JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 30 No. 9, pp. 99-109, September 2025
https://doi.org/10.9708/jksci.2025.30.09.099

Cybercrime Incident and Arrest Prediction Model Through Time Series Analysis: Prediction of Trends and Patterns

Ji-Hyeok Choi*, Kyu-Cheol Cho**

*Student, Dept. of Computer Science, Inha Technical College, Incheon, Korea
**Professor, Dept. of Computer Science, Inha Technical College, Incheon, Korea

[Abstract]

With the recent surge in cybercrime, effective response and prediction have become increasingly important. This study proposes a predictive model for predicting cybercrime incidents and arrest counts through time series analysis. In this research, we utilized Seasonal Autoregressive Integrated Moving Average (SARIMA) and Autoregressive Integrated Moving Average with Exogenous Variables (ARIMAX) models to analyze the trends and patterns of cybercrime occurrences, applying hyperparameter tuning to identify the optimal predictive variables. To evaluate the performance of each model, we used Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and Akaike Information Criterion (AIC) metrics to compare model accuracy and derive the optimal model. The results demonstrate that the proposed model can effectively predict cybercrime incidents and arrests, providing a valuable tool for anticipating future cybercrime risks and informing preventive strategy development.

▸ **Key words**: Cybercrime Prediction, Time Series Analysis, SARIMA, ARIMAX, Hyperparameter Tuning

[요 약]

최근 사이버 범죄는 급증하고 있으며 이에 대한 효과적인 대응과 예측이 중요해지고 있다. 본 연구는 시계열 분석을 통해 사이버 범죄 사건과 검거 건수를 예측하는 모델을 제안한다. 본 연구에서는 Seasonal Autoregressive Integrated Moving Average (SARIMA)와 Autoregressive Integrated Moving Average with Exogenous Variables (ARIMAX) 모델을 활용하여 사이버 범죄 발생 추세와 패턴을 분석하였으며, 최적의 예측 변수를 찾기 위해 하이퍼 파라미터 기법을 사용하였다. 각 모델의 성능을 평가하기 위해 Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), Akaike Information Criterion (AIC) 지표를 적용하여 모델의 정확성을 비교하고 최적의 모델을 도출하였다. 연구 결과, 사이버 범죄 사건과 검거 건수를 효율적으로 예측할 수 있는 모델을 제시하였으며, 이 모델은 미래의 사이버 범죄 발생 가능성을 예측하고, 예방 전략을 수립하는 데 유용하게 활용될 수 있을 것으로 기대한다.

▸ **주제어**: 사이버 범죄 예측, 시계열 분석, SARIMA, ARIMAX, 하이퍼 파라미터 튜닝

• First Author: Ji-Hyeok Choi, Corresponding Author: Kyu-Cheol Cho
*Ji-Hyeok Choi (chlwlgur0914@gmail.com), Dept. of Computer Science, Inha Technical College
**Kyu-Cheol Cho (kccho@inhatc.ac.kr), Dept. of Computer Science, Inha Technical College
• Received: 2025. 04. 29, Revised: 2025. 09. 16, Accepted: 2025. 09. 16.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

오늘날 디지털화가 가속화됨에 따라 사이버 공간에서 발생하는 범죄가 급증하고 있다. 국내 2023년 기준으로 사이버 범죄 발생 건수는 2022년에 비해 약 4.99% 증가하여 11,487건이 추가되었으며, 5년 전인 2018년과 비교하면 무려 80,951건, 즉 54.06%가 증가하였다[1]. 이러한 사이버 범죄는 개별 개인뿐 아니라 기업, 정부, 사회 전반에 걸쳐 심각한 위협을 초래하고 있으며, 그 심각성은 단순한 금전적 피해를 넘어 피해자들에게 정신적 질환[2]을 유발할 가능성까지 있어 이러한 심각성은 더욱 대두되고 있다.

더욱이, 사이버 범죄의 발생 건수가 급격히 증가하고 있음에도 불구하고 검거율이 지속해서 하락하고 있다는 점이다. 2023년 기준으로 사이버 범죄 검거율은 2022년보다 5.4% 낮아졌으며, 5년 전인 2018년과 비교하면 17.9%나 감소하였다[1]. 사이버 범죄는 급변하는 IT 트렌드와 기술 발전을 배경으로 더욱 정교해지고 있으며, 새로운 유형과 수법이 끊임없이 등장하고 있다. 이러한 상황에서 기존의 전통적 대응 방식으로는 증가하는 사이버 범죄에 효과적으로 대응하기 어렵다[3][4]. 사이버 범죄의 예방과 대응을 위해서는 기존의 수사 및 보안 체계를 강화할 뿐만 아니라, 예측 가능한 분석 모델과 AI 기술을 활용하여 범죄 발생 가능성을 사전에 식별하고 대응할 수 있는 능동적이고 혁신적인 접근이 필요하다.

본 연구에서는 사이버 범죄의 발생 건수와 검거 건수를 각각 예측하는 시계열 분석 모델을 구축하였다. 사이버 범죄 발생 건수 예측은 공격자의 행위 패턴과 범죄 동향을 이해하는 데 의미가 있으며, 검거 건수 예측은 수사기관의 대응 역량과 제도적 효율성을 이해하는데 의미가 있다. 본 연구는 2014년부터 2023년까지의 데이터를 바탕으로 사이버 범죄 발생 및 검거 추세와 패턴을 분석하고, 이를 토대로 미래의 범죄 발생 가능성을 예측하는 것을 목표로 한다.

사이버 범죄 데이터는 시간의 흐름에 따라 연속적으로 발생하며, 시계열 데이터 성격을 지니므로, Seasonal Autoregressive Integrated Moving Average (SARIMA)[5] 와 Autoregressive Integrated Moving Average with Exogenous Variables (ARIMAX)[6] 적용하였으며, 최적의 예측 정확도를 얻기 위해 하이퍼 파라미터 튜닝기법을 통해 모델을 최적화하였다.

특히, 각 모델의 예측 성능을 검증하기 위해 Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), Akaike Information Criterion (AIC) 등의 지표를 활용하여 성능을 비교하고 분석하였다.

본 논문의 구성은 다음과 같다. 2장에서는 본 연구와 관련이 있는 연구에 관하여 기술한다. 3장에서는 본 연구에 필요한 데이터 전처리 과정을 기술한다. 4장에서는 모델의 학습 과정과 학습시킨 모델의 결과를 비교하여 분석한 결과를 기술한다. 5장에서는 본 연구에 대한 결론을 기술한다.

II. Related Works

1. Time Series Analysis

수집한 사이버 범죄 데이터가 시간의 흐름에 따라 기록된 데이터이므로 이러한 특성을 반영하기 위해 시계열 분석을 채택하였다. 시계열(Time Series)이란 시간의 흐름에 따라 순서대로 수집된 데이터를 의미한다. 시계열 분석은 시간에 따라 변화하는 데이터를 분석하고, 그 안에서 발견되는 유의미한 정보를 바탕으로 미래값을 예측하는 통계적 기법이다[7]. 시계열 데이터는 연속적인 시간 순서로 구성되며, 경제, 금융, 기상, 사회 현상 등 시간의 종속성을 보이는 다양한 분야에서 폭넓게 활용된다. 특히, 일반적인 회귀 분석과 달리 시계열 분석은 데이터가 시간에 따라 상호 의존적이라는 가정을 포함하며, 데이터의 시간적 구조와 주기성을 고려하여 과거 데이터로부터 미래 값을 예측할 수 있다는 장점이 있다[8].

본 연구에서는 사이버 범죄 데이터의 월별 데이터에서 시간적 흐름과 계절적 패턴이 발견됨에 따라 시계열 분석 모델인 SARIMA 모델을 적용하였고, 또한 여러 세부 유형으로 분류된 데이터를 다루기 위해 시계열 분석 모델에 외생 변수를 추가할 수 있는 ARIMAX 모델을 추가로 채택하였다.

2. Seasonal Autoregressive Integrated Moving Average (SARIMA)

SARIMA(Seasonal Autoregressive Integrated Moving Average) 모델은 계절성을 반영한 시계열 데이터를 분석하고 예측하는 데 사용되는 모델이다. 일반 ARIMA 모델과 달리, 계절적 요소를 포함하여 데이터의 주기적인 변동 패턴을 효과적으로 모델링할 수 있는 특징이 있다[9].

SARIMA 모델은 비계절성과 계절성 구성 요소를 포함하여 모델링이 이루어진다. 비계절 부분은 일반 ARIMA 모델의 일반 자기회귀(p), 일반 차분(d), 일반 이동 평균(q)으로 구성되며, 계절성 부분은 계절 자기회귀(P), 계절 차분(D), 계절 이동 평균(Q), 그리고 계절 주기(s)로 정의된

다. 이를 통해 데이터의 비계절적 패턴과 계절적 패턴을 동시에 모델링하여 예측의 정밀도를 높인다[10].

SARIMA 모델의 구성요소와 수식은 아래와 같이 정의된다(Fig.1).

- **SARIMA** $(p,d,q) \times (P,D,Q,s)$:
  - **비계절적 부분** (p,d,q)
    - **p**: 비계절 자기회귀(AR) 부분의 차수
    - **d**: 비계절 차분 차수
    - **q**: 비계절 이동 평균(MA) 부분의 차수
  - **계절적 부분** (P,D,Q,s)
    - **P**: 계절적 자기회귀(AR) 부분의 차수
    - **D**: 계절적 차분 차수
    - **Q**: 계절적 이동 평균(MA) 부분의 차수
    - **s**: 계절 주기

Fig. 1. SARIMA Composition

또한, SARIMA 모델은 차분(Differencing)을 통해 데이터의 정상성을 확보하여 시간의 흐름에 따른 데이터의 변화 경향을 보다 안정적으로 예측할 수 있다. 일반 차분(d)과 계절 차분(D)을 적용하여 데이터에서 추세와 계절성을 제거함으로써 모델이 시계열 데이터를 정상화하여 더 안정적인 예측을 가능하게 한다[5].

3. Autoregressive Integrated Moving Average with Exogenous Variables(ARIMAX)

ARIMAX (Autoregressive Integrated Moving Average with Exogenous Variables) 모델은 외부 변수(외생 변수)를 포함하여 시계열 데이터를 분석하고 예측하는 모델로, 시계열 데이터와 외부 요인 간의 관계를 동시에 반영할 수 있는 특징이 있다[6].

ARIMAX 모델은 ARIMA 모델의 확장 형태로, 외부 요인이 시계열 데이터에 미치는 영향을 함께 고려함으로써 예측의 정밀도를 높인다. 이러한 특성 덕분에 ARIMAX 모델은 특정 요인(예: 경제 지표, 사회적 변수 등)이 데이터에 중요한 영향을 미치는 경우 효과적으로 적용될 수 있다.

ARIMAX 모델은 비계절적인 자기회귀(AR), 차분(I), 이동 평균(MA) 요소를 포함하여 시계열 데이터의 비계절적 패턴을 분석하는 기본적인 ARIMA 구조를 갖는다. 여기에 외생 변수(X)가 추가되어, 시계열 데이터 외부의 다양한 요인들이 데이터에 어떻게 영향을 미치는지 반영할 수 있다[6].

ARIMAX 모델의 구성 요소와 수식은 아래와 같이 정의된다(Fig.2).

- **ARIMAX** $(p,d,q) \times (X)$:
  - **비계절적 부분** (p,d,q)
    - **p**: 비계절 자기회귀(AR) 부분의 차수
    - **d**: 비계절 차분 차수
    - **q**: 비계절 이동 평균(MA) 부분의 차수
  - **외생 변수** (X)
    - **X**: 시계열 데이터에 영향을 미치는 외부 요인

Fig. 2. ARIMAX Composition

ARIMAX 모델은 데이터의 시계열적 패턴과 함께 외부 요인이 데이터에 미치는 영향을 포함하여 더 정교한 예측을 가능하게 한다. 예를 들어, 사이버 범죄 데이터에 외생 변수로 경제 지표나 사회적 변동성을 추가함으로써, 단순 시계열 분석보다 더 정확한 예측이 가능하다. 이를 통해 ARIMAX 모델은 외부 요인을 고려한 예측이 필요한 다양한 분야에서 유용하게 활용될 수 있다[6].

III. Data Preparation

본 연구는 급증하는 사이버 범죄 데이터를 분석하여 효과적인 예측 모델을 제안하는 것을 목표로 한다. 이를 위해 시계열 분석 기법인 SARIMA와 외생 변수를 포함한 ARIMAX 모델을 활용하여 범죄 발생과 검거 건수의 데이터를 분석하고 예측하고자 한다.

연구 과정에서는 데이터의 수집과 전처리, 계절성 및 외부 요인의 반영 등 다양한 분석 단계를 거쳐, 예측의 정확도를 높이는 데 중점을 두었다. 데이터 가공 및 전처리 절차를 아래에 소개한다.

1. Data Collection

본 연구에서는 사이버 범죄 발생 건수와 검거 건수를 예측하기 위해, 공공데이터 포털, 경찰청, 보안뉴스에서 주 데이터를 수집하였다[11][12][13].

먼저, 공공데이터 포털에서 제공하는 경찰청_월별 사이버범죄 발생건수 및 검거건수 현황 데이터를 활용하였다. 해당 데이터는 2014년부터 2021년까지의 연도별, 월별 발생건수와 검거건수가 정리된 자료로 Excel 파일 형식으로 제공되었다. 그 후, 경찰청에서 제공한 사이버 범죄 세부

유형 발생건수 및 검거건수에 관한 데이터를 수집하였다. 해당 데이터는 2014년부터 2022년까지 연도별로 구성되어 있으며, 각 범죄 유형별 정보를 포함하고 있다.

추가적으로, 최신 데이터를 보완하기 위해 보안뉴스에서 공개된 정태호 의원(더불어민주당, 서울 관악구을)이 경찰청으로부터 받은 자료[13]를 활용하였다. 해당 자료는 기사의 표 형식으로 제공되었으며, 연구 목적에 맞게 Excel 파일로 변환하여 다른 데이터와 통합하였다. 수집된 데이터는 발생건수와 검거건수를 중심으로 하며, 연구의 신뢰성을 확보하기 위해 출처별 데이터의 일관성을 검토한 후 분석에 활용하였다.

2. Data Analysis and Features

월별 사이버 범죄 발생 건수와 검거 건수를 데이터 분해(Decomposition)기법을 통해 발생 건수와 검거 건수를 오리지널 데이터(Original), 추세(Trend), 주기적 성분(Seasonal), 잔차(Residual)로 각각 분리하여 분석을 진행하였다. 분석 결과, 발생건수 그래프(Fig.3), 검거 건수 그래프 (Fig.4) 두 그래프 모두 연간 주기와 시간이 지남에 따라 점진적인 증가 추세를 보이고 있었다. 이는 데이터의 시간적 구조를 반영한 분석이 적합함을 보여준다.

Fig. 3. Decomposition of Incident Graph(Monthly)

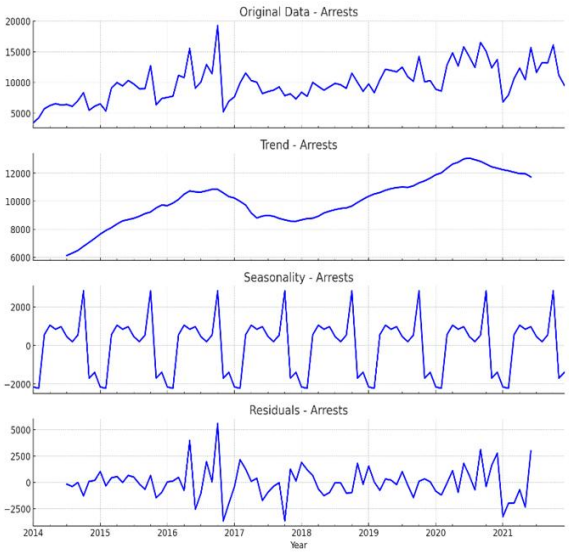

Fig. 4. Decomposition of Arrest Graph(Monthly)

특히 주목해 볼 점은 Fig.3, Fig.4의 Seasonality 그래프에서 1년 단위로 반복되는 주기적 계절성 패턴이 발견되었다. 이는 특정 시기에 범죄가 집중되는 경향을 설명하며, 이를 모델링 과정에 포함시키는 것이 필요하다고 생각되어 이러한 결과를 바탕으로 월별 데이터를 활용하여 SARIMA(Seasonal Autoregressive Integrated Moving Average) 모델을 적용하는 것이 모델의 정확도를 높일 수 있다고 판단한다. SARIMA 모델은 데이터의 계절성 요소뿐만 아니라 장기적인 추세를 효과적으로 반영할 수 있는 분석 기법으로, 발생 건수와 검거 건수의 변동성을 예측하기에 적합하다.

추가적으로 월별 사이버 범죄 발생 건수와 검거 건수의 분해 결과를 병합하여 하나의 그래프로 나타내 본 결과 (Fig.5) 발생 건수는 꾸준히 증가하고 있는 반면, 검거 건수는 발생 증가 속도를 따라가지 못하며 낮은 증가율을 보였다. 이는 범죄 발생과 검거 간의 불균형을 나타내며, 검거율 향상을 위한 심층적 분석의 필요성을 시사한다.

한편, 세부 유형별 사이버 범죄 데이터를 분석한 결과, 해킹, 사이버 사기 등 특정 유형의 범죄가 발생 건수와 검거 건수에 상이한 영향을 미친다는 점이 확인되었다. 관련된 세부 내용은 데이터 전처리 절에 기술하였다. 이에 따라, 외생 변수(Exogenous Variables)를 포함하여 발생 건수와 검거 건수 간의 관계를 정량적으로 분석하기 위해서 ARIMAX(Autoregressive Integrated Moving Average with Exogenous Variables) 모델을 추가적으로 적용하였다. ARIMAX 모델은 세부 유형 데이터를 활용하여 외생 변수의 효과를 반영하고, 발생 건수와 검거 건수 간의 상관성을 심층적으로 분석할 수 있는 강점을 가진다.

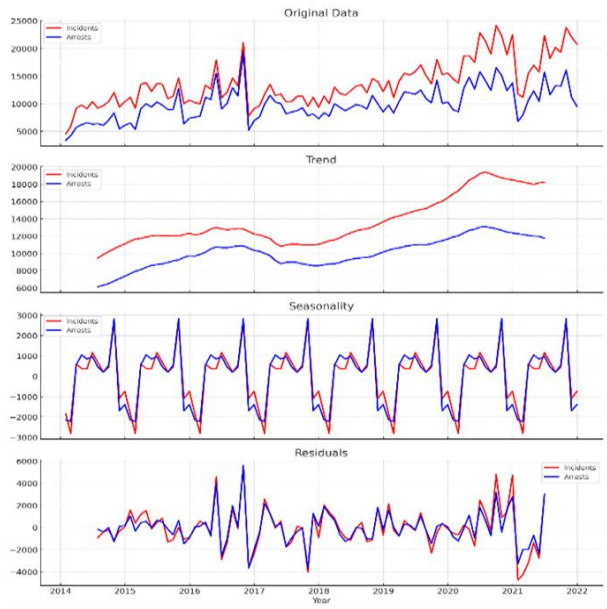

Fig. 5. Combined Analysis of Incident and Arrest

데이터의 특성과 분석 결과로서 본 연구에선 SARIMA 모델을 통해 월별 데이터의 변동성을 분석하고, ARIMAX 모델을 통해 세부 유형 데이터의 외생 변수 효과를 분석함으로써 사이버 범죄 데이터를 다각적으로 접근하기로 하였다. 두 모델의 조합은 데이터의 추세와 주기, 외생 변수 효과를 모두 반영하여 정확하고 다양한 예측을 가능하게 할 것으로 기대된다.

3. Data Preprocessing

데이터 전처리(Data Preprocessing)는 본 연구의 핵심 모델인 SARIMA와 ARIMAX에 적합한 데이터셋을 구성하기 위해 필수적인 과정이다. 전처리 과정은 수집된 데이터를 처리, 분류, 통합, 추가, 보완하여(Fig.6) 분석의 신뢰성과 정확성을 확보하는 데 중점을 두었다.

Fig. 6. Data Preprocessing Procedure

우선, 수집된 데이터를 필요하지 않은 데이터는 삭제하고, 모델 학습에 적합한 구조로 열과 행을 재구성하는 전처리 과정을 진행했다. 이후 사이버 범죄 월별 데이터와 세부 유형별 데이터로 나누어 분류하였다. 월별 데이터는 SARIMA 모델에 적용하기 위해 발생 건수와 검거 건수로 각각 분리하였으며, 이는 데이터의 계절성과 추세를 효과적으로 분석하고 예측하는 데 적합한 형태로 가공되었다.

세부 유형별 데이터는 ARIMAX 모델에서 외생 변수로 활용하기 위해 총 13개 유형(해킹, DDoS, 악성 프로그램, 사이버 사기 등)의 324개의 데이터를 발생 건수와 검거 건수를 독립적으로 나누어 정리하였다. 총합 데이터(종속 변수)와 각 외생 변수 간의 상관관계를 분석한 결과, 사이버 사기(0.9578), 사이버 명예훼손 모욕(0.8693), 해킹(0.7843), 사이버 금융범죄(0.7320)와 같이 일부 변수는 높은 상관관계를 보였으나, 나머지 변수들은 상대적으로 낮은 상관관계를 나타냈다.

이에 따라, 외생 변수를 보다 효과적으로 활용하기 위해 정보통신망 침해범죄, 정보통신망 이용범죄, 불법콘텐츠 범죄의 세 가지 대분류로 통합하였다. 이러한 통합 과정을 통해 외생 변수의 상관관계는 각각 A.0.9976, B.0.9374, C.0.9166으로 증가하였으며, 데이터의 활용도를 극대화할 수 있었다. 이로써 ARIMAX 모델에 적합한 외생 변수를 구성하여, 예측 모델의 정밀성을 높일 수 있는 기반을 마련하였다.

추가적으로, 데이터의 최신성을 확보하기 위해 2024년 10월 2일 보안뉴스[13]에서 발표된 자료를 통해 60건의 데이터를 기존 데이터셋에 통합하였다. 세부 유형 데이터는 연간 데이터 형식 그대로 통합하였으나, SARIMA에 적용할 월별 데이터로 활용하기 위해서는 연간 데이터 형식으로 제공된 2024년 8월까지의 데이터를 월별 데이터로 변환할 필요가 있다.

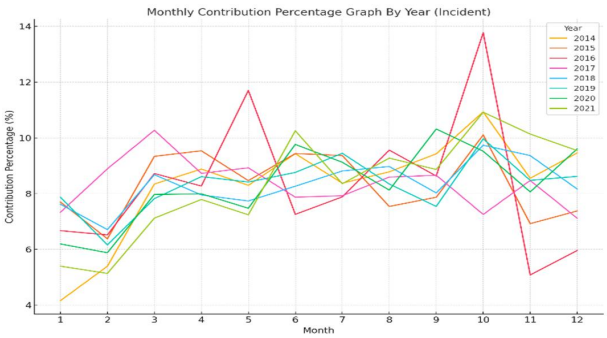

Fig. 7. Monthly Contribution Percentage Graph (Incident)

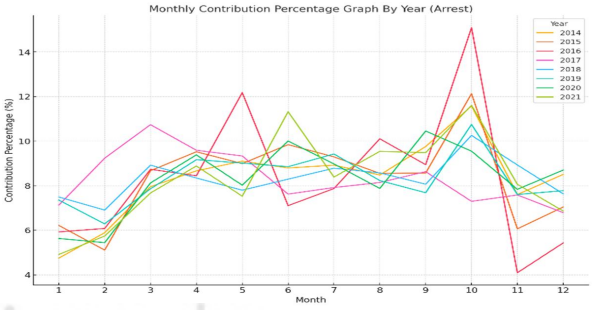

Fig. 8. Monthly Contribution Percentage Graph (Arrest)

위 그림은 연도별 월별 기여도를 나타낸 것으로, 약간의 편차는 존재하지만 대체로 일정한 패턴을 보여준다(Fig. 7, Fig.8). 이러한 관찰을 바탕으로, 본 연구에서는 월별 데이터를 보완하기 위해 연간 데이터를 월별로 분배하는 방식을 적용하였다. 월별 평균 기여도는 각 월이 연간 데이터에서 차지하는 비율을 기반으로 산출되었으며, 이는 연도 간의 패턴이 대체로 유사하다는 점에서 신뢰성을 확보할 수 있었다. 산출된 월별 평균 기여도를 활용하여 연간 데이터를 월별 데이터로 변환한 결과, 새로운 데이터는 기존 데이터와 비교했을 때 일관성이 유지되었으며, 데이터의 편차는 허용 범위 내에 안정적으로 분포하였다.

본 방식의 구조는 다음과 같다. 새로운 연간 데이터 총합(YearlyTotal)을 각 월별로 분배하여 새로운 연도의 월별 데이터를 생성하기 위해, 먼저 기존 연도들의 월별 데이터를 활용하여 각 월이 연간 데이터에 기여하는 비율을 의미하는 월별 평균 기여도(MonthlyContribution)를 산출하였다. 이후, 산출된 월별 평균 기여도를 추가하고자 하는 새로운 연도에 적용함으로써, 해당 연도의 월별 예상 건수를 계산하고 월별 데이터를 완성하였다. 월별 평균 기여도는 연간 총 사건 건수 대비 각 월의 사건 건수가 차지하는 평균 비율을 나타내며, (식.1)(식.2)에 제시된 수식을 통해 자세히 설명하였다.

$$MonthlyContribution_i = \frac{\sum_{y=1}^{n} MonthlyCount_{iy}}{\sum_{y=1}^{n} YearlyTotal_y} * 100 \quad (1)$$

- $MonthlyContribution_i$: 월 평균 기여도(백분율)
- $\sum_{y=1}^{n} MonthlyCount_{iy}$: 연도(y=1부터 n)동안 i월의 사건 건수 합
- $\sum_{y=1}^{n} YearlyTotal_y$: 여러 연도 동안의 연간 총 사건
- n : 분석에 사용된 연도의 수

위 식은 각 월의 사건 건수가 전체 연간 데이터에서 차지하는 평균 비율을 계산하며, 이를 통해 여러 연도에 걸친 데이터를 기반으로 월별 평균 기여도를 구하였다(식.1). 그다음, 산출된 월별 평균 기여도를 활용하여 새로 추가된 연도의 월별 사건 건수를 다음 식으로 계산하였다(식.2).

$$MonthlyCount_i = YearlyTotal * \frac{MonthlyContribution_i}{100} \quad (2)$$

- $MonthlyCount_i$: i월의 예상 사건 건수
- $YearlyTotal$: 새로 추가된 연도의 총 사건 건수
- $MonthlyContribution_i$: 월 평균 기여도(백분율)

위 식은 연간 데이터를 각 월별 기여도에 따라 분배하는 방식으로, 새로 추가된 데이터를 기존 데이터와 일관되게 변환하는 데 사용되었다(식.2). 또한, 월별 평균 기여도의 총합은 모든 월을 합산했을 때 항상 100%가 되도록 계산되어, 데이터의 신뢰성과 일관성을 유지하였다(식.3).

$$\sum_{i=1}^{12} MonthlyContribution_i = 100 \quad (3)$$

$\sum_{i=1}^{12} MonthlyContribution_i$: 월 평균 기여도 합계

이와 같은 방식을 통해 새로 추가된 연도의 데이터를 월별 데이터로 변환하고, 기존 데이터와 결합하여 결측치를 보완하였다. 이를 통해 데이터의 안정성을 유지하며 분석의 정확도를 확보하였다.

전처리 과정을 통해 SARIMA 모델은 월별 데이터의 계절성과 추세를 분석할 수 있었고, ARIMAX 모델은 세부 유형 데이터를 외생 변수로 활용하여 사이버 범죄 발생 건수와 검거 건수 간의 상관성을 분석할 수 있었다. 결과적으로, 본 연구는 데이터의 신뢰성과 예측 모델의 정확성을 모두 확보할 수 있었다.

VI. Experiment

본 실험에서는 3장에서 전처리를 마친 데이터를 이용하여 SARIMA 모델과 ARIMAX 모델을 통해 최적의 사이버 범죄 예측 모델을 구축하고, 그 실험 과정과 결과를 상세히 기술한다.

1. Experimental Environment

본 연구의 실험은 Visual Studio Code 환경에서 수행하였다. 데이터 분석과 예측 모델 구현을 위해 Python 3.12.2 버전을 사용하였다. 또한, 데이터 처리와 분석을 위한 pandas 2.2.1, numpy 1.26.4, 시계열 분석 모델 구현을 위한 statsmodels 0.14.1, 예측 정확도 평가를 위한 scikit-learn 1.4.1.post1, 데이터 시각화를 위한 matplotlib 3.8.4 버전의 라이브러리를 각각 활용하였다. 모델의 성능 평가는 Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), Akaike Information Criterion (AIC)을 사용하여 진행하였다.

RMSE는 실제값과 예측값의 차이를 제곱 평균한 뒤 제곱근을 취한 지표이며, MAPE는 실제값 대비 예측 오차의 비율을 평가한 지표로 두 평가 지표 모두 값이 낮을수록 예측 정확도가 높다. 아래에 두 평가 지표의 계산식은 다음과 같다.(식.4)(식.5)[14]

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \quad (4)$$

- n: 분석에 사용된 연도의 수
- t: 시점의 인덱스 (1부터 n까지)
- y_t: 시점 t에서의 실제값
- $\hat{y}_t$: 시점 t에서의 예측값

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \quad (5)$$

- n: 분석에 사용된 연도의 수
- t: 시점의 인덱스
- y_t: 시점 t에서의 실제값
- $\hat{y}_t$: 시점 t에서의 예측값

AIC는 통계적 모델의 적합도와 복잡성을 함께 고려하여 모델을 평가하는 지표로, 값이 낮을수록 모델의 설명력은 유지하면서 불필요한 복잡성이 적음을 의미한다. 아래에 AIC의 계산식은 다음과 같다.(식.6)[15]

$$AIC = 2k - 2\ln(\hat{L}) \quad (6)$$

- k : 모형의 자유모수 개수
- $\hat{L}$: 모형의 최대 우도값

2. SARIMA Parameter

SARIMA 모델의 실험에서는 전처리를 마친 월별 사이버 범죄 발생 건수 데이터를 사용하였다. 최적의 모델을 찾기 위해 단계적 하이퍼 파라미터 최적화(Sequential Hyperparameter Optimization) 기법을 적용하였으며, 이는 다양한 파라미터 조합을 효율적으로 탐색하여 모델 성능을 극대화할 수 있는 장점이 있다.

본 연구에서는 자기회귀(p), 차분(d), 이동평균(q)의 파라미터 값을 각각 p와 q는 0에서 3까지, d는 0에서 2까지 설정하고, 계절성을 보이는 월별 데이터의 특성에 맞게 계절성 파라미터 계절 주기(s)를 12개월로 설정하였다. 이후 itertools의 product 함수를 활용하여 가능한 모든 파라미터 조합을 생성한 결과, 총 2,034개의 모델이 도출되었다. 이 모델들은 하이퍼 파라미터 최적화 과정을 통해 최종적으로 선정되며, 부적절한 파라미터를 제거하는 과정이 필요하다. 해당 최적화 과정의 절차는 아래 그림과 같다.

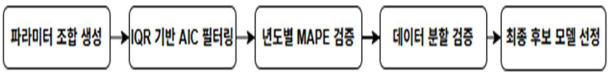

Fig. 9. sum of monthly average contributions

우선적으로, 모델의 성능과 복잡성 사이의 균형을 평가하기 위해 AIC를 활용하였다. AIC는 모델의 적합도를 측정하는 지표로[15], 극단적으로 낮거나 높은 값을 가지는 모델은 일반화 성능이 떨어질 가능성이 있으므로 이상치를 판별하여 제거하는 과정이 필요하다. 이를 위해 2,034개 전체 모델들에 대해 사분위수 범위(Interquartile Range, IQR) 방법을 적용하여 AIC 값이 극단적으로 벗어난 모델을 제거하였다. AIC 값의 제1사분위수(Q1)와 제3사분위수(Q3)를 기반으로 사분위 범위(IQR = Q3 - Q1)를 계산한 후, 이상치 탐지 기준을 Q1 - 1.5×IQR 및 Q3 + 1.5×IQR로 설정하였다[16]. 이 과정에서 각 모델의 AIC 값이 하한값(816.63) 미만이거나 상한값(2287.77) 초과인 모델은 이상치로 간주하고 제거하였다. 이러한 필터링 과정은 과적합(overfitting) 문제를 방지하고, 보다 일반화 성능이 우수한 모델을 찾기 위한 과정으로 수행되었다. AIC 필터링을 통과한 모델들은 이후 추가적인 하이퍼 파라미터 최적화 단계를 거쳐 최종 모델로 선정된다.

앞선 과정에서 AIC 기반 필터링을 통과한 모델 후보군에 대해 모델의 일반화 성능을 검증하기 위해 MAPE 지표를 활용해 MAPE 값이 일정 기준 이하인 모델만을 선별하는 과정을 거쳤다. MAPE는 모델의 상대적 예측 오류를 평가하는 데 유용하다. 특히, 단위 차이가 있는 데이터에서도 비교가 가능하다는 장점이 있어 본 연구에서는 모델 평가의 핵심 기준으로 활용하였다. 본 연구에서는 MAPE 값이 3% 이하인 모델을 신뢰할 수 있는 예측 모델로 간주하였다.

우선, 2021년까지의 데이터를 학습 데이터로 사용하여 2022년 발생 건수를 예측하고, 실제값(230,355건)과 비교하여 MAPE가 3% 이하인 모델을 선별하였다. 이후, 동일한 방법으로 2022년까지의 데이터를 학습 데이터로 사용하여 2023년 발생 건수를 예측하고, 실제값(241,842건)과 비교하여 MAPE 3% 이하를 만족하는 모델을 추가적으로 선별하였다. 이러한 과정을 통해 예측 정확도가 높은 모델 후보군을 도출한 후, 데이터 분할을 활용한 추가 검증을 수행하였다. 데이터 분할 추가 검증은 70:30, 80:20,

90:10의 세 가지 데이터 분할 방식을 적용하여 모델의 일반화 성능을 추가적으로 검증하였다. 데이터 분할 기법은 특정 학습 데이터에 과적합되는 문제를 방지하고, 모델이 다양한 데이터 환경에서도 일관된 예측 성능을 유지하는지를 평가하는 데 활용되었다.

먼저, 전체 데이터의 70%를 학습 데이터로, 30%를 테스트 데이터로 설정한 후 MAPE 값을 평가하였다. 이때, MAPE 값이 20% 이하를 만족하는 모델만을 선별하여 80:20 데이터 분할 검증 단계로 진행하였다. 이후, 80:20 분할에서는 18% 이하의 MAPE 값을 만족하는 모델만을 선별하여 90:10 분할 검증을 수행하였으며, 마지막으로 90:10 분할에서 MAPE 값이 15% 이하인 모델을 최종 모델 후보군으로 선정하였다. 최종 파라미터 후보군으로는 발생 건수의 모델의 경우 총 8개의 파라미터로 아래 표와 같다(Table.1).

Table 1. Final parameter for SARIMA (incident)

(p,d,q)	(P,D,Q,s)
(3, 1, 2)	(0, 1, 0, 12)
(2, 0, 0)	(0, 2, 1, 12)
(2, 0, 3)	(3, 1, 1, 12)
(1, 0, 3)	(3, 1, 1, 12)
(2, 0, 1)	(3, 1, 1, 12)
(1, 0, 2)	(3, 1, 1, 12)
(3, 0, 1)	(2, 1, 2, 12)
(0, 0, 3)	(3, 2, 3, 12)

검거 건수의 모델의 경우 총 5개의 파라미터로 아래 표와 같다(Table.2).

Table 2. Final parameter for SARIMA (arrest)

(p,d,q)	(P,D,Q,s)
(2, 0, 2)	(0, 2, 2, 12)
(1, 0, 1)	(2, 0, 1, 12)
(2, 2, 1)	(2, 0, 2, 12)
(1, 0, 2)	(1, 2, 2, 12)
(0, 0, 0)	(3, 1, 1, 12)

이러한 단계적 검증 방식은 점진적으로 더 적은 비율의 테스트 데이터에서 성능을 평가하는 구조로, 모델이 특정 학습 데이터에 과적합되지 않고, 전체 데이터에서 일관된 예측 성능을 유지하는지를 확인하는 데 초점을 맞추었다. 최종적으로, 모든 데이터 분할 단계를 통과한 모델을 최적 모델로 선정하여 예측 분석을 수행하였다. 최종 선별된 다수의 모델을 활용하여 2024년, 2025년, 2026년의 사이버 범죄 발생 건수를 예측하였다. 개별 모델이 가진 예측 오차를 보완하고, 안정적인 예측 결과를 도출하기 위해 앙상블 기법(Ensemble Method)[17]을 적용하였다.

앙상블 기법은 여러 개의 모델을 결합하여 단일 모델보다 신뢰성 높은 예측을 수행하는 방법으로, 본 연구에서는 단순 평균 방식(Simple Averaging)을 적용하였다. 이는 개별 모델이 산출한 예측값을 평균 내어 최종 예측값을 도출하는 방식으로, 특정 개별 모델의 편향을 줄이고 보다 일반화된 예측 성능을 확보하는 데 효과적이다[17]. 앞서 데이터 분할 검증을 통과한 최종 8개의 발생 건수 모델과 5개의 검거 건수 모델을 활용하여 2024년 9~12월 및 2025년, 2026년의 예측을 수행하였다. 각 모델별 개별 예측값을 산출한 후, 단순 평균 방식 기법을 적용하여 최종 예측값을 도출하였다. 최종 예측된 2024년, 2025년, 2026년의 사이버 범죄 발생 건수 및 검거 건수는 아래의 표와 같다(Table.3).

Table 3. Cybercrime prediction results with SARIMA

년도	2024	2025	2026
발생	314,773	335,962	350,137
검거	167,444	176,684	183,472

본 연구에서 적용한 앙상블 기법을 통해 개별 모델이 내포하는 불확실성을 줄이고, 안정적인 사이버 범죄 예측 모델을 제시하였다.

3. ARIMAX Parameter

ARIMAX 모델의 실험에서는 전처리를 마친 사이버 범죄 세부유형 데이터를 사용하였다. 데이터는 독립변수(X)에 사이버 범죄 세부유형, 종속변수(y)에 총합을 설정한 후, endog=y, exog=X로 구성하였다. ARIMAX 모델은 외생 변수와 종속 변수 간의 관계를 반영하여 정확한 예측을 가능하게 한다.

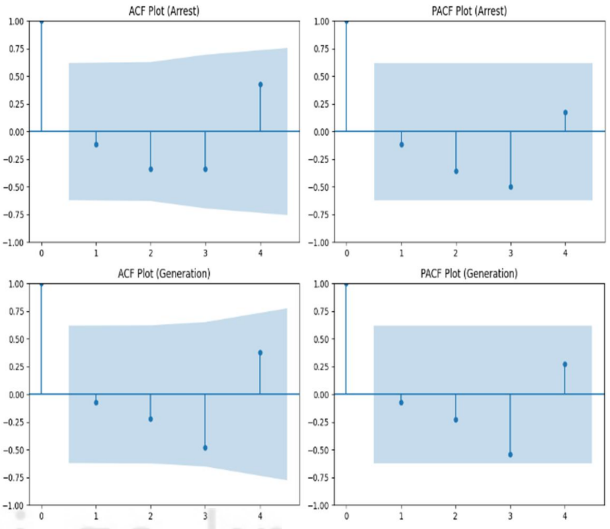

Fig. 10. ACF and PACF Plots for ARIMAX Parameter

ARIMAX 모델을 적용하기 위해서는 (p, d, q) 파라미터를 선정해야 하며, 본 연구에서는 데이터의 특성을 고려하여 이를 결정하였다. 그러나 샘플 수가 적어 ADF 검정을 통한 정상성 평가가 어려운 한계가 있어, 1차 차분(d=1)을 적용한 후 ACF/PACF 분석을 통해 최적의 (p, q) 값을 도출하였다. 분석 결과, ACF는 lag 1 이후 급격히 감소하고 PACF는 lag 1 이후 절단 패턴을 나타내어(Fig.10), 자기회귀(p)와 이동평균(q)의 차수를 각각 1로 설정하여 ARIMAX(1,1,1) 모델을 선정하였다.

ARIMAX 모델의 종속변수(총합)를 정확히 예측하기 위해서는 외생변수(독립변수)의 미래값이 미리 결정되어 있어야 한다.

이를 위해 SARIMA 모델을 사용하여 각 외생 변수의 2025년과 2026년 미래값을 독립적으로 예측하였으며, 예측된 결과를 새로운 데이터프레임으로 결합하여 ARIMAX 모델의 입력 데이터로 사용하였다. 외생 변수의 개별적 예측을 통해 ARIMAX 모델의 입력 데이터가 구성되었으며, 이를 통해 종속 변수인 사이버 범죄 총합의 정확한 예측이 가능하였다. ARIMAX 모델은 최종 구성된 외생 변수 데이터를 이용하여 2025년과 2026년의 총 사이버 범죄 발생 건수를 예측하였다. 월별 기여도 방법으로 추가된 2024년 사이버 범죄 발생과 검거 예측건수와 ARIMAX 모델을 통해 도출된 2025년과 2026년의 사이버 범죄 발생과 검거 예측 건수는 아래의 표와 같다(Table.4).

Table 4. Cybercrime prediction results with ARIMAX

년도	2024	2025	2026
발생	322,705	338,297	356,237
검거	168,354	175,001	177,523

본 연구에서 적용한 ARIMAX 모델은 외생 변수를 활용하여 종속 변수(총합)의 미래 값을 정확하게 예측할 수 있도록 하였으며, 이를 통해 단순한 시계열 모델보다 사이버 범죄 발생과 검거의 변화 양상을 더욱 정확하게 반영할 수 있었다.

4. Experiment Results

본 연구에서는 사이버 범죄 발생 및 검거 건수를 예측하기 위해 SARIMA와 ARIMAX 모델을 적용하였으며, 2014년부터 2023년까지의 데이터를 학습한 후 2024년부터 2026년까지의 발생 및 검거 건수를 예측하였다. 모델별 예측 결과를 비교하고 분석하기 위해 사이버 범죄 발생 건수, 검거 건수 및 검거율의 연도별 변화를 Fig.11, Fig.12에 나타내었다.

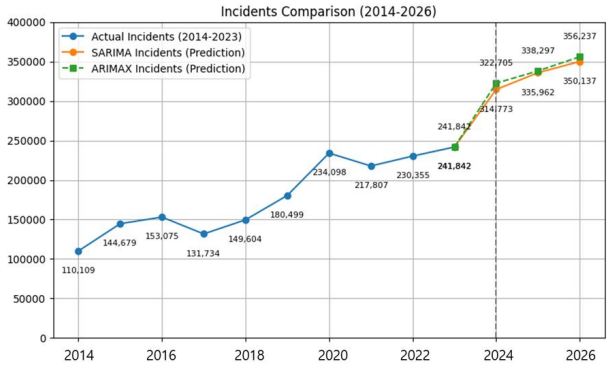

Fig. 11. Cybercrime incidents prediction results graph

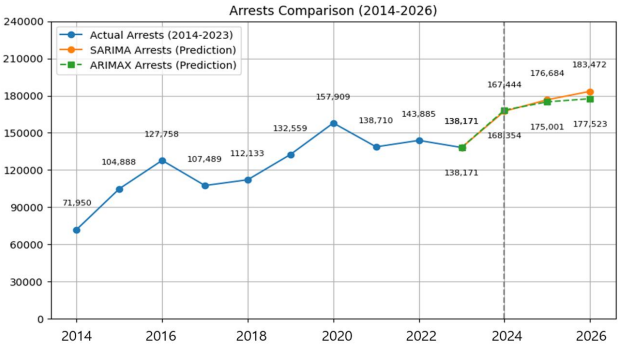

Fig. 12. Cybercrime arrest prediction results graph

사이버 범죄 발생 건수 예측 결과, 두 모델 모두 2024년 이후 지속적인 증가 추세를 보이는 것으로 나타났다. SARIMA 모델은 2024년 322,705건, 2025년 338,297건, 2026년 356,237건으로 예측되었으며, ARIMAX 모델은 2024년 314,773건, 2025년 335,962건, 2026년 350,137건으로 예측되었다. 두 모델 간 예측값의 차이는 존재하지만, 전반적인 증가 추세는 일관되게 나타났다. 사이버 범죄 검거 건수 예측 결과에서도 SARIMA와 ARIMAX 모델은 유사한 추세를 나타내었다. SARIMA 모델은 2024년 168,354건, 2025년 175,001건, 2026년 177,523건으로 예측되었으며, ARIMAX 모델은 2024년 167,444건, 2025년 176,684건, 2026년 183,472건으로 예측되었다. 예측 결과에 따르면 2024년 이후 검거 건수는 증가할 것으로 예상되나, 발생 건수의 증가폭에 비해 상대적으로 완만한 증가를 보였다.

사이버 범죄 검거율(Arrest Rate)은 검거 건수를 발생 건수로 나눈 값으로 계산되었으며, 모델별 예측 결과는 Fig.13에 제시하였다. SARIMA 모델은 2024년 52.2%, 2025년 51.7%, 2026년 49.8%로 예측하였으며, ARIMAX 모델은 2024년 53.2%, 2025년 52.6%, 2026년 52.4%로 예측되었다. 두 모델 모두 사이버 범죄 검거율이 점진적으로 감소하는 경향을 나타냈다.

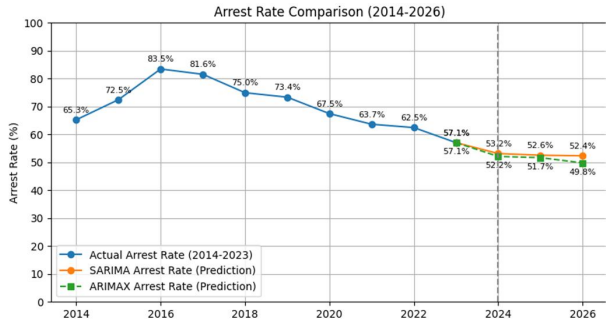

Fig. 13. Cybercrime arrest rate prediction graph

V. Conclusions

본 연구에서는 사이버 범죄 발생 및 검거 건수를 예측하기 위해 시계열 분석 모델인 SARIMA와 외생 변수를 포함한 ARIMAX 모델을 활용하였다. 분석 결과, 두 모델 모두 2024년부터 2026년까지 사이버 범죄 발생 건수가 지속해서 증가할 것으로 예측하였으며, 검거 건수 또한 증가할 것으로 나타났다. 그러나 검거율(Arrest Rate)의 경우, 발생 건수 증가 속도에 비해 완만한 증가를 보이며 점진적으로 감소하는 경향을 나타냈다. 이러한 연구 결과는 중요한 시사점을 제공한다. 본 연구의 예측에 따르면 사이버 범죄 발생이 지속적으로 증가할 것으로 예상되며, 이에 대한 대응 전략 마련이 필요하다. 특히, 예측 결과에서 검거율이 점진적으로 감소하는 경향을 보였다는 점은 현재의 사이버 범죄 대응 체계가 증가하는 범죄 발생을 효과적으로 억제하는 것이 쉽지 않음을 시사한다. 이에 따라, 수사 및 대응 역량 강화를 위한 정책적 고려가 필요하며, 예측 모델을 활용한 지속적인 모니터링이 요구된다.

본 연구에서 제시하는 2024년부터 2026년까지의 예측치는 미래에 대한 전망을 담고 있으므로, 해당 기간 사이버 경찰청 등 공신력 있는 기관에서 실제 발표될 데이터와 비교하여 모델의 최종적인 예측 정확도를 검증할 수 있을 것이다. 다만, 본 연구에서 활용한 모델은 기존 학습 및 테스트 데이터를 통해 파라미터의 유효성이 충분히 검증되었기에, 예측 정확도에 있어서 긍정적인 결과가 예상된다. 더욱이, 향후 데이터가 축적되면 예측 성능은 더욱 향상될 것으로 기대되며, 특히, 장기간의 데이터가 추가로 확보된다면 범죄 발생의 주기성과 장기적 추세를 더욱 명확하게 분석할 수 있어, 모델의 안정성과 신뢰성이 증가할 것으로 예상된다.

이를 바탕으로, 본 연구에서 제안한 예측 모델은 수사기관에서 향후 사이버 범죄의 발생 건수와 검거율의 추세를 미리 파악하고, 이를 바탕으로 보다 효과적인 인력 배치 및 선제적인 대응 전략 수립에 중요한 참고 자료로 활용될 수 있을 것으로 기대된다

REFERENCES

[1] Korean National Police Agency, Cybercrime Incidence and Arrest Statistics. e-Nara Indicators (e-나라지표). Available at: https://www.index.go.kr/unity/potal/main/EachDtlPageDetail.do?idx_cd=1608. Accessed March 7, 2025.

[2] Y.S. Kim and S. H. Byun, "Effects of Cyber Defamation Victims' Post-Traumatic Stress on Coping Behaviour: Focusing on the Theory of Reasoned Action," Journal of the Korea Society of Digital Industry and Information Management (KSDIM), vol. 15, no. 1, pp. 29-41, March 2019. DOI: 10.17662/ksdim.2019.15.1.029

[3] Korea Internet & Security Agency (KISA), "*2024 Cyber Threat Trend Report: First Half*", pp. 1-110, 2024.

[4] Deloitte Center for Integrated Research (Deloitte), "Different Types of Cyber Attacks and Response Strategies: Comparison of Cyber Threats and Considerations between 20 Countries in 3 Regions Around the World", pp. 1-20, April 2024.

[5] J.H. Kim and J.Y. Kim, "Prediction of Covid-19 Confirmed Number of Cases Using SARIMA Model," Journal of the Korea Institute of Information and Communication Engineering, vol. 26, no. 1, pp. 58-63, January 2022. DOI: 10.6109/jkiice.2022.26.1.58

[6] J.W. Won, B.C. Seong and I.S. Chang, "An Application of ARIMAX for Predicting Long-Term National Health Insurance Expenditure in Korea." The Korean Journal of Health Economics and Policy, vol. 22, no. 2, pp. 1-27, 2016.

[7] S. Yang, "Time Series Analysis: Understanding Changes in Indicators Over Time," HEARTCOUNT COMMUNITY, October 10, 2023. https://community.heartcount.io/ko/time-series-2/

[8] Woogong, "Differences Between Regression Analysis and Time Series Analysis," Tistory Blog, August 2, 2023. https://woogong80.tistory.com/285

[9] J.Y. Kim, S.B. Kang, E.H. Son, K.S. Choi, and S.M. Oh, "Predicting Behavioral Patterns of Elderly Living Alone Using Deep Learning (LSTM) and Statistical Methods (SARIMA): A Comparative Study," Proceedings of the 2024 Fall Conference of the Korea Institute of Information Technology, pp. 526-529.

[10] D.C. Han, D.W. Lee, and D.Y. Jung, "A Study on the Traffic Volume Correction and Prediction Using SARIMA Algorithm," Journal of the Korean Society of Intelligent Transport Systems, vol. 20, no. 6, pp. 1-13, December 2021. DOI: 10.12815/kits.2021.20.6.1

[11] Korean National Police Agency, "Monthly Cybercrime Incidence

and Arrest Statistics.", Available at: https://www.data.go.kr/data/15053884/fileData.do?recommendDataYn=Y. Accessed January 7, 2025.

[12] Korean National Police Agency, "Cybercrime Statistics by Detailed Type.", Available at: https://www.police.go.kr/www/open/publice/publice0204.jsp. Accessed May 2, 2024.

[13] BoanNews, "Increase in Police Cyber Investigation Personnel, but Cybercrime Cases and Arrest Rates Decline.", Available at: https://m.boannews.com/html/detail.html?idx=133306&page=1&kind=2. Accessed March 2025.

[14] H.J. Moon and N.W. Cho, "Text Mining Based Price Prediction in Public Auctions using Land Appraisal Reports. "Journal of Real Estate Analysis, Vol.11, No.1, pp.109-133. April 2025. DOI: 10.30902/jrea.2025.11.1.109

[15] H.C. Kim, "Development of a Regression-Based Tool Life Prediction Model in Manufacturing Environments," Journal of the Korean Society for Precision Engineering, vol. 42, no. 3, pp. 247-252, March 2025. DOI: 10.7736/JKSPE.024.131

[16] Oracle Help Center, "IQR (Interquartile Range) - Insights Metrics Definition.", Available at: https://docs.oracle.com/cloud/help/ko/pbcs_common/PFUSU/insights_metrics_IQR.htm.

[17] S. Kumar, "Ensemble Methods in Machine Learning.", Medium. Available at: https://medium.com/@shashank25.it/ensemble-methods-in-machine-learning-2d4cc7513c77.

Authors

Ji-Hyeok Choi received the Associate of Science (A.S.) degree from the Department of Computer Science at Inha Technical College, Korea, in 2024. He received the Bachelor of Science (B.S.) degree in the same department in 2025. Mr. Choi has practical experience in the IT field and is interested in Data Analysis and Data Science.

Kyu-Cheol Cho received the B.S., M.S. and Ph.D. degrees in Computer Science and Information Engineering from Inha University, Korea, in 2005, 2007 and 2013, respectively. Dr. Cho joined the faculty of the Department of Computer Science at Inha Technical College, Incheon, Korea, in 2016. He is currently an assistant professor in the Department of Computer Science, Inha Technical College. He is interested in cloud computing, green IT and web programming.