

Optimizing Stock Price Prediction via Macro context and Filter-DQN Framework

Jean-Ho Kim*, Eun-Hong Park*, Ha-Young Kim**

*Student, Graduate School of Information, Yonsei University, Seoul, Korea

**Associate Professor, Graduate School of Information, Yonsei University, Seoul, Korea

[Abstract]

The Korea Composite Stock Price Index exhibits high sensitivity to global market conditions, as net exports form a cornerstone of the economic structure in Korea. However, most existing studies on stock price prediction limited in two key aspects: First, they primarily focus on indicators within the domestic market, thereby omitted the influence of global market dynamics. Second, they lack of thorough consideration of feature selection methods, which hinders performance and efficiency of deep prediction models in data-hungry financial time series. In this paper, we propose extended stock data that integrates stock market indicators from major export destinations along with macroeconomic variables and Filter-DQN, a novel feature selection method that incorporates mutual information theory and reinforcement learning approach. The proposed method enhance both prediction performance and computational efficiency by effectively filtering out redundant variables. Experimental results demonstrate that the Filter-DQN method improved performance of baseline models by 8.28%, 45.28%, compared to full variables and naive mutual information based selection, respectively. Futhermore, the reduced dimensionality of variables leads to 2.6% reduction in computational cost.

▶ **Key words:** Feature selection, Filter methods, Reinforcement Learning, Stock price prediction, Deep Learning

[요 약]

한국 종합주가지수는 수출 중심의 경제 구조로 인하여 글로벌 시장 변동성에 민감하게 반응한다. 그러나 대부분의 기존 주가 예측 연구는 두 가지 한계점이 있었다. 첫째, 국제와 국내 지표를 통합적으로 충분히 활용하지 못하였으며, 둘째, 다량의 변수 중 중요한 정보를 효과적으로 선별하지 못해 데이터가 부족한 금융 분야에서 딥러닝 기반 모델의 예측 성능과 효율성에서 그 한계가 있었다. 본 연구는 이러한 한계점들을 개선하기 위해 주요 수출 대상국의 시장 동향과 국제적인 거시경제 변수를 포함한 확장된 데이터 셋과 상호 정보 이론과 강화 학습 알고리즘을 결합한 Filter-DQN 방법을 제안한다. 제안된 방법은 전체 변수 집합 중에 유의미한 변수들만을 선별함으로써 계산 비용을 절감하고 예측 성능을 극대화시키고자 한다. 실험 결과, 제안된 방식은 전체 변수를 사용하여 예측하였을 때 대비 8.28%, 기본적인 상호 정보량 기반 방식 대비 45.28%의 성능 향상을 달성했으며, 변수 차원의 축소로 계산 비용 또한 2.6% 절감하는 효과를 보였다.

▶ **주제어:** 변수 선택법, 필터 기법, 강화 학습, 주가 예측, 딥러닝

-
- First Author: Jean-Ho Kim, Eun-Hong Park, Corresponding Author: Ha-Young Kim
 - *Jean-Ho Kim (jeanhokim@yonsei.ac.kr), Graduate School of Information, Yonsei University
 - *Eun-Hong Park (pehhh03@yonsei.ac.kr), Graduate School of Information, Yonsei University
 - **Ha-Young Kim (hayoung.kim@yonsei.ac.kr), Graduate School of Information, Yonsei University
 - Received: 2024. 12. 05, Revised: 2024. 12. 30, Accepted: 2024. 12. 30.

I. Introduction

금융 시장은 주식 가격, 환율, 금리뿐 아니라 국내외 경제, 정치, 지정학적 요인 등 다양한 요소가 복합적으로 작용하여 시장의 움직임을 결정한다. 다변량 변수들이 유기적으로 얽혀있기 때문에 시장의 변동성을 반영한 시계열 예측은 투자전략 수립, 리스크 관리, 시장의 변동성 분석 그리고 잠재적 위협에 대비하기 위한 필수적인 도구로 자리 잡고 있다. 최근 국제 금융 시장의 상호 의존성이 심화되면서 특정 국가의 주식 시장 변동성이 다른 국가에 미치는 영향력이 점점 커지고 있다[1]. 특히 한국의 KOSPI 지수는 미국, 유럽, 일본 등 주요 국제 시장의 움직임에 민감하게 반응하며[2], 이러한 상호 연관성은 KOSPI 지수 예측의 정확도를 높이는 데 있어 최우선으로 고려해야 할 사항이다.

기존의 주가 예측 연구는 주로 기술적 지표를 활용하여 과거 주가 데이터를 기반으로 미래의 주가를 예측하는 데 중점을 두었다[3]. 그러나 이러한 접근 방식은 대개 KOSPI 지수의 자체적인 패턴 분석에 국한되거나, 한정된 국내 거시 경제 변수만을 활용한 시계열 분석에 그치는 경우가 많았다[4]. 기술적 주가 분석 지표들만을 이용하여 분류 회귀 나무모형, 서포트 벡터(Support Vector Machine, SVM) 모형, 딥러닝 모형을 활용한 KOSPI 지수의 등락을 예측하는 연구를 진행했지만, 국제 시장과의 복잡한 상호작용을 충분히 반영하지 못한 채 주식 시장 내재적인 변동성만을 설명하거나 상승과 하락의 분류로 접근한 것에 기인하여 예측 성능을 극대화하는 데 있어 한계가 존재했다[5]. 1997년 외환위기 이후 자본 시장의 개방으로 미국, 중국, 일본 등 국제 거시 경제 변수가 한국 주식 시장에 미치는 영향에 대한 연구가 활발히 이루어졌지만[6], 국제 시장의 급변하는 동향을 실시간으로 반영하는 주가 예측 모델은 여전히 미비한 실정이다.

최근 인공지능 기술의 발전과 함께 딥러닝 기반의 시계열 예측 모델은 뛰어난 성능을 보여주며 금융 분야에 접합되어 활발히 연구되어 왔다[7]. 그러나 데이터가 상대적으로 부족한(data-hungry) 금융 분야에서[8], 고차원적이고 이질적인 변수들이 포함될 경우 불필요하거나 상관성이 낮은 변수로 인하여 모델의 복잡성이 증가하고 예측 성능이 저하되는 문제가 있다[9]. 이로 인하여 주요 변수의 하위 집합만을 선택하여 모델의 입력 변수로 사용하는 딥러닝 기반의 변수 선택(Feature Selection) 기법은 모델의 과적합(overfitting) 현상을 방지하는 데 필수적이다.

본 논문은 국제 거시경제 지표와 주요 수출 대상국의 주식시장 지표를 포함해 다양한 변수로 구성되어 있는 데이터 셋에서 유의미한 변수를 선별하여 KOSPI 지수를 효율적으로 예측하는 모델을 제안한다. 주어진 다변량 금융 시계열 데이터에서 상호 정보량(Mutual Information, MI)[10]을 기반으로 가장 유의미한 변수의 하위 집합을 강화 학습(Reinforcement Learning) 방식으로 찾아내는 Filter-DQN을 제안한다.

본 논문에서는 다변량 금융 시계열 데이터에서 변수 선택과 예측 모델링을 결합하여 KOSPI 지수를 효과적으로 예측하는 새로운 접근법을 제안한다. 제안된 방법론은 두 단계로 구성된다. 첫 번째 단계에서는 변수 간의 상호 정보를 DQN(Deep Q-Network)[11]의 에이전트(Agent)에게 보상(Reward)으로써 부여하며 학습시키는 Filter-DQN 기법을 통해 다변량 데이터에서 유의미한 변수의 하위 집합을 선별한다. 두 번째 단계에서는 선별된 변수들을 입력으로 활용하여 Transformer 기반 시계열 예측 모델[12-14]을 학습시키고, 이를 통해 KOSPI 종가 지수의 변동성을 정교하게 예측한다. 이러한 통합적 접근은 변수 선택을 통해 모델의 복잡성을 줄이는 동시에, Transformer 모델의 강력한 학습 능력을 활용하여 예측 성능을 극대화할 수 있다.

본 연구의 차별점은 다음과 같다.

- 국제 금융 시장의 변동성을 효과적으로 반영하고 확장된 변수 차원으로 인한 과적합 문제를 최소화하기 위하여 상호 정보 이론과 DQN을 결합한 새로운 Filter-DQN 방식을 제안한다.

- 연구의 타당성을 검증하기 위해 약 40개의 내재적, 기술적, 국제적 지표를 포함한 데이터 셋을 기반으로 Transformer 기반 시계열 예측 모델을 활용하여 실험을 진행하였다. 실험 결과, 전체 변수를 입력한 모델 대비 평균적으로 예측 정확도와 안정성 측면에서 MSE(Mean Square Error) 18.6%, MAE(Mean Absolute Error)는 11.5% 성능이 향상됨을 보였으며 각 모델의 훈련 속도 역시 평균 2.6% 단축되었다.

본 논문의 구성은 다음과 같다. 2장에서 관련 연구 동향을 분석하고, 3장에서는 Filter-DQN 프레임워크 기반의 변수 선택 방법론을 설명한다. 4장에서는 사용된 데이터 셋과 실험 결과를 제시하며, 마지막으로 5장에서 결론과 향후 연구 방향에 대하여 논의한다.

II. Preliminaries

1. Related works

1.1 Stock Prediction

금융 시장 예측은 과거 주가 데이터와 다양한 경제 지표를 활용한 시계열 분석에 의존해 왔다. 기존 연구는 상대강도지수(Relative Strength Index, RSI), 이동평균 수렴발산(Moving Average Convergence Divergence, MACD), 이동평균(Moving Averaging, MA)과 같은 기술적 지표를 활용하여 시장의 방향성을 예측하고 투자 전략을 수립하는 데 중점을 두었다[15]. 이러한 전통적 예측 모델들은 단기적인 시장 패턴 분석에는 효과적이지만, 장기적인 추세 변화나 국제 경제 변수와의 상호작용을 반영하는 데는 한계가 있었다. 과거 데이터의 규칙성에 의존하기 때문에 새로운 경제 상황이나 국제 경제 변수나 지정학적 요인과 같은 외생적 충격을 고려하지 못해 국제 금융 시장과의 상호작용을 고려한 예측에서는 정확도가 떨어지는 한계를 보였다.

이러한 한계를 극복하기 위하여 최근 연구들은 미국, 중국, 한국 등의 주요 국가의 국제 거시경제 지표가 KOSPI 지수의 변동성에 미치는 영향을 분석하며 국제 시장과의 상호작용을 실증적으로 확인하고자 했다[16]. 다요인 모형(Multi Factor Model)은 각국의 금리, 생산지수, 물가지수와 같은 주요 경제 지표가 한국 주식 시장에 미치는 영향을 분석하는데 활용되었지만[17], 대부분 연구가 정적 변수 선택 기법에 의존하고 있어 실시간으로 움직이는 시장 환경에 유연하게 대응하기는 어려웠다.

이를 해결하기 위하여 최근에는 LSTM(Long Short Term Memory)과 Transformer 기반 딥러닝 모델을 활용하여 비선형적인 변수 간 관계를 학습하고 장기 의존성을 효과적으로 처리하는 접근 방식이 제안되었다. 순환 신경망(Recurrent Neural Network, RNN) 기반 모델은 시계열 데이터의 순차적인 의존성을 단기적으로 학습할 수 있었으나, 장기적인 의존성을 포착하는 데는 한계가 있었고[18], LSTM과 Transformer 기반의 모델이 문제를 해결하며 금융 시계열 데이터의 예측 성능을 크게 개선하였다[19,20].

그러나 이러한 딥러닝 기반 접근법도 고차원 데이터에서 변수 선택이 이루어지지 않을 경우 과적합 위험이 존재하며, 모델의 복잡성 증가로 인하여 학습 속도가 저하될 가능성이 있다. 특히, 금융 데이터는 변수 간의 복잡한 상호작용과 시계열 데이터의 길이가 충분치 못하다는 데이터 부족[8] 문제를 가지므로, 변수 선택을 적절히 수행하지

못할 경우 예측 성능이 제한될 수 있다. 국제 경제 지표를 포함하여 변수의 차원을 효과적으로 압축하거나 선별하여 모델에 입력하는 것이 주가 예측의 정확도를 향상시키는 핵심이다.

1.2 Feature Selection

다변량 정보의 손실을 최소화하면서 변수 차원을 축소하기 위하여 기존 연구에서는 필터(filter) 기법, 주성분 분석(Principal Component Analysis, PCA), 오토인코더(Autoencoder) 등을 활용하여 변수 선택 문제를 해결하고자 했다[21-23]. 오토인코더 기반 접근법은 입력 데이터를 인코더(Encoder)를 통하여 저차원 잠재 공간으로 압축하고, 디코더(Decoder)를 통하여 데이터를 다시 재구성(reconstruct)하여 주요 정보를 보존하면서도 불필요한 잡음(noise)을 제거하는 방법이다. 주성분 분석 기법은 데이터의 분산을 최대화하는 주성분을 찾아 데이터 간 중복성을 최소화하고, 데이터의 구조를 가장 잘 반영하는 기저(bases)를 추출하는 차원 축소 기법이다. 위 두 방식 모두 다양한 시계열 연구에서 활용되어 성능이 입증되었으나[24], 각각 선형 결합(Linear Combination)과 비선형 신경망에 기반하기 때문에 결과값의 해석 가능성이 제한적이라는 한계를 가진다.

필터 기법은 예측 모델과는 독립적으로 계산된 통계량을 기반으로 상관성이 높은 변수를 선별하는 방식으로, 계산이 빠르고 다양한 모델에 일반화가 용이하다는 장점이 있다[21]. 하지만 훈련 데이터 셋에 대하여 일괄적으로 계산되므로 시간의 흐름에 따른 변동성을 반영하지 못한다는 단점이 있다. 상호 정보량은 두 변수 간 상호 의존성을 측정하는 지표로, 변수 선택을 최적화하고 예측 성능을 향상시키는 데 활용되어 왔다[10]. 관련성과 중복성을 동시에 고려한 변수 선택 방법론으로, 예측 성능을 개선에 효과적이었으나 고차원 데이터에서는 변수 간 모든 조합을 계산하기에 계산 비용이 급격히 증가하는 문제가 있었다. Peng et al. (2005)은 상호 정보량을 기반으로 다차원 데이터에서 변수들 간의 중복 정보를 줄여 예측 모델의 효율성을 높이는 방법을 제안하였다[25]. Pablo et al. (2009)의 연구 결과에서는 정규화된 상호 정보량을 활용하여 노이즈에 민감한 기존 상호 정보량의 단점을 극복하고 예측 정확도를 향상시키는 데 성공하였다[26].

1.3 Deep Q-Network

DQN은 행동 가치학습(Q-Learning) 알고리즘을 심층 신경망(Deep Neural Network)으로 확장한 강화 학습 모

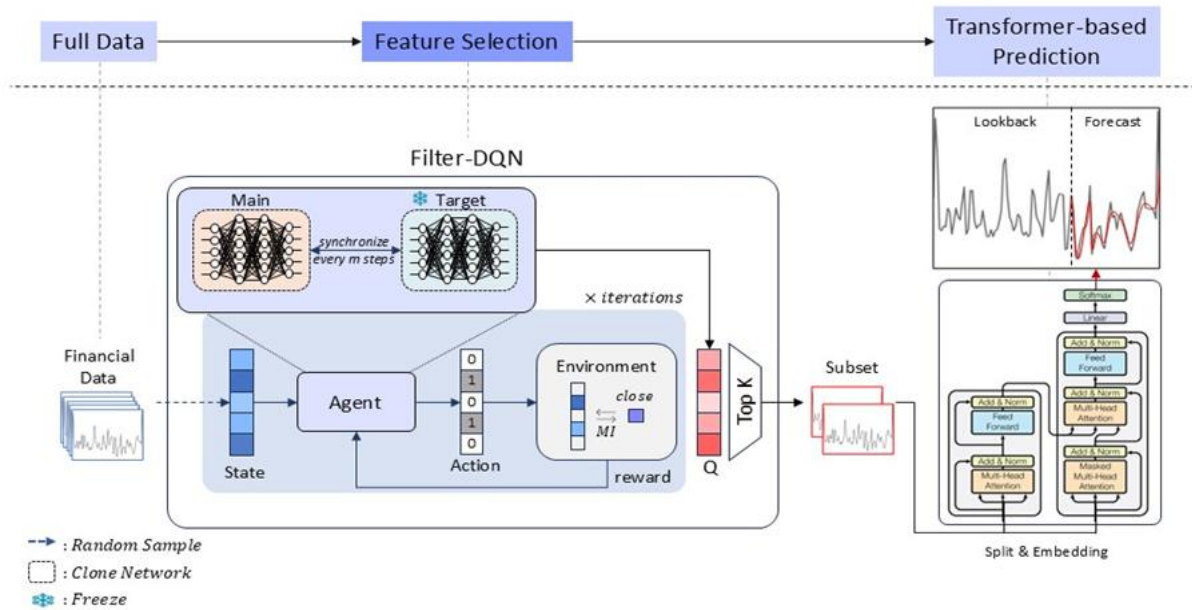


Fig. 1. Architecture of proposed Filter-DQN with transformer-based predictor

델로, 에이전트가 주어진 환경(Environment)에서 행동(Action)과 그에 따른 상태 천이(State Transition), 보상을 반복 학습하여 누적 보상을 극대화하는 최적 정책(Policy)을 산출하도록 설계되었다[11]. DQN은 MDP(Markov Decision Process)로 정의된 이산 상태 공간에서 순차적 의사결정 문제를 모델링하고, 고차원 상태 공간에서의 효율적인 학습을 가능하게 함으로써 강화 학습 문제를 효과적으로 해결하였다. 행동 가치학습은 행동 가치(Q-value) 함수를 학습하여 주어진 상태에서 행동을 선택했을 때 기대되는 미래 보상을 예측한다. 행동 가치학습은 벨만 방정식(Bellman Equation)을 반복적으로 업데이트해 최적의 행동 가치를 학습하는 알고리즘으로 전통적인 방법은 탐색 공간이 크고 상태와 행동의 조합이 많은 복잡한 문제에서는 학습이 어렵다. 이를 해결하기 위해 DQN은 경험 재플레이(Experience Replay)와 타깃 네트워크(Target Network)를 설정하는 기법을 활용했다.

DQN 방식의 최적화는 이산적인 차원의 그리드 탐색(Grid Search)에 해당하는 Atari 게임에서 인간 수준의 수행 성능을 보여주었으며 금융 시계열 예측에도 적용 가능하다. 또한 Xu et al. (2021)은 DQN을 사용해 자동화된 변수 선택 알고리즘을 제안했다[27]. 이 연구는 변수 선택 문제를 MDP로 모델링하여 에이전트가 특성 공간과 상호 작용하면서 최적의 특성을 선택하도록 설계하였다. 경험 재플레이와 타깃 네트워크를 활용하여 학습의 안정성과 예측 성능을 개선할 수 있었다. 이러한 연구는 DQN이 변수 간의 상호 의존성을 학습하고 시간에 따라 변화하는 변수의 중요도를 반영할 수 있는 가능성을 시사했다.

본 연구는 기존 필터 기법에서 활용되었던 상호 정보 이론과 DQN을 결합한 새로운 프레임워크를 제안하여 금융 데이터의 복잡한 변수 간 상관관계를 포착하여 변수를 선별함으로써 예측의 정확도와 효율성을 향상시키고자 한다.

III. The Proposed Scheme

1. Filter-DQN Framework

딥러닝 모델에 국제적 거시경제 지표를 포함한 다량의 변수 정보를 입력할 시 발생하는 과적합 현상[28]과 차원의 저주[29]를 완화하기 위해 본 연구에서는 Filter-DQN 프레임워크를 제안한다. 제안된 프레임워크는 Fig. 1[30]에 제시한 바와 같이 훈련 데이터의 무작위 시점 시계열 정보를 상태로써 입력받아, 에이전트는 어떤 변수를 선택할지 행동을 결정한다. 환경에서 행동에 의해 선택된 변수와 목적 변수와의 상호 정보량을 기반으로 보상을 부여하며, 이는 반복되며 목적 변수와 높은 상호 정보를 가지게 되는 변수들에 대한 행동 가치를 에이전트가 학습한다. 학습이 완료된 에이전트의 행동 가치를 기반으로 상위 k 개의 변수를 필터링하고, 이를 통해 생성된 하위 변수 집합을 출력한다. 최종적으로 선별된 하위 변수 집합은 Transformer 기반 예측 모델[12-14]에 입력되어 시간 흐름에 따른 변동성과 변수 간의 상관관계를 학습함으로써 예측값을 도출한다. Transformer 기반 예측 모델은 멀티헤드 어텐션 매커니즘(Multi-Head Attention Mechanism)을 통해 경쟁력 있는 성능을 보여주었으나,

파라미터 수 증가로 인해 과적합 위험이 존재한다. 이를 해결하기 위해 제안된 방법은 변수의 차원을 축소하여 모델의 학습 안정성과 일반화 성능을 개선하는 데 중점을 두었다. 또한, 이 방법은 변수를 새로운 차원으로 추출하거나 변형을 가하는 방식이 아닌 필터 방식으로써 해석의 용이성이라는 강점을 갖는다.

우선 강화학습 요소에 대하여 다음과 같이 정의한다. 상태 $s_t \in R^F$, 여기서 F 는 변수의 총 개수를 의미한다. 행동 $a_t \in R^F$, 어떤 특성을 선택할지에 대한 에이전트의 행동을 의미하며, 보상 r_t 는 행동을 통해 고른 각각의 변수들과 목적 변수의 상호 정보량을 계산하여 각 변수를 선택한 행동에 대해 보상을 부여한다. 후행하는 예측 모델의 존재에 기인하여 테스트 기간의 정보 누설을 방지하고자, 전체 데이터의 길이에서 훈련과 테스트 기간으로 나누어 훈련 데이터들에 대해서만 Filter-DQN에 입력한다.

$$s_t, a_t \in R^F$$

$$t \sim \text{unif}(0, L-1)$$

$\text{unif}(\cdot)$ 은 주어진 범위 내에서 균등한 확률 분포를 가진다는 것을 의미하며, 입력 데이터는 $x^{F \times L}$ 로 특성의 개수와 길이를 가진 시계열 정보로 표현된다. 이때 L 은 훈련 데이터의 길이이다. 시간 축 L 에 대해서 랜덤 샘플링을 진행하여 무작위 t 시점에 대한 s_t 와 a_t 를 추출한다.

$$a_t \in \{0, 1\}$$

$$a_t \sim \text{Bernoulli}(0.5)^F, p \in \epsilon$$

$$a_t = \text{argmax}_a Q(s_t, a_t), p \in (1 - \epsilon)$$

$\text{Bernoulli}(\cdot)$ 는 베르누이 분포를 의미하며, 전체 변수 개수(F)에 대해 0.5의 확률로 1를, 나머지 확률로 0을 갖는 분포를 의미한다. Q 는 Q-네트워크를 의미한다. 에이전트는 상태 정보를 입력받아 Q-네트워크를 통해 행동 가치를 계산하며 Q-네트워크는 정책 신경망을 의미한다. 탐색 확률 ϵ 에 대해서는 무작위 행동을, 탐욕적 정책 확률 $(1 - \epsilon)$ 에 대해서는 행동 가치를 기반으로 최대의 보상을 얻는 행동 a_t 를 출력한다. 이때의 a_t 는 이진 벡터로, 선택한 변수들의 자릿수에 대해 1, 선택되지 않은 변수의 자릿수들에 대해 0으로 나타낸다.

행동으로 선택된 변수 부분집합(F')로 구성된 데이터 $x^{F' \times L}$ 와 $x^{F_{\text{target}} \times L}$ 과의 상호 정보량 값을 보상으로써 부여한다. 여기서 $D_{KL}(\cdot)$ 는 Kullback-Leibler Divergence를 의미한다.

$$r_t = MI(x^{F' \times L}, x^{F_{\text{target}} \times L})$$

$$MI = D_{KL}(P_{(x,y)} \| P_x \otimes P_y)$$

정책의 학습 과정은 DQN[24]의 방식과 유사하게 진행된다. Filter-DQN은 보상을 최대화하는 Q-네트워크의 정책을 학습하는 것을 목표로 한다. 이 과정에서 안정적인 학습을 위해 메인 네트워크(Q_{main})과 이를 복제한 타겟 네트워크(Q_{target})로 이루어지며, 타겟 네트워크는 일정 스텝마다 메인 네트워크와 동기화 되도록 설계한다. 최종적으로 Filter-DQN은 메인 네트워크와 타겟 네트워크의 MSE 손실 함수를 최소화하며 최적 정책을 학습한다.

$$y_{\text{target}} = r_t + \gamma \max_{a_{t+1}} Q_{\text{target}}(s_{t+1}, a_{t+1})$$

$$L(\theta) = E[(y_{\text{target}} - Q_{\text{main}}(s_t, a_t))^2]$$

γ 는 할인율로 미래 보상에 대한 가중치이며, θ 는 정책 파라미터를 의미한다. 학습된 Q-네트워크를 통해 나온 각각의 변수들에 대한 행동 가치에서 상위 k 개의 변수를 선택하여 출력한다.

2. Action Grouping

변수들의 개수 F 가 증가함에 따라 탐색 공간이 기하급수적으로 확장되는 문제를 해결하고자 본 연구에서는 변수들을 성격이 유사한 집단으로 분할하고 탐색 시 각 집단 내에서 무작위로 m 개 이상의 변수를 선택하도록 학습시켰다. 이를 통해 탐색 공간을 효율적으로 축소하면서도 다양한 변수 조합을 학습할 수 있는 효과를 얻었다. Table.1에 제시된 바와 같이 데이터를 세 가지 주요 집단으로 분류하였으며, 이러한 변수 군집화 방식은 탐색의 효율성을 높이고 Filter-DQN 방식이 국제 시장 변동성에 유연하게 대응할 수 있도록 미세 조정하는 효과가 있다.

IV. Empirical Results

1. Data Description

Table 1. Features of the Dataset

Groups	Features
Intrinsic	Open, High, Low, Close, Volume
Technical Indicators	MA (10,20), MACD (Signal, History), SAR, SAREXT, Bullinger Band (Upper, Middle, Lower), ATR, STD, RSI, CCI, ROC, Williams %R, ADOSC, OBV, Stochastic MOM (Slow %K, Slow %D)
Global Context	S&P 500, NASDAQ, DJIA, RUSSELL, VIX, PHLX Semiconductor Sector Index, Dollar, Gold, WTI, Nikkei 225, HSI, SSEC, STOXX, Treasury Bill(10Y-2Y)

한국의 주식 시장은 주요 수출국의 경제 상황, 외국인 자금 유입, 환율 변동과 같은 국제적인 요인에 민감하게 반응한다. 기존 주가 예측 연구에서는 주로 가격 정보 (Open, High, Low, Close, Volume)나 기술적 지표와 같이 특정 국가에 국한되는 지표들을 이용하여 주식 예측 모델을 설계했다는 점에서 한정적이었다[3,4]. 본 연구에서는 기존의 한정된 데이터 범위를 넘어 유의미한 국제적 시장 맥락을 반영하기 위하여 Table 1.의 Global Context에 해당하는 변수 데이터를 추가 구성하였다. 이를 통해 예측 모델이 국제 경제의 변화와 KOSPI 가격 간의 관계를 긴밀히 포착할 수 있도록 설계했다.

미국 증시 데이터로는 미국 경제 전반을 대표하는 S&P(Standard & Poor), 하이테크 중심의 NASDAQ, 안정적인 대형주들로 구성된 DJIA, 중소기업들로 이루어진 Russell, 시장의 불확실성을 나타내는 VIX, 반도체 업계를 대표하는 SOX, 원-달러 환율인 Dollar, 미국채의 장단기 스프레드를 나타내는 지표로 구성하였다.

또한 일본을 대표하는 주식으로 Nikkei 225, 홍콩의 금융 및 부동산 비중이 높으며 중국 본토와도 밀접한 연관을 가진 HSI, 상하이종합지수인 SSE, 유럽 17개국의 주요 기업들로 구성된 STOXX로 구성하였다.

추가적으로 경제 안정성과 성장 상황을 파악하는 데 도움이 되는 금(Gold)과 원유 가격(West Texas Intermediate, WTI) 데이터들도 포함하였다.

기술적 지표로는 다음과 같은 요소들을 포함하였다: 10일과 20일의 이동평균선(MA), 이동평균 수렴 • 발산 지수(MACD)와 신호선(signal)과 히스토그램(histogram), 반전 시점을 예측하는 Parabolic Stop and Reverse(SAR)와 Extended SAR(SAREXT), 볼린저 밴드(Bollinger Band)의 상단, 중단, 하단 밴드, 평균 실제 변동폭(Average True Range, ATR), 표준편차(standard deviation, STD), 상대 강도 지수(Relative Strength Index, RSI), 상품 채널 지수(Commodity Channel Index, CCI), 변화율(Rate of Change, ROC), 윌리엄스(Williams) %R, 누적분배 오실레이터(ADOSC), 거래량 분석 지표(On-Balance Volume, OBV), 확률적 모멘텀(Stochastic Momentum)의 Slow %K와 Slow %D.

전체 데이터 셋의 변수 구성은 Table.1에 제시한 바와 같으며, 데이터를 온전히 수집할 수 있는 최대 기간과 변수들을 고려하여 전체 기간은 1996년 1월 4일부터 2024년 8월 30일까지의 40개의 변수로 이뤄진 데이터를 구성하였다. 가격 지표에 해당하는 변수들은 인베스팅닷컴(Investing.com)에서, 기술적 지표는 가격 정보를 기반으

로 산출하였다. 또한 국제 시장 정세를 나타내는 거시 경제 변수들은 FRED(Federal Reserve Economic Data)를 사용하였다. 전체 데이터는 Filter-DQN과 예측 모델에서 모두 훈련(80%), 검증(10%), 테스트(10%) 기간으로 분리하여 실험하였으며, 훈련 기간은 1996년 1월 4일부터 2018년 4월 24일까지 5651일의 데이터를, 검증 기간은 2018년 4월 24일부터 2021년 1월 6일까지 663일의 데이터를, 테스트 기간은 2021년 1월 6일부터 2024년 8월 30일까지 663일의 데이터를 사용하였다.

2. Experimental Setting

Table 2. Hyper-parameter Setting

Model	Hyper-parameter	Value
Filter-DQN	ϵ	0.9
	ϵ decay	0.99
	gamma	0.99
	hidden_layer	128
	epoch	1000
	k	20
Prediction Model	lookback length	60
	predict length	60
	Model dimension	256
	Feed Forward layer size	128
	number of heads	4
	number of workers	3
	dropout ratio	0.2
	patience	3

예측 모델에서는 60일의 관측치로 60일 후의 예측값을 계산하였고, 다변량 시계열 정보를 입력받아 시장의 흐름을 분석하기에 용이한 종가를 최종 예측한다. 하이퍼 파라미터(hyper-parameter)는 Table.2에 제시한 바와 같으며, 과적합 위험을 최소화하기 위하여 모델의 차원과 순방향 신경망(Feed Forward Network)의 크기를 간소화하였으며, 드롭아웃(Dropout)과 조기 종료 역치(Early Stopping Patience)를 설정하였다. 공정한 성능 평가를 위해 예측 모델에 모두 동일하게 적용하였다. 모든 실험은 동일한 하드웨어에서 진행하였고, 사양은 다음과 같다. CPU는 AMD Ryzen 7 3700X, GPU는 Nvidia RTX A4000, RAM은 32GB이다.

평가지표로 MSE, MAE, MAPE(Mean Absolute Percentage Error)를 사용하여 모델의 예측 성능을 확인하였다. MSE는 예측값인 \hat{Y} 와 실제 값인 Y 의 차인 오차에 대한 제곱 평균이며, MAE는 오차의 절댓값에 대한 평균, MAPE는 오차를 실제 값으로 나눈 값들의 평균이다. 각 평가 지표에 대한 수식은 다음과 같다.

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - Y_i)^2$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{Y}_i - Y_i|$$

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{\hat{Y}_i - Y_i}{Y_i} \right|$$

즉 MSE는 제곱 연산으로 인하여 오차의 범위가 클 시 이를 더욱 반영되는 효과를 가지며, MAE는 오차에 크기 그 자체를 반영하는 효과를, MAPE는 오차의 크기를 원본 수치의 백분율로 나타낸 것을 의미한다.

3. Experimental Results

Table 3. Accuracy table

Method		MSE	MAE	MAPE
Autoformer	w/ Full	<u>0.063</u>	0.193	0.102
	w/ MI	0.067	0.201	0.106
	w/ Ours	0.062	<u>0.195</u>	<u>0.103</u>
Crossformer	w/ Full	0.048	0.173	0.087
	w/ MI	<u>0.042</u>	<u>0.156</u>	<u>0.085</u>
	w/ Ours	0.041	0.159	0.081
Nonstationary	w/ Full	<u>0.068</u>	<u>0.202</u>	<u>0.111</u>
	w/ MI	0.780	0.207	0.115
	w/ Ours	0.053	0.173	0.093

본 연구의 예측 모델로는 시계열 예측 부분에서 우수한 성능을 보인 Nonstationary-Transformer, Autoformer, Crossformer를 선정하였다[12-14]. Nonstationary-Transformer는 비정상적인 데이터의 복잡한 패턴을 잘 학습할 수 있도록 설계되어 있는 모델로, 금융 데이터와 같이 변동성이 큰 시계열 데이터를 예측하는 데 있어 강점을 보이는 모델이다[12]. Autoformer는 장기적 추세와 계절성을 효과적으로 학습하도록 설계된 모델로, 인코더-디코더 구조와 Auto-correlation 메커니즘으로 계산의 효율성을 극대화하였다[13]. Crossformer는 다차원 시계열 데이터 간의 상호관계를 학습하는데 최적화된 모델이다[14].

실험의 결과는 Table. 3과 같으며, Filter-DQN의 방법론의 효용성을 보여준다. Autoformer를 베이스라인으로 했을 시에는 가장 낮은 MSE 수치를 보였으며, Table.3에 제시한 바와 같이, 계산 효율성 측면에서 더 효율적일 수 있음을 제시한다. Crossformer와 Nonstationary-Transformer와 각각 접합했을 때, 성능 향상이 가장 컸다. 전체 데이터를 예측 모델의 입력으로 넣은 것 대비,

MSE에서 각각 약 14.57%, 22.8% 성능 향상이 있었고, MAE에서는 각 8.4%, 14.7% 성능 향상이 있었다.

Naive MI는 훈련 데이터와 테스트를 분리한 상태에서 일괄적으로 훈련 데이터 셋에서 각 변수들과 타겟 변수와의 MI를 계산한 뒤, 높은 MI 값을 가지는 상위 k 개 만큼을 고른 필터 방법이다. Filter-DQN 방식은 naive MI 필터 방법 대비 모든 모델의 평가 지표에서 우세한 성능을 보였다. Naive MI는 훈련 데이터와 테스트를 분리한 상태에서 일괄적으로 훈련 데이터 셋에서 각 변수들과 타겟 변수와의 MI를 계산한 뒤, 높은 MI 값을 가지는 상위 k 개 만큼 고른 필터 방법이다. Filter-DQN은 MI 값을 기반으로 보상을 주는 시스템이지만, 기존의 MI 방법 대비 우수한 성능을 보여 Filter-DQN 프레임워크의 효과를 입증하였다.

4. Additional Analysis

Table 4. Training speed comparison (sec/iter)

Model	Full	Selected (Ours)	change(%)
Autoformer	0.5729	0.5644	1.5(%)
Crossformer	0.5858	0.5570	4.9(%)
Nonstationary	0.5400	0.5332	1.3(%)

또한 Table 4.에 제시된 바와 같이, 각각의 베이스라인에 적용하였을 때 훈련 속도를 비교하였다. Autoformer를 예측 모델로써 사용하였을 때는 전체 데이터 셋을 입력한 것 대비 1.5% 정도 훈련 속도의 향상이 있었다. Crossformer를 예측 모델로 사용하였을 때 훈련 속도의 차이는 4.9%, Nonstationary-Transformer에서는 1.3% 차이를 보였다. 훈련에서는 모델의 전체 학습 과정이 포함되며 실제 Nonstationary-Transformer와 같이 Transformer 기반의 예측 모델에서 Attention을 통해 상관도를 계산할 시 계산 복잡도는 $O(T^2 \times D)$ 인 것을 고려하면 D (Dimension)에 해당하는 변수의 개수가 절반으로 준 것은 내부적으로 계산의 복잡도를 향상시켰을 것으로 사료된다. Autoformer의 경우 계산의 복잡도는 $O(T \log T \times D)$ 이며, Crossformer의 복잡도는 $O(\frac{1}{T^2} \times T^2 D)$ 이다. 두 모델의 경우에도 위 서술한 바와 마찬가지로이다. 공정한 성능 평가를 위해 MI 방식과 Filter-DQN을 사용하여 선별한 변수의 개수를 동일하게 맞추어 진행하였으며, 이때 두 방식은 속도 측면에서 동일한 수치를 보인다.

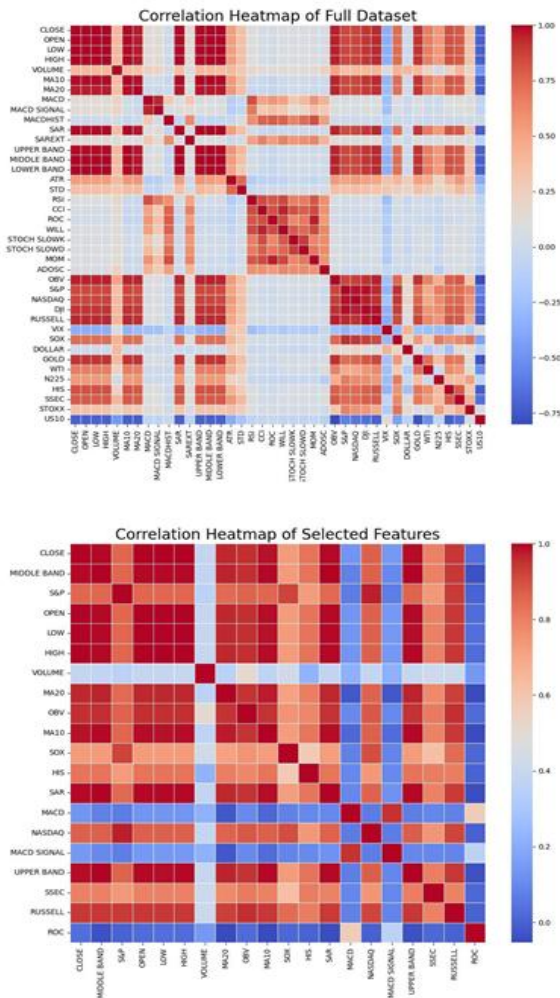


Fig. 2. Spearman's rank correlation heatmap of full data(top) and features selected with Filter-DQN(bottom)

전체 데이터와 개별 특징 선택법을 적용한 데이터 셋의 각 특징별 스피어만 상관계수를 측정한 결과 Fig. 2와 같았다. 스피어만 상관계수는 각 변수의 값을 순위로 변환한 후, 두 변수의 순위 차이로 계산되는 상관계수이며, 비선형적인 단조(monotonic) 관계를 포착할 수 있다는 특징이 있다. 0에 가까워질수록 두 변수 간의 상관관계가 미비하다는 것을 의미한다.

Filter-DQN을 거치지 않은 전체 데이터 셋은 Fig. 2의 위와 같이 상관관계가 뚜렷하지 않은 변수들이 포함되어 있음을 확인할 수 있다. 그러나 제안 방법을 통해 선택된 변수들은 높은 양 또는 음의 상관관계를 가지는 변수들로 선별되었으며, 이는 Fig. 2의 아래 그림에서 확인할 수 있었다. 또한, Fig. 3의 유색 점선은 전체 데이터 셋을 학습하였을 시, 유색 선은 Filter-DQN으로 선별된 변수들로 학습되었을 시, 검정 선은 실제 값을 의미한다. 시각화를 통하여 관찰한 결과, Table. 3에서 확인된 바와 같이, 전체적으로 실제값(Ground Truth)를 세 모델에서 점선 대

비 실선이 추세나 패턴을 더욱 잘 반영하고 있는 모습을 보이며, 특히 Nonstationary-Transformer에서 제안된 방법의 효과가 뚜렷한 것을 관측할 수 있었다.

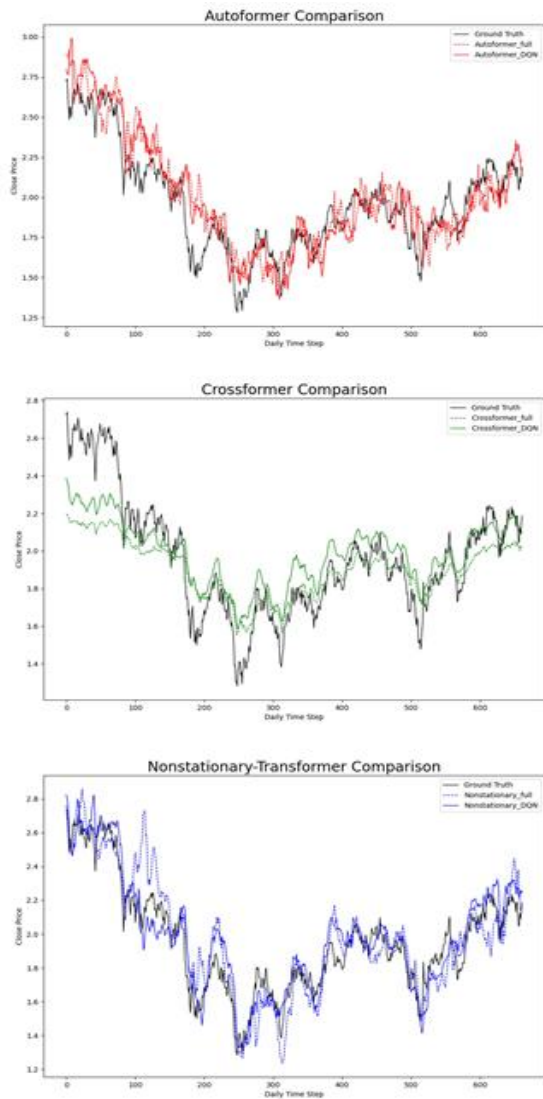


Fig. 3. Visualization of time series forecasting for baseline models compared with and without proposed method.

다양한 국제 거시 경제 지표 중에 Filter-DQN으로 선별된 S&P, NASDAQ, SOX, Russell, HSI, SSEC은 KOSPI의 주요 산업 및 미국, 중국 등 주요 교역국과의 경제적 의존 관계를 반영하고 있다. 특히 반도체 기업의 주가를 추정하는 지수인 SOX는 삼성전자, SK하이닉스와 같은 한국의 주요 대기업과의 산업적 연관성이 매우 높다. SOX 지수의 강세는 국제 반도체 수요 증가를 의미하며, 이는 KOSPI에서 높은 비중을 차지하는 반도체 관련 종목들의 주가 상승으로 이어질 가능성을 시사한다. 또한 S&P, NASDAQ, SSEC, HSI는 각각 미국과 중국을 대표하는 지수로, 2023년 기준 한국의 두 국가에 대한 수출 두

국가에 대한 한국의 수출 비율은 38%에 달한다. 이러한 변수들을 한국 경제의 국제 시장 의존성과 산업 구조를 반영하고 KOSPI 지수의 변동성을 설명하는 데 있어 주요한 지표로 작용할 수 있다.

또한, Granger 인과관계 분석을 통해 선별된 변수들에 대한 신뢰성을 추가 실험하였으며, 그 결과, S&P의 P-value는 0.000018로 KOSPI 지수와의 매우 높은 인과관계를 입증하였다. 이 외에도 Russell, SOX, NASDAQ 등 다른 변수들 또한 Granger 인과관계 테스트에서 KOSPI 지수와의 유의수준 5% 수준에서 유의미한 관계임을 입증하는 결과를 보였다.

V. Conclusion

본 연구는 KOSPI 지수를 예측하기 위해 가격 정보와 기술적 지표에만 의존했던 기존 접근 방식의 한계점을 극복하고자, 국제 거시 경제 지표와 시장 지수를 통합적으로 활용한 확장된 데이터 셋을 고안하였다. 확장된 변수 중에서 딥러닝 모델의 예측 성능 향상에 기여할 수 있는 하위 변수 집합을 찾아내기 위해 Filter-DQN을 제안하여 효율적인 변수 선택을 가능하게 했다. 이를 Transformer 기반 세 가지 SOTA (State-of-the-Art) 시계열 예측 모델에 적용한 결과, 전체 변수를 입력한 경우와 기본 MI 방식을 사용한 변수 선택 대비 Filter-DQN 방식이 각각 평균 8.28% 및 45.28%의 성능 개선을 보였다. 추가적으로, Table.4에 제시된 실험 결과에 따르면 Filter-DQN 방식은 전체 훈련 속도를 평균 2.6% 단축하는 효율성을 보여 주었다. 본 연구의 결과는 국제 시장 지표와 기술적 지표 간의 상호작용을 효과적으로 반영한 변수 선택 전략이 시계열 예측의 정확성과 계산 효율성을 동시에 향상시킬 수 있음을 입증한다. 이러한 접근은 향후 금융 분야에서 국제 시장 지표와 기술적 지표 간의 상호작용을 반영한 주가 예측 모델을 통해 특정 종목이나 지수의 중단기 가격 변동을 예측하고, 이를 기반으로 포트폴리오 리밸런싱 시점을 최적화 하거나 헷지(hedge) 전략을 강화할 수 있어 펀드 관리나 ETF 운용 등 맞춤형 투자 상품 개발에 중요한 기여를 할 수 있을 것으로 기대된다.

향후 후속 연구에서는 외생 변수를 더욱 확장하여 국제 시장 지표와 KOSPI 간의 상호작용을 엄밀히 분석하고자 한다. OPEC 원유 수출과 천연가스 데이터를 활용하여 에너지 및 원자재 지수와 한국 시장 간의 연관성을 살펴 보거나, 경제 정책의 불확실성과 지정학적 위험이 투자 심리

에 미치는 영향을 고려하여 GPR(Geopolitical Risk Index)와 EPU(Economic Policy Uncertainty Index)를 활용한 시장 변동성 기반의 KOSPI 예측 연구의 확장 가능성을 제시한다. 또한, 글로벌 금융 위기나 코로나19 팬데믹과 같은 특정 위기 시점들로 구분하여 각 상황에 선택되는 변수와 그 이유에 대한 심층적 분석 가능성도 시사한다. 이러한 연구를 통해 시장 상황별 변수의 선택 패턴이 KOSPI 주가 예측에 미치는 영향을 상세히 규명하고, 이를 바탕으로 모델의 예측 성능을 더욱 향상시키는 데 기여하고자 한다.

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.2023R1A2C200337911 and No. RS-2023-00220762).

REFERENCES

- [1] Wan Soo Choi. "The Volatility Spillovers and Volatility Impulse Response: Centered on U.S. and East Asian Stock Markets." *Journal of Industrial Economics and Business* 30.2 (2017): 565-588. DOI: 10.22558/jieb.2017.4.30.2.565
- [2] Tae Hyuk Kim and Suk Gyu Kang. "The Influence of Nasdaq Stock Market on the Price Volatility in Korean Stock Market." *Asia-Pacific Journal of Financial Studies* 30.1 (2002): 363-390. UCI: G704-000202.2002.30..009
- [3] Md Masum Billah et al. "Stock price prediction: comparison of different moving average techniques using deep learning model." *Neural Computing and Applications* 36.11 (2024): 5861-5871. DOI: 10.1007/s00521-023-09369-0
- [4] Woo Sik Lee. "A deep learning analysis of the KOSPI's directions." *Journal of the Korean Data and Information Science Society* 28.2 (2017): 287-295. DOI: 10.7465/jkdi.2017.28.2.287
- [5] Sang Youn Jee and Kyung Soo Kim. "Estimating Volatility between Global Financial Markets." *Korea International Accounting Review* 90 (2020): 1-24. DOI: 10.21073/kiar.2020.90.001
- [6] Jong Hyup Shin. "The Effects of Macroeconomic Variables on KOSPI200 Sectoral Stock Indices." *Journal of Korean National Economy* 41.4 (2023): 83-98. DOI: 10.30776/JES.41.4.4
- [7] Ga Yeon Jung et al. "Kospi Index Prediction using a Financial-specific Sentiment Analysis and Deep Learning-based

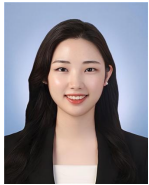
- Time Series Prediction Model." *Industrial Engineering & Management Systems* 50.4 (2024): 240-250. DOI: 10.7232/JKIIE.2024.50.4.240
- [8] Kentaro Imajo et al. "Deep portfolio optimization via distributional prediction of residual factors." *Proceedings of the AAAI conference on artificial intelligence* 35.1 (2021). DOI: 10.1609/aaai.v35i1.16095
- [9] Young In Yoon and Hye Young Jung. "A Comparison of Uncertainty Quantification of Deep Learning models for Time Series." *Journal of The Korean Data Analysis Society* 26.1 (2024): 163-174. DOI: 10.37727/jkdas.2024.26.1.163
- [10] Roberto Battiti. "Using mutual information for selecting features in supervised neural net learning." *IEEE Transactions on neural networks* 5.4 (1994): 537-550. DOI: 10.1109/72.298224
- [11] Volodymyr Mnih. "Playing Atari with deep reinforcement learning." *arXiv preprint arXiv:1312.5602* (2013). DOI: 10.48550/arXiv.1312.5602
- [12] Yong Liu et al. "Non-stationary transformers: Exploring the stationarity in time series forecasting." *Advances in Neural Information Processing Systems* 35 (2022): 9881-9893. DOI: 10.48550/arXiv.2205.14415
- [13] Haixu Wu et al. "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting." *Advances in Neural Information Processing Systems* 34 (2021): 22419-22430. DOI: 10.48550/arXiv.2106.13008
- [14] Yunhao Zhang and Junchi Yan. "Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting." *The Eleventh International Conference on Learning Representations* (2023). DOI: 10.48550/arXiv.2205.09336
- [15] Manish Agrawal, Asif Ullah Khan and Piyush Kumar Shukla. "Stock price prediction using technical indicators: a predictive model using optimal deep learning." *Learning* 6.2 (2019): 7. DOI: 10.35940/ijrteB3048.078219
- [16] Jae Eung Lee and Ji Hyeong Han. "Layer-wise Relevance Propagation based technical and macroeconomic indicator impact analysis for an explainable deep learning model to predict an increase and decrease in KOSPI." *Journal of Korean Institute of Information Scientists and Engineers* 48.12 (2021): 1289-1297. DOI: 10.5626/JOK.2021.48.12.1289
- [17] Jung Hoon Moon and Gyu Sik Han. "Analysis of the Impact of US China and Korea Macroeconomic Variables on KOSPI and VKOSPI." *Asia Pacific Journal of Business* 15.1 (2024): 209-223. DOI: 10.32599/apjb.15.1.202403.209
- [18] Sepp Hochreiter. "Long Short-term Memory." *Neural Computation* MIT-Press (1997). DOI: 10.1162/neco.1997.9.8.1735
- [19] Sumin Nam Jieun Kim and ZoonKy Lee. "Time-Invariant Stock Movement Prediction After Golden Cross Using LSTM." *Journal of The Korea Society of Computer and Information* 28.8 (2023): 59-66. DOI: 10.9708/jksci.2023.28.08.059
- [20] Min Seob Song and Junghye Min. "Comparison of Stock Price Prediction Using Time Series and Non-Time Series Data." *Journal of the Korea Society of Computer and Information* 28.8 (2023): 67-75. DOI: 10.9708/jksci.2023.28.08.067
- [21] Isabelle Guyon and Andre Elisseeff. "An introduction to variable and feature selection." *Journal of Machine Learning Research* 3.Mar (2003): 1157-1182. DOI: 10.1162/153244303322753616
- [22] Svante Wold, Kim Esbensen and Paul Geladi. "Principal component analysis." *Chemometrics and Intelligent Laboratory Systems* 2.1-3 (1987): 37-52. DOI: 10.1016/0169-7439(87)80084-9
- [23] Geoffrey E Hinton and Ruslan R Salakhutdinov. "Reducing the dimensionality of data with neural networks." *Science* 313.5786 (2006): 504-507. DOI: 10.1126/science.1127647
- [24] Hang Zhao et al. "Multivariate time-series anomaly detection via graph attention network." *2020 IEEE International Conference on Data Mining IEEE* (2020). DOI: 10.1109/ICDM50108.2020.00093
- [25] Hanchuan Peng, Fuhui Long and Chris Ding. "Feature selection based on mutual information criteria of max-dependency max-relevance and min-redundancy." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27.8 (2005): 1226-1238. DOI: 10.1109/TPAMI.2005.159
- [26] Pablo A Estevez et al. "Normalized mutual information feature selection." *IEEE Transactions on Neural Networks* 20.2 (2009): 189-201. DOI: 10.1109/TNN.2008.2005601
- [27] Ruohao Xu et al. "Dynamic feature selection algorithm based on Q-learning mechanism." *Applied Intelligence* (2021): 1-12. DOI: 10.1007/s10489-021-02257-x
- [28] Mustafa Abdul Salam et al. "The effect of different dimensionality reduction techniques on machine learning overfitting problem." *International Journal of Advanced Computer Science and Applications* 12.4 (2021): 641-655. DOI: 10.14569/IJACSA.2021.0120480
- [29] Adolfo Crespo Marquez. "The curse of dimensionality." *Digital Maintenance Management: Guiding Digital Transformation in Maintenance*. Cham: Springer International Publishing (2022): 67-86. DOI: 10.1007/978-3-030-97660-6_7
- [30] Ashish Vaswani et al. "Attention is all you need." *Advances in Neural Information Processing Systems* (2017). DOI: 10.48550/arXiv.1706.03762

Authors



Jean-Ho Kim received the bachelor's degree in the Underwood International College in Yonsei University in 2023. He is a master student at the Yonsei University Graduate School of Information since 2023.

His current research interests include deep learning and financial engineering.



Eun-Hong Park received the bachelor's degree in the Department of Business Administration from Soongsil University in 2021. She is a master student at the Yonsei University Graduate School of Information

since 2024. Her current research interests include deep learning and financial engineering.



Ha-Young Kim is an Associate Professor at Graduate School of Information, Yonsei University, Korea. She received her Ph.D. degree at department of Mathematics, Purdue University, USA.

From 2011 to 2016, she was a research staff member in Samsung Advanced Institute of Technology (SAIT) of Samsung Electronics, Korea, working on various recognition systems with deep learning. Her primary research areas are deep learning and computational Finance. She has published in leading journals, including Information Fusion, Applied Soft Computing, Expert Systems with Applications, Stochastic and Dynamics, Computers in Biology and Medicine, PLoS ONE, automation in construction, Journal of Computing in Civil Engineering and Annals of Finance. She is the inventor of 8 patents and 13 patent applications.