ILI 예측모델 개발 프로젝트**:** 인플루엔자 **B**의 예측을 통한 의약품 재고 최적화

ILI Predictive Model Development Project: Optimizing Drug Inventory by Predicting Influenza B

중앙대학교 SCM 분석 학회 씀씀이 B팀
강다훈 김희연 김태형 이정우 유현승 윤설리 한예호 전주현

Abstract

This study aims to develop a predictive model for forecasting the prevalence of influenza B and, based on these forecasts, optimize the supply chain management (SCM) of pharmaceutical companies. Influenza B has a low mutation rate but tends to spread intensively during specific periods, primarily affecting school-age children and adolescents. To address this, an AI-based model was constructed to predict the timing and intensity of influenza B outbreaks using influenza-like illness (ILI) data. The analysis applied a multivariate time series forecasting method that combined ILI data from Korea with seasonal patterns observed in Argentina. The main variables included past Korean ILI data, past Argentine ILI data, average temperature, absolute humidity, and daily temperature variation.

The ultimate goal of the model is to accurately forecast the timing and intensity of influenza B outbreaks and provide actionable information for pharmaceutical companies' inventory management and supply chain planning. The prediction results are expected to be closely integrated with monthly and weekly production schedules and inventory operations, thereby reducing inventory losses and preventing supply chain risks in advance. Furthermore, this model is anticipated to enhance the financial stability and operational efficiency of pharmaceutical companies, contributing to SCM innovation and strengthening the global competitiveness of the domestic pharmaceutical industry.

Keywords: influenza B type, predictive model, supply chain management (SCM), ILI data, LSTM, time series prediction, inventory management, AI

요약

본 연구는 인플루엔자 B형의 유행을 예측하고 이를 바탕으로 제약기업의 공급망 관리(SCM)를 최적화하기 위한 예측 모델을 개발하는 것을 목적으로 한다. 인플루엔자 B형은 변이 발생 빈도가 낮지만 특정 시기에 집중적으로 유행하며, 주로 학령기 아동과 청소년에서 발생하는 특징을 가진다. 이를 위해 ILI(influenza-like illness) 데이터를 기반으로 인플루엔자 B형 유행의 시기와 강도를 예측하는 AI 모델을 구축하였다. 분석은 한국의 ILI 데이터와 아르헨티나의 계절 패턴을 결합한 다변량 시계열 예측 기법을 적용하였으며, 주요 변수로 과거 한국 ILI, 과거 아르헨티나 ILI평균 기온, 절대습도, 일교차를 사용하였다.

본 모델의 궁극적인 목표는 인플루엔자 B형의 유행 시기와 강도를 정밀하게 예측하여 제약기업의 재고 관리와 공급망 계획에 활용 가능한 정보를 제공하는 것이다. 예측 결과는 제약기업의 월별 및 주별 생산 계획과 재고 운영에 밀접하게 연계되어 재고 손실을 줄이고 공급망 리스크를 사전에 방지할 수 있을 것으로 기대된다. 나아가, 본 모델은 제약기업의 재무 안정성과 운영 효율성을 높이며, 국내 제약 산업의 SCM 혁신과 글로벌 경쟁력 강화에도 기여할 수 있을 것으로 전망된다.

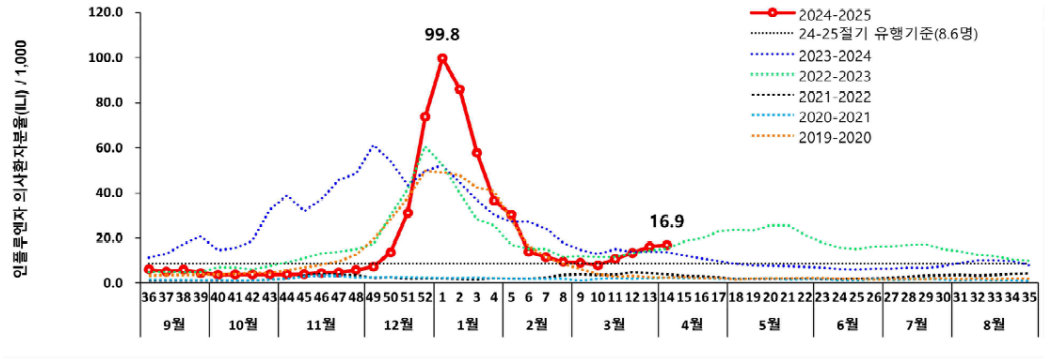
핵심어**(Keywords):** 인플루엔자 **B**형**,** 예측 모델**,** 공급망 관리**(SCM), ILI** 데이터**, LSTM,** 시계열 예측**,** 재고 관리**, AI**

---

1. 인플루엔자 **B**형 유행 예측의 필요성

전 세계적으로 계절성 인플루엔자(influenza)는 매년 약 10억 명 이상의 발병 사례가 보고되고 있다. 2025년 기준, 외래환자 1,000명당 인플루엔자 의사환자(ILI: influenza-like illness) 비율은 평균 8.6명을 상회하며 유행 수준을 지속적으로 초과하였다. 2025년 14주차에는 ILI 비율이 16.9명으로 4주 연속 증가세를 보였고, 이 중 B형

인플루엔자 검출률이 21.1%로 가장 높은 비중을 차지하였다(질병관리청, 2025). 이러한 유행은 보건 의료 시스템뿐만 아니라 의약품 공급망과 제약 산업 전반에 직접적인 부담을 초래할 수 있다..



【 인플루엔자 의사환자 분율(2016~2025년 14주차) 】

▲질병관리청 보도자료_인플루엔자 의사환자 분율 ***(2016~2025년 14주차)***

예를 들어, S제약은 인플루엔자 유행으로 인한 공급망 혼란으로 생산시설 내 과다 재고를 시가로 재평가하면서 2020년 3분기까지 약 31억 원의 손실을 기록했다. H약품 역시 재고자산 가치 하락으로 155억 원 규모의 손실을 처리했는데, 이는 같은 기간 영업이익(71억 원)의 두 배를 초과하는 수준이었다.

이처럼 인플루엔자 유행을 사전에 예측하지 못할 경우, 백신 및 의약품 재고 조절 실패로 인해 기업 재무구조에 심각한 영향을 미칠 수 있다. 이에 글로벌 제약사들은 이미 유행 예측 모델을 SCM(Supply Chain Management)과 연계해 수요 기반 생산 전략을 운영하고 있다. 예를 들어, GSK는 WHO FluNet 및 CDC 데이터를 바탕으로 바이러스 변이 정보를 분석하고, 머신러닝 기반 계절 예측 모델을 통해 6~9개월 전부터 백신 생산을 조율하여 수요 불균형을 최소화하고 있다.

반면, 국내 제약사는 과거 판매 실적과 단순 계절성 패턴에 의존하는 경향이 강해, 인플루엔자처럼 비선형적이고 변동성이 큰 감염병 특성을 반영하기 어렵다. 특히 인플루엔자 B형은 A형과 달리 유행 시기가 늦거나 이원화되는 경향이 있으며, 학령기 아동 및 청소년을 중심으로 특정 시기에 집중적으로 확산된다. 최근 국내 통계에서도 B형 인플루엔자는 봄철(3~4월)에 확산되고, A형과 유행 시점이 분리되는 특징이 확인되었다. 따라서 B형 인플루엔자에 대한 독립적인 예측은 단순 보건 정책을 넘어, 제약사의 백신 수급 전략에서 핵심 지표로 기능할 수 있다.

---

2. 분석 목표 및 예측모델 구축 전략

세계보건기구(WHO)와 미국 질병통제예방센터(CDC)는 머신러닝 기반 예측 시스템의 활용을 공식적으로 권고하고 있으며, 캐나다·미국·유럽 등 여러 국가에서는 인플루엔자 유행 예측 모델의 유효성이 이미 수차례 입증되었다. 이러한 흐름 속에서, ILI 데이터를 활용한 예측 모델은 단순한 보건 정보 제공을 넘어, 제약기업의 공급망 관리(SCM) 전략을 결정짓는 핵심 도구로 기능할 수 있다.

본 프로젝트는 이러한 흐름을 반영하여, ILI 데이터를 바탕으로 인플루엔자 B형 유행 강도와 시기를 정밀하게 예측하는 AI 모델을 구축하고, 이를 제약사의 SCM 체계에 적용하는 것을 목표로 한다.

구체적인 분석 단계는 다음과 같이 구성된다:

- 데이터 수집 및 시각화
- 예측 모델링(SARIMAX, LSTM)
- 모델 성능 평가: 회귀(R^2, MSE, MAE), 분류(Accuracy, F1, Recall)
- 결과 해석: 변수 간 상관관계 분석, 주요 인자 도출, 정책 적용 가능성 평가

본 모델은 검색 트렌드, 백신 접종률, 절대습도, 사회적 방역 정책 등 다양한 외생 변수를 반영함으로써 예측의 정밀도를 극대화하고자 하였다. 구체적으로는 시계열 기반 데이터 전처리, 변수 상관성 분석, 예측 모델링, 성능 평가 및 해석 단계로 구성되며, 궁극적으로는 예측 정확도를 기반으로 한 실용적 정책 대안 도출까지를 지향한다. 해당 모델이 구축될 경우, 국내 제약사는 매년 약 100억 원 이상으로 추산되는 재고 리스크를 줄이고, 적시 생산·공급 전략을 통해 백신의 낭비와 부족을 예방할 수 있을 것으로 기대된다.

---

3. 선행 연구 고찰 및 적용 가능성 평가

본 프로젝트에서는 인플루엔자 B형 유행 예측을 위한 분석 모델을 설계함에 있어, 국내외의 대표적인 선행 연구 세 편을 검토하였다. 각각의 연구는 예측모델의 변수 구성, 분석 프레임, 적용 가능성 등에서 의미 있는 시사점을 제공하였으나, 동시에 본 프로젝트의 목적에 부합하지 않는 한계점도 함께 나타났다. 다음은 각 논문에 대한 요약과 본 프로젝트에의 적용 가능성 분석이다.

연구 ***1.*** 캐나다 토론토 지역 기반 인플루엔자 ***B***형 분석

해당 연구는 2010년부터 2015년까지 6년간 캐나다 토론토 지역에서 수집된 기후 및 인플루엔자 데이터를 활용하여 B형 인플루엔자 유행과 주요 기상 변수 간의 관계를 분석하였다.

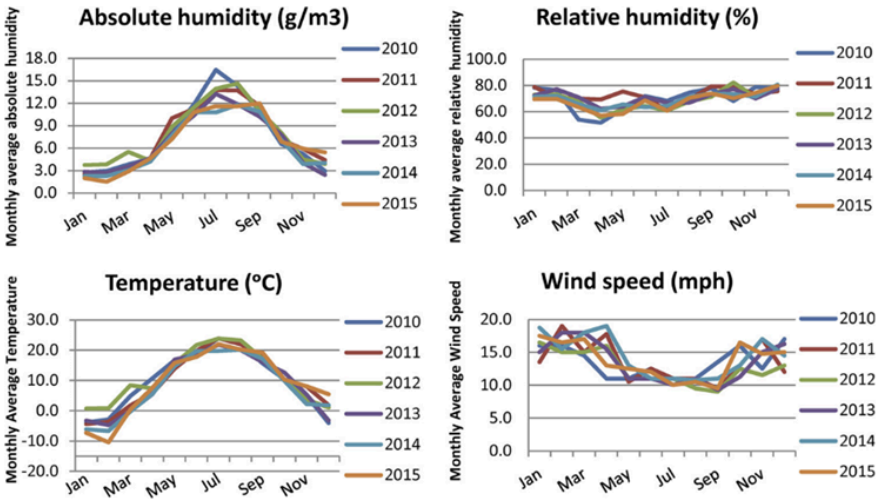


FIG 2 Absolute humidity, relative humidity, temperature, and wind speed in the Toronto area, January 2010 to December 2015. AH and temperature were the highest in the summer and the lowest in the winter, whereas RH and WS did not vary considerably between months.

▲토론토 절대습도, 상대습도, 기온 및 풍속

TABLE 1 Daily, weekly, and monthly values for environmental factors in Toronto area, January 2010 to December 2015[a]

Time interval	AH (g/m³)	RH (%)	Temp (°C)	WS (km/h)
Daily	6.2 (0.5 to 20.7)	71.1 (31.5 to 99.5)	9.4 (−20.7 to 31.1)	13 (4.01 to 42.1)
Weekly	6.0 (1.0 to 17.8)	70.2 (40.0 to 90.0)	9.3 (−16.9 to 27.3)	13 (6.0 to 28.0)
Monthly	6.2 (1.4 to 16.4)	70.8 (51.5 to 82.1)	9.9 (−10.4 to 23.8)	13 (9.0 to 19.0)

[a]The values represent the median (range). AH, absolute humidity; RH, relative humidity; WS, wind speed.

▲토론토 기후

위 연구에서는 4개 기상 관측소에서 수집한 기온, 절대습도, 상대습도, 풍속 데이터와 실험실 확진 인플루엔자 환자 데이터를 연계하여, 상관분석, 선형 회귀, 스플라인(spline) 기반 비선형 회귀 분석을 수행하였다.

이 연구에서 도출된 핵심 변수는 ▲온도 변동, ▲절대 및 상대습도, ▲기온, ▲연령이었다. 분석 결과, 일교차가 클수록 인플루엔자 B형 감염자 수가 유의하게 증가하였으며(온도 변동), 절대습도 10.5g/m³, 상대습도 60%를 초과하면 확진자 수가 급격히 감소하는 임계 패턴이 나타났다. 또한 기온이 15°C를 넘는 구간부터 B형 확진자 수가 급감하였으며, 이는 A형이 0°C를 기준으로 감소하는 것과 상반되는 양상이다. 연령별로는 5~19세 학령기 아동 및 청소년의 상대 위험도(IRR: 4.32)가 가장 높았으며, 풍속은 유의미한 영향이 없는 것으로 나타났다.

TABLE 6 Adjusted RH nonlinear negative binomial regression models exploring the relationship of environmental factors and nonlinearity of RH with influenza A and B viruses[a]

	All influenza viruses			Influenza A virus			Influenza B virus		
Demographic or climatic factor	IRR (95% CI)	*P* value	Nonlinearity *P* value	IRR (95% CI)	Association *P* value	Nonlinearity *P* value	IRR (95% CI)	*P* value	Nonlinearity *P* value
Age group (yr)									
65+	1.00	NA	NA	1.00	NA	NA	1.00	NA	NA
<1	0.67 (0.53-0.85)	<0.0001*	NA	0.69 (0.50-0.95)	0.0024*	NA	0.73 (0.49-1.10)	<0.0001	NA
1–4	1.32 (1.07–1.63)	<0.0001*	NA	1.18 (0.88–1.59)	<0.0001	NA	1.89 (1.33–2.67)	<0.0001*	NA
5–19	2.42 (1.95–3.01)	<0.0001*	NA	1.82 (1.33–2.48)	0.005*	NA	4.32 (3.07–6.09)	0.0037*	NA
20–64	1.19 (1.01–1.42)	0.0012*	NA	1.22 (0.96–1.54)	0.0247	NA	1.33 (0.99–1.78)	0.1316	NA
Outbreak status									
Yes	1.00	NA	NA	1.00	NA	NA	1.00	NA	NA
No	0.27 (0.23–0.32)	<0.0001*	NA	0.23 (0.19–0.29)	<0.0001*	NA	0.30 (0.22–0.39)	<0.0001*	NA
RH[b]	NA	0.0013*	0.0056*	NA	<0.0001*	0.4923	NA	<0.0001*	<0.0001*
Temp	NA	<0.0001*	<0.0001*	NA	<0.0001*	<0.0001*	NA	<0.0001*	<0.0001*
WS	1.00 (0.98–1.02)	0.8200	NA	1.00 (0.97–1.03)	0.8879	NA	0.99 (0.95–1.02)	0.6061	NA
Temp fluctuation	1.03 (1.01–1.05)	<0.0001*	<0.0001	0.99 (0.97–1.01)	<0.0001	NA	1.09 (1.06–1.11)	<0.0001*	NA

[a]RH, relative humidity; WS, wind speed; IRR, incidence relative risk, CI, confidence interval; NA, the measurement is not applicable for that variable. The RH nonlinear regression model explored the association of environmental factors with influenza activity as well as the nonlinearity of the association for RH and temperature with influenza A and B viruses and all influenza viruses. The left column lists independent/predictable variables for which this model was adjusted. The total weekly numbers of positive influenza A and B virus counts were used as dependent variables. A significant result (*) for association is considered when the 95% confidence interval does not cross 1 and the *P* value is <0.05. A significant result for nonlinearity is considered when the *P* value is <0.05. The incidence relative risk of 1.00 indicates the category used for reference/comparison. AH, temperature, and WS were measured by the use of weekly median measurements. Influenza A virus and influenza B virus represent the total weekly numbers of positive specimens. All influenza viruses represent the sum of influenza A and B virus-positive specimens.
[b]RH was also examined for a nonlinear association with influenza A and B viruses.

▲비선형성 회귀 모형 결과

해당 연구는 B형 인플루엔자의 활동성이 A형과 상이한 계절적·기상학적 특성을 가진다는 점을 명확히 보여주며, 본 프로젝트의 변수 구성과 비선형 회귀 모델 설계에 유의미한 참고자료가 되었다. 다만, 해당 연구는 캐나다 지역을 기반으로 한 분석이기 때문에, 대한민국과의 기후·생활환경 차이를 고려한 지역별 조정(local adaptation)이 필요하다는 한계도 존재한다.

연구 ***2. LightGBM***을 사용한 ***SHAP*** 기반의 설명 가능한 인플루엔자 발생 예측***(2020)***[9]

두 번째 검토 대상은 국내 데이터를 바탕으로 LightGBM 및 SHAP(Shapley Additive Explanations)를 활용해 인플루엔자 발생을 예측한 연구이다. 이 논문은 타국 ILI 환자수와 머신러닝 알고리즘을 기반으로 높은 예측

정확도를 추구하면서, SHAP 분석을 통해 각 변수의 중요도를 직관적으로 해석할 수 있도록 설계되었다. 또한, 국내의 실존 데이터를 활용해 로컬라이징이 용이하다는 장점이 있다.

해당 연구는 인플루엔자 발병에 영향을 미치는 기상 요인 및 사회적 요인을 반영하였다는 점에서 벤치마킹할 가치가 높으나, 타겟 변수로 A형과 B형을 구분하지 않고 전체 인플루엔자 및 유사 질환을 통합하여 분석했다는 점에서 직접적인 적용에는 제약이 있다. 결과적으로 B형 고유의 계절성이나 민감한 환경 반응을 반영한 모델 구성은 불가능하며, B형 특성에 대한 근본적인 통찰을 얻기에는 한계가 있었다.

연구 ***3. Forecasting seasonal influenza-like illness in South Korea after 2- and 30-weeks using Google Trends and influenza data from Argentina***[10]

해당 연구는 ARMAX 및 SARIMAX와 같은 통계 기반 시계열 기법을 통해, 아르헨티나의 인플루엔자 데이터를 외부 변수로 삼아 대한민국 내 인플루엔자 발생을 장기(30주 후) 예측하는 모델을 설계하였다.

본 프로젝트의 궁극적인 목적이 중장기 수요 기반 약품 생산 전략 수립에 있다는 점에서 해당 연구는 프레임 설계와 데이터 흐름 관점에서 참고할 부분이 있었다. 특히 Rolling Window 기법 등 시계열 모델링 측면에서 유용한 아이디어를 제공하였다. 그러나 이 연구 역시 전체 인플루엔자 대상이며, B형 단독 변수로 분리하여 분석하지 않았다는 점에서 직접적인 적용에는 한계가 존재한다. 또한, Google Trends 등 해외 검색 데이터를 활용하였다는 점은 국내 적용 시 데이터 해석 및 구조 차이로 인해 재구성이 필요하다.

---

4. 예측 모델 설계 및 변수 구성 전략

본 장에서는 인플루엔자 B형 유행 예측을 위한 모델링의 개요, 예측 대상 및 기준, 예측 단위, 모델링 방식, 그리고 주요 변수의 선정 및 제외 사유를 정리한다. 실험 결과나 성능 평가는 이후 5장에서 다룬다.

4.1 예측 목표 및 정의

본 프로젝트의 1차 목적은 주 단위 INF_B 환자수를 예측하여, 이를 기반으로 인플루엔자 B형의 유행 여부 및 시기를 사전에 판단하는 데 있다. 궁극적으로는 해당 예측값을 의약품 재고 및 생산 계획에 반영함으로써 제약기업의 재고 손실과 공급망 리스크를 최소화하는 것이 목표이다.

유행 여부 판단 기준은 미국 CDC의 권고안을 준용하되, 본 연구에서는 인플루엔자 B형의 특성을 반영하여 아래와 같이 정의의하였다.

- 유행 기준 (주차 t) : 최근 3년간 주별 INF_B 평균 + (1 × 표준편차)
  → 예측된 INF_B가 이 기준을 초과할 경우 “유행”으로 간주

CDC의 공식 기준은 인플루엔자 A형과 B형을 모두 포함한 통합 기준이다. 그러나 B형 인플루엔자는 A형에 비해 전체 발생 규모가 작고, 계절별 변동 폭도 상대적으로 낮아 동일한 임계치를 적용할 경우 실제 유행 국면을 포착하기 어렵다. 이에 본 연구에서는 인플루엔자 B형의 특성에 맞게 수치를 조정하여 공식에 반영하였다.

4.2 분석 단위 및 기간

항목	내용
예측 단위	주 단위 (Weekly)
학습 기간	2014년 1월 ~ 2024년 4월(코로나 기간 제외)
평가 기간	2024년 5월 ~ 2025년 4월

4.3 모델링 접근 전략

예측 방식은 시계열 기반 모델을 중심으로 설계되며, 다음 두 가지 모델이 실험에 포함된다:

- SARIMAX: 계절성과 외생 변수 처리가 가능한 전통 통계 기반 모델
- LSTM: 시계열 내 장기 의존성과 비선형적 특성을 반영할 수 있는 딥러닝 기반 모델

두 모델의 성능 및 한계는 실험 후 비교 분석하며, 최종적으로 더 높은 정확도를 보인 모델을 활용할 예정이다.

4.4 변수 구성 및 데이터 선택 기준

4.4.1 최종 사용 변수

변수명	설명
한국 ILI	국내 INF_B 주간 환자수
타국 ILI	선행 시그널
평균 기온	국내 서울 기준 주간 기온
일교차	평균 일일 기온 차
상대 습도	주간 평균 상대습도

4.4.2 제외된 변수 및 사유

변수	제외 사유
강수량	기존 문헌에서 상관성 낮음, 예측 타당성 부족
연령별 독감 데이터	인플루엔자 A/B형 미구분 데이터로 예측 정확도 저하 우려
백신 접종률	연 단위 데이터만 존재하여 주단위 예측과 시간 불일치
검색량 지표(Google Trends, Naver)	주 단위 데이터 누락 및 불연속성으로 인한 분석 정확도 저하 우려

5. 모델링

본 장에서는 인플루엔자 B형 양성률(Influenza B, INF_B)을 주간 단위로 예측하기 위한 전체 모델링 흐름을 설명한다. 사용 데이터의 유형과 전처리 방식부터, 시계열 예측 기법인 SARIMAX와 LSTM의 모델 설계, 성능 비교, 최종 변수 조합에 따른 예측 결과까지의 전 과정을 포함하며, 외생 변수를 구성하기 위해 시행한 국가 간 시계열 유사도 분석 역시 포함한다.

5.1 사용 데이터 설명

본 프로젝트는 총 두 가지 주요 범주의 데이터를 활용하여 예측 모델을 구성하였다.

(1) INF_B 전처리 데이터 **(WHO FluNet** 기반**)**

- 대상: 전 세계 국가별 B형 인플루엔자 주간 발생률
- 주요 컬럼 :
    - COUNTRY: 국가명
    - DATE, YEAR, WEEK, YEAR_WEEK: 날짜 정보
    - INF_B: 해당 주차의 INF_B 건수
    - occur_t-0 ~ occur_t-52: 시점 t 기준 직전 0~52주의 lag 값

(2) 국내 기상 변수

- 대상: 주간 단위로 집계된 기후 정보
- 주요 컬럼:
    - 일자: 주차 날짜 (YYYY-MM-DD)
    - 평균 기온, 일교차, 상대습도

5.2 데이터 전처리

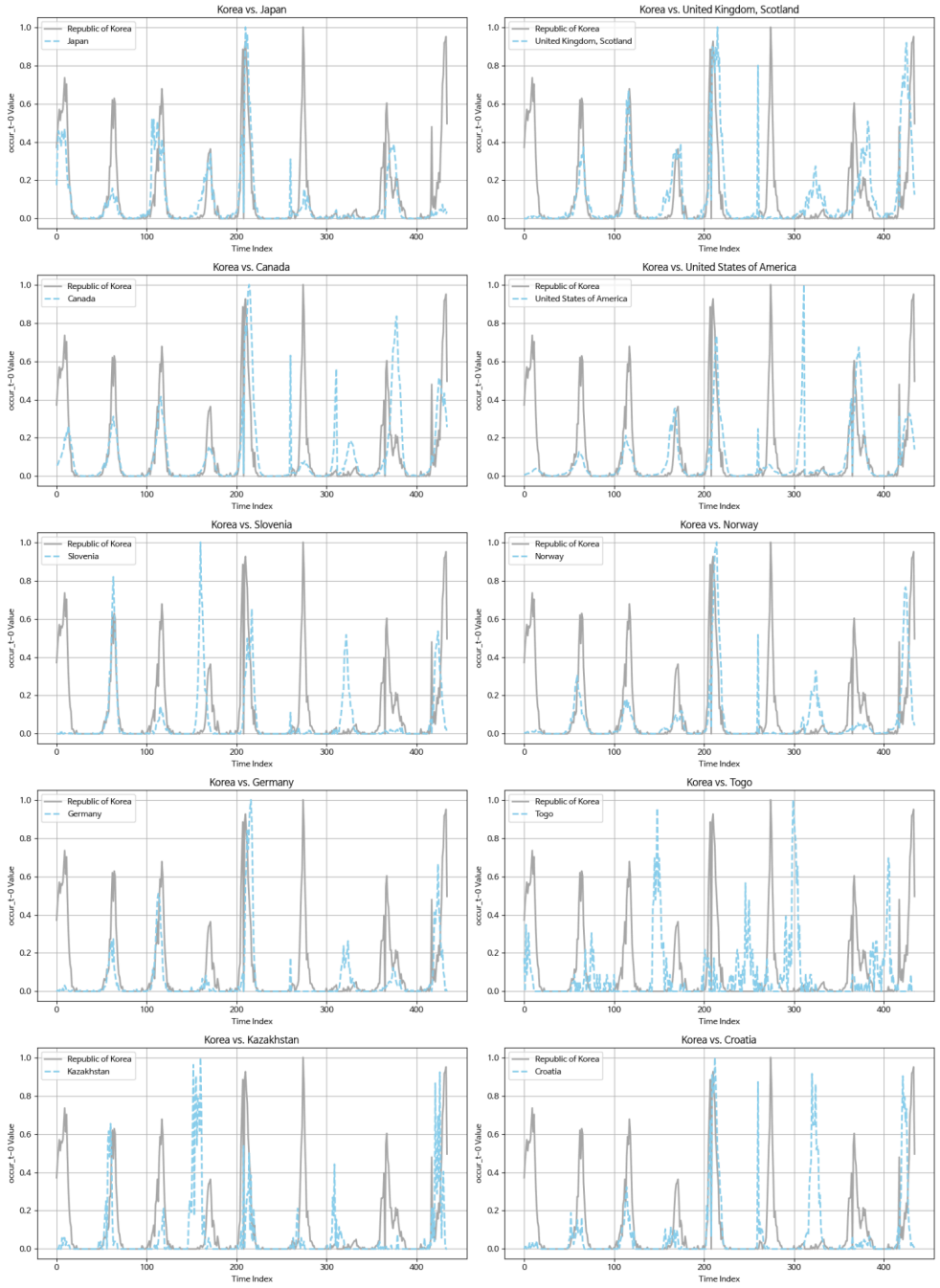
- 사용 기간: 2014.01 ~ 2025.05
- 제외 기간: 2020~2022년 (COVID-19의 구조적 비정상성 제거)
- 주요 전처리 항목:
    - 모든 시계열에 대해 YEAR_WEEK 생성
    - INF_B의 결측은 0으로 대체
    - 국가별 row 수를 동일하게 맞추기 위해 padding 수행
    - occur_t-0부터 occur_t-52까지의 lag 변수 생성
    - 모든 시계열 변수는 MinMax Scaling을 통해 정규화하여 비교 가능하도록 조정

5.3 국가 간 **ILI** 유사도 분석

한국 시계열과 유사한 국가를 식별하기 위해 코사인 유사도(Cosine Similarity)와 동적 시계열 거리(Dynamic Time Warping, DTW)를 이용해 각 국가별 유사도를 정량적으로 측정하였다.

(1) 분석 방식

- 코사인 유사도: 한국의 INF_B 시계열을 기준으로, 각 국가의 occur_t-0부터 occur_t-52까지의 lag 시계열과의 유사도를 계산하였다. 각 국가별로 0~52주 사이의 lag 중 가장 높은 유사도를 대표값으로 채택하였다.
    - 코사인 유사도는 시계열의 변화 방향을 직관적이고 신속하게 비교할 수 있는 방법으로, 이를 통해 한국과 동일한 시점에서 유사한 추세를 보이는 국가를 식별하고 예측 모델의 외생 변수로 활용하였다.
- **DTW** 거리: 한국 시계열과 각 국가의 시계열 간 시간 왜곡을 허용한 상태에서 전체 패턴의 유사성을 측정하였다. 시계열 길이를 일치시킨 후, DTW 거리 값을 계산하고 이를 0~1 범위로 정규화하였다.
    - DTW는 시점이 어긋난 경우에도 전체적인 곡선 구조의 유사성을 비교할 수 있는 방법으로, 이를 통해 시차를 두고 유사한 패턴을 보이는 국가를 선별하여 선행 예측 신호로 활용 가능한 외생 변수 후보를 식별하였다.



▲ **DTW** 상위 **10**개국 **vs** 한국 **ILI** 시계열 플롯

- 최종 점수 **(final_score)**:
  - final_score = cosine_similarity - scaled_dtw
  - 유사도가 높고 거리 차가 적은 국가가 우선 선별됨

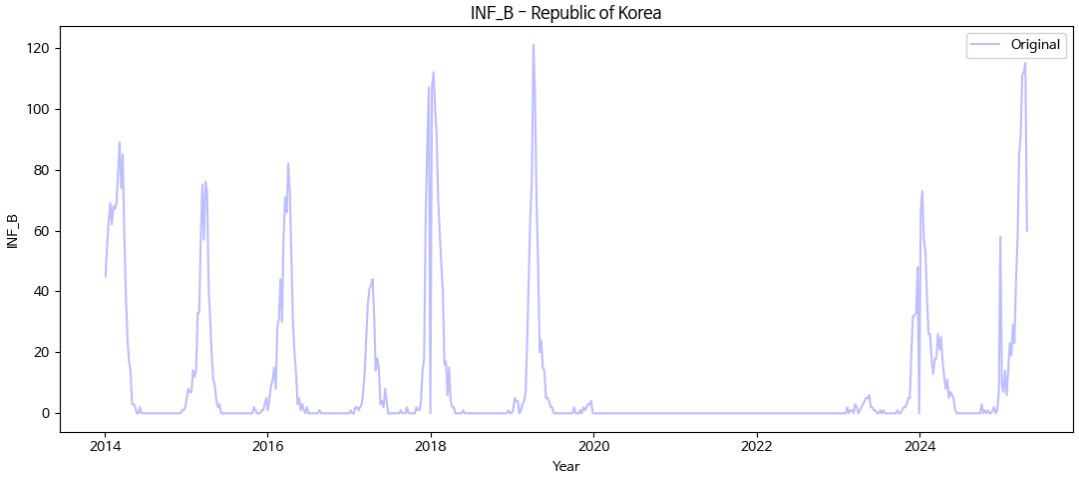
(2) 주요 결과 **(**상위 **10**개국**)**

순위	국가명	**final score**
1	Japan	0.7367

2	Canada	0.6218
3	UK, Scotland	0.5778
4	USA	0.4252
5	China	0.3819
6	Norway	0.3606
7	Turkey	0.3582
8	Slovenia	0.3288
9	Argentina	0.2996
10	Germany	0.2895

5.4 SARIMAX 예측 모델

5.4.1. ADF 정상성 검정 및 계절성 분석



▲ 대한민국 **INF_B** 시계열 그래프

(A) 정상성*(Stationary)* 검정

SARIMAX 모델은 시계열의 평균, 분산, 자기상관 구조가 시간에 따라 변하지 않는 정상성을 전제로 한다. 만약 정상성이 없으면 모델의 계수 추정이 왜곡되고 예측 정확도가 낮아질 수 있기 때문에, 모델링 전에 반드시 이를 확인해야 한다.

본 연구에서는 Augmented Dickey-Fuller(ADF) 검정을 통해 정상성을 평가하였다. ADF 검정은 시계열에 단위근(unit root)이 존재하는지 확인하는 방법으로, 단위근이 있으면 비정상, 없으면 정상으로 본다. 귀무가설(H_0)은 “단위근이 존재한다(비정상)”이며, 대립가설(H_1)은 “단위근이 존재하지 않는다(정상)”이다.

검정 결과는 다음과 같다.

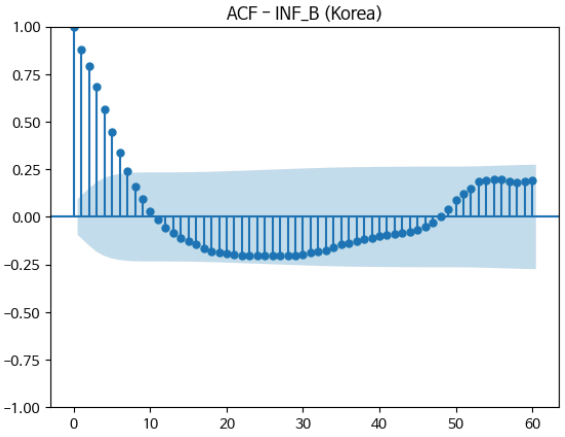
- ADF Test Statistic: -5.8286
- p-value: 0.0000
- 임계값(Critical Values):
  - 1%: -3.4456
  - 5%: -2.8683
  - 10%: -2.5704

검정 결과, ADF 통계량은 -5.8286, p-value는 0.0000으로 나타났고, 이는 모든 유의수준(1%, 5%, 10%)에서 임계값보다 작아 귀무가설이 기각됨을 의미한다. 따라서 한국의 INF_B 시계열은 통계적으로 정상성을 만족한다고 볼 수 있다.

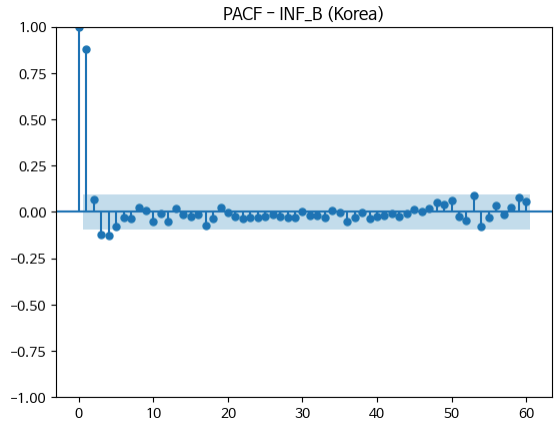
즉, 본 연구의 데이터는 SARIMAX 모델 적용을 위한 기본 요건을 충족하며, 이후 계절성 및 모형 차수 분석의 기초로 활용될 수 있다

(B) 계절성 분석

또한 시계열의 계절성 존재 여부를 분석하기 위해 자기상관 함수(ACF) 및 부분 자기상관 함수(PACF) 그래프를 시각화하였다.

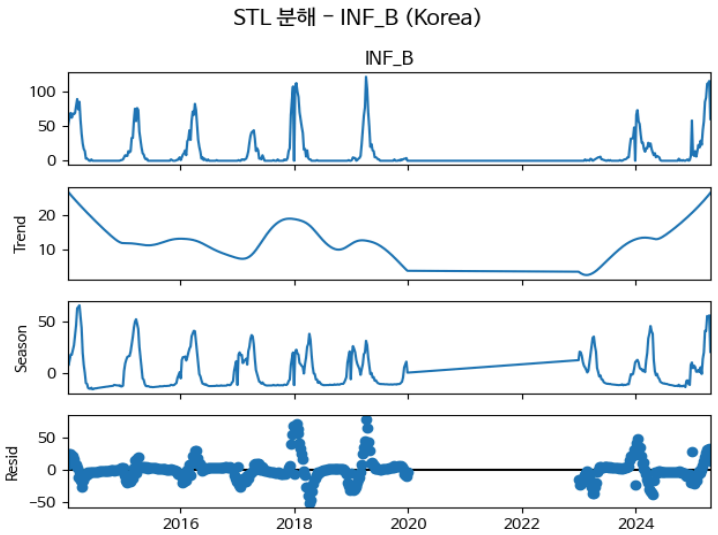


▲ 자기상관함수**(ACF)**



▲부분자기상관함수**(PACF)**

ACF에서는 약 52주 시점에서 다시 상승하는 모습이 나타나, 1년 단위의 계절 주기가 존재할 가능성을 보여주었다. PACF에서는 1차 시점에서 뚜렷한 spike가 나타난 뒤 빠르게 감소하는 패턴을 보였는데, 이는 AR(1) 구조가 적합할 수 있음을 시사한다.



▲ **STW(Seasonal-Trend decompostion using Loess)**분석

STL(Seasonal-Trend decomposition using Loess) 분석 결과, INF_B 시계열은 매년 겨울 정점 이후 급격히 감소하는 뚜렷한 계절적 패턴을 반복하는 것으로 나타났으며, 이는 Seasonal 성분에서도 일관되게 확인되었다

OLS Regression Results

Dep. Variable:	INF_B	R-squared:	0.404
Model:	OLS	Adj. R-squared:	0.323
Method:	Least Squares	F-statistic:	4.973
Date:	Thu, 10 Jul 2025	Prob (F-statistic):	7.98e-21
Time:	17:16:46	Log-Likelihood:	-1911.6
No. Observations:	435	AIC:	3929.
Df Residuals:	382	BIC:	4145.
Df Model:	52		
Covariance Type:	nonrobust		

▲ **OLS** 회귀분석 결과

계절성을 수치적으로 검증하기 위해 주차(week)를 기준으로 52개의 더미 변수를 투입해 OLS 회귀분석을 수행한 결과, 절반 이상에서 p<0.05로 유의성이 나타났다. 또한 F-통계량도 유의수준을 통과하여, 인플루엔자 발생에 주차가 중요한 계절 요인임을 통계적으로 입증하였다.

이상의 결과를 종합하면, INF_B 시계열은 정상성과 뚜렷한 계절성을 동시에 보유하고 있으며, 이는 SARIMA 또는 SARIMAX 모델이 적용될 수 있는 기반 요건을 충족한다. 다만, 이후 분석에서 확인되는 바와 같이 팬데믹 기간의 구조적 변동성 및 외부 충격의 영향으로 인해, 고정된 계절성과 선형 추세를 가정하는 SARIMA의 적용에는 일정한 한계가 존재할 수 있다.

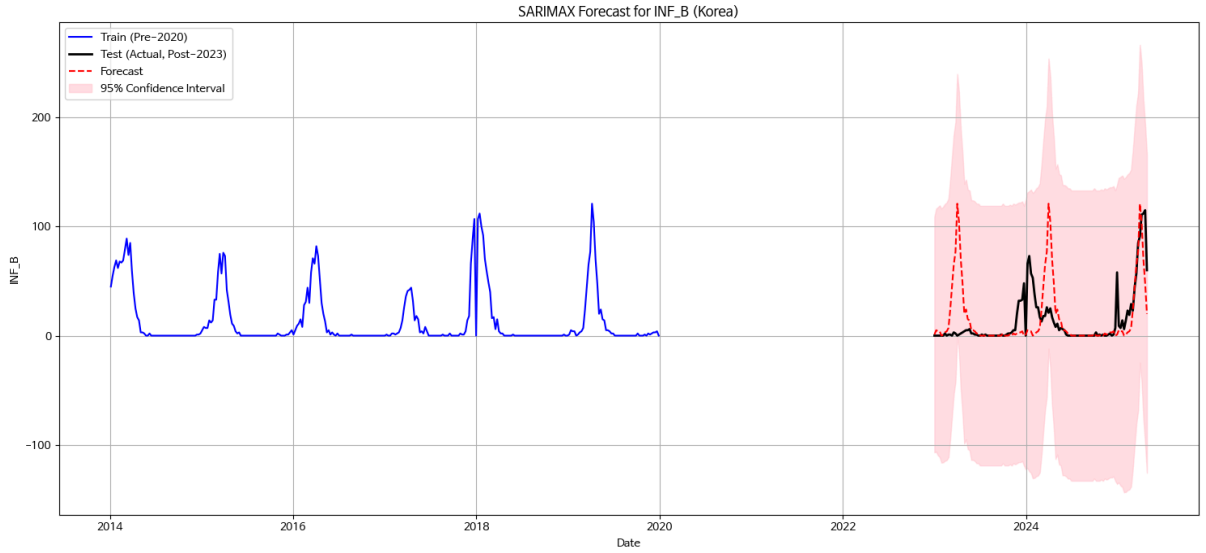
5.4.2 (p,d,q,S) 설정 및 모델(한국 단독) Fitting

앞선 정상성 및 계절성 검정 결과, 한국의 INF_B 시계열은 안정적이고 계절성이 뚜렷한 것으로 확인되었다. 이를 바탕으로, 본 연구에서는 시계열 예측에 널리 사용되는 ARIMA 기반 모델 중 계절성과 외생 변수를 반영할 수 있는 SARIMAX(Seasonal ARIMA with eXogenous variables)를 활용하였다.

모형 설정 과정에서 비계절 모수(p, d, q)는 ACF와 PACF 분석 결과를 참고하여 (1,0,0)으로 결정하였다. 특히 PACF에서 1차 시점의 강한 spike 이후 빠른 감소가 나타나, 자기회귀(AR) 성분이 1차까지만 유효함을 시사하였다. 계절 모수(P, D, Q)는 주기 s=52(1년 단위)로 고정한 뒤 Grid Search를 통해 평가한 결과, (0,1,0)이 최적 조합으로 도출되었다. 이는 계절적 변화를 설명하기 위해 1차 차분이 필요하며, 추가적인 계절 AR이나 MA 요소는 크지 않음을 의미한다.

따라서 최종 모델은 SARIMAX(1,0,0) × (0,1,0,52)로 구성되었으며, 학습 구간은 코로나19 이전의 2014~2019년, 검증 구간은 2023년 1월부터 2025년 4월까지로 설정하였다.

성능 평가 결과, MAE는 16.13, MSE는 875.61, RMSE는 29.59로 나타났고, R^2는 -0.4437을 기록하였다. 특히 R^2가 음수라는 것은 모델이 단순 평균값으로 예측하는 것보다 설명력이 낮다는 뜻으로, 예측력이 매우 제한적임을 보여준다. 이는 SARIMAX가 고정된 계절성과 선형 구조를 전제로 하기 때문에, 팬데믹 이후의 불규칙하고 비선형적인 유행 패턴을 충분히 반영하지 못했기 때문으로 해석된다



▲ **SARIMAX** 모델 예측 결과

이러한 결과는 단일 시계열만을 활용한 예측의 한계를 보여준다. 특히 국외에서 먼저 발생하는 독감 유행과 같은 외부 요인을 고려하지 못한 점이 성능 저하의 원인일 수 있다. 이에 따라 본 연구에서는 성능 개선을 위해, 한국과 유사한 국가의 시계열을 외생 변수로 추가하는 모델 확장을 수행하였다.

5.4.3 외생변수**(**타국 **ILI)** 추가 **SARIMAX**

외생 변수의 후보 국가는 앞서 수행한 국가 간 유사도 분석 결과를 바탕으로 선정하였다. 구체적으로는 코사인 유사도 및 DTW 거리 기반의 종합 점수(final_score)를 기준으로, 유사도가 높은 국가 순으로 시계열을 차례로 반영하였다. 각 국가별로 최적의 시차(lag)를 고려하여 시계열을 이동시킨 후, 해당 변수를 exogenous variable로 추가하였다. 외생 변수의 선택 방법은 다음과 같다:

(1) final score 기반 랭킹 상위 국가를 대상으로 순차적으로 모델 외생변수에 추가

(2) 국가별 최적 lag를 적용하여 시계열 shift

(3) 외생 변수 계수의 통계적 유의성($p < 0.05$) 확인

(4) 예측 성능 지표 RMSE 개선 여부를 기준으로 채택 여부 결정

실험에 사용된 데이터는 기존과 동일하게 2014년 1월부터 2019년 12월을 훈련구간, 2023년 1월부터 2024년 4월까지의 테스트 구간으로 설정했으며, 팬데믹 기간인 2020~2022년은 제외하였다. 모델 구조는 SARIMAX(1,0,0) × (0,1,0,52)로 고정하였다.

외생 변수로 첫 번째로 추가된 국가는 일본(Japan)이었으며, lag 없이 동시 시점(t)에서의 INF_B 데이터를 적용하였다. 그 결과 RMSE는 29.59에서 29.17로 약 1.44% 감소하였으며, 변수 계수 역시 p=0.000으로 통계적으로 유의하였다. 반면, 두 번째로 추가된 국가는 캐나다(Canada)였으나, 해당 변수는 예측 성능을 오히려 저하시켰고(RMSE 32.40), 통계적으로도 유의하지 않아 최종 모델에는 포함되지 않았다.

<실험 결과 요약**>**

단계	추가 국가 **(lag)**	모델 **RMSE**	**R^2 Score**	성능 개선 여부	외생변수 유의성**($p<0.05$)**

0	(Baseline)	29.5906	-0.4437	-	-
1	Japan (0)	29.1655	-0.4025	개선 (↓1.44%)	p = 0.000 < 0.05 (유의)
2	Canada (0)	32.4021	-0.7310	성능 저하	추가하지 않음

5.4.4 SARIMAX 결론

이러한 실험 결과는 일부 국가의 ILI 발생 패턴이 한국의 유행과 일정한 상관성을 보이며, 선행 지표로서의 가능성을 지닌다는 점을 시사한다. 그러나 전체적으로는 SARIMAX 구조 자체의 한계가 명확히 드러났다. 우선, 팬데믹에 따른 시계열 연속성의 훼손으로 인해 전체 패턴의 일관성이 약화되었고, 이로 인해 모델의 예측력이 크게 떨어졌다. 또한 SARIMAX는 본질적으로 선형 추세와 고정된 계절성을 가정하기 때문에, 최근 수년간 급격하게 변화한 유행 양상이나 예외적 패턴을 반영하는 데 한계가 있었다. 실제로 예측 성능을 나타내는 결정계수(R^2)는 외생 변수 유무와 관계없이 음수(-0.44 이하)를 기록하여, 단순 평균값 예측보다도 낮은 수준의 설명력을 보였다.

결과적으로, SARIMAX 모델은 본 연구의 목적에 부합하는 예측 정확도와 실용적 설명력을 확보하기 어렵다는 결론에 도달하였으며, 보다 유연하게 비선형성과 복합적인 요인을 반영할 수 있는 LSTM 기반 딥러닝 모델로 분석을 전환하게 되었다.

5.5 LSTM X 예측 모델

5.5.1 모델 선정의 이유

이번 프로젝트를 진행하는 과정에서, 궁극적인 목표는 Influenza-B의 국내 발병률을 예측하는 것이었다. 바이러스의 주기성, 계절성을 반영하는 것이 중요하다고 판단하여 SARIMAX(Seasonal Auto-Regressive Integrated Moving Average) 모델을 통한 수요예측을 시도하였다. SARIMAX는 주기성이 뚜렷한 시계열 데이터를 기반으로 일정한 패턴을 예측하는 데 강점을 가지며, 외생 변수를 함께 포함할 수 있다는 점에서 유연한 분석이 가능하기 때문이다.

그러나, 데이터 수집 및 전처리 전처리 과정에서, Influenza-B의 최근 유행 양상은 과거와는 상이한 비선형적 변화가 존재하며, 특히 2020~2022년 사이의 COVID-19 팬데믹 기간 동안 시계열의 연속성이 크게 훼손되었던 점을 확인할 수 있었다. 또한, 유행구간의 정점 시기나 강도의 변동성 또한 기존 패턴과 일치하지 않는 현상을 확인하였다. 이러한 비선형성과 구조적 변화는 고정된 계절성과 선형 추세를 전제로 설계된 SARIMAX 모델이 충분히 반영하기 어려운 요인이며, 실제 예측 성능에서도 그 한계가 확인했다. 모델의 결정계수(R^2)는 -0.44로 음수를 기록하여 단순 평균값을 예측하는 것보다도 설명력이 낮은 수준이었고, RMSE(성능 평가의 지표, 낮을수록 높은 성능) 또한 29.6으로 매우 높게 나타났다.

이와 같은 문제점을 해결하고 보다 복잡한 패턴을 반영하기 위해, 딥러닝 기반의 LSTM(Long Short-Term Memory) 모델을 이용한 분석으로 방향을 전환하였다. LSTM은 시계열 데이터 내의 장기 의존성과 시점 간 시간 지연 효과를 효과적으로 학습할 수 있는 구조로, 기존의 선형 기반 모델이 다루기 어려운 비정형적 패턴이나 외부 요인의 비선형 결합 등을 학습하는 데 적합한 모델로 평가된다. 특히 본 프로젝트의 경우, 해외 국가의 ILI 발생 패턴이나 기후 변수 등의 외생 정보를 함께 반영할 필요가 있었기 때문에, LSTM 구조의 적용이 더욱 타당하다고 판단하였다.

5.5.2 Influenza-B의 유행 여부 동적 라벨링 함수

LSTM 기반 예측 모델의 출발점으로는 먼저, 예측 대상 변수인 Influenza-B를 정량 수치뿐만 아니라 이진 분류 형태로도 활용할 수 있도록 유행 여부*(유행 vs 비유행)*를 라벨링하는 절차를 도입하였다. 이는 기존의 고정 임계값 방식 대신, 각 시점에서 동적으로 유행 기준선을 설정하는 방식으로 구현하였다.

1) 각 시점의 직전 *156*주*(*약 *3*년*)*의 *INF_B* 데이터를 참조*,* 해당 구간의 평균값*(μ□)*과 표준편차*(σ□)*를 계산

2) 결과값에 기반하여 유행 기준선*(Threshold□ = μ□ + σ□)*을 설정

3) 해당 주의 Influenza-B 수치가 이 기준선을 초과할 경우 *'유행(1)',* 이하일 경우 '비유행*(0)'*으로 라벨링

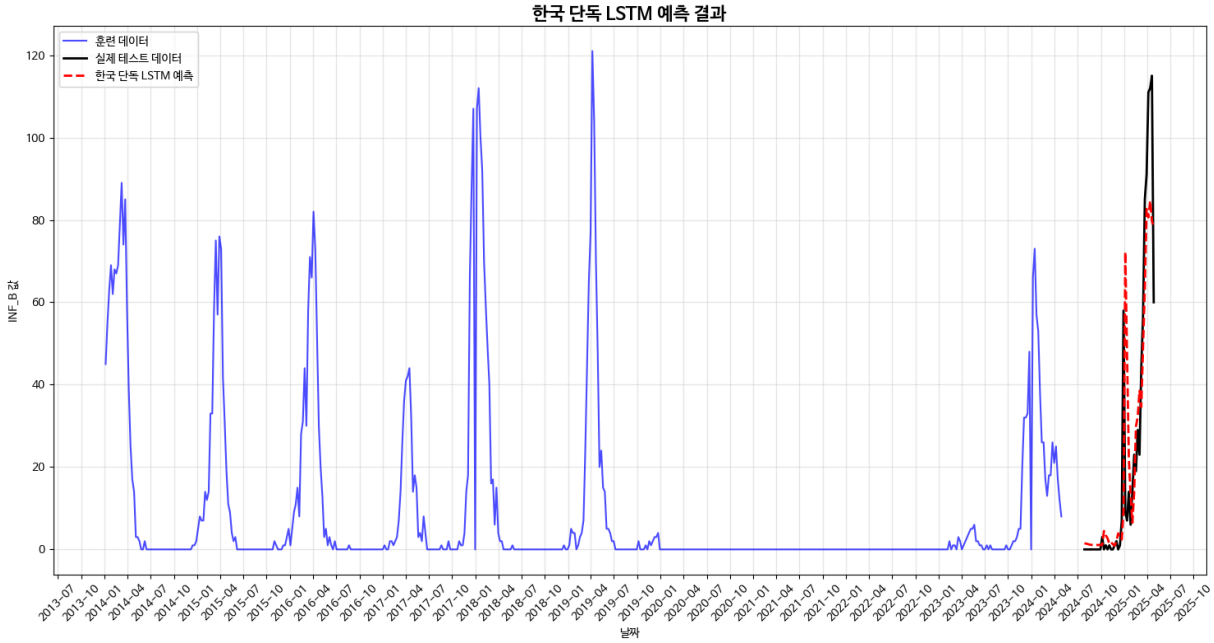
구체적으로 도식화하면 다음과 같다. 이때 최소 156주의 과거 데이터를 필요로 하므로, 시계열 초반인 2014~2016년은 라벨 산출이 불가능하며, 최종적으로는 2017년 이후 구간에 대해 유효한 유행 여부 라벨이 생성되었다. 이 라벨은 후속 LSTM 모델의 예측 결과를 평가할 때 이진 분류 지표(정확도/Accuracy, 재현율/Recall, F1-score)로 활용된다.

5.5.3 한국 **ILI** 기반 **LSTM** 모델

이러한 구조적 전환 이후, 본격적인 LSTM 모델을 구축하기 위해 가장 먼저 수행한 실험은 한국 단독 Influenza-B 시계열 데이터만을 입력 변수로 사용하는 단일 변수 기반 LSTM 모델이었다. 입력 시퀀스는 12주로 구성하여, 해당 기간의 정보를 바탕으로 다음 1주의 Influenza-B 값을 예측하도록 설계하였다. 기간 설정에 대한 근거는 Influenza-B의 자체 특성인 계절성 신호의 반영 정합성이다. 3개월(12주)는 시작/정점/소강으로 이뤄지는 한 단위의 추세를 반영할 수 있으며, 이를 더 늘리거나 줄일 경우 과적합/장기의존성 등의 문제를 일으킬 수 있다. 모델 구조는 LSTM(64unit) → Dense(1)의 단순한 구조로 설계되었으며, 최적화에는 RMSProp 옵티마이저를, 손실 함수에는 평균제곱오차(MSE)를 적용하였다. 학습 데이터는 2014년부터 2024년 4월까지이며, 이후의 데이터를 테스트 데이터로 활용하였다.

설정 항목	값
입력 시퀀스 길이	12주
예측 대상	INF_B
모델 구조	LSTM(64) → Dense(1)
최적화 및 손실 함수	RMSProp (optimizer), MSE (loss)
epochs	80
batch size	8
Validation split	20%

이 기본 모델의 예측 성능은 다음과 같다. RMSE는 17.7684, MAE는 10.2214, 결정계수 R^2는 0.7406으로 기록되었다. 이는 기존 SARIMAX 모델 대비 크게 향상된 수치로, 모델이 단순히 평균 추세만을 학습하는 것이 아닌, 유행 시점과 강도 등 변동성을 일정 수준 이상 반영하고 있음을 보여준다. 특히 R^2 지표가 양수로 전환되며 0.74를 기록한 점은 모델의 설명력이 강력하며, 통계적으로 유의하다는 신호로 해석된다.



▲ 한국 단독 데이터 반영 **LSTM** 예측 결과

현재까지의 결과는 LSTM이 인플루엔자 B형의 비선형적 유행 양상과 시계열 내 복합적인 신호를 효과적으로 학습할 수 있는 적절한 대안 모델임을 보여주고 있다. 이후 외생 변수(해외 ILI, 기후 등)의 단계적 추가를 통해 예측 정확도를 더욱 향상시키는 방향으로 모델을 확장해 나갈 것이다.

5.5.4 한국 **+** 아르헨티나**(Lag 32) ILI** 기반 **LSTM** 모델

LSTM 기반 예측 모델 구축 과정에서, 기존 SARIMAX와 마찬가지로 한국과 시계열 패턴이 유사한 다수 국가의 ILI 데이터를 외생 변수로 추가하는 방식을 시도하였다. 앞서 SARIMAX와 동일하게 Final score가 높은 북반구 국가들(일본, 캐나다, 영국 등)을 대상으로 다변량 LSTM 모델을 구축하였다. 그러나 실험 결과, 이들 국가 데이터를 추가한 모델의 예측 성능은 오히려 한국 단독 모델보다 낮은 R^2 값을 기록하며 하락하는 양상을 보였다.

성능 저하의 원인은 두 가지로 해석될 수 있다. 첫째, 한국과 시계열이 유사한 국가 대부분이 동일한 북반구에 위치하고 있어, 구조적으로 비슷한 시점에 동일한 패턴이 나타나는 데이터(lag ≈ 0)를 입력 변수로 사용하는 셈이 되었다. 이로 인해 모델이 중복된 정보를 과도하게 학습하게 되었고, 유효한 시간 지연 신호(Lead signal)를 반영하지 못한 채 차원만 증가한 과적합 구조로 이어진 것이다. 둘째, LSTM 구조상 입력 피처 수가 늘어날수록 상대적으로 품질이 낮은 피처가 노이즈로 작용할 가능성이 높아지는데, 여기서 한국과 지나치게 유사한 북반구 국가 데이터는 유익한 정보보다는 오히려 불필요한 분산을 유발하는 결과로 나타났다.

구조적 한계를 극복하기 위해 Lead Time의 문제를 해결하였다. 남반구 국가 중 한국과의 시계열 유사도 상위 국가에 주목하였는데, 남반구의 경우 계절이 반대이므로 시계열 상 몇 개월 앞선 시점에서 유사한 패턴이 나타날 가능성이 있다. 실제로 코사인 유사도와 DTW 점수 기반 유사도 분석 결과, 아르헨티나는 약 32주 앞선 시점에서 한국, 일본과 비슷한 유행 양상을 보이는 국가로 분석되었다. 이에 따라, 아르헨티나의 Influenza-B 시계열 데이터를 32주 시차를 적용해 재정렬하여 한국 시계열과 병합하여 예측 모델의 입력 변수로 사용하였다.

모델은 기존과 동일한 LSTM(64) → Dense(1) 구조를 유지하였고, 입력 시퀀스는 12주, 예측 대상은 한국의 Influenza-B 수치로 설정되었다. 학습 구간은 2014년부터 2024년 4월까지, 테스트 구간은 2024년 5월 이후로 설정하였다.

설정 항목	값
입력 시퀀스 길이	12주
예측 대상	INF_B
모델 구조	LSTM(64) → Dense(1)
최적화 및 손실 함수	RMSProp (optimizer), MSE (loss)
epochs	80
batch size	8
Validation split	20%

성능 평가 결과 RMSE는 16.7174, MAE는 10.6196, R^2는 0.7926으로 나타났으며, 이는 단일 변수 모델(R^2 = 0.7406) 대비 유의미한 성능 향상을 보여주었다. 아르헨티나의 시차 적용된 시계열이 한국 Influenza-B의 유의미한 선행 지표 역할을 할 수 있음을 시사한다.

▲ 한국 + 아르헨티나 데이터 반영 **LSTM** 예측 결과

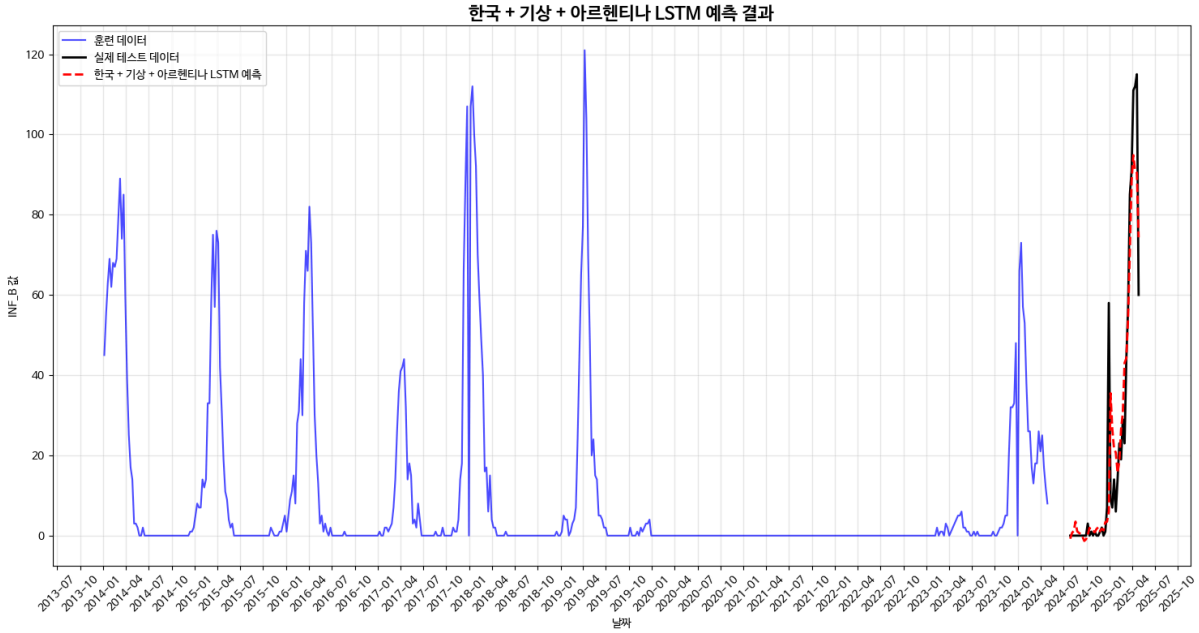
5.5.5 한국 **+** 아르헨티나**(Lag 32) ILI +** 국내 기상 변수 기반 **LSTM** 모델

앞선 실험에서 확인한 바와 같이, 아르헨티나의 독감 시계열은 한국의 Influenza-B 예측에 선행 신호로써 긍정적인 영향을 미치는 것으로 나타났다. 이에 본 연구에서는 국내 기상 변수(평균 기온, 일교차, 상대습도)를 추가하여 다변량 LSTM 모델을 확장하였다. 이 세 변수는 참고문헌과 선행 연구에서 Influenza-B형 발생과 높은 상관성을 보인 요인들로, 추가 변수로의 적합성을 입증하였다.

결과적으로 입력 시계열은 총 다섯 가지(한국 Influenza-B, 아르헨티나 32-lagged Influenza-B, 국내 기온, 일교차, 상대습도)로 구성되며, 시퀀스 길이, 모델 구조, 학습 조건 등은 이전 실험과 동일하게 유지하였다. 학습 구간은 COVID-19 펜데믹 기간(2020~2022)을 제외한 2014년부터 2024년 4월이며, 테스트 구간은 2024년 5월 이후로 설정하였다.

설정 항목	값
입력 시퀀스 길이	12주
입력 변수	INF_B, ARG_INF_B, 기온, 일교차, 습도
모델 구조	LSTM(64) → Dense(1)
최적화 및 손실 함수	RMSProp (optimizer), MSE (loss)
epochs	80
batch size	8
Validation split	20%

해당 모델의 성능은 RMSE 12.5929, MAE 7.0014, R^2 0.8697로, 기존 모든 실험 대비 가장 높은 예측 정확도를 보였다. 특히 RMSE 기준으로는 한국 단독 모델 대비 약 **29%,** 아르헨티나 추가 모델 대비 약 **25%** 이상 감소하는 성과를 보였으며, 이는 기후 변수들이 예측 성능 향상에 실질적으로 기여했음을 의미한다. 이러한 결과는 독감 예측 모델에 있어 선행 해외 시계열 데이터와 기후 외생변수의 조합이 가장 효과적인 입력 구조임을 시사하며 이후 정책적 적용 가능성 또한 높인다.



▲ 한국 **+** 아르헨티나 **+** 기상데이터 반영 **LSTM** 예측 결과

5.5.6 한국 **+** 아르헨티나 **+** 남반구 다국가 기반 **LSTM** 모델

마지막으로, 아르헨티나를 제외한 시계열 유사도가 높은 남반구 국가들(남아공, 칠레, 호주, 뉴질랜드, 브라질)을 추가 변수로 포함시키는 다국가 병합 모델을 실험하였다. 각국의 lag 값은 유사도 기반 사전 분석 결과에 따라 개별 적용되었으며, 적용한 lag는 밑의 표와 같다.

국가	**lag**	**cosine** 유사도	최종 점수
South Africa	31	0.447921	0.054122
Chile	19	0.443564	-0.393376
Australia	23	0.393162	0.015700
New Zealand	30	0.381004	0.010079
Brazil	43	0.276682	-0.255194

그러나 예상과는 달리 다국가 병합 모델의 성능은 기존 **2**개국 모델 대비 하락하였다. 뉴질랜드가 포함된 모델의 경우 RMSE는 17.7562, R^2는 0.7661로 측정되었으며, 이는 아르헨티나 단독 추가 모델(R^2 = 0.7926)보다 낮은 수치이다. 남아공, 호주, 칠레, 브라질 등 다른 국가들을 병합한 모델에서도 유사한 경향이 반복되었다.

이와 같은 결과는 다국가 병합 시 발생하는 입력 변수의 과도한 증가와 데이터 품질의 이질성에서 기인한 것으로 분석된다. 특히 COVID-19 기간 데이터의 제거로 인해 학습 데이터가 상대적으로 제한된 상황이며, 여기서 입력 차원이 급격히 늘어나면 모델이 노이즈에 과도하게 반응하는 Overfitting의 가능성이 높아진다. 또한, 동일한 lag 구조로 병합하더라도, 일부 국가의 독감 유행 패턴은 국지적인 패턴을 갖고있으며, Domestic Variance가 유효한 패턴을 학습하기 어렵게 한다는 한계점이 작용했다고 볼 수 있다.

결론적으로, 본 실험에서는 아르헨티나 단일국가와 기후 변수를 조합한 모델이 예측 효율성과 실용성 측면에서 가장 우수하다는 결론을 도출하였으며, 다국가 병합 전략은 일반화 성능을 저해한다는 시사점을 얻어내었다.

5.5.7 이진 분류 평가

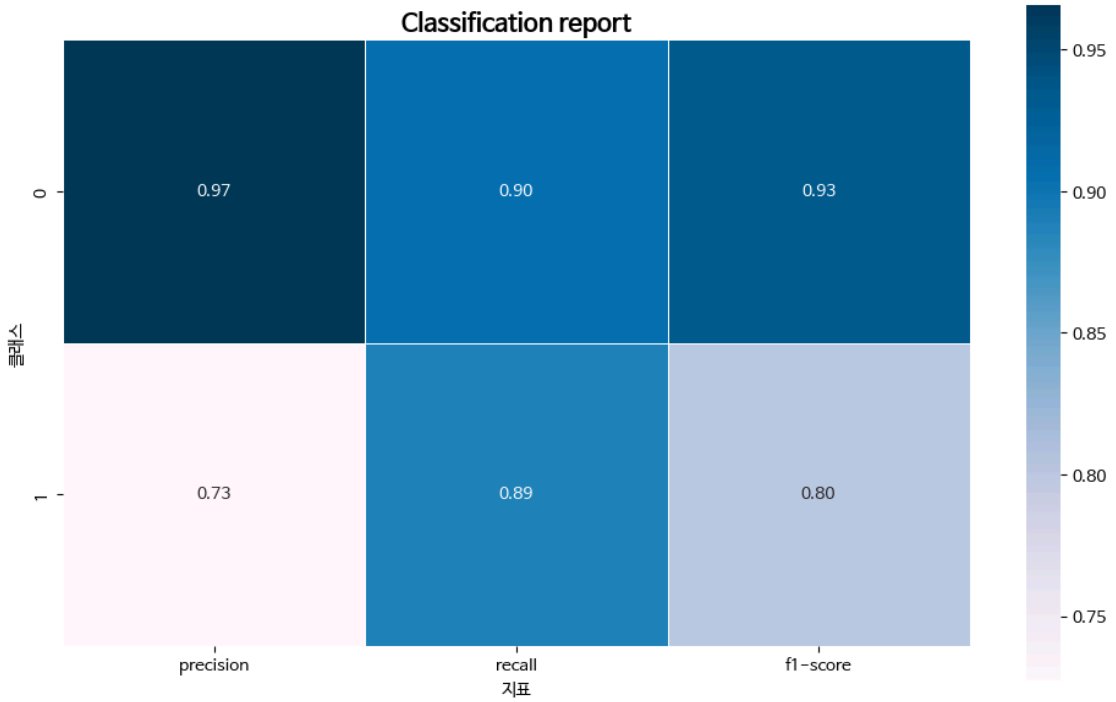
앞선 일련의 실험을 종합적으로 분석한 결과, 한국 단독 모델과 다양한 외생 변수 조합 간의 성능을 비교하였을 때, **[한국 + 아르헨티나(lag=32) + 국내 기상 변수(평균 기온, 일교차, 상대습도)]** 조합이 가장 높은 설명력과 안정적인 일반화 성능을 포함한 우수한 예측 성능을 보였다. 본 연구에서는 해당 변수 구성을 최종 LSTM 예측 모델의 입력으로 채택하였으며, 이후 유행 여부의 판단을 위한 이진 분류 파이프라인에도 동일하게 적용하였다.

이진 분류 판단은 정량 예측된 Influenza-B 값이 일정 기준을 초과하는지를 기준으로 이루어진다. 기준선은 고정값이 아닌 해당 시점 기준 과거 156주(약 3년)의 Influenza-B 평균값에 표준편차를 1배수 가산한 동적 임계값(Threshold)으로 설정하였다.

Threshold_t*(임계값) = 평균*(156*주*) + 1× σ*(표준편차*)

구체적으로는 위와 같이 정의되며, 예측값이 이 값을 초과하면 유행 발생(1), 이하지 않으면 비유행(0)으로 분류한다. 이 방식은 각 연도의 계절성 및 감염 강도를 동적으로 반영할 수 있어, 유행 정의 기준의 현실성과 민감도를 모두 확보할 수 있다는 장점이 있다.

이와 같은 기준 하에서 최종 예측 결과를 이진 분류 문제로 변환하고 평가한 결과, 전체 정확도(Accuracy)는 0.900, F1 점수는 0.800으로 나타났다. 특히, 실제 유행 발생(Class-1)을 탐지하는 데 중요한 Recall(재현율)은 0.889에 달하여, 본 모델이 유행 발생 시기를 놓치지 않고 사전에 탐지할 수 있는 가능성을 시사한다. 이는 감염병 발생의 조기 예측과 의약품 공급망 선제적 대응의 가능성을 의미한다.

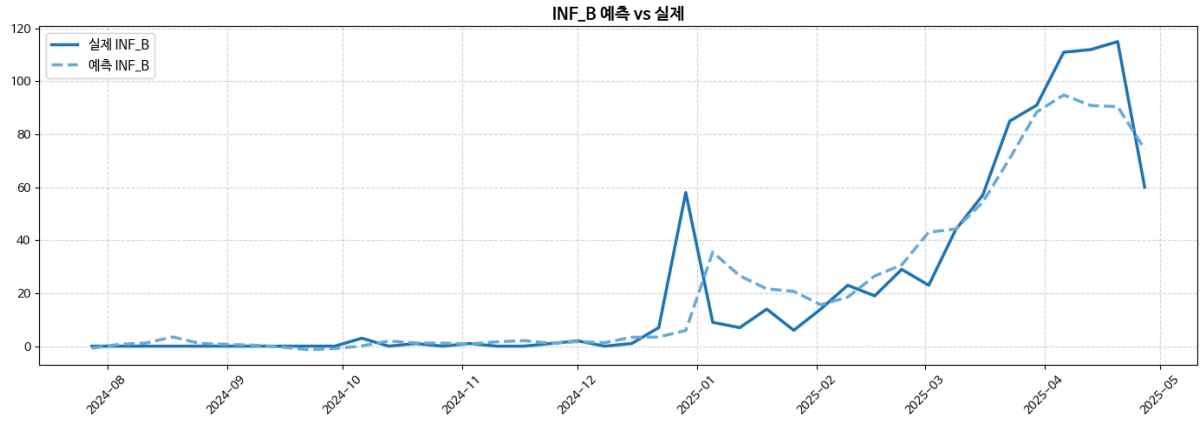


▲ 이진 분류 결과값 히트맵

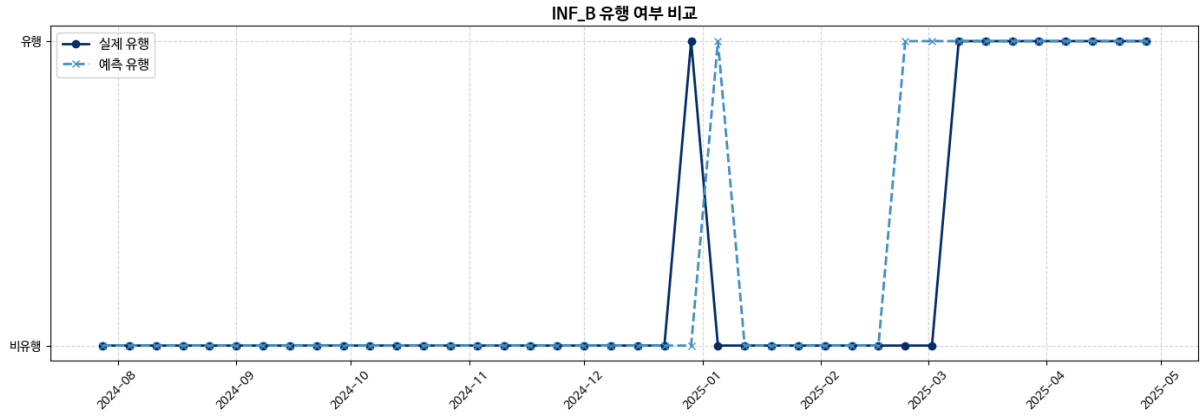
- 전체 정확도(Accuracy): 0.900
- 전체 F1 점수: 0.800

이러한 정량 지표 외에도, 시계열 예측값과 실제 Influenza-B 값의 시각적 비교 결과에서도 일관된 패턴 대응이 확인되었으며, 유행 여부 이진 분류 결과 또한 실제 라벨과 대부분 일치하는 양상을 보였다. 예측 RMSE는

12.5929, MAE는 7.0014, R^2는 0.8697로 정량 예측 성능 자체도 우수하며, 이를 기반으로 도출된 이진 판단 결과 역시 신뢰도 높은 경보 지표로 기능할 수 있음을 검증하였다.



▲ **Influenza-B** 예측 **vs** 실제 값



▲ 유행 여부**(Class-1)** 비교

5.5.8 모델 해석 및 결론

요약하면, 본 연구에서 개발한 LSTM 기반 인플루엔자 B형 주간 예측 모델은 감염병 유행의 비선형성과 계절성, 다중 외생 변수 간의 상호작용이라는 복합적 구조를 효과적으로 학습하여, 기존 전통 모델인 SARIMAX에 비해 현저히 높은 예측 성능을 달성하였다.

먼저, 한국 단독의 인플루엔자 B형 데이터만을 활용한 모델에서도 일정 수준의 예측력은 확보되었으나, 아르헨티나의 Influenza-B 시계열 데이터를 32주 선행된 형태로 외생 변수로 추가하였을 때 예측 정확도가 뚜렷하게 향상되었다.

아르헨티나는 한국과 계절이 반대인 남반구 국가로, 한국보다 약 32주 앞서 바이러스 유행이 시작되는 경향성이 존재한다. 본 연구에서는 이 계절 차이를 반영해 아르헨티나의 Influenza-B 시계열 데이터를 32주 선행된 형태로 모델에 입력하였으며, 그 결과 한국 내 유행 시기를 사전에 포착할 수 있는 유의미한 조기 경고 신호로의 역할을 할

수 있었다. 한국에서 Influenza-B가 본격적으로 유행하기 전에 아르헨티나에서 유사한 양상의 유행이 이미 나타나고 있었고, 이를 기반으로 예측에 활용하면 더 빠르고 정확한 대응 전략 수립이 가능하다는 것이다..

이처럼 실무적으로는 해외 선행 국가의 감시 데이터를 분석에 포함시키는 것이 유행 예측 정밀도를 높이는 데 핵심적인 역할을 할 수 있다. 본 연구에서는 아르헨티나 외에도 다수 국가의 시계열 데이터를 실험적으로 병합해보았으나, 구조적 유사성이 낮은 국가의 경우 오히려 예측 정확도가 저하되었다.

- 한국 **Influenza-B(**기준 시계열**)**: 예측 모델의 중심축이 되는 기준 시계열 데이터
- 아르헨티나 **Influenza-B (lag 32 weeks)**: 한국 유행의 조기 탐지 변수
- 기상 변수 **(**기온·습도·일교차**)**: 유행 강도와 지속성에 직접적 관련이 있는 학술적 변수

이 세 가지 변수 조합을 기반으로 학습된 최종 LSTM 모델은 R^2 0.8697, RMSE 약 12.6 수준의 우수한 예측 성능을 기록하였고, 예측된 ILI 수치를 기반으로 산출한 유행 여부 이진 분류에서도 F1-score 0.800, 재현율(Recall) 0.889로 높은 실효성을 보였다. 실제 백신 수요 예측, 생산 일정 조정, 재고 운영 등 의사결정 실무에서 활용 가능한 수준의 정량적 근거가 될 수 있는 당위성을 확보하였다.

그러나 동시에 본 연구는 외생 변수의 선택과 구조적 적합성에 대한 문제도 확인하였다. 다수의 국가를 무분별하게 추가하거나 구조적 유사성이 낮은 국가의 데이터를 병합한 경우 오히려 모델의 과적합 위험이 증가하고 예측 성능이 저하되는 현상이 관찰하였다. 이는 LSTM이 다차원 입력에 민감한 모델인 동시에, 데이터 품질과 구조적 유사성이 확보되지 않을 경우 학습 안정성이 크게 흔들릴 수 있다는 것이다.

따라서 향후 감염병 예측 시스템의 실무 적용 시, 외부 데이터를 무조건적으로 확장하기보다는 다음과 같은 조건을 충족하는 경우에 한해 변수 추가가 고려되어야 한다:

- 시계열 구조적 유사도 확보 **(cosine, DTW** 기반 사전 검증**)**
- 적절한 시차 보정 적용 **(**최적 **lag** 분석 후 **shift)**
- 실질적 예측 성능 개선이 확인된 경우에만 채택 **(RMSE, R^2** 기준**)**

결론적으로 본 프로젝트는 구조적으로 일관된 해외 감시 시계열과 기후 요인을 통합한 비선형 딥러닝 기반 감염병 예측 전략이 실용성과 예측력을 모두 갖춘 유의미한 접근 방식임을 실증하였다. 특히 제약 산업 및 보건의료 분야에서 초기 의사결정 및 자원 배분에 활용 가능한 정량적 예측 인프라를 마련할 수 있다는 점에서 의의를 갖는다.

---

6. 기대 효과 및 한계

6.1 기대효과

1. 유행 사전 예측을 통한 재고 손실 최소화
   주 단위 예측 결과를 기반으로, 제약기업은 유행 도래 수 주 전에 생산 및 유통 계획을 조정할 수 있다. 이는 반복적으로 발생해온 백신 및 의약품의 초과 재고 문제를 완화하며, 유통기한 만료에 따른 자산 손실을 줄이는 데 기여할 수 있다.

2. 공급망**(SCM)** 민첩성 향상 및 운영 리스크 사전 차단
   주간 예측 정보를 단기 재고 운영 및 월간 생산계획에 반영함으로써, 공급 지연, 생산 병목, 과잉 납품 등

공급망 리스크 요인을 사전 제어할 수 있다. 이는, 공급망 탄력성 강화라는 전략적 가치를 제공한다.

3. 경영 의사결정의 정량화 및 재무 안정성 확보
   감염병 유행 여부를 정량적 예측 변수로 확보함으로써, 마케팅·생산·물류 부문 간의 협업과 전략 조정이 가능해진다. 이는 기업의 자본 운용 효율성에 직결되며, 재무 리스크를 줄일 수 있다.

4. 글로벌 수준의 **SCM** 전략 도입 기반 마련
   본 모델은 GSK, Sanofi 등 글로벌 제약사들이 실제로 도입하고 있는 감염병 기반 SCM 전략과 유사한 구조를 보유하고 있다. 향후 모델을 지속 개선하고 예측 API 형태로 통합할 경우, 글로벌 수준의 수요 예측 및 공급 최적화 시스템을 내재화할 수 있는 기반을 갖추게 될 것이다.

6.2 한계 및 향후 보완점

1. 도메인 지식 기반 변수 설계의 한계
   본 연구에서 사용된 일부 외생 변수(예: 기상 변수)는 기존 문헌에서의 언급과 통계적 성능 향상에 기반해 선택되었으나, 변수 간 인과 구조나 생물학적 메커니즘에 대한 도메인 해석은 상대적으로 부족하였다. 예를 들어, “기온 15도 이하에서의 독감 증가” 등 정성적 인사이트를 모델링에 명시적으로 반영하지 못한 부분은 향후 보완될 필요가 있다.

2. 선행 연구 및 학술적 근거 기반의 제약
   한국과 해외 국가 간의 독감 시계열 동조 현상 및 선행 신호 활용에 대한 정량적 연구는 아직 축적된 레퍼런스가 많지 않아, 실증적 결과를 해석하는 데 한계가 존재했다. 이는 모델 선택이나 변수 조합에 있어 ‘경험적 최적화’에 의존하는 측면을 일부 남겼으며, 향후 관련 연구가 확대된다면 보다 이론적으로 정교한 설계가 가능할 것이다.

3. 학습 데이터의 제약 및 불균형
   COVID- 19 기간(2020~2022년)을 구조적 이상치로 판단해 이 구간을 제외함에 따라 학습에 활용 가능한 데이터가 제한되었고, 특히 강한 유행과 비유행 간의 데이터 비율 불균형 문제도 존재하였다. 이는 이진 분류 평가(Recall: 0.889, F1: 0.800 등)에서 영향을 줄 수 있으며, 향후 장기적 데이터 확보 및 보정 기법의 도입이 필요하다.

4. 모델 확장성의 한계 및 과적합 위험
   외생 변수의 수가 증가할수록 오히려 예측 성능이 저하되는 경우가 발생하였다. 이는 LSTM 구조가 데이터 양에 비해 고차원 입력에 민감하며, 적절한 변수 선택 및 차원 축소 기법(PCA 등)의 부재 시 과적합 가능성이 커진다는 점을 시사한다.

참고문헌

1. ***Chen, X., Tao, F., Chen, Y., Cheng, J., Zhou, Y., & Wang, X. (2025). Forecasting influenza epidemics in China using transmission dynamic model with absolute humidity. Infectious Disease Modelling, 10, 50–59.***
2. ***Caini, S., & Kusznierz, G. (2019). The epidemiological signature of influenza B virus and its B/Victoria and B/Yamagata lineages in the 21st century. PLOS ONE, 14(9), e0222381.***
3. ***Ashraf, M. A., & Raza, M. A. (2024). A comprehensive review of influenza B virus, its biological and clinical aspects. Frontiers in Microbiology, 15, 1467029.***

4. *Moon, J., Jung, S., Park, S., & Hwang, E. (2021). Machine learning-based two-stage data selection scheme for long-term influenza forecasting. Computers, Materials & Continua, 68(3), 2945–2959.*
5. *Kandula, S., & Shaman, J. (2019). Near-term forecasts of influenza-like illness. Epidemics, 27, 41–51.*
6. *Lee, K. C.-Y., Lin, L. C. Y., Leung, C. T., Yau, D. S.-W., Chan, J. Y. N., Ip, D. K. M., & Lau, E. H. Y. (2024). An adaptive weight ensemble approach to forecast influenza activity in an irregular seasonality context. Nature Communications, 15, Article No. 4040.*
7. *O'Donnell MJ, Fang J, Mittleman MA, Kapral MK, Wellenius GA; Investigators of the Registry of Canadian Stroke Network. Fine particulate air pollution (PM2.5) and the risk of acute ischemic stroke. Epidemiology. 2011 May;22(3):422-31. doi: 10.1097/EDE.0b013e3182126580. PMID: 21399501; PMCID: PMC3102528.*
8. *Choi SB, Ahn I (2020) Forecasting seasonal influenza-like illness in South Korea after 2 and 30 weeks using Google Trends and influenza data from Argentina. PLoS ONE 15(7): e0233855. https://doi.org/10.1371/journal.pone.0233855*
9. *Sungwoo, P. (2020). SHAP-based explainable influenza occurrence forecasting using lightGBM*
10. *Soo Beom Choi. (2020). Forecasting seasonal influenza-like illness in South Korea after 2 and 30 weeks using Google Trends and influenza data from Argentina*

ILI 예측모델 개발 프로젝트**:** 인플루엔자 **B**의 예측을 통한 의약품 재고 최적화

1. 인플루엔자 **B**형 유행 예측의 필요성

전 세계적으로 계절성 인플루엔자는 매년 10억 건 이상의 발병 사례가 보고된다. 국내 제약사들은 인플루엔자 유행 시기 예측 불일치로 인한 과잉 재고 평가손실과 재고자산 가치 하락으로 수십억~수백억 원 규모의 손실을 경험한 바 있다. 따라서 A형과 구분되는 B형 인플루엔자의 독립적 예측은 백신 수급 및 공급망 안정성 확보의 핵심 과제로 대두된다.

2. 분석 목표 및 예측모델 구축 전략

본 연구의 목표는 B형 인플루엔자의 유행을 사전에 탐지할 수 있는 예측모델을 수립하는 것이다. 이를 위해 데이터 수집 단계에서 인플루엔자 B형 데이터와 연령별·시기별 발생 특성을 반영할 수 있는 대체 데이터를 확보하는 전략을 채택하였다.

단계:

1. 데이터 수집·시각화
2. 예측 모델링 (SARIMAX, LSTM)
3. 성능 평가 회귀: R^2, MSE, MAE / 분류: Accuracy, F1, Recall
4. 결과 해석: 상관관계·주요 인자 도출·정책 적용 가능성 평가

3. 예측 모델 설계 및 변수 구성 전략

- 목표: 주 단위 INF_B 환자수 예측 → 유행 시기 사전 파악, 재고·생산 계획 반영.

3.1 유행 기준

- 유행 기준 (주차 t) : 최근 3년간 주별 INF_B 평균 + (1 × 표준편차)
  → 예측된 INF_B가 해당 기준을 초과할 경우 “유행”으로 간주

3.2 분석 단위 및 기간

- 분석 단위: 주 단위(Weekly)
- 학습 구간: 2014년 1월 ~ 2024년 4월 (코로나 기간 제외)
- 평가 구간: 2024년 5월 ~ 2025년 4월

해당 구간 설정은 코로나 팬데믹이라는 비정상적 변동을 배제하고, 최근 10년간의 안정적 패턴을 반영하기 위함이다.

3.3 예측 모델

본 연구에서는 두 가지 접근을 비교·평가한다.

1. SARIMAX: 계절성과 외생 변수 처리가 가능한 전통 통계 기반 모델

2. LSTM(Long Short-Term Memory): 시계열의 장기 의존성과 비선형성을 처리할 수 있는 딥러닝 기반 모델

두 모델의 성능을 비교한 뒤, 재현성과 설명력이 높은 모델을 고정해 최종 활용할 계획이다.

3.4 변수 구성

3.4.1 최종 사용 변수

- 국내 ILI (INF_B 주간 환자수): 기본 종속 변수.
- 타국 ILI: 선행 시그널 역할 가능.
- 평균 기온(서울 기준): 계절성 반영.
- 일교차: 선행 연구에서 B형 증가와 연관성 확인.
- 상대습도(주간 평균): 바이러스 전파 억제 임계치 검증 목적.

3.4.2 제외 변수 및 사유

- 강수량: 상관성 낮음, 설명력 부족.
- 연령별 독감 데이터: 인플루엔자 A/B형 미구분으로 활용 불가.
- 백신 접종률: 연 단위 자료로 주 단위 분석과 시간 불일치.
- 검색량 지표: 주 단위 데이터 누락, 불연속성 문제.

4.모델링

4.1 사용 데이터

본 연구에서는 WHO FluNet 자료를 기반으로 2014년 1월부터 2025년 5월까지의 국가별 주간 B형 인플루엔자(INF_B) 발생률을 수집하였다. 국내 기상 자료(평균 기온, 일교차, 상대습도)를 결합하였으며, 코로나19의 비정상적 패턴이 반영된 데이터는 제외하였다. 각 변수는 lag 52주 형태로 변환하여 시계열 예측에 활용하였다.

4.2 데이터 전처리

연·주차(YEAR_WEEK) 변수를 생성하고, 결측값은 0으로 처리하였다. Padding을 적용하여 국가별 row수를 동일하게 확보하고, MinMax Scaling으로 변수를 정규화 하였다.

4.3 국가 간 **ILI** 유사도 분석

한국과 타국 간의 시계열 유사도를 평가하기 위해 코사인 유사도와 DTW 거리를 산출하고 이를 종합 점수(final score)로 환산하였다. 분석 결과, 일본(0.7367), 캐나다(0.6218) 등에서 높은 유사도가 확인되었다.

4.4 SARIMAX 예측모델

여러가지 시계열 검정 결과(ADF, ACF, PACF, STL), 한국의 INF_B 시계열은 정상성을 충족하여 SARIMAX에 적용할 수 있었다. 그러나 팬데믹 기간의 구조적 변동성 및 외부 충격의 영향으로 인해, 고정된 계절성과 선형 추세를 가정하는 SARIMA의 적용에는 일정한 한계가 존재할 수 있다. 실제로 한국 ILI 단독 모델과 한국과 타국

ILI까지 반영한 모델을 실험해 보았음에도 불구하고, 예측 성능을 나타내는 결정계수(R^2)가 음수가 나오는 등, 만족할 만한 성과를 낼 수 없었다.

따라서, SARIMAX 모델의 한계로 인해 예측 정확도와 설명력을 확보하기 어렵다는 결론이 나왔다. 이에 따라, 좀 더 다양한 요인을 반영할 수 있는 LSTM기반 딥 러닝 모델로 분석을 전환하게 되었다.

4.5 LSTM 모델

LSTM의 경우 비선형, 복합 시계열 데이터의 장기 의존성 학습 및 외생 변수 반영에 유리하다.

최초 모델 구축에서는 한국 Influenza-B 시계열 데이터를 기반으로 예측하였고 이는 기존 SARIMAX 기반 모델과 비교해 향상된 성능을 확인하였다.

이후, 남반구의 계절 반대성에 의한 Lead Time에 주목해 아르헨티나의 32주 lag 데이터를 추가해 모델을 구축하였고, 이는 추가적인 성능향상으로 이뤄졌으며, 아르헨티나 데이터가 예측에 있어 효과적인 선행지표라는 시사점을 얻었다.

여기에 더하여 국내 기상변수를 추가하였는데, 이는 Influenza-B 유행에 직접적 영향을 주는 변수로 학술자료에 근거한 것이다. 여기서 최고 성능을 확인하였으며 결론적으로 이 모델을 채택하였다.

남반구 다국가 기반의 LSTM 모델 구축도 추가로 시도했지만, 품질 이질성 및 Overfitting에 의해 예측 성능이 하락하며 채택을 기각하였다.

평가를 위해 “유행”과 “비유행”을 구분하는 라벨링 작업을 진행하였으며, 이는 직전 3년 간의 평균, 표준편차를 활용한 동적 임계값을 기반으로 하여 진행되었다.

최종적으로, 채택한 모델에 입각해 이진 분류를 진행한 결과, 정확도 0.900, 재현율 0.899(F1-Score: 0.800)을 기록하며 높은 신뢰도를 기록하여, Influenza-B 예측 모델의 구축에 성공하였다.

5. 결론 및 제언

제약사의 SCM 및 생산계획 구축을 위한, ‘한국 Influenza-B 유행 가능성 예측 모델’에 필요한 변수는 해당 연구에서 아르헨티나의 Influenza-B(lag 32 weeks) 선행 시계열 데이터, 국내 기상 변수로 판단이 되었다. 이를 통해 제약사는 재고 효율화와 선제적 대응을 할 수 있을 것으로 전망된다. 다만, 해당 연구 결과를 토대로 재고 예측에 아르헨티나 단일 국가를 사용하는 것은 일반화의 오류가 될 수 있다.

추가적으로, 인과관계에 대한 도메인 지식 부족과 COVID-19 기간으로 인해 데이터 확보에 제약이 존재하였으므로, 향후 데이터 수집 단계에서의 보완이 요구된다. 또한 무분별한 외생변수 추가 시 모델 과적합으로 인한 성능 저하의 가능성이이 존재한다는 것도 시사점이다. 따라서, 이후 추가 연구 진행 및 실제 업무용 모델 구축 시에는 시계열의 구조적 유사도를 사전 검증하고, 적절한 시차 보정을 통한 정제된 변수의 사용이 요구된다.