델의 해석가능성과 성능 사이에는 반비례 관계가 존재하는 것으로 알려져 있다(Došilović *et al.*, 2018). 실제로, 선형 회귀모델과 같이 파라미터가 적게 요구되는 저수준 예측모델의 경우, 요인에 대한 가중치를 통해 모델의 결과물을 해석할 수 있으나, 영향변수와 요인변수 사이에 선형관계를 가정하기 때문에, 모델을 적용할 수 있는 문제는 한정적이다. 반대로, 대규모 파라미터를 통해 예측을 수행하는 DNN (Deep Neural Network), CNN(Convolutional Neural Network) 등 고수준의 딥러닝 모델들은 이미지 분류, 음성인식 등의 복잡한 문제에서 높은 성능을 보이지만, 모델의 결과를 해석하기에는 무리가 있다. 해석 가능한 인공지능에 대한 개념적 접근은 Arrieta *et al.*(2020)의 연구에 자세히 설명되어 있기에, 본 연구에서는 해석 가능한 인공지능의 기술적 측면을 설명하고자 한다.

해석 가능한 인공지능 방법은 내재적 해석 가능성(Intrinsic interpretability)에 기반하는 방법과, 사후 해석 가능성(Post-hoc interpretability)에 기반하는 방법으로 구분될 수 있다(Kim *et al.*, 2022). 먼저, 내재적 해석 가능성이란, 통계적 접근을 통해 데이터가 특정 분포를 따른다는 가정을 세우고, 모델의 복잡성을 제한하여 해석 가능성을 높이는 방법이다(Vollert *et al.*, 2021). 선형 분류모델에서, 모델이 갖는 파라미터 가중치를 입력변수가 결과에 미치는 영향도로 해석하는 방법이 그 예시이다(Vollert *et al.*, 2021). 반대로 사후 해석 가능성은, 모델이 예측한 개별의 결과값에 모델의 입력 값이 미치는 영향을 판단하

는 방법으로, 복잡한 모델에도 적용이 가능한 장점이 존재한다. 대표적으로, 모델의 국소적 예측 결과값에 대한 해석을 수행하는 LIME(Local Interpretable Model-agnostic Explanations) 알고리즘과, Anchor 알고리즘이 개발되었으며(Ribeiro *et al.*, 2016; Ribeiro *et al.*, 2018), 모델의 입력 값의 변화가 모델 예측치에 미치는 영향 정도인 Shapley value를 활용하는 SHAP (SHapley Additive exPlanation)알고리즘이 개발되었다(Lundberg and Lee, 2017).

본 연구는 바이오화학 기술에 대한 비즈니스 잠재성 평가를 목표로 하기 때문에, 국소적 해석 알고리즘인 LIME을 활용하고자 한다. LIME 알고리즘은 모델의 개별 예측 값에 대한 판단 근거를 제시하기 때문에, 개별 기술의 비즈니스 잠재성 평가결과에 대한 근거를 파악할 수 있을 것으로 기대한다.

3. 데이터 및 연구절차

본 장에서는 연구에 활용되는 데이터와 연구절차를 설명하고자 한다. 바이오화학 특허의 수명예측을 통한 바이오화학 기술의 비즈니스 잠재성 평가를 위해서는 1) 바이오화학 특허의 수집, 2) 서지정보 기반의 예측모델 학습 및 모델 해석이 수행되어야 한다. 따라서, 3.1장에서는 바이오화학 특허의 수집방법과 수집된 데이터의 통계정보를 설명한 뒤, 3.2장에서는 특허의 서지정보를 활용하여 특허지표를 추출하는 방법과 추출된 지표를 활용한 예측모델 학습 및 모델 해석방법을 구체적으로 설명한다.

3.1 데이터

바이오화학 특허를 수집하기 위해 본 연구는 미국 화학 학회의 CAS(Chemical Abstract Service)가 제공하는 CAS 특허 데이터베이스를 활용하였다. CAS 특허 데이터베이스는 특허의 제목, 요약문, 분류코드 등의 정보에 기반하여 화학물질을 다루고 있는 특허를 선별해 놓은 특허 데이터베이스다. CAS 특허 데이터베이스를 통해 EPO(European Patent Office), KPO (Korea Patent Office), USPTO(United States Patent and Trademark Office) 등 전 세계 64개 특허청에 출원된 특허들 중 화학물질을 다루고 있는 특허를 파악할 수 있다.

다양한 화학물질 특허들 중, 바이오화학 기술을 다루는 특허를 식별하기 위해, CAS 특허 데이터베이스가 갖는 분류코드와 색인 규칙을 활용하였다. CAS 특허 데이터베이스는 특허검색의 편리성을 위해 화학물질의 용도에 따른 분류코드, 색인 규칙을 제공한다. 본 연구에서는 바이오화학 기술을 의미하는 세 가지 분류코드(BIO10; Microbial, Algal, and Fungal Biochemistry, BIO11; Plant Biochemistry, BIO16; Fermentation and Bioindustrial Biochemistry)와 색인규칙(BPN; Biosynthetic preparation, BMF; Bioindustrial manufacture)을 사용하였다.

CAS 특허 데이터베이스에서 바이오화학 특허 67,859건을 수집하였으며, 수집된 특허 데이터 예시는 <Table 1>과 같다. <Table 1>에서, “Pat num”은 특허 번호를 의미하며 “STN pat num”은 CAS 특허 데이터베이스에서 해당 특허를 구분하는 고유번호다. 또한, “Chemical name”은 해당 특허에서 나타난 화학물질의 이름을 의미하며, “Cas number”는 해당 화학물질에 대한 CAS 고유번호를 의미한다. 예를 들어, EP3257945 특허는 EPO에 출원된 특허로, “1,3-Propanediol”, “2,3-Butanediol”, “Glycols” 물질과 관련된 바이오화학 특허다. 또한, US2015353611 특허는 USPTO에 출원된 특허로, “preparation”, “Nisin A”와 관련된 특허임을 알 수 있다.

수집된 67,859 건의 특허의 서지정보는 EPO가 제공하는 특허 데이터베이스인 PATSTAT(Worldwide Patent Statistical Database)을 활용하여 파악할 수 있다. PATSTAT은 다른 특허청에서 제공하는 특허 데이터베이스와 다르게, EPO를 포함한 다른 국가의 특허청에 출원된 특허정보를 모두 제공한다. 수집된 바이오화학 특허는 64개국 특허청에 출원된 특허를 대상으로 한다. 따라서, 다국적 특허 데이터를 제공하는 PATSTAT을 활용하는 것이 적합하다. PATSTAT이 제공하는 특허의 공개번호 정보와 수집한 바이오화학 특허의 Pat num을 통해 두 데이터베이스를 매칭할 수 있으며, 수집된 67,859건의 특허 중, 62,026건의 특허(91.4%)가 PATSTAT 정보와 매칭되었다.

Table 1. Example of bio-based Chemical Patents Identified from CAS Patents

Pat num	STN pat num	nation	Cas number	Chemical name
EP3257945	5830	EP	504-63-2P	1,3-Propanediol
P3257945	5830	EP	513-85-9P	2,3-Butanediol
EP3257945	5830	EP	-	Glycols
JP2017066082	3740	JP	106096-93-9P	Basic fibroblast growth factor
US2015353611	4474	US	50-21-5P	preparation
US2015353611	4474	US	1414-45-5P	Nisin A
KR20180027496	4626	KR	64-17-5P	Ethanol, preparation
KR20180027496	4626	KR	71-36-3P	1-Butanol, preparation
KR20180027496	4626	KR	67-63-0P	2-Propanol, preparation
KR20120082673	4680	KR	89-00-9P	2,3-Pyridinedicarboxylic acid

Table 2. Top 10 Nations with Registered Patents

Rank	Nation	Number of registered patents
1	US (United States)	8,301
2	JP (Japan)	3,899
3	EP (Europe)	3,691
4	KR (Korea)	3,121
5	CN (China)	2,673
6	AU (Australia)	1,883
7	ES (Spain)	1,696
8	CA (Canada)	1,489
9	RU (Russia)	601
10	MX (Mexico)	579

매칭된 62,026건의 특허는 등록된 특허와 출원상태인 특허가 모두 포함되어 있다. 특허의 수명 예측모델을 학습하기 위해서는 등록되었으며, 수명이 결정된 특허가 필요하다. 따라서, 62,026건의 특허 중 등록된 특허 30,069 건을 식별하였다. <Table 2>를 통해 바이오화학 특허가 등록된 상위 10개국을 확인할 수 있다. 미국에 등록된 바이오화학 특허가 8,301개로 가장 많았으며, 일본, 유럽, 한국이 그 다음이었다.

3.2 연구절차

본 연구의 절차는 <Figure 1>의 네 단계를 따른다. 먼저, 등록된 바이오화학 특허의 수명정보를 정의한다. 다음으로, 각 특허의 서지정보를 활용하여 특허지표를 정의 및 추출한다. 다음으로, 수명 예측모델을 구축한다. 최종적으로, 해석가능 인공지능 알고리즘을 활용하여, 구축된 모델을 해석하여 비즈니스 잠재성에 영향을 미치는 요소들을 분석한다.

(1) 바이오화학 특허의 수명 정의

수집된 바이오화학 기술의 특허수명을 정의하기 위해서는 특허수명의 시작 시점과 종료 시점을 정의해야 한다. 특허수명의 시작 시점은 PATSTAT이 제공하는 서지정보를 활용할 수 있으며, 종료 시점은 PATSTAT이 제공하는 행정정보를 활용할 수 있다. PATSTAT은 개별 특허에 대한 등록일자를 제공하지 않기 때문에, 본 연구는 등록공개 일자를 특허의 등록일자로 활용하였다. 따라서, 특허수명의 시작 시점은 등록된 특허의 등록일자로 정의한다. 또한, PATSTAT은 1997년부터 각 국가의 행정정보를 21개의 클래스로 구분하여 제공하고 있다 (<Table 3>).

A class는 특허의 출원신청이 접수되었음을 의미하는 행정정보이며, B class는 특허 출원 이후, 일련의 사건으로 인해 출원과정이 중지되었음을 의미하는 행정 정보다. 본 연구에서는 특허의 권리가 말소되었음을 의미하는 H class를 통해 특허의 수명이 종료되었음을 파악한다. 구체적으로, 각 특허의 마지막 행정 이벤트가 H class에 해당할 경우, 해당 특허의 수명이 종료된 것으로 정의하며, <Table 4>는 특허 KR101043190에 대한 행정정보 예시이다. <Table 4>에서, 특허의 권리가 말소되었음을 의미하는 H class가 2017년 6월 16일에 발생하였기 때문에, 해당 특허의 수명 종료 시점은 2017년으로 정의할 수 있다.

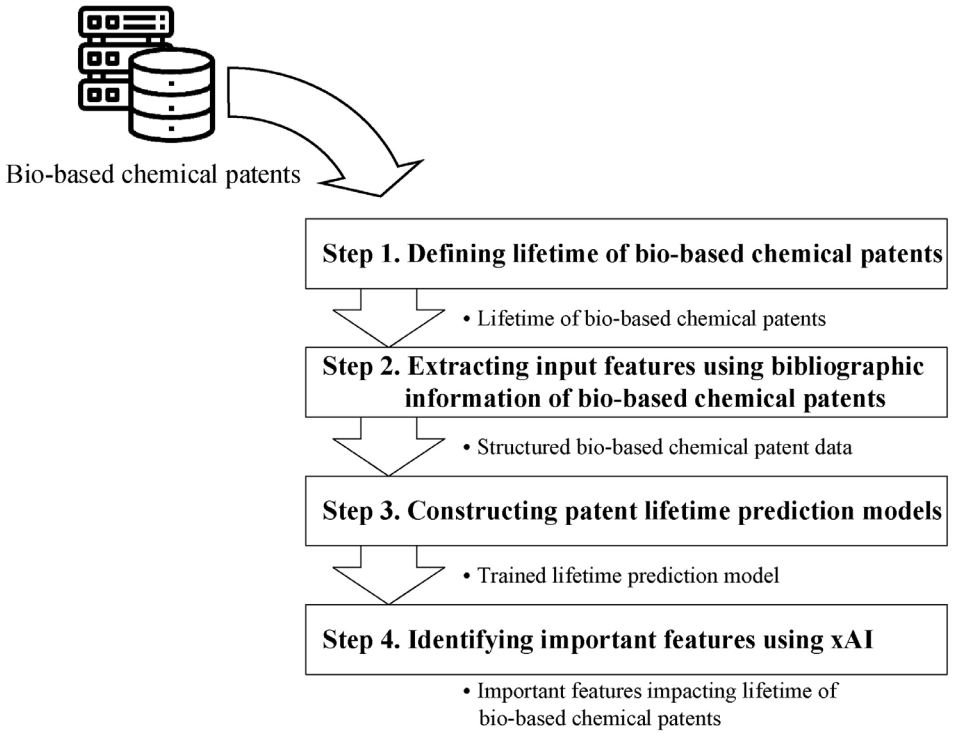

Figure 1. Overall Research Process

Table 3. List of 21 Legal Event Classes and Descriptions

Class	Description	Class	Description
A	Application filling	P	Re-publication of document after modification
B	Application discontinuation	Q	Document publication
C	Application revival	R	Party data change
D	Search and examination	S	Information on licensing and similar transactions
E	Pre-grant review request	T	Administrative procedure adjustment
F	IP right grant	U	Payment
G	Protection beyond IP right term	V	Appeal
H	IP right cessation	W	Other
K	IP right revival	Y	Correction and deletion of event information
L	IP right review request	Z	Classification pending
M	IP right maintenance		

Table 4. Example of Legal Events of a Patent (KR101043190)

Event code	Event class	Description	Event date
KR A201	D: Search and examination	REQUEST FOR EXAMINATION	2008-05-07
KR E902	B: Application discontinuation	NOTIFICATION OF REASON FOR REFUSAL	2010-05-18
KR E701	F: IP right grant	DECISION TO GRANT OR REGISTRATION OF PATENT RIGHT	2011-05-19
KR GRNT	F: IP right grant	WRITTEN DECISION TO GRANT	2011-06-15
KR FPAY	U: Payment	ANNUAL FEE PAYMENT	2014-06-05
KR FPAY	U: Payment	ANNUAL FEE PAYMENT	2015-06-04
KR FPAY	U: Payment	ANNUAL FEE PAYMENT	2016-06-02
KR LAPS	H: IP right cessation	LAPSE DUE TO UNPAID ANNUAL FEE	2017-06-16

(2) 바이오화학 특허의 서지정보를 활용한 특허지표 추출

본 연구는, 특허의 수명예측을 통해 바이오화학 특허의 비즈니스 잠재성을 평가하는 것을 목표로 한다. 따라서, 바이오화학 특허의 비즈니스 잠재성에 영향을 줄 수 있는 요소들을 정의하기

Table 5. Extracted Input Features Using Patent Bibliographic Information

Dimension	Input feature	Description
Geographical scope (Fischer and Leidinger, 2014; Squicciarini *et al.*, 2013; Ko *et al.*, 2019)	Member info (EU member)	1 if the applicant's nation is an EU member; otherwise, 0
	Member info (EPO member)	1 if the applicant's nation is an EPO member; otherwise, 0
	Member info (OECD member)	1 if the applicant's nation is an OECD member; otherwise, 0
	Member info (organization)	1 if the applicant type is an organization; otherwise, 0
	Member info (continent)	The applicant nation's continent
	Number of family patents	Number of family patents already applied for technology preoccupation
	Number of priority nations	Number of unique countries of prior patents
	Number of priority patents	Number of priority patents already applied for technology preoccupation
Technological scope (Lanjouw and Schankerman, 2004; Zhang *et al.*, 2016; Kim *et al.*, 2020)	Number of claims	Scope of overall technological claims a patent has
	Number of abstract words	Degree to which a patent describes an invention in depth
	Number of IPCs	Number of main classes to which a patent belongs
	Number of IPC sections	IPC sections to which a patent belongs
Cooperation degree (Guellec and Potterie, 2000; Lai and Che, 2009; Choi *et al.*, 2020)	Number of applicants	Number of patent applicants
	Number of applicant nations	Number of unique countries of patent applicants
	Number of inventors	Number of patent inventors
	Number of inventor nations	Number of unique countries of patent inventors
Novelty (Schoenmakers *et al.*, 2010; Verhoeven *et al.*, 2016)	Number of backward citations	Number of backward citations of a patent
	Median gap of backward citations	Median of differences in application date between a patent and its backward patents

위해 특허의 서지정보를 활용한다. 특허의 서지정보를 통해 기술개발에 투입된 인적자원의 규모, 기술의 사업화 범위, 기술 수명주기 등을 간접적으로 파악할 수 있으며, 본 연구가 특허기반의 선행연구를 참고하여 정의한 지표는 <Table 5>와 같다.

<Table 5>에서 Dimension은 특허지표를 통해 파악할 수 있는 기술의 특징을 의미한다. 예를 들어, Geographical scope는 특허가 보호받을 수 있는 국가정보를 의미하는데, 특허는 특허가 등록된 국가에서만 권리를 보장받을 수 있기 때문에, Geographical scope를 통해 기술사업화가 가능한 지역적 범위를 파악할 수 있다. 특히, PATSTAT이 글로벌 특허정보를 제공하기 때문에, 특허가 등록된 국가정보를 식별할 수 있으며, 유럽국가 여부, 국가가 위치한 대륙정보, OECD(Organization for Economic Cooperation and Development) 국가 여부 등을 파악할 수 있다. 다음으로, 특허의 Family 정보와 우선권 정보를 활용한 지표를 정의하였다. Family와 우선권 정보를 통해 출원국가의 범위를 확인할 수 있으며, 다양한 선행연구에서 Family와 우선권 지표는 기술의 사업화 범위를 식별하기 위해 활용되었다(Jürgens and Herrero-Solana, 2017; Kabore and Park, 2019). 다음으로, 기술개발에 투입된 인적 자원을 파악하기 위해 출원인, 발명자 정보를 활용한 지표를 정의하였다. 특허의 발명자가 많을수록 많은 인력이 투입되었음을 의미하며, 발명자의 국가가 다양할수록 특허의 가치가 높은 것으로 알려져 있다(Ferrucci and Lissoni, 2019). 다음으로, 특허 초록의 길이와 특허의 IPC 정보를 통해 특허의 기술적 범위를 식별할 수 있는 지표를 정의하였다. 특허는 기술에 대한 권리를 보장하는 제도이기 때문에, 특허가 다루는 기술의 범위가 넓을수록 보장받을 수 있는 권리가 넓음을 의미한다(Ko *et al.*, 2019).

최종적으로, 본 연구는 18개의 지표를 정의하였으며, PATSTAT에서 지표를 추출하여 바이오화학 특허의 수명 예측모델 구축에 활용한다.

(3) 바이오화학 특허의 수명 예측모델 구축

본 단계에서는 정의된 18개의 지표와, 특허의 수명 정보를 활용하여 비즈니스 잠재성이 높은 특허를 예측하는 모델을 구축한다. 비즈니스 잠재성이 높은 특허는 상대적으로 수명이 오래 유지된 특허를 의미하며, 비즈니스 잠재성이 낮은 특허는 상대적으로 수명이 짧게 유지된 특허를 의미한다. 따라서, 본 연구가 구축하는 바이오화학 특허의 수명 예측모델은 18개의 입력지표를 통해 비즈니스 잠재성이 높은 특허를 예측하는 이진 분류모델이다(1: 잠재성 높음, 0: 잠재성 낮음).

이진 분류모델 구축을 위해, 1) 다중 의사결정 트리인 DRF (Distributed Random Forest), 2) 데이터의 비선형 관계를 식별할 수 있는 인공 신경망인 DNN(Deep Neural Network), 3) 여러 예측모형의 잔차를 통해 새로운 모형을 적합시키는 방법인 GBM(Gradient Boosting Machine), 4) 랜덤 포레스트 모델의 변종으로, 트리의 각 후보 특성을 분석하는 XRT(Extremely Randomized Trees), 5) 다차원 데이터를 선형 모델을 이용하여 분류하는 LMC(Linear Model Classifier)를 활용하였으며, 각 모델에 대한 최적의 하이퍼 파라미터를 정의하기 위해 최적의 하이퍼 파라미터를 제공해주는 AutoML(Automated Machine Learning) 기법을 활용한다. AutoML은 하이퍼 파라미터 탐색 공간을 설계하고 최적화하는 과정을 통해 하이퍼 파라미터에 따른 모델의 성능을 제공해 준다(He *et al.*, 2021). 본 연구는 Python 기반의 AutoML 라이브러리인 H2O를 사용하여 각 모델의 하이퍼 파라미터를 도출하며, 최적의 하이퍼 파라미터를 갖는 5개 모델의 성능을 비교한다.

학습된 예측모델들의 성능을 비교하기 위해, 데이터를 5개의 동일한 크기로 분할하고, 학습과 검증을 5번 반복하는 5-fold cross-validation 방법을 활용한다. 각 반복 과정에서 4fold는 모델의 학습에 활용되고, 1fold는 검증에 활용된다. 최종적으로, 5번의 검증결과 평균을 통해 모델의 최종 성능을 산출하게 된다. 본 연구는 모델의 성능을 평가하기 위해, 모델이 True로 분류한 데이터 중 실제 True의 비율을 의미하는 Precision(EQ. 1)과 실제 True인 데이터 중 모델이 True로 예측한 비율을 의미하는 Recall(EQ 2)을 산출한 뒤, 이들의 조화평균으로 계산되는 F1 score를 활용하였다(EQ. 3). F1 Score를 활용하기 때문에, 모델의 Precision과 Recall이 모두 높은 모델을 선정할 수 있다.

$$Precision = \sum \frac{TP}{FP+TP} \tag{1}$$

$$Recall = \sum \frac{TP}{FN+TP} \tag{2}$$

$$F_1 = 2^* \frac{precision^* recall}{precision + recall} \tag{3}$$

EQs. 1-3에서, TP는 True Positive로, 실제 True인데 예측모델이 True라고 판단한 경우를 의미하며, TN은 True Negative로, 실제 False인데 모델이 False로 예측한 경우를 의미한다. 예를 들어, 비즈니스 가치가 높은 특허를 모델이 높게 예측한다면 TP, 비즈니스 가치가 낮은 특허를 모델이 낮게 예측한다면 FP이다. FP는 False Positive로 실제 False인데 모델이 True로 판단한 경우를 의미하며, FN은 False Negative로 실제 True인데 모델이 False로 예측한 경우를 의미한다. 즉, 비즈니스 가치가 높을 것으로 예측했지만 실제로는 낮은 경우 FP, 비즈니스 가치가 낮을 것으로 예측했지만 실제로는 높은 경우가 FN이다.

(4) 해석가능 인공지능 알고리즘을 활용한 주요변수 분석

마지막으로, 해석가능 인공지능 알고리즘인 LIME을 활용하여 모델의 예측결과를 해석한다. LIME은 학습된 예측모델의 결정경계 주변으로, 해석이 용이한 선형 분류모델을 학습하여 선형 분류모델의 회귀계수를 통해 결과를 해석한다(<Figure 2>). 이때 새로 학습된 선형 분류모델을 대리모델(Surrogate model)이라 하는데, 대리모델은 학습된 예측모델을 모방하는 해석가능한 모델이다. <Figure 2>에서, 실선은 학습

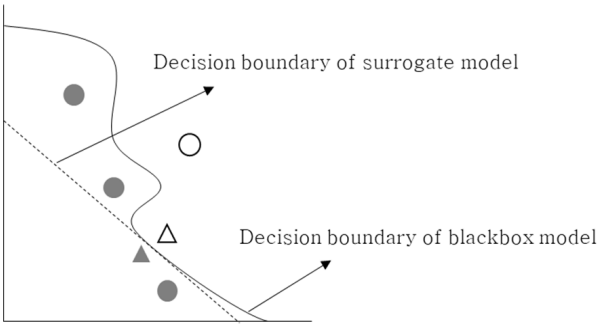

Figure 2. Schematic of the LIME Algorithm

된 예측모델의 결정경계를 나타내며, 비선형적 패턴을 가지고 있다. 하얀색 세모가 LIME을 통해 해석하고자 하는 데이터라면, 하얀색 세모와 주변의 데이터를 활용하여 해당 지점에서의 대리모델을 학습하게 된다. 즉, 학습된 대리모델은 하얀색 세모와 회색 세모를 분류하는데 높은 성능을 보이지만, 다른 데이터를 분류하지는 못한다.

고차원의 결정경계를 갖는 예측모델도 결국, 지역적으로 선형으로 표현할 수 있다는 가정을 적용하기 때문에, 직관적으로 사용 가능한 알고리즘이다. 본 연구에서는, 바이오화학 기술 특허의 비즈니스 잠재성 예측모델을 해석할 대리모델로 릿지 회귀모델(Ridge regression model)을 활용하였다(McDonald, 2009). 릿지 회귀모델의 계수를 통해 모델의 예측에 대한 변수의 영향도를 파악할 수 있어, 바이오화학 특허의 수명예측 결과에 대한 이유를 파악할 수 있다.

4. 연구결과

본 장에서는, 3장에 기술한 데이터와 연구절차에 기반한 연구결과를 서술한다. 먼저, 바이오화학 특허의 수명을 정의하였으며, 수명이 확정된 특허를 대상으로 특허지표를 추출하였다. 다음으로, 바이오화학 특허의 수명 예측모델을 구축한 뒤, 해석가능 인공지능 알고리즘인 LIME을 활용하여 예측모델의 결과를 해석하였다.

4.1 바이오화학 특허의 수명 정의

바이오화학 특허의 수명을 정의하기 위해, CAS 특허 데이터베이스와 PATSTAT을 활용하여 선별된 등록특허를 대상으로, 행정정보를 분석하였다. 중국특허의 경우 PATSTAT에서 제공되는 행정정보가 정확하지 않아 수명정보를 파악할 수 없었다. 따라서, 등록된 30,069건의 특허 중, 중국에 등록된 2,673건의 특허를 제외한 27,396건에 대한 행정 정보를 분석하였고, 마지막 이벤트가 H class인 특허 9,298건을 식별하였다. 특허의 등록 공개정보와 행정정보를 활용하여 정의된 특허 수명 예시는 <Table 6>과 같다. Start date는 특허의 등록공보가 공개된 시점을 의미하며, End date는 H class에 해당하는 행정이벤트가 발생한 시점을 의미한다. 즉, End date와 Start date의 차이를 특허의 Lifetime으로 정의하게 되며, 예를 들어 EP1322775 특허의 수명은 2011-05-11부터 2017-11-30까지인 7년으로 정의할 수 있다.

4.2 바이오화학 특허의 서지정보를 활용한 특허지표 추출

수명이 확정된 9,298건의 특허의 서지정보를 활용하여 예측모델 구축에 활용할 18개의 특허지표를 추출하였다. <Table 7>은 추출된 특허 지표와, 예측모델에 활용하기 위한 one-hot encoding 예시를 보여준다.

Member info 지표의 경우, 해당 여부를 의미하기 때문에 0과 1로 표현되지만, 대륙정보를 의미하는 continent의 경우, “Europe”, “Asia”, “North America”, “South America”, “Africa”, Australia and Oceania”, “None” 값을 가지고 있기 때문에, 0과 1을 갖는 7bit로 변환하였다. 또한, 청구항 수와 같은 숫자형 지표들은 결측값이 존재하는 경우가 있는데, 결측값들은 Median 값으로 표현하였다. 결과적으로, 학습에 활용되는 바이오화학 특허는 31차원으로 표현되었으며, 31차원의 입력지표를 통해 특허 수명 예측모델을 구축하였다.

4.3 바이오화학 특허의 수명 예측모델 구축

특허수명 예측모델의 학습집합을 구성하기 위해, 본 연구는 Stanine score를 활용하였다. Stanine score는 주어진 데이터를 백분율에 따라 점수를 할당하는 방법으로, 백분율을 9등급으로 구분하여 주어진 데이터에 등급을 부여하는 방법이다(Hills, 1983). Stanine score는 평가대상 데이터의 분포와 관계없이 사용 가능하며, 국내에서는 교육부의 수능등급평가, 한국발명진흥원의 특허 가치평가에 활용되고 있다. Stanine score

Table 6. Examples of Determined Lifetime of Patents

Patent number	Start date	End date	Lifetime
EP1322775	2011-05-11	2017-11-30	7
EP1352965	2008-03-05	2021-12-19	14
US2004253678	2004-12-16	2012-08-28	8
US2002039580	2002-04-04	2021-11-02	20
KR101021789	2011-03-17	2019-03-08	8

Table 7. Examples of Input Features and Their Encoded Values

Input feature	Original value	Encoded value
Member info (EU member)	0	0
Member info (EPO member)	0	0
Member info (OECD member)	1	1
Member info (organization)	0	0
Member info (continent)	Asia	0100000
Number of family patents	20	20
Number of claims	0	Replacing missing values with the median
Number of priority nations	1	1
Number of priority patents	1	1
Number of applicants	1	1
Number of applicant nations	1	1
Number of inventors	1	1
Number of inventor nations	1	1
Number of abstract words	514	514
Number of IPCs	8	8
Number of IPC sections	B, C	01100000
Number of backward citations	2	2
Median gap of backward citations	2533.5	2533.5

는 1~3등급까지를 평균 이상으로, 4~6등급을 평균으로, 7~9등급을 평균 이하로 정의하고 있으며, 아동 읽기능력 평가(Hindmarsh *et al.*, 2021), 학업 성취도 분석(Wang *et al.*, 2020) 등의 연구에서 활용되었다.

본 연구에서, 수명이 확정된 9,298건의 바이오화학 특허의 수명과 그에 따른 Stanine 점수는 <Figure 3>과 같다. <Figure 3>에서, Stanine 상・하위 1등급은 상위 4%, 하위 4%로 정의할 수 있으며, 상위 1등급의 경우 특허수명이 14년 이상인 특허들을 의미하며 하위 1등급의 경우 특허수명이 2년 미만인 특허들을 의미한다. 본 연구에서는 특허수명이 상위 3등급 이내인 특허들(수명이 9년 이상 유지된 특허들로, 비즈니스 잠재성이 높은 특허들) 2,199건, 특허수명이 하위 3등급 이내인 특허들(수명이 4년 미만으로 유지된 특허들로, 비즈니스 잠재성이 낮은 특허들) 2,274건을 학습집합으로 활용하였다. 즉, 본 연구가 구축하는 모델은 수명의 유지기간이 평균 이상인 특허(상위 3등급 이내에 포함될 특허)와 평균 이하인 특허(하위 3등급 이내)에 포함될 특허를 구분하는 모델이다. 따라서, 구축된 모델을 통해 비즈니스 잠재성이 상대적으로 높은 특허와 낮은 특허를 선별할 수 있다.

구성된 학습집합을 활용하여, DRF, DNN, XRT, GBM, LM 모델의 학습을 진행하였다. 모델 학습을 위해 데이터를 8:2:2 (학습, 검증, 테스트)로 구분하였으며, 학습은 5-fold cross validation 기법을 활용하여 진행되었다. 각 모델의 최적 하이퍼파라미터를 정의하기 위해 Python AutoML 라이브러리인 H2O를 활용하였고, 학습된 모델의 최종 성능은 <Table 8>과 같다.

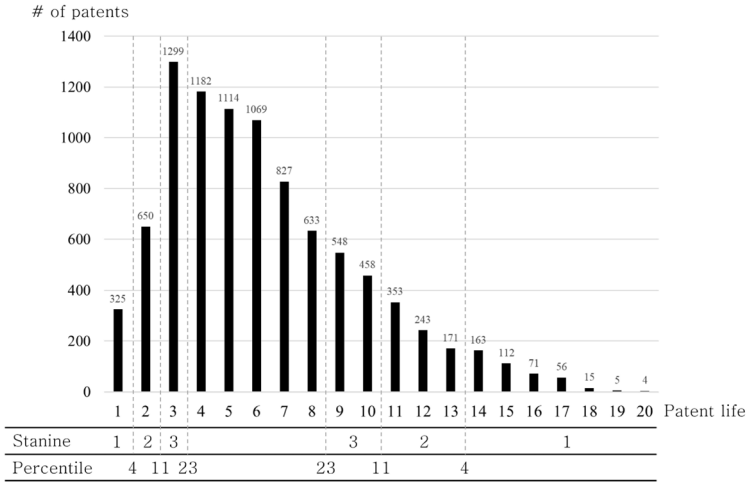

Figure 3. Patent Lifetime Distribution and Stanine Score

Table 8. Performances of the Trained Models

Model	Threshold	F1 score	AUC	Accuracy	Precision	Recall
Distributed Random Forest (DRF)	0.490	0.813	0.8863	0.8147	0.8072	0.818
Deep Neural Network (DNN)	0.435	0.810	0.8808	0.8058	0.7806	0.841
Gradient Boosting Machine (GBM)	0.265	0.799	0.879	0.784	0.734	0.877
Extremely Randomized Tree (XRT)	0.383	0.798	0.871	0.779	0.723	0.891
Linear Model (LM)	0.319	0.768	0.844	0.728	0.660	0.918

Table 9. Examples of Prediction Results of the DRF Model

Patent number	Real value	Predicted value	Prediction score
US2011097772	1	1	0.999
US2007207108	1	1	0.999
EP1123386	1	1	0.976
HK1038244	1	1	0.958
AU2001250318	1	1	0.955
…			
JP2013006799	0	0	0.172
DE102013009145	0	0	0.074
KR20150060221	0	0	0.068

모델이 비즈니스 잠재성이 높을 것으로 예상한 특허들 중, 실제로 비즈니스 잠재성이 높은 특허가 많음을 의미하는 Precision을 통해 살펴보면, 가장 합리적인 모델은 DRF 모델이다. 하지만, 실제로 비즈니스 잠재성이 높은 특허들을 잘 선별해 낼 수 있는 모델은 Recall이 가장 높은 LM 모델이다. 이처럼, 다양한 성능지표를 활용하여 적합한 모델을 선택할 수 있다. 본 연구에서는, Precision과 Recall을 조화평균을 통해, 두 관점을 모두 반영할 수 있는 F1 score가 가장 높은 DRF 모델을 가장 좋은 성능의 모델로 선정하였다. DRF 모델로 예측한 결과의 예시는 <Table 9>와 같다.

<Table 9> 에서, Real value는 실제로 특허의 수명이 상위 3등급, 즉, 비즈니스 잠재성이 높은 특허를 의미하며, Predicted value는 모델이 비즈니스 잠재성이 높은 특허로 예측한 결과이다. 즉, 0이면 특허의 비즈니스 잠재성이 낮음을 의미하며, 1이면 특허의 비즈니스 잠재성이 높음을 의미한다. Prediction score는 모델의 출력 값을 의미하는데, <Table 8>을 통해 확인할 수 있는 것과 같이 DRF의 Threshold인 0.490 이상의 값은 비즈니스 잠재성이 높은 특허로, 미만의 값은 비즈니스 잠재성이 낮은 특허로 판별하게 된다.

4.4 해석가능 인공지능 알고리즘을 활용한 주요변수 분석

최종적으로, 학습된 DRF 모델에 LIME 알고리즘을 적용하였다. 모델의 결과에 대한 해석을 수행하기 위해, 모델이 1로 예측하여 실제 1인 경우, 1로 예측하였으나 실제 0인 경우, 0으로 예측하였으나 실제 1인 경우, 0으로 예측하여 실제 0인 경우에 대한 사례를 선택하여 수행하였다(<Figure 4>~<Figure 7>).

먼저, <Figure 4>는 비즈니스 잠재성이 높은 특허를 옳게 판단한 경우로, IPC section H가 0이면서, 출원인이 1명 초과이고, North America 대륙에 등록된 특허이면 비즈니스 잠재성이 높다고 판단하였다. 출원인이 많을수록 비즈니스 이해관계자가 많이 연관되어 있는 특허로 해석할 수 있으며, 미국의 시장규모를 고려한다면, 합리적인 판단으로 생각할 수 있다.

다음으로, <Figure 5>는 실제로 비즈니스 잠재성이 높은 특허이지만 비즈니스 잠재성이 낮은 특허로 판단된 경우이다. 실제 AU2008360491 특허는 2012년 등록 이후, 2021년까지

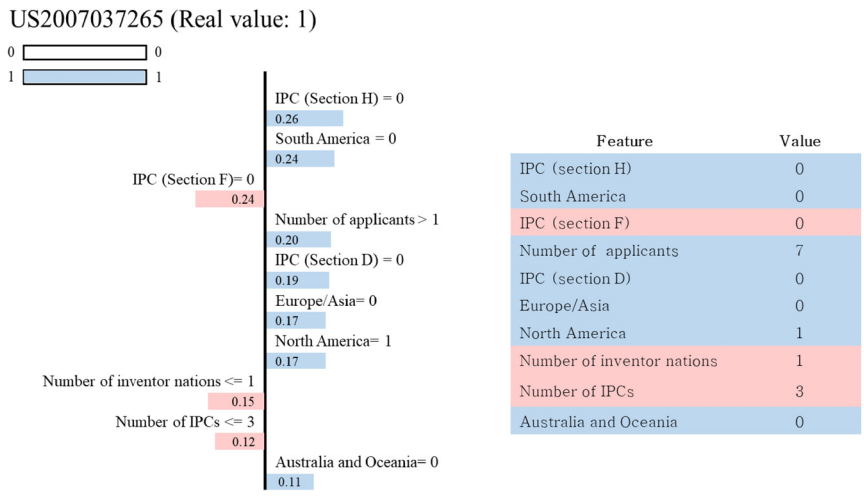

Figure 4. Explanation of US2007037265 Prediction Result (real value 1, prediction 1)

권리가 유지된 특허로, 9년의 특허수명을 갖는 Stanine 3등급에 해당하는 특허이다. 하지만 비즈니스 잠재성이 낮게 판단된 이유는, 출원인이 한 명이며, South·North America 대륙에 출원된 특허가 아니며, 발명자의 수가 적기 때문이다. 즉, 특허

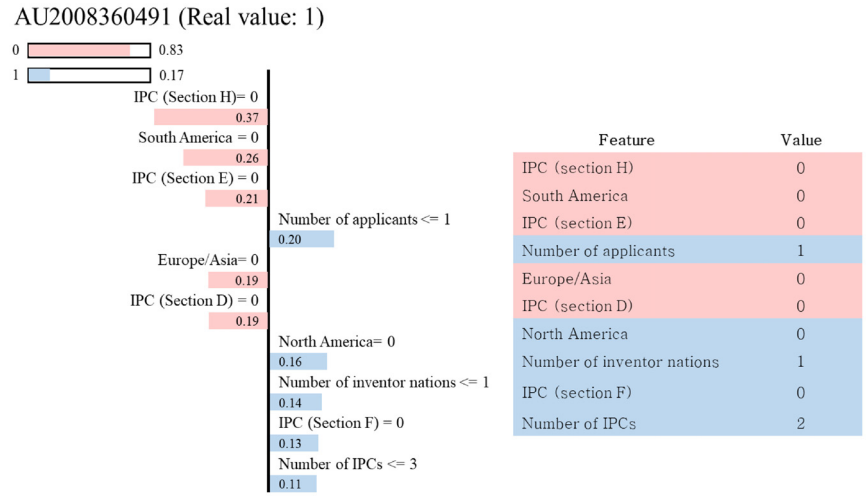

Figure 5. Explanation of AU2008360491 Prediction Result (real value 1, prediction 0)

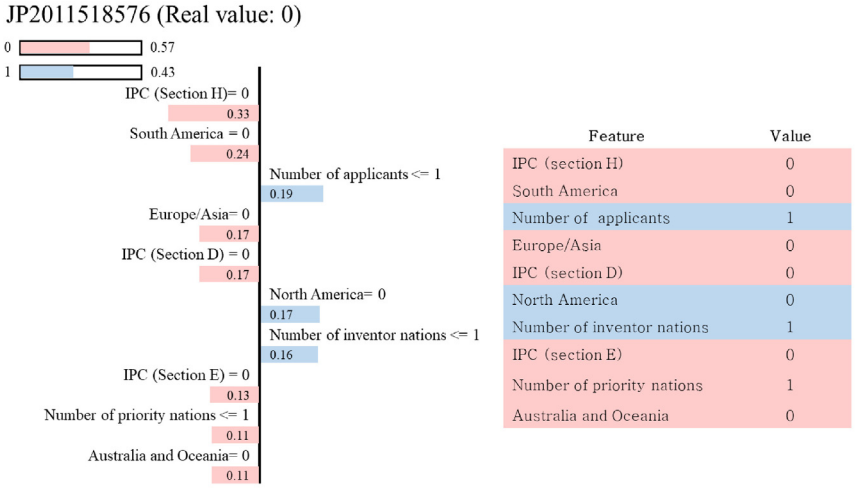

Figure 6. Explanation of JP2011518576 Prediction Result (real value 0, prediction 0)

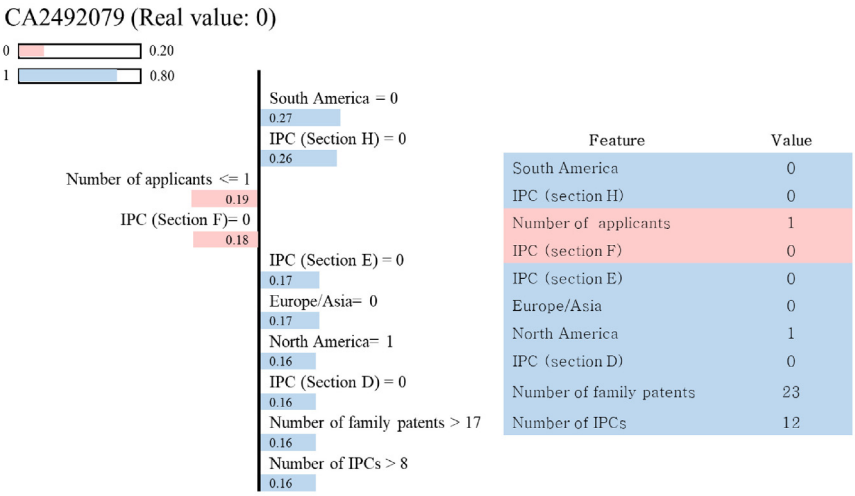

Figure 7. Explanation of CA2492079 Prediction Result (real value 0, prediction 1)

를 통해 비즈니스를 영위할 수 있는 시장의 규모가 상대적으로 작으며, 기술개발에 투입된 인적자원이 적어, 비즈니스 잠재성이 낮은 특허로 판단된 것으로 해석할 수 있다.

다음으로, <Figure 6>은실제 비즈니스 잠재성이 낮은 특허를 비즈니스 잠재성이 낮다고 예측한 경우를 보여준다. AU2008360491 특허와 유사하게, South·North America 대륙에 출원된 특허가 아니고, 출원인이 한 명에 불과하기 때문에 비즈니스와 관련된 이해관계자가 적고, 시장의 규모가 작은 것으로 판단되었다. 이와 같은 결과를 통해, 시장의 규모와 출원인의 수가 기술의 비즈니스 잠재성을 판단하는데 큰 영향을 주는 것을 확인할 수 있다.

마지막으로, <Figure 7>은 실제로는 비즈니스 잠재성이 낮은 특허임에도 불구하고, 비즈니스 잠재성이 높은 것으로 판단된 경우를 보여준다. 출원인 수가 적어 비즈니스와 관련된 이해관계자가 적은 것으로 해석될 수 있지만, North America 대륙에 출원된 특허이고, 특허의 Family 특허수가 17개를 초과하며, 보유한 IPC 수가 8개 초과이기 때문에, 비즈니스 잠재성이 높은 특허로 판단되었다. 특허의 Family 수가 해당 특허가 해외 여러 국가에서도 권리를 보장받을 수 있는 정도를 의미한다는 점에서, 특허를 통해 다양한 시장에서 비즈니스를 영위할 수 있음을 파악할 수 있다. 또한, IPC수가 12인 특허로, 특허를 통해 보호받을 수 있는 기술범위가 넓은 것을 알 수 있다. 즉, 모델의 예측 결과가 틀렸음에도 불구하고, 결과를 도출한 이유가 합리적임을 확인할 수 있었다.

본 장에서는 해석가능 인공지능 알고리즘인 LIME을 통해 모델의 예측 결과를 해석하였다. 네 가지의 사례를 대상으로 모델을 해석한 결과, 비즈니스 대상 지역의 규모, 투입된 인적 자원의 규모, 비즈니스와 관련된 이해관계자의 규모, 특허를 통해 보호받을 수 있는 기술범위 등이 모델의 결과에 영향을 주는 것으로 나타났다. 이러한 결과를, 특허 전문가 그룹과 논의하였는데, 모델의 예측 과정이 매우 합리적이라는 공통된 의견을 얻을 수 있었다. 따라서, 본 연구가 제시한 모델을 통해, 비즈니스 잠재성이 높은 바이오화학 기술을 선별할 수 있으며, 결과에 대한 합리적 이유를 확인할 수 있기 때문에, 기술사업화를 위한 투자과정에 합리적인 근거를 제시할 수 있을 것으로 기대한다.

5. 결론

본 연구는 바이오화학 특허의 수명예측을 통해 바이오화학 기술의 비즈니스 잠재성을 평가하였다. 또한, 해석가능 인공지능인 LIME을 활용하여, 모델의 결과를 해석하였다. 이를 위해, CAS 특허 데이터베이스와 PATSTAT을 활용하여 바이오화학 특허를 선별하였으며, 선별된 특허의 서지정보를 통해 입력지표를 추출하였다. 다음으로, 5개의 머신러닝 모델을 학습하였고, 성능이 가장 높은 예측모델인 DRF 모델을 구축하였다. 최종적으로, DRF 모델이 예측한 결과를 해석가능 인공지능 알고리즘인 LIME을 통해 해석하였다.

본 연구의 기여는 다음과 같다. 먼저, 본 연구가 구축한 바이오화학 특허의 특허수명 예측모델은 기술의 비즈니스화가 어려운 바이오화학 기술분야에서 비즈니스 잠재성이 높은 특허를 식별할 수 있기 때문에, 기술투자에 대한 불확실성을 낮추는데 활용될 수 있을 것이다. 다음으로, 모델의 학습과정에 범국가적 특허정보가 활용되었기 때문에, 바이오화학 기술에 대한 세계적 기술동향이 학습에 반영되었다. 따라서, 본 연구가 제시한 예측모형을 통해 글로벌 시장을 고려한 기술의 비즈니스 잠재성을 평가할 수 있을 것으로 기대한다. 마지막으로, 모델의 예측에 대한 근거를 LIME 알고리즘을 통해 파악하고 제시하였다. 제시된 근거는 바이오화학 기술의 비즈니스화를 위한 투자과정에서, 투자에 대한 기초자료로 활용될 수 있을 것으로 기대한다.

그럼에도 불구하고, 본 연구는 다음과 같은 한계를 갖는다. 먼저, 학습에 활용된 데이터 규모가 적다. 추후, 확장된 데이터를 활용하여 특허수명 예측에 최적화된 모델을 개발이 필요할 것이다. 다음으로, 특허지표를 산출하는 과정에서 특허의 텍스트 정보를 활용하지 못하였다. 특허의 텍스트를 통해 기술의 세부 내용을 파악할 수 있다는 점에서, 추후 연구를 통해 텍스트 기반의 지표추출을 시도할 필요가 있다. 마지막으로, 지표의 해석과정에서 지표간 상관관계를 고려하지 못하였다. 따라서, 추후 지표간 상관관계를 고려할 수 있는 해석 알고리즘을 활용하여 연구를 수행할 필요가 있다.

참고문헌

Alemu, M. (2020), Trend of biotechnology applications in pest management: A review, *International Journal of Applied Sciences and Biotechnology*, **8**(2), 108-131.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D. and Benjamins, R. (2020), Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, **58**, 82-115.

Castelvecchi, D. (2016), Can we open the black box of AI? *Nature News*, **538**(7623), 20.

Cheng, S. and Lee, H. (2008), R&D resource allocation and linkage with business strategies-a technology lifecycle perspective, *2008 IEEE International Conference on Industrial Engineering and Engineering Management*, 47-52.

Cho, H. R. (2019), Is the duration of the patent right appropriate? *The Magazine of the Society of Air-Conditioning and Refrigerating Engineers of Korea*, **48**(10), 72-73.

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, Y. H. (2007), 2020 vision and strategy of the bio industry, KIET.

Došilović, F. K., Brčić, M., and Hlupić, N. (2018), Explainable

artificial intelligence: A survey, *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, 0210-0215.
Ferrucci, E. and Lissoni, F. (2019), Foreign inventors in Europe and the United States: Diversity and patent quality, *Research Policy*, **48**(9), 103774.
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.
Hall, B. H., Jaffe, A., and Trajtenberg, M. (2005), Market value and patent citations, *RAND Journal of Economics*, 16-38.
He, X., Zhao, K., and Chu, X. (2021), AutoML: A survey of the state-of-the-art, *Knowledge-Based Systems*, **212**, 106622.
Hills, J. R. (1983), Interpreting Stanine Scores, *Educational Measurement: Issues and Practice*, **2**(3), 18-27.
Hindmarsh, G. P., Black, A. A., White, S. L., Hopkins, S., and Wood, J. M. (2021), Eye movement patterns and reading ability in children, *Ophthalmic and Physiological Optics*, **41**(5), 1134-1143.
Jun, S.-P., Park, H.-W., and Yoo, J. Y. (2012), The development of the method of determining remaining cited-patent life time using the survival curve analysis, *Journal of Korea Technology Innovation Society,* **15**(4), 745-765.
Jürgens, B. and Herrero-Solana, V. (2017), Patent bibliometrics and its use for technology watch, *Journal of Intelligence Studies in Business*, 7(2), 17-26.
Kabore, F. P. and Park, W. G. (2019), Can patent family size and composition signal patent value? *Applied Economics*, **51**(60), 6476-6496.
Kim, M. and Geum, Y. (2020), Predicting Patent Transactions Using Patent-Based Machine Learning Techniques, *IEEE Access*, **8**, 188833-188843.
Ko, N., Jeong, B., Seo, W., and Yoon, J. (2019), A transferability evaluation model for intellectual property, *Computers & Industrial Engineering*, 131, 344-355.
Lai, Y.-H. and Che, H.-C. (2009), Modeling patent legal value by Extension Neural Network, *Expert Systems with Applications*, **36**(7), 10520-10528.
Lanjouw, J. O. and Schankerman, M. (2004), Patent quality and research productivity: Measuring innovation with multiple indicators, *The Economic Journal*, **114**(495), 441-465.
Lee, J. H., Kim, J. H., Jo, M. H., and Jeong, H. M. (2012), Product and Technology Lifecycle Utilization Trends, *Information and Communication Industry Promotion Agency Weekly Technology Trend*, **1550**, 1-13.
Lee, J. R. (2011), S&T Policy Agenda and Options for the BioEconomy, *Policy Research*, 1-340.
Lee, K. H. and Yoon, B. S. (2006), The Effects of Patents on Firm Value: Venture vs. non-Venture, *Technology Management Economics Society*, 1-109.
Lundberg, S. M. and Lee, S.-I. (2017), A unified approach to interpreting model predictions, *Advances in Neural Information Processing Systems* 30.
McDonald, G. C. (2009), Ridge regression, *Wiley Interdisciplinary Reviews: Computational Statistics*, **1**(1), 93-100.
Nam, J. K. (2015), Policy Tasks for the Biochemical Industry as a New Growth Engine Focusing on Bioplastics, KIET.
Preece, A., Harborne, D., Braines, D., Tomsett, R., and Chakraborty, S. (2018), Stakeholders in explainable AI. arXiv preprint arXiv:1810.00184.
Ribeiro, M. T., Singh, S., and Guestrin, C. (2016), "Why should i trust you" Explaining the predictions of any classifier, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144.
Ribeiro, M. T., Singh, S., and Guestrin, C. (2018), Anchors: High-precision model-agnostic explanations, *Proceedings of the AAAI Conference on Artificial Intelligence.*
Russell, S. J. (2010), *Artificial intelligence a modern approach*, Pearson Education, Inc.
Schoenmakers, W. and Duysters, G. (2010), The technological origins of radical inventions, *Research Policy*, **39**(8), 1051-1059.
Seo, J. I. and Young, Y. J. (2010), Biotechnology Industry and Development Trends of SME, KISTI
Shrestha, A. and Mahmood, A. (2019), Review of deep learning algorithms and architectures, *IEEE Access*, **7**, 53040-53065.
Squicciarini, M., Dernis, H., and Criscuolo, C. (2013), Measuring patent quality: Indicators of technological and economic value, OECD Science, No. 2013/03.
Verhoeven, D., Bakker, J., and Veugelers, R. (2016), Measuring technological novelty with patent-based indicators, *Research Policy*, **45**(3), 707-723.
Vollert, S., Atzmueller, M., and Theissler, A. (2021), Interpretable Machine Learning: A brief survey from the predictive maintenance perspective, *2021 26th IEEE international conference on emerging technologies and factory automation (ETFA)*, 01-08.
Wang, S., Rubie-Davies, C. M., and Meissel, K. (2020), The stability and trajectories of teacher expectations: Student achievement level as a moderator, *Learning and Individual Differences*, **78**, 101819.
Yoo, S.-H. (2004), A Study on the Forecasting Model of Technology Life Cycles by Analysis of US Patent Citation, *Information Management Research*, **35**(1), 93-112.
Yoo, S.-H., Lee, Y.-H., and Won, D.-K. (2006), A study on estimation of technology life span using analysis of patent citation, *Journal of the Korean Operations Research and Management Science Society* **31**(4), 1-11.
Yoon, S.-H. (2004), A Study on the Forecasting Model of Technology Life Cycles by Analysis of US Patent Citation, *Information Management Research*, **35**(1), 93-112.
Yu-Heng, C., Chia-Yon, C., and Shun-Chung, L. (2010), Technology forecasting of new clean energy: The example of hydrogen energy and fuel cell, *African Journal of Business Management*, **4**(7), 1372-1380.
Zhang, Y., Shang, L., Huang, L., Porter, A. L., Zhang, G., Lu, J. and Zhu, D. (2016), A hybrid similarity measure method for patent portfolio analysis, *Journal of Informetrics*, **10**(4), 1108-1130.

저자소개

이지호 : 건국대학교 산업공학과, 컴퓨터공학과에서 2019년 학사학위를 취득하고 건국대학교 산업공학과 석박사통합과정에 재학 중이다. 주요 연구관심분야는 Machine learning-based prediction/decision system, Computational customer analysis, Computational patent analysis, Natural language processing for business in-

telligence이다.

이승현 : 건국대학교 산업공학과에서 2021년 학사학위를 취득하고 건국대학교 산업공학과 석박사통합과정에 재학 중이다. 주요 연구관심분야는 Social media mining for business opportunities, Data-driven prognostics and health management이다.

손은수 : 숙명여자대학교 약학대학에서 학사, 석사 학위를 취득한 후, ㈜녹십자에서 3년 간 재직하였다. 1995년부터 한국과학기술정보연구원에서 근무하고 있으며 글로벌R&D분석센터에서 책임연구원으로 재직 중이다. 주요 연구분야는 과학계량학(Scientometrics), S&T indicator, Network Analysis이며, 특히 바이오 과학기술분야의 데이터 기반 분석에 주력하고 있다.

윤장혁 : POSTECH 산업공학과에서 학사, 석사 학위를 취득한 후, LG CNS에서 4년 간 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 intelligence이다.

이재민 : 서울대학교 물리학과에서 학사, 박사(석박사 통합과정) 학위를 취득한 후, 삼성전자 반도체연구소에서 2년 간 재직하였다. 2010년부터 한국과학기술정보연구원에서 근무하고 있으며 현재는 미래기술분석센터 기술지능연구팀 팀장(책임연구원)으로 재직 중이다. 주요 연구분야는 데이터 분석 기반의 Technology Intelligence, Natural language processing이다.