JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 30 No. 2, pp. 33-43, February 2025
https://doi.org/10.9708/jksci.2025.30.02.033

Prediction of Voice Phishing Victimization by Age Groups Using Principal Component Analysis: A Comparison of Regression Analysis and VAR Model Performance

Jungwoo Bae*, Byunghong Choi**

*Student, Gyeonggi Science High School, Suwon, Korea
**Teacher, Gyeonggi Science High School, Suwon, Korea

[Abstract]

Voice phishing victimization data exhibit distinct temporal patterns. This study compares the predictive performance of VAR and multiple linear regression models, considering their ability to capture time-series volatility. Principal component analysis was applied to voice phishing data (Jan 2018-Sep 2024), identifying three patterns explaining 99.31% of variation. An optimal window size of 12 months was determined through AIC testing to avoid overfitting. Model validation controlling for seasonal variability using Jul-Sep 2024 data showed the VAR model (MAE 33.06, MAPE 14.72%, RMSE 56.41, R^2 0.94) outperformed multiple linear regression by 2-3 times across all metrics. The VAR model was then used to predict victimization for February 2025, with 95% confidence intervals accounting for data volatility. Results predicted decreased victimization across all age groups, leading to suggestions for age-specific prevention strategies.

▸ **Key words**: Voice phishing, Principal Component Analysis, VAR model, Time series analysis, Regression analysis

[요 약]

보이스피싱 피해 데이터는 시기별로 상이한 패턴을 보인다. 본 연구는 시계열 변동성을 반영하는 VAR 모델과 다중선형회귀 모델의 예측 성능을 비교하였다. 2018년 1월부터 2024년 9월까지의 데이터에 주성분분석을 적용하여 전체 변동의 99.31%를 설명하는 세 가지 패턴을 도출하였다. AIC 검정으로 최적 이동 창 크기를 12개월로 결정하여 과적합을 방지하였다. 계절적 변동성을 통제하기 위해 선정한 2024년 7-9월 데이터로 검증한 결과, VAR 모델(MAE 33.06, MAPE 14.72%, RMSE 56.41, R^2 0.94)이 다중선형회귀보다 모든 평가척도에서 약 2-3배 우수한 성능을 보였다. VAR 모델로 2025년 2월 연령대별 피해를 예측하고 95% 신뢰구간을 산출한 결과, 모든 연령대에서 피해 감소가 예측되어 연령별 맞춤형 예방 정책을 제안하였다.

▸ **주제어**: 보이스피싱, 주성분분석, VAR 모델, 시계열 분석, 회귀분석

• First Author: Jungwoo Bae, Corresponding Author: Byunghong Choi
*Jungwoo Bae (gs23056@gs.hs.kr), Gyeonggi Science High School
**Byunghong Choi (cbh0706@snu.ac.kr), Gyeonggi Science High School
• Received: 2025. 01. 08, Revised: 2025. 02. 20, Accepted: 2025. 02. 21.

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

I. Introduction

보이스피싱은 비대면 사기 수법으로, 피해자에게 심리적 압박을 가하여 금전적 피해를 입히는 범죄이다. 이에 따라 사전에 피해 규모를 예측하고, 이를 바탕으로 예방 정책을 수립하는 과학적 접근이 필요한 실정이다[1].

지금까지의 보이스피싱 예측 연구는 주로 단기적 탐지와 예방에 초점을 맞추어 왔다. 머신러닝과 딥러닝을 활용한 이상거래탐지시스템[2]과 자연어 처리 기반의 실시간 탐지 연구[3, 4]는 개별 사례의 즉각적 탐지에는 성과를 보였으나, 장기적인 피해 패턴 예측에는 한계를 보였다. 선형회귀분석을 활용한 범죄 수사 기술 예측[5]이나 장단기 메모리(Long Short-Term Memory, LSTM), 게이트 순환 유닛(Gated Recurrent Unit, GRU) 등 신경망 모델을 활용한 비교 연구[6] 역시 단순 선형 관계나 복잡한 모델 구조에 의존하여, 실질적인 정책 수립에 필요한 해석 가능한 결과를 제시하지 못했다. 또한 횡단보도 신호 시스템에 다중회귀 모델을 적용하여 보행환경 개선을 시도한 연구[7]에서는 구체적인 예측 모델과 정책적 함의를 효과적으로 연계시키지 못했다는 한계가 있다.

국제적으로도 딥러닝을 이용하여 문제를 분석하거나 예측을 하려는 여러 시도가 있었다. 비지도 학습을 통한 패턴 분석[8], 딥러닝 기반 스팸 예측[9], 교통 사고 등을 예방하기 위하여 합성곱 신경망(Convolutional Neural Network, CNN)과 심층 신경망(Deep Neural Network, DNN)의 성능을 결합한 사고 감지기를 도입한 경우도 있다[10]. 이 연구들은 대부분 방대한 학습 데이터를 요구하거나 모델의 복잡성으로 인해 실용적 적용에 제약이 있었다.

이 연구에서는 2018년 1월부터 2024년 9월까지의 경찰청 보이스피싱 피해 데이터를 활용하여 시계열 예측 모델을 구축한다. 데이터의 효율적인 분석을 위해 주성분분석(Principal Component Analysis, PCA)[11]을 적용하여 차원을 축소하고, 과적합 방지를 위해 12개월의 이동 창 기법을 도입한다. 이를 바탕으로 벡터자기회귀(Vector Autoregression, VAR) 모델[12]과 다중선형회귀 모델을 구축하여 계절적 변동성을 통제하고 안정적인 비교가 가능한 2024년 7월부터 9월까지의 피해 규모를 예측하고 두 모델의 성능을 비교 분석한다. 분석 결과를 통해 보이스피싱 피해 예측에 더 적합한 모델을 선정하고, 2025년 2월의 연령대별 피해 규모를 예측한다. 이러한 결과는 정책 수립에 실증적 근거를 제공할 것이다.

이 연구는 선행연구들과 대비하여 세 가지 차별화된 접근을 제시한다. 첫째, 데이터의 다수 변수 간 상관관계를 고려하여 정보 손실을 최소화하면서 데이터 특성을 효과적으로 설명하기 위해 주성분분석을 실시하였다. 이 방법은 각 변수의 영향력을 개별적으로 수치화할 수 있어 실증적 분석에 적합하며, 모델 구축과 검증이 비교적 용이하면서도 신뢰성 높은 결과를 제공한다는 장점이 있다.

둘째, 두 가지 주요 통계적 방법론인 VAR 모델과 다중선형회귀 모델을 사용하여 예측 성능을 비교하고자 한다. 다양한 시계열 분석 방법 중 VAR 모델을 선택한 근거는 다중선형회귀 모델과의 데이터 처리 방식에서 중요한 비교 가치가 있기 때문이다. 다중선형회귀 모델은 여러 설명변수가 종속변수에 미치는 영향을 동시에 분석할 수 있어 연령대별 피해 특성을 종합적으로 이해하는 데 유용하다. 반면 VAR 모델은 시계열 데이터의 특성과 연령대별 피해 패턴을 분석하는 데 더 적합하다고 판단하였다. 기존 연구에서 VAR 모델이 주로 경제 데이터 분석에 활용된 반면, 본 연구는 범죄 예측 모델링에 VAR을 적용한 점에서 차별성을 갖는다.

셋째, 2025년 2월의 연령대별 보이스피싱 피해를 시계열 분석을 통해 예측하고, 이를 바탕으로 연령대별 특성을 고려한 맞춤형 예방 정책 수립을 위한 실증적 근거를 제공한다. 이러한 접근은 기존 연구들이 시도하지 않았던 예측의 정확성과 정책적 활용성의 균형적 달성을 가능하게 한다.

본 논문의 구성은 다음과 같다. 2장에서는 연구의 이론적 배경으로 주성분분석, VAR 모델, 다중선형회귀 모델을 설명한다. 3장에서는 데이터셋 구성, 차원 축소를 위한 주성분분석, 보이스피싱 피해 예측, 성능 평가 비교 분석을 다룬다. 마지막으로 4장에서는 결론을 제시한다.

II. Preliminaries

1. PCA Modeling

이 연구에서 주성분분석은 연령대별 보이스피싱 발생 데이터의 구조적 특성을 파악하기 위해 사용되었다. 주성분분석은 다차원 데이터의 변동 패턴을 효과적으로 분석하는 다변량 통계기법[13]으로, 6개 연령대(20대 이하, 30대, 40대, 50대, 60대, 70대 이상)의 시계열적 변동 패턴을 이해하는 데 적합하다.

연령대별 보이스피싱 발생 데이터를 행렬X로 표현하면, 각 열은 특정 연령대를, 각 행은 시간(2018년 1월부터 2024년까지 9월까지 81개월)을 나타낸다. 주성분분석은 이러한 다차원 데이터에서 원래 변수들의 선형 결합을 통해 새로운 변수(주성분)를 생성한다. 첫 번째 주성분은 전

체 데이터 변동을 가장 잘 설명하는 방향을 나타내며, 두 번째 주성분은 첫 번째 주성분과 직교하면서 남은 변동을 최대한 설명한다. 이후의 주성분도 동일한 원리로 구성된다. 이 연구에서 각 주성분 Y_k는 식 (1)로 표현된다.

$$Y_k = a_{k1}X_1 + a_{k2}X_2 + \ldots + a_{k6}X_6 \quad (1)$$

식 (1)에서 X_1부터 X_6은 각 연령대의 보이스피싱 발생 데이터를 나타내며, a_{k1}부터 a_{k6}은 각각 k번째 주성분의 계수를 의미한다. 이 계수들은 각 연령대가 해당 주성분에 기여하는 정도를 나타내며, 보이스피싱 발생의 연령별 패턴을 해석하는 데 중요한 정보를 제공한다.

데이터의 시계열적 특성을 고려할 때, 각 주성분의 시간에 따른 변화는 보이스피싱 발생의 주요 변동 패턴을 나타낸다. 특히, 주성분 점수의 시계열적 변화는 보이스피싱 발생의 시간적 추세와 주기성을 파악하는 데 활용된다. 각 주성분이 설명하는 분산의 비율(설명분산비율-explained variance ratio)은 식 (2)와 같이 계산된다. 식 (2)는 i번째 주성분이 전체 분산 중 차지하는 비율을 백분율로 나타낸다.

$$\text{설명분산비율}_i(\%) = (\lambda_i / \sum_{k=1}^{n} \lambda_k) \times 100 \quad (2)$$

식 (2)에서 λ_i는 i번째 주성분의 고유값(Eigenvalue)이고 n는 주성분의 수를 나타낸다. 이 설명분산비율은 각 주성분이 전체 보이스피싱 발생 패턴의 변동을 얼마나 잘 설명하는지를 나타내는 지표가 된다.

이 연구에서는 연령대 간 척도 차이를 고려하여 데이터를 정규화한 뒤 분석을 수행하였다. 이 과정은 보이스피싱 피해 건수의 절대적 규모가 상이한 연령대들 간에 상대적 변동 패턴을 공정하게 비교할 수 있도록 하기 위함이다.

2. Theoretical background of the VAR model

VAR 모델은 다변량 시계열 자료의 동적 관계를 분석하고 예측하는 통계적 방법으로, 연령대별 보이스피싱 피해의 상호 연관성과 시간적 변화를 동시에 고려할 수 있다. 여러 시계열 분석 방법 중 자기회귀통합이동평균(Autoregressive Integrated Moving Average, ARIMA)과 Prophet은 단변량 분석 모델로서 단일 시계열 변수만을 고려하여 예측을 수행하며, 계절성 자기회귀통합이동평균(Seasonal ARIMA with eXogenous variables, SARIMAX)는 외생변수를 포함할 수 있으나 단일 종속변수에 대한 예측에 초점을 맞추는 한계를 가진다[14]. 반면, VAR 모델은 다중선형회귀 모델과 마찬가지로 다변량 분석이 가능하며 변수들 간의 선형적 관계를 가정한다는 공통점이 있어, 두 모델의 예측 성능을 직접적으로 비교할 수 있다는 장점을 갖는다. 따라서 이 연구에서는 여러 독립변수의 영향력을 분석하는 다중선형회귀 모델과의 성능 비교를 위해 VAR 모델을 분석 대상으로 선정하였다.

2.1 Estimating and analyzing VAR models

VAR 모델은 현재 시점의 변수값이 모든 변수의 과거 값에 의해 영향을 받는다는 기본 전제를 바탕으로 한다. 이 관계는 식 (3)으로 표현된다[15].

$$Y_t = c + A_1 Y_{t-1} + A_2 Y_{t-2} + \ldots + A_p Y_{t-p} + \epsilon_t \quad (3)$$

식 (3)에서 Y_t는 t시점의 내생변수 벡터, c는 실제 모델의 상수항 벡터, A_i는 i번째 시차의 계수행렬, ε_t는 t시점의 오차항 벡터를 나타낸다.

VAR 모델에서 적절한 시차 선택은 모델의 예측력을 좌우하는 중요한 요소이다. 시차가 너무 작으면 변수 간의 동적 관계를 충분히 포착하지 못하며, 반대인 경우 추정해야 할 파라미터 수가 증가해 모델의 효율성이 저하된다. 이에 이 연구에서는 최적 시차를 선택하기 위해 AIC(Akaike Information Criterion)와 BIC(Bayesian Information Criterion)의 정보이론 기반 지표를 활용하였다.

최종 예측값은 식 (4)로 정의되며 p차 VAR 모델(VAR(p))의 구조를 반영한다. VAR(p)는 현재 시점(t)으로부터 h시점 이후의 값을 과거 p개의 시점 정보를 기반으로 예측하는 방식이다.

$$Y_{t+h} = \hat{c} + \widehat{A}_1 Y_{t+h-1} + \widehat{A}_2 Y_{t+h-2} + \ldots + \widehat{A}_p Y_{t+h-p} \quad (4)$$

식 (4)에서 Y_{t+h}는 t시점부터 h시점 이후의 예측값, $\hat{c}$는 추정된 상수항 벡터, $\widehat{A}_i$는 계수 행렬, 그리고 $Y_{t+h-1},\ldots,Y_{t+h-p}$는 미래 시점$t+h$에서 과거 $1,\ldots,p$시점만큼 떨어진 값을 나타낸다.

2.2 Time series tests for normality and unit roots

VAR 모델을 적용하기 위해서는 시계열 자료가 정상성(stationarity)을 충족해야 한다. 정상성이란 시계열의 확률적 특성이 시간에 따라 변하지 않는 성질로, 평균이 일

정하고 분산이 유한하며, 공분산이 시간과 무관하게 시차에만 의존하는 특성을 의미한다[16].

정상성 여부를 검증하기 위해 일반적으로 ADF(Augmented Dickey-Fuller) 검정이 사용된다. ADF 검정의 귀무가설은 "시계열이 단위근을 가진다(비정상적이다)"이며, 대립가설은 "시계열이 정상적이다"이다. 검정통계량의 p값이 유의수준(일반적으로 0.05)보다 작을 경우, 귀무가설을 기각하고 해당 시계열이 정상적이라고 판단한다[17].

3. Multiple linear regression model framework

다중선형회귀 모델은 여러 독립변수와 종속변수 간의 선형관계를 모델링하는 통계적 방법으로, 식 (5)와 같이 일반적으로 표현된다[12].

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \quad (5)$$

식 (5)에서 Y는 종속 변수, $X_1, X_2, \ldots, X_n$은 독립 변수를 나타낸다. 그리고 β_0는 y절편이며 $\beta_1, \beta_2, \ldots, \beta_n$는 각 독립변수에 대한 회귀 계수로, 독립 변수가 종속 변수에 미치는 영향을 수치화하여 나타낸다. 이 모델은 Y와 독립 변수 $X_1, X_2, \ldots, X_n$간의 선형 관계를 모델링하며, 회귀 계수 $\beta_1, \beta_2, \ldots, \beta_n$는 추정된 값을 기반으로 종속 변수의 예측에 활용된다. 이 연구에서의 $X_1, X_2, \ldots, X_n$은 연령대별 피해 건수를 나타내며, n은 주성분분석 결과로 도출된 주성분의 수를 의미한다. 회귀 계수는 최소제곱법(Ordinary Least Squares, OLS)을 통해 추정되며, 이는 실제 관측값과 예측값의 차이 제곱합을 최소화하는 방식으로 계산된다.

III. Voice Phishing Prediction

1. Dataset configuration

1.1 Victims of voice phishing

Table 1은 경찰청에서 제공한 자료로, 2018년 1월부터 2024년 9월까지의 보이스피싱 월별 현황을 나타낸다. Table 2는 경찰청에서 제공하는 보이스피싱 피해자의 연령별 현황을 나타낸다.

Table 1. Monthly Voice Phishing Statistics

Year	Month	Voice Phishing Incidents
2018	1	3035
2018	2	2254
...	...	...
2024	7	1682
2024	8	1709
2024	9	1203

Table 2. Voice Phishing Victims by Age Group

Year	20s and under	30s	40s	50s	60s	70s and above
2016	3209	3735	4542	3834	1261	459
2017	5273	4887	6473	5412	1807	407
2018	4480	6483	9842	9313	3389	625
2019	3855	6041	10264	11825	4617	1065
2020	5323	4406	7704	9217	4188	843
2021	5459	3299	6755	9564	4778	1127
2022	6805	1821	3413	5378	3462	953
2023	8886	1621	2325	3149	2144	777

Table 2에서는 2016년부터 2023년까지의 데이터를 20대 이하, 30대, 40대, 50대, 60대, 70대 이상으로 구분하여 작성하였다.

1.2 Configuration of datasets for forecasting

이 연구는 2018년 1월부터 2024년 9월까지의 경찰청 보이스피싱 피해 데이터를 기반으로 분석을 진행하였다. 데이터는 연령대별(20대 이하, 30대, 40대, 50대, 60대, 70대 이상) 및 월별(1월부터 12월) 피해 건수로 구성되었으며, 최종적으로 총 81개월의 시계열 데이터를 Table 3에 정리하였다. Table 3에서 Table 2의 보이스피싱 연령대별 현황 데이터는 2023년 12월까지만 공개되어 있어, 2024년 1월부터 9월까지의 데이터는 결측치로 간주하고, 연령대별 비율을 2018년~2023년 데이터를 기준으로 산출하고, 평균값을 적용하여 2024년 데이터를 추정하였다.

Table 3. Monthly Voice Phishing Incidents by Age Group

Year	Month	20s and under	30s	40s	50s	60s	70s and above
2018	1	398	576	875	828	301	56
2018	2	296	428	650	615	224	41
2018	3	392	568	862	816	297	55
2018	4	383	554	841	796	290	53
...	...	...	...	...	...	...	...
2024	6	760	139	199	270	184	67
2024	7	791	144	207	280	191	69
2024	8	803	147	210	285	194	70
2024	9	565	103	148	200	137	49

2. PCA for dimension reduction

2.1 Explained variance ratio of PCA

Table 3의 데이터를 기반으로 주성분분석을 진행한 결과, Fig. 1과 같이 총 6개의 주성분이 도출되었다. Fig. 1은 각 주성분의 설명분산비율(%)과 누적설명분산비율(%)을 시각적으로 나타내고 있다.

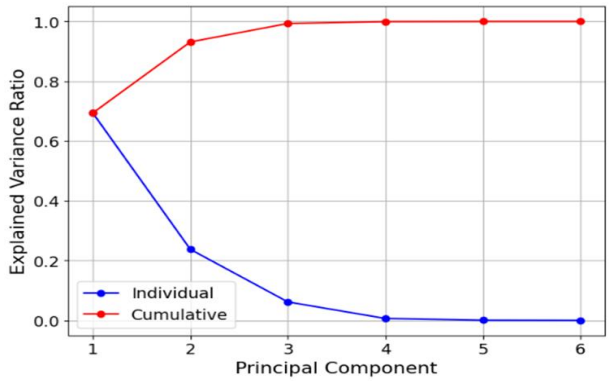

Fig. 1. Explained Variance Ratio of Principal Components (%)

Fig. 1에서 첫 번째 주성분은 전체 분산의 69.38%를 설명하며 가장 높은 설명분산비율을 나타냈고, 두 번째 주성분은 23.76%를, 세 번째 주성분은 6.17%를 설명하였다. 이 외에 네 번째 주성분은 0.64%, 다섯 번째 주성분은 0.05%의 설명분산비율을 보였으며, 여섯 번째 주성분의 설명분산비율은 0.00%로 나타났다.

주성분 수 결정을 위해 카이저 기준(Kaiser criterion)과 누적 설명분산비율 95% 이상이라는 두 가지 기준을 검토하였다. 카이저 기준에 따르면 첫 번째(고유값: 4.2150)와 두 번째(고유값: 1.4435) 주성분만이 1 이상의 고유값을 가졌다. 여기서 고유값은 해당 주성분이 설명하는 분산의 크기를 의미한다. 그러나 누적 설명분산비율 95% 이상이라는 기준을 고려하여 세 번째 주성분까지 포함하였으며, 이들의 누적 설명분산비율은 99.31%로 나타났다. 이러한 통계적 기준을 통해 선택된 세 개의 주성분이 실제로 어떤 의미를 가지는지 적재값(Factor loading) 분석을 통해 확인할 수 있다.

2.2 Analysis of the characteristic patterns of each principal component

선정된 세 개 주성분의 특성을 파악하기 위해 각 연령대가 주성분에 미치는 영향을 분석하였으며, 그 결과는 Fig. 2와 같다. Fig. 2의 적재값 분석을 통해 각 주성분의 특성을 확인할 수 있다. 첫 번째 주성분(설명분산비율 69.38%)은 경제활동 연령대(30대~50대)의 전반적인 피해 패턴을 반영한다.

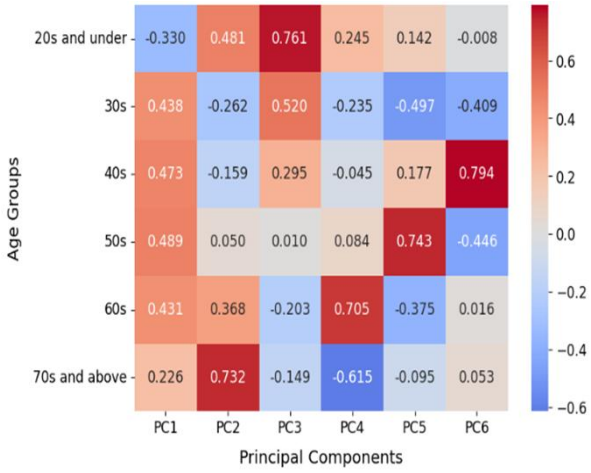

Fig. 2. Factor Loadings of Age Groups on Principal Components

50대는 0.489, 40대 0.473, 30대 0.438, 60대 0.431의 높은 양의 적재값과 20대 이하 -0.330의 음의 적재값은 경제활동이 가장 활발한 연령대가 보이스피싱의 주요 표적임을 시사한다. 두 번째 주성분(설명분산비율 23.76%)은 금융취약 계층(20대 이하 및 70대 이상)의 특성을 반영한다. 70대 이상 0.732과 20대 이하 0.481의 강한 양의 적재값, 30대 -0.262의 음의 적재값은 금융 경험이 부족하거나 디지털 취약성이 있는 계층과 경제활동이 활발한 연령층 간의 차이를 보여준다. 세 번째 주성분(설명분산비율 6.17%)은 디지털 친숙도에 따른 세대 간 차이를 반영한다. 20대 이하 0.761와 30대 0.520의 높은 양의 적재값, 60대 -0.203와 70대 이상 -0.149의 음의 적재값은 디지털 네이티브 세대와 고령층 간의 피해 패턴 차이를 명확히 보여준다.

이러한 차원 축소는 VAR 모델의 파라미터 수를 36개(6×6)에서 9개(3×3)로 감소시킴으로써 모델의 효율성을 크게 향상시켰다. 주성분분석을 VAR 모델과 다중선형회귀 모델에 적용한 주요 장점은 다음과 같다. 첫째, 차원 축소를 통해 모델의 복잡도를 줄이면서도 높은 설명분산비율을 유지할 수 있다. 둘째, 주성분별로 시계열 특성을 분석하고 예측을 수행할 수 있다. 셋째, 연령대 간 상관관계로 인해 발생할 수 있는 다중공선성 문제를 효과적으로 해결할 수 있다.

2.3 Time series normality test results

이 연구에서는 주성분분석을 통해 도출된 주성분들의 정상성을 검증하기 위해 단위근 검정(Augmented Dickey-Fuller, ADF)을 수행하였다. ADF 검정은 시계열

데이터의 확률적 특성을 분석하고 정상성을 확인하는 데 널리 사용되는 통계적 기법이다. 분석 결과, 세 개의 주성분(PC1, PC2, PC3) 모두 통계적으로 유의미한 정상성을 확보한 것으로 확인되었으며, 구체적인 결과는 Table 4에 제시되어 있다.

Table 4. Results of ADF Test

	PC1	PC2	PC3
ADF Statistic	-8.7483	-5.4098	-4.672
p-value	1.354e-12 (< 0.01)	3.465e-05 (< 0.01)	7.793e-04 (< 0.01)

Table 4의 결과에 따르면, 각 주성분의 p-value가 0.01보다 현저히 작아 귀무가설(단위근 존재)을 기각하고 대립가설(정상 시계열)을 채택할 수 있었다. 이는 이 연구에서 사용된 시계열 데이터가 통계적으로 안정적이며, VAR 모델과 다중선형회귀 모델에 적합함을 의미한다[18].

정상성 확보는 시계열 분석의 중요한 전제 조건으로, 데이터의 통계적 특성이 시간에 따라 일정하게 유지됨을 보장한다. 이 연구의 결과는 연령별 보이스피싱 데이터의 주성분들이 안정적인 시계열 특성을 가지고 있음을 확인해 주며, 이를 기반으로 신뢰성 있는 분석과 예측이 가능함을 시사한다.

3. Prediction of voice phishing damage

3.1 Optimization of time series prediction models

이 연구에서는 2018년 1월부터 2024년 9월까지의 데이터를 사용한다. 예측 모델의 정확도 검증을 위해 2024년 7월부터 9월까지의 3개월을 선택하였는데, 이는 단일 계절 내 연속된 기간을 선택함으로써 계절적 변동성을 통제하고 안정적인 비교가 가능하기 때문이다. 그러나 전체 데이터의 수가 제한적이어서 모델이 데이터의 패턴을 과도하게 학습하는 과적합 문제가 발생할 수 있다. 이를 방지하기 위해 이동 창(Moving Window) 접근 방식을 채택하였다. 여기서 이동 창은 전체 데이터셋을 한 번에 사용하는 대신, 가장 최근의 일정 기간 데이터만을 순차적으로 사용하는 방식이다. 예를 들어, 12개월 이동 창의 경우 2024년 7월을 예측할 때는 2023년 7월부터 2024년 6월까지의 최근 12개월 데이터를 사용하는 방식으로 예측 시점에 따라 학습에 사용되는 데이터가 순차적으로 이동하게 된다. 각 이동 창 크기는 서로 다른 특성을 가지고 있는데, 12개월의 경우 가장 최근의 트렌드를 반영하지만 장기적 패턴을 놓칠 수 있고, 24개월의 경우 장기적 패턴은 잘 포착하지만 최근의 변화에 덜 민감할 수 있다. 이동 창은 전체 데이터셋 대신 일정 기간의 최근 데이터만을 사용하여 학습함으로써 과적합을 줄이고 최신 트렌드를 더 잘 반영할 수 있는 장점이 있다. 최적의 이동 창 크기를 결정하기 위해 AIC(Akaike Information Criterion) 검정을 활용하였다. AIC는 모델의 적합도와 복잡성 사이의 균형을 평가하는 통계적 도구로, 식 (6)의 수식으로 표현한다.

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

식 (6)에서 k는 모델의 파라미터 수이고, L은 모델의 최대 우도값(maximum likelihood value)이다. L은 실제 데이터와 예측값 사이의 잔차의 평균제곱오차(Mean Squared Error, MSE)를 사용하여 식 (7)과 같이 표현할 수 있다.

$$L = -n/2log(MSE) - n/2*log(2\pi) - n/2 \tag{7}$$

식 (7)에서 n은 관측치의 수, π는 원주율을 나타낸다. AIC 값이 낮을수록 더 적합한 모델을 의미하며, 모델이 불필요하게 복잡해지는 것을 방지함으로써 과적합을 억제할 수 있다.

이 연구에서는 12개월, 18개월, 24개월의 세 가지 이동 창 크기를 대상으로 정확도를 비교 분석하였다. Fig. 3은 2018년부터 2024년도의 데이터를 대상으로 이동 창 크기를 12개월, 18개월, 24개월로 지정하고, 2024년 7월, 8월, 9월의 보이스피싱 피해 예측에 대해 AIC 검정을 실시하였을 때의 결과이다.

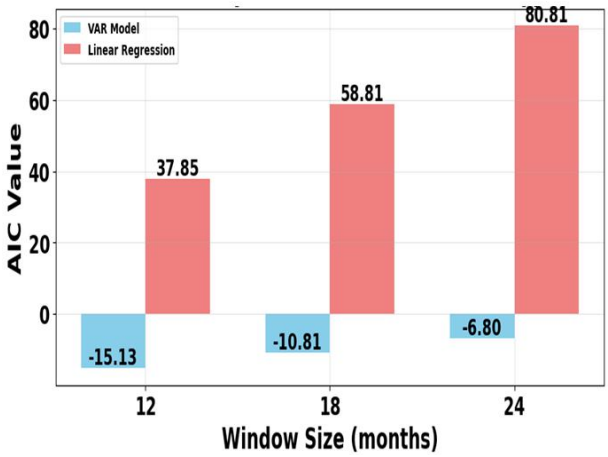

Fig. 3. Comparison of AIC Values by Moving Window

Fig. 3의 AIC 검정 결과, 이동 창 크기가 12개월일 때 VAR 모델은 -15.13, 다중선형회귀 모델은 37.85의 AIC 값을 보였다. 18개월의 경우 VAR 모델은 -10.81, 다중선

형회귀 모델은 58.81을, 24개월의 경우 각각 -6.80과 80.81을 기록하였다. 두 모델 모두 12개월의 이동 창 크기에서 가장 낮은 AIC 값을 보여 최적의 성능을 나타냈다. 이에 이 연구는 제한된 데이터셋에서도 과적합을 방지하면서 예측 정확도를 최적화하기 위하여 이동 창 크기를 12개월로 결정하였다.

3.2 Comparison of VAR model and linear regression model

이 연구에서는 2018년 1월부터 2024년 9월까지의 연령대별 보이스피싱 피해 데이터를 대상으로 주성분분석을 통해 데이터의 차원을 축소하였고 과적합을 방지하기 위해 이동 창 기법을 적용하였다. AIC 검정을 통해 최적의 이동 창 크기를 12개월로 결정하였으며, 6개 연령대 변수와 이동 창 기법을 적용하여 모델의 복잡성을 제어하였다. 이를 바탕으로 VAR 모델과 다중선형회귀 모델을 구축하여 2024년 7월부터 9월까지의 피해 규모를 Table 5와 Fig. 4 같이 예측하였다. LR(Linear Regression)은 다중선형회귀 모델을 나타낸다.

Table 5. Comparison of Model Predictions (July-September 2024)

Year-Month	Category	Model	20s and under	30s	40s	50s	60s	70s and above
2024.7	Predicted	VAR	751	135	194	264	190	65
		LR	851	178	289	415	269	86
	Actual		791	144	207	280	191	69
2024.8	Predicted	VAR	767	133	188	256	188	66
		LR	684	102	117	141	126	49
	Actual		803	147	210	285	194	70
2024.9	Predicted	VAR	775	133	186	255	188	66
		LR	936	172	258	364	251	86
	Actual		565	103	148	200	137	49

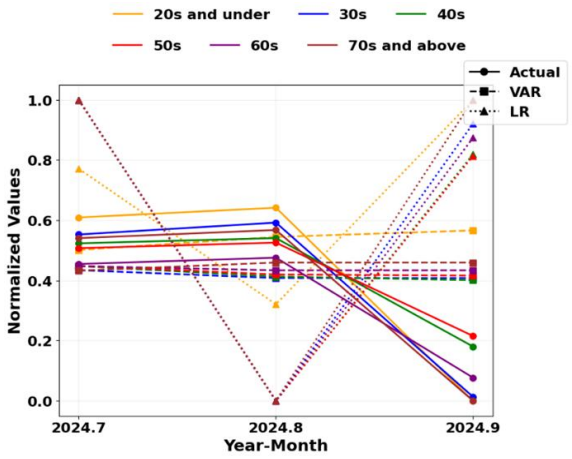

Fig. 4. VAR Model versus Linear Regression – Predicted versus Actual Values by Month (2024)

Fig. 4는 Table 5와 같이 나타난 피해건수를 0부터 1까지의 값으로 스케일링한 정규화된 데이터를 바탕으로 두 모델의 시계열 패턴을 비교하였다. Fig. 4에서 연령대별 예측값과 실제값의 시계열 패턴을 통해 두 모델의 예측 정확도를 정량적·시각적으로 분석할 수 있다. VAR 모델(불연속선)은 실제값의 추세를 잘 반영하는 반면, 다중선형회귀 모델(연속선)은 실제값과의 편차가 크며, 시계열 특성을 반영하는 데 한계가 있음을 확인할 수 있다.

4. Comparative analysis of performance evaluation

4.1 Metrics for performance evaluation

이 연구에서는 다중선형회귀 모델과 VAR 모델의 예측 성능 평가를 위해 Table 6에 제시된 4가지 평가 척도[19]를 활용하였다.

Table 6. Performance Evaluation Metrics for Prediction Models

Metrics	Description	Formula
MAE (Mean Absolute Error)	·Average absolute difference between predicted and actual values ·Provides intuitive measure of prediction accuracy	$MAE = (1/n)\sum_{i=1}^{n}\lvert y_i - \widehat{y_i}\rvert$
RMSE (Root Mean Square Error)	·Square root of mean squared errors ·More sensitive to large prediction deviations	$RMSE = \sqrt{(1/n)\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}$
MAPE (Mean Absolute Percentage Error)	·Relative error expressed as percentage ·Useful for comparing errors across different scales	$MAPE = (100\%/n)\sum_{i=1}^{n}\lvert(y_i - \widehat{y_i})/y_i\rvert$
R^2 (R^2-Coefficient of Determination)	·Indicates model's explanatory power ·Closer to 1 means better predictive performance	$R^2 = 1 - [\sum_{i=1}^{n}(y_i - \widehat{y_i})^2]/[\sum_{i=1}^{n}(y_i - \bar{y})^2]$

각 척도는 예측 성능의 다양한 측면을 평가하는데, MAE는 예측값과 실제값 간의 평균 절대 차이를 나타내어 예측 정확도를 직관적으로 보여준다. RMSE는 오차의 제곱 평균의 제곱근으로 큰 오차에 더 민감하게 반응하며, MAPE는 실제값 대비 상대적 오차를 백분율로 표현하여 단위가 다른 경우에도 비교가 용이하다. R^2은 모델의 전반적인 설명분산비율을 나타내며, 값이 1에 가까울수록 예측 성능이 우수함을 의미한다.

Table 6에서 n은 2024년 7월부터 9월까지의 예측 기간을 의미하며, 총 3개월($i = 1,2,3$) 동안의 월별 예측 결

과를 나타낸다. 즉, $i=1$은 7월, $i=2$는 8월, $i=3$은 9월을 의미하며, 각 월별 피해 규모를 독립적으로 예측한 값이다. y_i는 i번째 실제 피해건수, $\hat{y}_i$는 i번째 예측값, $\bar{y}$는 y의 평균 피해 건수를 나타낸다.

이러한 다면적 평가 지표를 통해 각 모델의 예측 성능을 종합적으로 분석할 수 있으며, 이를 바탕으로 보이스피싱 예측이라는 실제 문제에서 더 적합한 모델을 판단할 수 있다[20].

4.2 Comparison of prediction performance analysis results

주성분분석을 적용한 다중선형회귀 모델과 VAR 모델의 성능 비교 결과는 Fig. 5에 제시되어 있다.

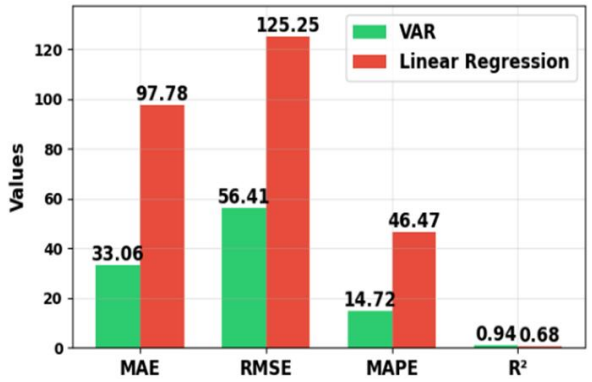

Fig. 5. Model Performance Metrics Comparison

Fig. 5의 결과에 따르면, VAR 모델은 MAPE 14.72% 및 결정 계수(R^2) 0.94를 기록하여 높은 예측 정확도를 보였다. 반면, 다중선형회귀 모델은 MAPE 46.47% 및 R^2 0.68로, 상대적으로 낮은 예측 성능을 나타냈다. VAR 모델이 높은 R^2 값을 기록한 것은 시계열 데이터에서 변수 간 상호작용을 효과적으로 반영했기 때문으로 해석할 수 있다. 특히, MAE는 VAR 모델이 33.06, 다중선형회귀 모델이 97.78로, VAR 모델의 오차가 약 3배 더 작았다. 또한, RMSE 값에서도 VAR 모델(56.41)이 다중선형회귀 모델(125.25)보다 약 2배 낮아, 전반적인 예측 성능이 더 우수한 것으로 확인되었다.

이는 시계열 데이터의 특성을 반영하는 VAR 모델의 장점이 효과적으로 작용했음을 시사한다. 특히, VAR 모델은 실제값과 예측값 간의 오차가 작고, 데이터의 변동성을 보다 정확하게 설명하는 것으로 나타나, 보이스피싱 피해 예측에 있어 다중선형회귀 모델보다 더 적합한 모델임을 시사한다.

4.3 Predicting voice phishing damage by age group and suggesting customized prevention policies

주성분분석 기반 VAR 모델이 높은 예측 정확도를 보였으므로, 이를 활용하여 2025년 2월의 연령별 보이스피싱 피해를 예측하고자 한다. 예측을 위해 2018년 1월부터 2024년 9월까지의 보이스피싱 피해 데이터를 사용하였다.

2018년부터 2024년도까지의 과거 데이터의 추세를 분석한 결과, 연령대별로 서로 다른 특징적인 패턴이 관찰되었다. 20대 이하 연령층은 2021년 이후 가파른 증가세를 보이며 2024년에는 562건에서 929건까지 큰 변동성을 나타냈다. 반면 30~50대는 2018년부터 2021년까지 비교적 안정적인 추세를 보이다가 2022년을 기점으로 급격한 감소세로 전환되어 낮은 수준을 유지하고 있다. 60대 이상 연령층도 유사한 패턴을 보이나 상대적으로 변동성이 작았다. 특히 2022년을 기점으로 20대 이하를 제외한 모든 연령대에서 뚜렷한 구조적 감소가 관찰되었다.

예측의 불확실성을 정량화하기 위해 표준편차에 기반한 95% 신뢰구간 추정 방법[21]을 사용하였다. 구체적으로, 예측값($\hat{y}$)과 표준편차(σ)를 계산한 후 ($\hat{y} \pm 1.96\sigma$) 구간을 95% 신뢰구간으로 설정하였다. 이는 예측값의 분포가 정규분포를 따른다는 가정 하에서, 실제 값이 95%의 확률로 이 구간 내에 존재함을 의미한다. 보이스피싱 피해건수는 음수가 될 수 없으므로 신뢰구간의 하한을 0으로 제한하였다. Fig. 6은 이러한 방법으로 2025년 2월의 연령대별 보이스피싱 피해건수를 예측한 결과이다. 신뢰구간은 음영 영역(shaded area)으로 표시하였으며, 이는 예측값의 95% 신뢰수준을 나타낸다.

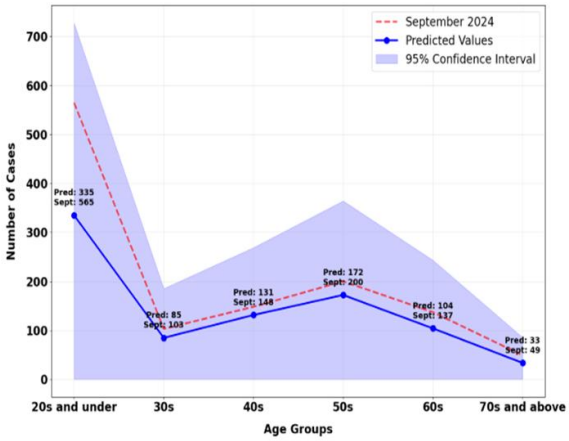

Fig. 6. Voice Phishing Damage Prediction by Age Group Comparison between 2024.09 and 2025.02

Fig. 6에 나타난 2024년 9월의 실제 데이터와 2025년 2월의 예측값을 비교한 결과, 모든 연령대에서 보이스피싱

피해가 감소하는 추세를 보였다. 20대 이하 연령층의 경우 실제값 565건에서 335건으로 가장 큰 폭의 감소가 예측되었으며, 다른 연령대에서도 소폭의 감소세가 나타났다. 구체적으로 30대는 실제값 103건에서 85건으로, 40대는 148건에서 131건으로, 50대는 200건에서 172건으로, 60대는 137건에서 104건으로, 70대 이상은 49건에서 33건으로 감소할 것으로 예측되었다. 이러한 전반적인 감소 추세는 최근 강화된 보이스피싱 예방 정책과 지속적인 홍보 효과가 반영된 것으로 해석된다. 다만, 모든 연령대에서 신뢰구간이 상당히 넓게 나타났다. 이는 보이스피싱 범죄의 특성상 사회적 요인과 범죄 수법의 변화에 따라 실제 피해 건수가 크게 변동될 수 있음을 의미한다. 예측 결과의 분석을 바탕으로, 연령대별 특성을 고려한 맞춤형 대응 정책을 Table 7과 같이 제안한다.

Table 7. Proposed Age-Specific Prevention Policies for Voice Phishing

Age group	Vulnerability Factors	Suggested Prevention Policies
20s and under	• Limited financial experience • High digital platform usage • Job-seeking related risks[22]	• Digital platform fraud detection • ob scam prevention programs • Financial education[23]
30s-50s	• Active financial activities • Multiple financial responsibilities • Investment-related risks[24]	• Workplace security training • Investment fraud detection • Enhanced transaction verification[25]
60s and above	• Limited digital literacy • Health/retirement concerns • Susceptibility to emotional manipulation [22, 24]	• Simplified security guidelines • Enhanced withdrawal verification • Senior-focused fraud detection[25]
All age groups	• Common vulnerabilities • Evolving fraud techniques	• Integrated response system • Enhanced prevention education • International cooperation

제안된 정책들은 예측된 피해 증가 추세를 선제적으로 차단하고, 각 연령대별 특성을 고려한 효과적인 대응을 통해 보이스피싱 피해를 억제할 수 있을 것으로 기대된다. 특히 신뢰구간이 가장 넓게 나타난 20대 이하와 50대 연령층에 대해서는 더욱 강화된 정책적 접근이 필요하다. 20대 이하의 경우 디지털 플랫폼을 통한 취업 사기나 대출 사기에 취약하므로 실시간 의심 메시지 탐지 시스템과 같은 기술적 대응이 시급하다. 50대의 경우 투자 사기와 같은 고액 피해에 노출될 위험이 크므로 금융기관의 대규모 거래 검증 절차 강화가 중요하다.

IV. Conclusions

이 연구에서는 경찰청의 보이스피싱 피해 데이터를 활용하여 2018년 1월부터 2024년 9월까지의 연령별 및 월별 데이터를 분석하였다. 다중선형회귀 모델과 VAR 모델의 예측 정확도를 비교하기 위해 주성분분석으로 전체 변동의 99.31%를 설명하는 세 가지 주요 패턴을 도출하여 데이터의 복잡성을 줄였다. 과적합 방지를 위해 이동 창 기법을 적용하였으며, AIC 검정을 통해 이동 창의 최적 크기를 12개월로 결정하였다. 이를 적용하여 계절적 변동성을 통제하기 위해 선정한 기간인 2024년 7월부터 9월까지의 보이스피싱 피해 예측 모델을 구축한 결과, VAR 모델이 다중선형회귀 모델보다 월등히 우수한 성능을 보였다. MAE와 RMSE 모두에서 VAR 모델이 현저히 낮은 오차를 보여, 연령대별 보이스피싱 피해의 시간적 특성과 연령대 간의 상호작용을 효과적으로 포착할 수 있음을 입증하였다. VAR 모델을 활용하여 2025년 2월의 보이스피싱 피해 건수를 예측하였으며, 표준편차에 기반한 95% 신뢰구간을 산출하였다.

분석 결과, 모든 연령대에서 피해 감소가 예측되었으며, 20대 이하 연령층에서 가장 큰 폭의 감소가 예측되었다. 50대, 40대, 60대, 30대, 70대 이상 순으로 감소폭이 나타났으며, 이는 최근 강화된 보이스피싱 예방 정책과 지속적인 홍보 효과가 반영된 것으로 해석된다. 20대 이하의 큰 폭 감소는 디지털 네이티브 세대의 보안 의식 향상을 시사한다. 다만, 20대 이하와 50대 연령층에서 특히 넓은 신뢰구간이 관찰되었는데, 이는 사회적 요인과 범죄 수법의 변화에 따른 피해 건수의 변동 가능성을 의미한다. 따라서 정부 차원의 통합 대응 시스템 구축과 실시간 메신저 기반 사기 탐지 시스템 도입 등 체계적인 예방 대책의 수립이 시급하다.

이 연구의 학술적 기여도는 다음과 같다. 첫째, 시계열 데이터 분석에서 VAR 모델과 다중선형회귀 모델의 성능을 실제 데이터로 비교 분석하여, 다양한 평가 지표를 통해 VAR 모델의 우수성을 실증적으로 검증하였다. 둘째, 주성분분석과 이동 창 기법을 결합한 체계적인 분석 프레임워크를 제시하여 시계열 데이터 분석의 방법론적 발전에 기여하였다. 셋째, 검증된 모델의 예측 결과를 바탕으로 연령대별 맞춤형 예방 정책을 제시함으로써 학술적 분

석과 정책적 활용을 효과적으로 연계하였다.

이 연구의 한계점은 다음과 같다. 연령대별 피해 건수의 시계열 분석에만 초점을 맞추어 경제상황 등 외부 환경요인을 고려하지 않았으며, 성별, 직업, 지역 등 다른 인구통계학적 특성을 포함하지 않았다. 또한 제한된 데이터로 인해 예측값의 신뢰구간이 상당히 넓게 나타나 예측의 정확도에 제약이 있었다.

향후 연구에서는 보이스피싱 수법 유형별 예측 모델을 개발하고, 경제지표 등 외부 요인을 포함하는 통합적 예측 모델을 구축하고자 한다. 다양한 인구통계학적 특성을 고려한 다차원 분석을 통해 더욱 정교한 예측이 가능할 것으로 기대된다. 특히 시계열 데이터의 특성을 고려한 다양한 가중치 부여 방법과 신뢰구간 추정 기법을 실험하여 예측의 불확실성을 줄이는 연구가 필요하다. 이를 통해 예측의 정확도를 향상시키고 보다 신뢰성 있는 결과를 도출할 수 있을 것으로 기대된다.

REFERENCES

[1] V. Mandalapu, L. Elluri, P. Vyas and N. Roy, "Crime Prediction Using Machine Learning and Deep Learning: A Systematic Review and Future Directions," in IEEE Access, Vol. 11, 2023. DOI: 10.1109/ACCESS.2023.3286344

[2] Lee, Seoungyong , Julak Lee, “A Study on the Prediction Method of Voice Phishing Damage Using Big Data and FDS,” Korean Security Management Association, No. 62, pp. 185-204, 2022. DOI: 10.36623/KSSR.2020.62.8

[3] Jae-Kyung Lee, Jin-Beom Seo, and Young-Bok Cho, “A Deep Learning-Based Smartphone Phishing Attacks Countermeasures,” Proceedings of the Korean Society of Computer Information Conference, Jeju, 2022. 7. 14.

[4] Sang-Min Kim, Seung-Min Rho, “Voice Phishing Detection Using Deep Learning-based NLP and Knowledge Distillation Techniques,” The Journal of Society for e-Business Studies, Vol. 29, No. 4, pp. 139-148, 2024. DOI: 10.7838/jsebs.2024.29.4.139

[5] Sang-Yong Choi, “The Trends and Prospects of Mobile Forensics Using Linear Regression,” Journal of the Korea Society of Computer and Information, Vol. 27, No. 10, pp. 115-121, 2022. DOI: 10.9708/jksci.2022.27.10.115

[6] Min-Seob Song, Junghye Min, “Comparison of Stock Price Prediction Using Time Series and Non-Time Series Data,” Journal of the Korea Society of Computer and Information, Vol. 28, No. 8, pp. 67-75, 2023. DOI: 10.9708/jksci.2023.28.08.067

[7] No, Wonjun & Lee, David & Noh, Byeongjoon & Kim, Youngchul, “How do crosswalk delays affect pedestrian access in zoning areas? Walking access reduction by signalized crosswalks in Seoul,” Applied Geography, Vol. 156, 2023. DOI: 10.1016/j.apgeog.2023.102975

[8] Marco Tumaini, Jiannong Cao, and Milos Stojmenovic, “Unsupervised pregnancy and physical activity detection in mammals using circadian rhythms,” In Proceedings of the 10th International Symposium on Information and Communication Technology, pp. 350–356, 2019. DOI: 10.1145/3368926.3369663

[9] Kumar Pradeep, "Predictive analytics for spam email classification using machine learning techniques," International Journal of Computer Applications in Technology, pp. 282-296, 2020. DOI: 10.1504/ijcat.2020.111844

[10] Jin, Zhixiong and Byeongjoon Noh, "From Prediction to Prevention: Leveraging Deep Learning in Traffic Accident Prediction Systems," Electronics, Vol. 12, No. 20, 2023. DOI: 10.3390/electronics12204335

[11] Jeon, Dongha, “A Light-weight Android Malware Classification Model through Feature Dimension Reduction Using PCA,” Master's thesis, Korea National Defense University, 2023. http://www.riss.kr/link?id=T16684529

[12] Kim, Hyun Seo, Choi, Won Seok, and Zhang, Byoung-Tak, “Predicting Health Indicators Using Vector Autoregression (VAR),” Proceedings of the Korean Information Science Society Conference, Gangwon, Korea, 2019. http://www.kiise.or.kr

[13] Kim, Changsoo, “A Study on the Characteristics of Influent Quality of Industrial Wastewater Treatment Plants using Statistical PCA(Principal Component Analysis),” Doctoral Dissertation, Korea National University of Transportation, 2021. http://www.riss.kr/link?id=T15902021

[14] Song, Geun-Won, Regression Analysis and ARIMA Time Series Analysis, Korean Studies Information, 2013. ISBN:9788926846438

[15] Kwon-Soon, Moon, “A understanding of Vector Autoregressive Model,” Journal of the Korean Official Statistics, Vol. 2, No. 1, pp. 23-57, 1997. DOI: 10.13000/jfmse.2016.28.1.198

[16] PARK, Jeasung, KIM, Byung Jong, KIM, Wonkyu, & JANG, Eunhyuk, “The Development of Econometric Model for Air Transportation Demand Based on Stationarity in Time-series,” Journal of Korean Society of Transportation, Vol. 34, No. 1, pp. 95-106, 2016. https://doi.org/10.7470/jkst.2016.34.1.095

[17] Shin, Woongjae, “Detecting Bubbles in Cryptocurrency Markets: Using the Generalized Supremum Augmented Dickey-Fuller (GSADF) Test,” Master's Thesis, Seoul National University, 2019. http://www.riss.kr/link?id=T15051671

[18] Okyoung Na, “Determining the existence of unit roots based on detrended data,” The Korean Journal of Applied Statistics, Vol. 34, No. 2, pp. 205-223, 2021. DOI: 10.5351/KJAS.2021.34.2.205

[19] Jae-Hyun Han, Chi-Hyun Jo, Eun-chong Koh, Do-Hyung Kim, & Soo-Wook Lee, “Proposal of a Stacked SARIMAX-GRU to

Improve Time Series Prediction Performance," Journal of the Korea Institute of Information and Communication Engineering, Vol. 28. No. 10, pp, 1131-1137, 2024. DOI: 10.6109/jkiice.2024.28.10.1131

[20] Safat, Wajiha & Asghar, Soahail & Gilani, Saira, "Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques," IEEE Access, 2021. DOI: 10.1109/ACCESS.2021.307811

[21] Lee, Dong Kyu, Junyong In, and Sangseok Lee. "Standard deviation and standard error of the mean." Korean journal of anesthesiology, Vol. 68, No. 3, pp. 220-223, 2015. DOI: 10.4097/kjae.2015.68.3.220

[22] Choi, J.H., "Voice Phishing Damage Increases Among MZ Generation: Need for Pan-governmental Measures," EDAILY, 2024. 7. 22. http://www.edaily.co.kr/

[23] Teh-Jen Sun, HyeonKi Jo, Yuri Seo, Seol Roh, HakHo Kim, Eui-Nam Huh, "Voice Cat : Design and Implementation of a Metaverse-based Voice Phishing Experience and Prevention Platform Service," Proceedings of the Korea Software Congress, Korean Institute of Information Scientists and Engineers, pp. 1267-1269, 2023.

[24] Moon, K.M., "Here's Your Card Delivery: Analysis of Voice Phishing Victims by Age and Gender," Maeil Business News Korea, 2025. 1. 21.

[25] Shin, S.W., "A Study on the Actual Situation and Countermeasures of Voice Phishing," Korean Journal of Public Safety and Criminal Justice, Vol. 19, No. 4, pp. 165-186, 2022. DOI: 10.25023/kapsa.19.4.202211.165

Authors

Jungwoo Bae is a student at Gyeonggi Science High School, Korea. His research interests include statistics, data analysis, artificial intelligence, and mathematical modeling.

Byunghong Choi received the B.S. degree in Mathematics Education from Kangwon National University, Korea, in 2015, and the M.S. degree in Mathematics Education from Seoul National University, Korea, in 2023. Mr. Choi joined the faculty of Gyeongi Science High School, Korea, in 2019, where he is currently a Mathematics Teacher. He is interested in mathematics education, statistical modeling, and AI-based assessment methods.