시계열 딥러닝 모델을 이용한 삼성전자 주가예측

**윤기웅**

Yoon Ki-woong

**서강대학교 컴퓨터공학과**

**(발표자 연락처:** [**rldnddbs@naver.com**](mailto:rldnddbs@naver.com) **)**

**요 약**

대한민국은 코로나 팬데믹 이후 경제와 사회 및 일상의 여러 변화를 맞이했다. 그 중 하나는 바로 주식 투자에 대한 관심이 증가했고 주식 투자 인구 또한 크게 늘어났다는 것이다. 하지만 코로나 시절에 활발하게 시행된 세계 주요 국가들의 금리인하와 양적완화 정책들은 서서히 사라지고 어느 종목에 투자해도 수익을 얻을 수 있었던 ‘파티’가 끝나면서 앞으로는 주가 상승이 예측되는 종목을 고르는 안목이 요구되었다. 이에 최근 활발하게 진행되고 있는 딥러닝 모델들을 이용하여 코스피(Korea Compoiste Stock Price Index, KOSPI)에 상장된 기업 중 가장 시가 총액이 높은 삼성전자의 주가 예측 알고리즘을 만들었다. 이 때 해당 회사의 주가 데이터와 주식 거래량 그리고 개인, 외국인, 기관의 코스피 순매수 데이터를 사용하였다. 앞으로의 40거래일 동안의 주가를 예측하기 위해 이전의 60, 90, 120, 150, 180 거래일 동안의 주가, 매매 데이터를 사용하고 가장 적절한 window size를 탐색한다. 실험에 사용된 모델은 LSTM, Dlinear, Nlinear, SegRNN으로 총 4개이고 이 중 SegRNN의 예측 성능이 가장 뛰어났다. 또한, Nlinear, SegRNN의 경우는 미래 주가의 추세를 추정 가능하여 주가 흐름 예측에 활용이 가능하다.

키워드: 시계열, 딥러닝, 주가 예측

**1. 서 론**

코로나 19는 전 세계 국가들에 사회적 경제적으로 큰 변화를 가져왔다. 특히, 대한민국을 포함한 여러 국가들의 금융시장에도 큰 영향을 미쳤다. 한국 주식시장의 변동성은 이전보다 커졌으며 개인 투자자들의 투자 참여도가 향상되었다. 국내 주식시장의 개인 투자자 비율은 코로나 이전 2019년 64%에 비해 2021년에는 73%으로 상승하고 거래대금 또한 급증했다[1].

하지만 국내 한 증권사에서 시행한 수익률 분석 결과를 살펴보면 2020년 11월 기준 20대 투자자들의 수익률이 다른 연령대보다 낮았고 20대 남성은 동일 연령대 여성의 수익률뿐만 아니라 다른 연령대에 비해도 3.8%로 매우 낮았다.[2] 이는 코로나 이후 강한 상승장세로 인한 FOMO(fearing of missing out, 유행에 뒤쳐지는 것에 대한 공포심리, 소외에 대한 불안감)증후군에 의한 것이라 볼 수 있다. 높은 투자 수익률을 기록했다는 주변인들의 이야기를 들으며 자신 또한 높은 수익을 얻고 싶은 마음에 신중하지 못한 주식 투자를 감행하고 투자의 실패로 이어지기 때문이다.

본 연구에서는 위와 같은 문제 해결을 위해 주식 투자 시 의사 결정에 도움이 되도록 딥러닝 주가 예측 모델을 만들고자 한다. 투자 종목은 코스피(Korea Compoiste Stock Price Index, KOSPI)에 속한 삼성전자(005930)이고 코스피의 외국인, 기관 그리고 개인의 매매 데이터를 추가로 활용하여 시계열 예측을 수행하고자 한다. 삼성전자는 반도체, 전자제품, 디스플레이와 전자 제품 제조 업체로 대한민국을 대표하는 기업이고 삼성전자의 보통주와 우선주의 시가총액을 합치면 전체 코스피 시가총액의 20%가량을 차지할 정도로 코스피에서 큰 비중을 차지하고 있다. 두 번째로 코스피 시가총액 비중이 높은 SK하이닉스(000660)의 시가 총액 비중이 5.5%임을 고려했을 때 대한민국 전체 주식시장에서 삼성전자의 영향력이 얼마나 높은 지 알 수 있다. 그만큼 삼성전자(005930)는 다른 종목들과 비교했을 때 가격 변동성이 적으며 안정적이다.

시계열 데이터는 크게 단변량 시계열 데이터와 다변량 시계열 데이터로 나눠진다. 단변량 시계열 데이터는 시간에 따라 하나의 변수만 존재하는 경우이다. 기상 데이터가 있다고 가정할 때 일일 기온 수치만 존재하는 경우 단변량 데이터이다. 이에 반해 다변량 시계열 데이터는 시간에 따라 여러 변수의 데이터가 존재하는 경우이다. 기상 데이터에 기온뿐만 아니라 일일 습도, 풍속 등 여러 변수가 있는 경우 다변량 시계열 데이터에 속한다. 본 연구에서는 주식의 종가 데이터와 거래량 데이터를 함께 사용하여 주가 예측을 진행하기에 다변량 시계열 데이터를 다룬다 할 수 있다.

시계열 데이터는 오늘날의 데이터 중심 세계에서 어디에나 존재한다. 교통량 예측, 금융 투자, 수요 예측 등 다양한 분야에서 시계열 예측이 사용된다[3]. 딥러닝 모델 사용 이전에는 ARIMA와 같은 전통적인 통계 방법과 머신러닝 모델들을 주로 사용했었다. 하지만, 순차 데이터(sequence data) 처리에 효율적인 LSTM와 GRU 등 RNN 계열의 모델들의 출현 이후로 딥러닝 모델들이 시계열 예측에 주로 사용되었다[4]. 최근에는 자연어 처리 분야에서 뛰어난 성능을 보이는 Transformer 기반의 변형 모델들과 MLP 기반의 비교적 속도가 빠른 Linear 모델들이 시계열 데이터 예측 분야에서 좋은 성능을 나타내어 활발한 연구가 이루어지고 있다.

**2. 사용 모델 소개**

LSTM(Long short-term memory)은 1997년 제안된 모델로써 RNN의 장기 종속성 학습에 한계를 보완하기 위해 고안되었다[5]. Cell state, gate 등의 구조를 이용해서 정보를 선택적으로 읽고 저장하여 데이터의 장기 종속성 특성을 효과적으로 처리한다.

LTSF-Linear모델은 트랜스포머 기반 모델을 이용한 시계열 예측에 대해 문제점을 제기하며 구상된 모델이다. 트랜스포머의 self-attention 메커니즘은 데이터 간의 시간 정보를 온전히 추출하기 힘들고 시간이 많이 소요된다는 약점이 있다. 이에 고안된LTSF-Linear는 단층 선형 모델의 간단한 모델인 동시에 9개의 벤치마크에서 트랜스포머 기반 모델의 성능을 뛰어넘는 우수한 성능을 보인다[6]. 관련 논문에서 제시한 모델들 중 Nlinear와 Dlinear 모델을 이용하여 주가예측을 진행했다.

SegRNN은 장기 시계열 예측 분야에서 RNN의 단점인 여러 차례의 반복(iteration)을 보완하고자 고안된 모델이다. 해당 모델은 장기 시계열 예측 처리에 있어 RNN의 반복을 줄이기 위해 구간별 반복(Segment-wise Iterations)과 병렬 다단계 예측(Parallel Multi-step Forecasting,PMF) 방법을 도입하여 예측 정확도와 추론 속도에 있어 현저한 성능 향상을 이끌어 냈다[7]. 최근 트랜스포머 모델이 시계열 예측 분야에서도 좋은 성능을 보여주고 있는 상황에서 RNN 기반 접근 방식 또한 성능의 우수성을 입증했다.

**3. 방법**

실험의 데이터를 얻기 위해서 python을 이용한 크롤링을 진행했다. 주식뿐만 아니라 여러 통계 데이터를 가져올 수 있는Pandas-datareader 라이브러리를 통해서 삼성전자(005930)의 2005년 1월 3일부터 2024년 6월 26일까지의 일마다의 거래량과 종가 데이터를 수집했다. 또한, 네이버 파이낸스 사이트 크롤링으로 2005년 1월 3일부터 2024년 6월 26일까지의 코스피 거래 데이터를 수집했다. 해당 코스피 거래 데이터에는 개인과 외국인, 기관계 그리고 기타법인의 각각 거래일별 순매수 금액이 저장되어 있다. 즉, 주가예측 모델 생성에 사용할 데이터는 주말과 공휴일을 제외한 총 4815 거래일의 삼성전자의 종가와 거래량 데이터, 코스피의 개인 외국인 기관계 그리고 기타법인의 거래량 데이터이다.

실험에 사용할 모델은 LSTM, Dlinear, Nlinear, SegRNN으로 총 4개이다. 시계열 데이터 예측에 필요한 window size를 각 모델마다 다르게 부여하고 각 모델들의 window size의 변화에 따른 예측 성능을 확인한다. 여기서 window size란 앞으로의 미래 데이터를 예측할 때 사용할 이전 데이터의 기간이다. 이번 실험에서는 향후 40 거래일 동안의 주가 예측을 진행하는데 다양하게 이전의 40, 60 , 90, 120, 150 , 180거래일 데이터를 이용해 실험을 진행한다.

또한, 실험에서는 RMSE(Root Mean Squared Error)를 사용해서 모델 예측 정확성을 평가했다. Dlinear, Nlinear 모델에 대해서는 주사 데이터만 사용하는 단변량 시계열 예측을 진행하고 LSTM, SegRNN에 대해서는 거래량과 주식 매매 데이터를 추가로 사용했다. Dlinear, Nlinear 모델들은 단변량 시계열 예측에 좋은 성능을 보여주고 SegRNN은 다변량 시계열 예측에 좋은 성능을 보이기 때문이다.

딥러닝 모델은 초기 가중치 설정과 데이터 셔플링 등에 대해 상당한 랜덤성을 내포하고 있다. 그렇기 때문에 코드의 seed에 따라서 시계열 예측 결과가 매번 변하게 된다. 이러한 이유로 각 모델마다 3개의 seed를 이용하여 3번의 예측을 시행하고 산술평균을 구하여 이를 최종 예측 결과로 사용한다. Seed란 난수 생성기의 초기값을 의미하며 컴퓨터 프로그램은 난수를 생성 시 주어진 seed에 기반한 난수들을 사용한다. 그래서 seed의 설정을 진행하여 모델의 재현성과 성능 비교에 있어 불확실성을 감소시켰다.

**4. 실험 결과**

2024년 4월 26일부터 2024년 6월 26일까지 총 40거래일의 삼성전자 종가를 예측한 결과와 실제 종가 데이터를 통해 모델의 예측 정확도를 측정했다. RMSE 평가식을 사용한 결과 아래 그림1.과 같다.

가장 최신 모델인 SegRNN이 나머지 3개의 모델들에 비해 좋은 예측 정확도를 보이고 있다. SegRNN은 현재 다변량 시계열 예측 부분에서 제일 좋은 성능을 갖고 있는 SOTA(State of the Art) 모델이고 주가 데이터에 대해서도 역시 다른 모델들에 비해 뛰어난 성능을 보여준다.

또한, 주목할 점은 window size가 크다고 해서 모델의 예측 성능이 향상되지 않는다는 것이다. 실험 결과를 보면 Nlinear의 경우를 제외하고 나머지 모델들의 경우 window size가 60 혹은 90인 경우 예측 오차가 최소가 된다. 이는 주가 예측 시 모델이 불필요한 데이터나 연관성이 떨어지는 데이터를 학습하면 정확도가 떨어진다는 것을 확인할 수 있다.

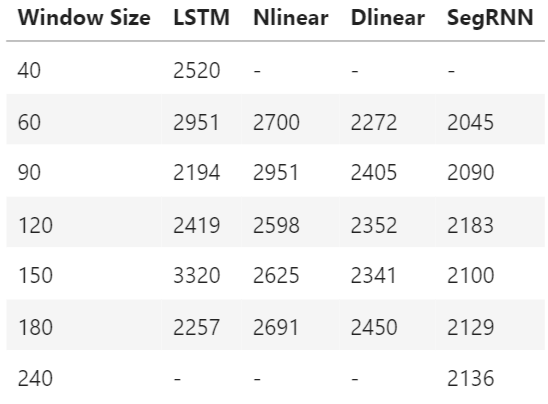


그림1. RMSE를 이용한 모델별 예측 성능

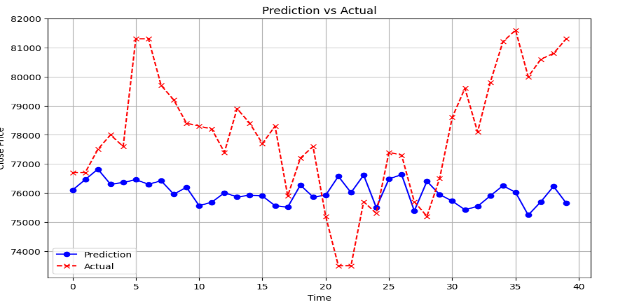


그림 2. LSTM(window size 60)

위 그림은 LSTM 모델, windowsize 60을 사용한 주가예측 그래프와 실제 주가 그래프이다. 다른 window size를 적용해도 비슷한 예측 그래프를 보이며 주가 등락과 흐름을 제대로 파악하지 못하고 학습한 데이터에 과적합된 경향을 나타낸다. 이러한 이유들로 주가 예측 모델로 LSTM 은 적합하지 않다.

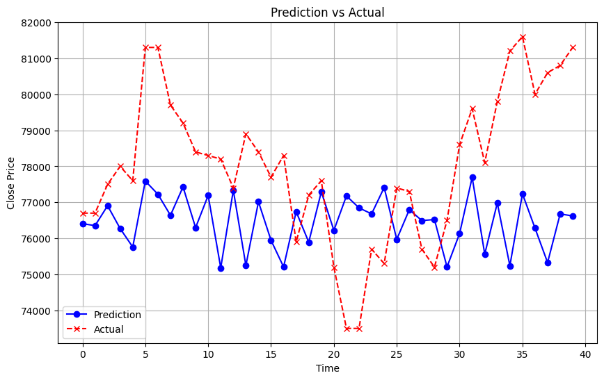


그림3. Nlinear(window size 60)

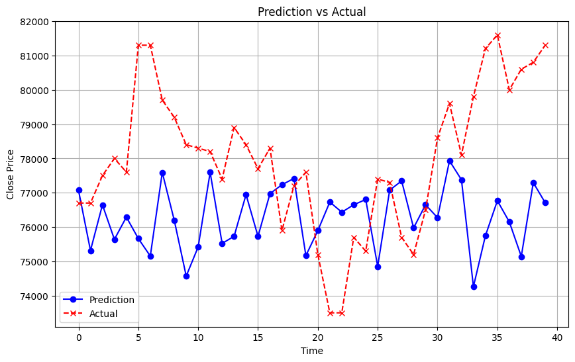


그림4. Nlinear(window size 90)

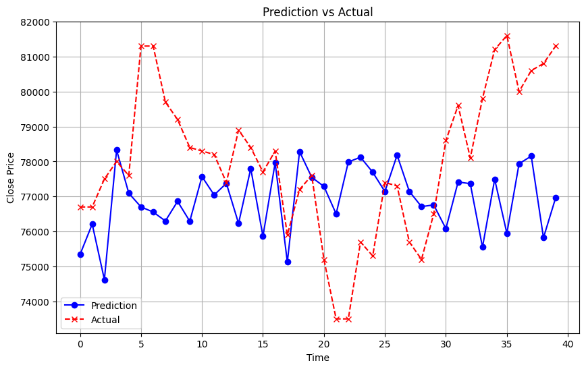


그림5. Nlinear(window size 120)

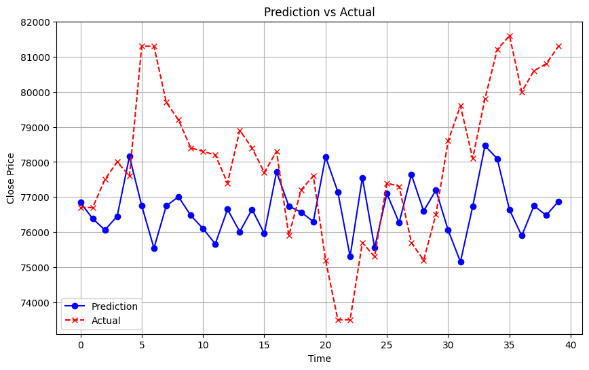


그림6. Nlinear(window size 150)

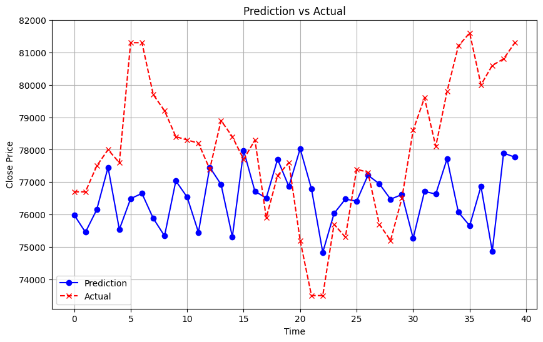


그림7. Nlinear(window size 180)

위의 그림3,4,5,6,7은 Nlinear모델의 window size 변화에 따른 예측 결과를 나타낸다. LSTM모델에 비해 예측 성능이 다소 향상되었고 주가 흐름도 조금 파악하고 예상 가능하다. 특히 window size 150을 적용한 그림6.을 살펴보면 주가를 정확히 예상하진 못 하지만 어느 정도의 상승화 하락추세를 예측한다.

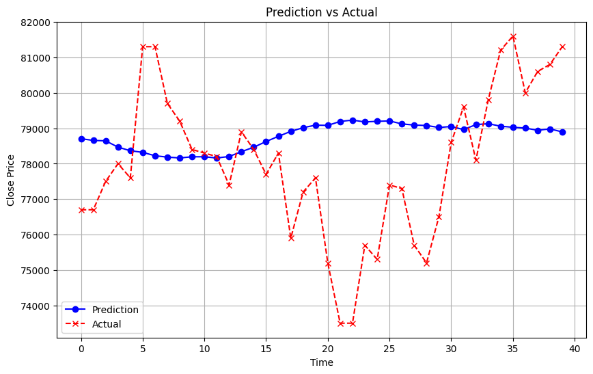


그림 8. Dlinear(window size 120)

그림8은 Dlinear모델의 120 window size를 사용하여 예측한 주가 그래프이다. LSTM 모델의 결과와 유사하게 주가 변동과 흐름을 예측하는데 취약하다. 비록 Nlinear에 비해 RMSE로 측정한 성능이 뛰어나도 실용성이 높지 않다. 이는 같은 모델 다른 window size를 적용해도 크게 양상이 변하지 않는다.

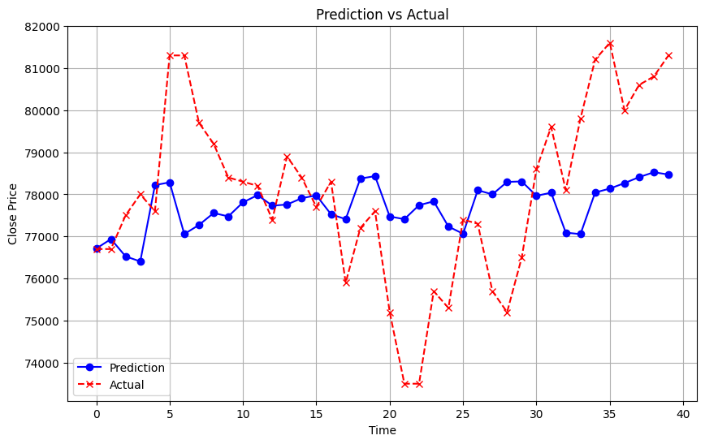


그림 9. SegRNN(window size 60)

**텍스트, 그래프, 라인, 도표이(가) 표시된 사진

자동 생성된 설명**

그림 10. SegRNN(window size 90)

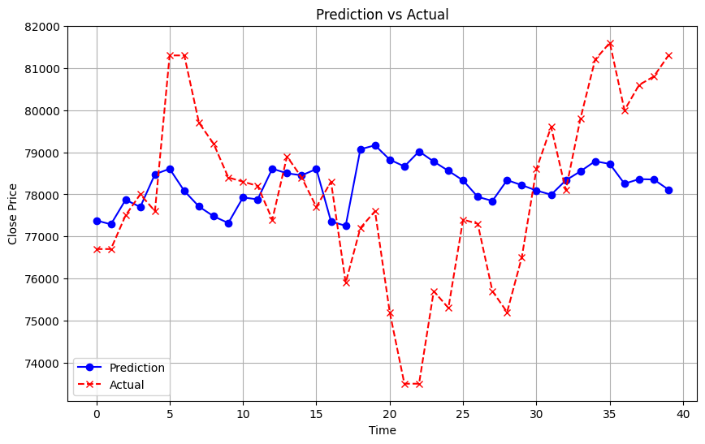


그림 11. SegRNN(window size 120)

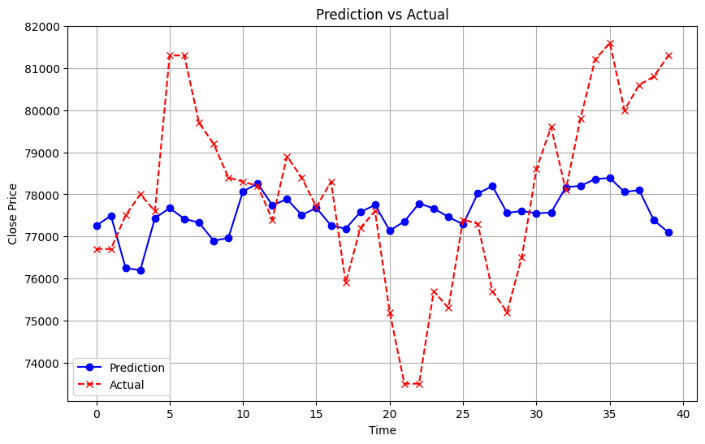
****

그림 12. SegRNN(window size 150)

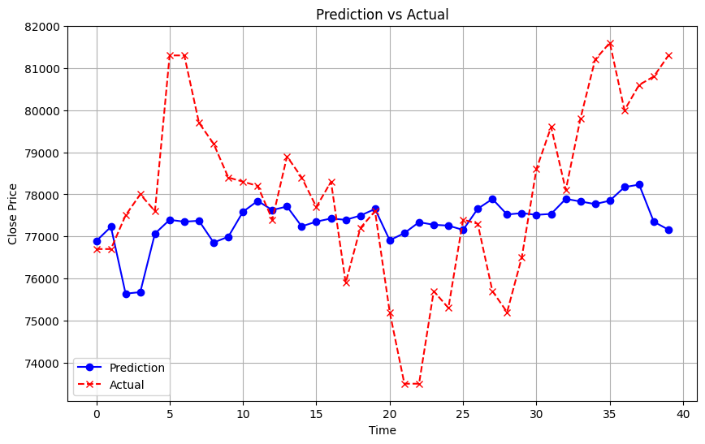
****

그림 13. SegRNN(window size 180)

위는 차례대로 SegRNN의 window size 변화에 따른 주가 예측 그래프이다. Nlinear의 예측과 유사하게 주가 변동의 흐름을 잘 예측한다. 비록 정확한 가격 예측을 성공하지 못하더라도 주가 상승과 하락 예측이 가능하기에 투자 시 참고 자료 정도로 사용이 가능하다.

**5. 결론**

사실 어떤 한 종목의 주가나 코스피 지수, 나스닥 지수들은 규칙적이지 않고 수많은 변수들과 사람들의 의사결정에 의해 결정되는 시계열 데이터를 예측 하기란 거의 불가능에 가깝다. 본 연구에서도 삼성전자(005930)의 주가 예측 알고리즘을 만들기 위해 여러 모델들을 사용해 보았으나 SegRNN, Nlinear 모델만 대략적인 주가의 상승, 하락을 예측하는데 성공했다. 또한, 연구 진행을 위해 코스피의 개인, 외국인, 기관의 순매수 데이터를 이용했는데 만약 코스피가 아닌 삼성전자(005930) 종목에 대한 순매수 데이터를 활용했다면 더욱 정교한 알고리즘을 만들 수 있을 것이다.

본 연구에 사용한 딥러닝 모델들은 각 모델들의 하이퍼 파라미터들에 대해 많은 영향을 받는다. 그렇기 때문에 하이퍼파라미터 최적화 프레임워크인 Optuna 등을 추가로 활용하여 최적의 수치를 찾게 된다면 더욱 정확한 알고리즘을 만들 수 있을 것이다.

**­**

**참고문헌**

|  |  |
| --- | --- |
| ISO 690 |  |

[1] 송경우; 정지영. 코로나 19 이후 개인투자자의 투자패턴에 대한 연구. *무역연구*, 2024, 20.1: 213-231.

[2] 김제림, “주식 열심히 사고파는데 …. 수익률 꼴찌 ‘20대 남’, 1등은 ‘30대 여’[스물스물]”, 매일경제,2021.01.09

[3] Aslam, S.; Herodotou, H.; Mohsin, S. M.; Javaid, N.; Ashraf, N.; and Aslam, S. 2021. A survey on deep learning methods for ­­power load and renewable energy forecasting in smart microgrids. Renewable and Sustainable Energy Reviews, 144: 110992.

[4] Yamak, Peter T., Li Yujian, and Pius K. Gadosey. "A comparison between arima, lstm, and gru for time series forecasting." *Proceedings of the 2019 2nd international conference on algorithms, computing and artificial intelligence*. 2019.

[5] Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. Neural Computation, 9(8): 1735–1780.

[6] ZENG, Ailing, et al. Are transformers effective for time series forecasting?. In: *Proceedings of the AAAI conference on artificial intelligence*. 2023. p. 11121-11128.

[7] Lin, Shengsheng, et al. "Segrnn: Segment recurrent neural network for long-term time series forecasting." *arXiv preprint arXiv:2308.11200* (2023).