

Weighted Ensemble을 활용한 다중 예측 기간 다중 자산 실현분산 예측

박상훈*¹, 성아영**², 하유진***³, 김건우*²

Multi-Horizon Realized Variance Forecasting across Asset Classes with a Weighted Ensemble

Sang Hoon Park*¹, Ayeong Seong**², Yu Jin Ha***³, and Gun-Woo Kim*²

본 논문은 교육부와 경상남도의 재원으로 지원받은 경상남도 지역혁신중심 대학지원체계(RISE) 사업,
정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구(RS-2026-25476256) 및 산업통상자원부 및
한국산업기술기획평가원(KEIT)의 지원을 받아 수행된 연구(RS-2025-02633048)의 결과임

요 약

본 연구는 기계학습 기반 피쳐 설계와 가중 앙상블을 결합하여 다중 자산·다중 기간 RV 예측 시스템을 설계하고 검증한다. 주식·채권·원자재 11개 ETF를 대상으로, 일별 OHLCV와 VIX 기간구조에서 도출한 35개 피쳐를 HAR 기반·HF Proxy·IV Surface·Alternative의 4개 그룹으로 체계화하고, Ridge와 XGBoost를 결합한 WEns 알고리즘을 제안한다. 예측 기간은 거래일 기준 1, 5, 22, 60, 90, 120, 180, 252의 8개 구간이다. 동일한 피쳐와 검증 절차 하에서 선형·트리·신경망 모델을 비교한 결과, 1~22거래일과 180~252거래일에서는 WEns가, 60~120거래일에서는 Ridge가 최고 성능을 보이며 8개 기간 중 1위를 차지하였다. 피쳐 중요도 합의 분석, 통계적 유의성 검정, 만기 불일치·결측 처리·자산별 지표 대체 강건성 실험을 통해 제안 시스템의 신뢰도와 해석 가능성을 점검하였다.

Abstract

This study designs and evaluates a machine learning system for multi-horizon RV forecasting across asset classes. We organize 35 features from daily OHLCV and the VIX term structure into four groups (HAR-based, HF Proxy, IV Surface, Alternative), and propose WEns, a weighted ensemble of Ridge regression and XGBoost, applied to 11 ETFs spanning equities, bonds, and commodities. Forecasting horizons cover eight trading-day intervals: 1, 5, 22, 60, 90, 120, 180, and 252. Under identical feature sets and validation procedures, WEns achieves the best performance over 1 - 22 and 180 - 252 trading days and Ridge over 60 - 120 trading days, ranking first in the eight horizons. Reliability and interpretability are examined through feature importance consensus, statistical significance tests, and robustness experiments on maturity mismatch, missing-value handling, and asset-specific indicator substitution.

Keywords

financial forecasting, volatility modeling, multi-asset analysis, ensemble learning, quantitative finance

* 경상국립대학교 컴퓨터공학과(*² 교신저자)

- ORCID¹: <https://orcid.org/0009-0002-0430-9745>

- ORCID²: <https://orcid.org/0000-0001-5643-4797>

** 성균관대학교 인공지능융합학과 석사과정

- ORCID: <https://orcid.org/0009-0005-6909-5006>

*** 경상국립대학교 AI융합공학부 석사과정

- ORCID: <https://orcid.org/0009-0004-8963-1913>

• Received: Apr. 14, 2026, Revised: May 08, 2026, Accepted: May 11, 2026

• Corresponding Author: Gun-Woo Kim

Dept. of Computer Science and Engineering, College of IT Engineering,

Gyeongsang National University, JinJu, Korea

Tel.: +82-55-772-3323, Email: gunwoo.kim@gnu.ac.kr

1. 서 론

최근 금융시장은 팬데믹, 지정학적 충격, 통화정책 전환과 같은 사건을 거치며 높은 변동성 국면을 반복적으로 경험해 왔다. 예를 들어 Yahoo Finance에서 수집한 코로나19 팬데믹이 본격화된 2020년 3월 16일 변동성지수(VIX, Volatility Index)는 82.7 수준까지 상승하였고, 러시아-우크라이나 전쟁 발발 전후인 2022년 3월 초에도 36.5 수준까지 상승하였다. 그림 1은 2010년부터 2024년까지의 VIX 시계열을 나타낸 것으로, 시장 충격이 발생할 때마다 변동성이 급등하는 패턴이 반복적으로 관찰된다. 이러한 사례는 변동성이 단순한 가격변동 폭을 넘어 시장의 불확실성 평가와 위험회피 수요를 반영하는 핵심 변수임을 보여준다.

미래 변동성 예측은 포트폴리오 위험 관리, 옵션 헤징 비용 산정, 자산배분 전략 설계의 핵심 입력값이다. 실현분산(RV, Realized Variance)는 실제 시장에서 사후적으로 관측되는 변동성 측도로서, 내재분산(IV, Implied Variance)나 일반화 자기회귀 조건부 이분산(GARCH, Generalized Autoregressive Conditional Heteroskedasticity) 조건부 분산과 달리 모형 가정 없이 계산되는 특성을 가진다. RV의 예측 가능성은 HAR-RV[1]이래 광범위하게 검증되어 왔으며, 이를 다중 자산·다중 기간으로 확장하는 것이 본 연구의 목적이다.

금융 예측에서 가장 직관적인 목표는 수익률 또

는 초과수익률이지만, 수익률은 신호 대 잡음 비율이 낮고 구조 변화의 영향을 크게 받기 때문에 안정적인 예측이 어렵다. Gu, Kelly, and Xiu[2]는 다양한 기계학습 방법을 비교하며 자산가격 예측에서 기계학습의 가능성을 제시하였다. 반면 변동성은 군집성과 지속성이 강하므로 수익률보다 예측 가능성이 상대적으로 높다. Corsi[1]의 HAR-RV 모형은 일·주·월 수준의 변동성 정보를 결합하여 이러한 특성을 간명하게 반영한 대표적 접근이며, 이후 Audrino and Knaus[3], Christensen et al.[4]은 정규화 회귀와 기계학습을 활용한 변동성 예측의 확장 가능성을 보였다.

본 연구는 미래 RV를 직접 예측 대상으로 삼는다. RV 예측은 변동성 리스크 관리의 핵심 과제이며, 풍부한 피쳐 세트와 다중 자산군 구조 아래에서 예측 가능성과 모형 간 성능 차이를 체계적으로 분석하기에 적합하다. 추가로, 예측된 RV를 관측 가능한 IV와 결합하면 분산 위험 프리미엄(VRP, Variance Risk Premium)를 파생 산출할 수 있으며[5][6], 이는 변동성 위험 가격화의 보조 참고 신호로 활용될 수 있다. 본 연구에서 사용하는 IV, RV, VRP의 구체적 정의와 산출 절차는 3장에서 제시한다.

RV 예측 연구는 GARCH 계열 및 HAR-RV 모형을 중심으로 축적되어 왔으나, 다중 자산·다중 기간을 함께 분석한 체계적 비교 연구는 상대적으로 제한적이다.

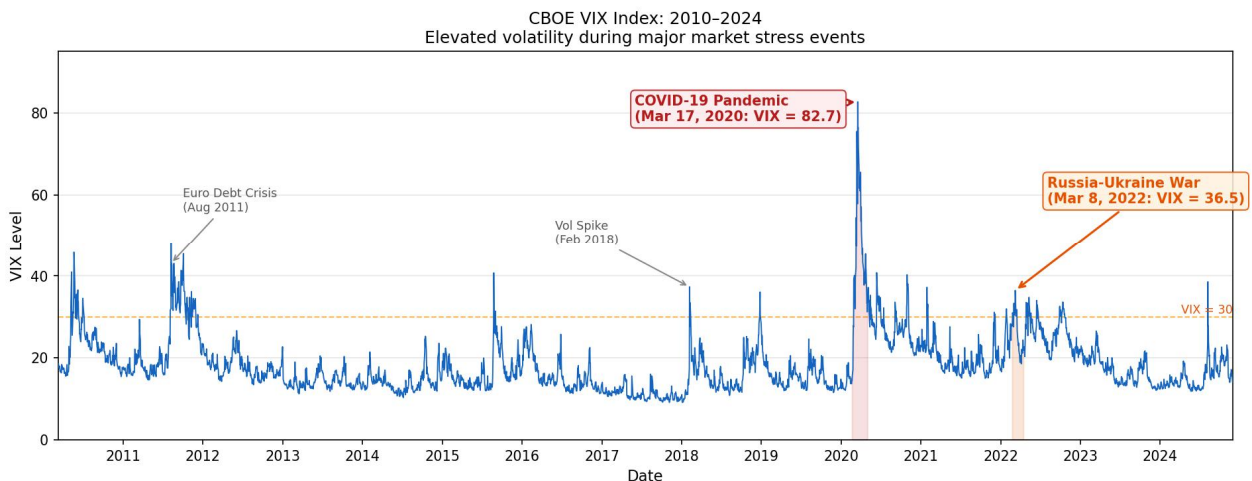


그림 1. VIX 지수 (2010~2024)

Fig. 1. VIX Index (2010~2024)

첫째, 다중 자산 RV 예측 연구는 대부분 자산별 옵션 자료 또는 고빈도 데이터에 의존하므로, 실제 분석 환경에서 주식·채권·원자재 전체에 대해 장기간 데이터를 확보하기 어렵다. 둘째, 기존 연구는 상대적으로 월간 또는 30일 안팎의 만기에 정렬된 분석이 많아, 1거래일부터 252거래일까지 예측 기간을 확장할 때 예측 구조가 어떻게 달라지는지를 체계적으로 비교한 연구는 제한적이기 때문에 예측 기간에 따라 어떤 모델이 적합한지에 대한 질문에 답하기 어렵다. 셋째, 동일한 피처와 동일한 검증 절차 아래에서 선형 모델, 트리 기반 모델, 시계열 신경망 모델을 예측 기간별로 비교한 결과도 충분하지 않다.

본 연구의 기여는 다음과 같다. 첫째, 공개적으로 이용 가능한 일별 OHLCV 데이터와 VIX 기간구조 지표로부터 HAR 기반·HF Proxy·IV Surface·Alternative의 4개 그룹, 35개 피처를 체계적으로 설계하는 피처 설계 파이프라인을 제안하였다. 둘째, Ridge와 XGBoost를 내부 검증 기반 가중치 선택으로 결합하는 가중 앙상블(WEns, Weighted Ensemble) 알고리즘을 설계하고, 동일한 피처·검증 절차 아래에서 선형·트리 기반·신경망 모델을 체계적으로 비교하여 예측 기간에 따라 상대적으로 우수한 모델 구조가 달라짐을 실증하였다. 셋째, 피처 그룹별 제거 실험과 3개 방법론(Ridge 계수, XGBoost gain, permutation importance) 합의 분석을 통해 IV Surface 및 HF Proxy 피처군이 RV 예측 성능에 실질적으로 기여함을 정량적으로 확인하였다.

넷째, WEns가 HAR-3, GARCH, IV-only 등 주요 베이스라인 대비 통계적으로 유의한 성능 우위를 가짐을 다중 비교 보정 DM(Diebold-Mariano) 검정으로 검증하였다.

이후 장의 구성은 다음과 같다. 2장에서는 HAR-RV, GARCH, 기계학습 기반 변동성 예측 관련 선행연구를 정리한다. 3장에서는 분석 대상 자산과 데이터 전처리, 피처 구성, 모델 설계 및 검증 절차를 기술한다. 4장에서는 기간별 모델 성능 비교, 피처 기여 분석, 고변동성 구간 분석, 통계적 유의성 검정 등 실증 결과를 제시한다. 5장에서는 주요 발견을 요약하고 본 연구의 한계와 향후 연구 방향을 논의한다.

II. 관련 연구

변동성 예측 연구는 크게 전통적 통계 모형과 기계학습 기반 접근으로 발전해왔다. 두 흐름의 특성과 한계를 순서대로 살펴보고, 본 연구의 필요성과 연결한다.

전통적 통계 모형 흐름에서 출발점은 Engle[7]의 자기회귀 조건부 이분산(ARCH, Autoregressive Conditional Heteroskedasticity)와 Bollerslev[8]의 GARCH(1,1)이다. GARCH 계열은 수익률의 분산이 과거 충격과 조건부 분산에 의해 결정된다는 가정 아래, 변동성의 군집성과 지속성을 체계적으로 포착한다. 그러나 GARCH는 일별 종가 정보만을 이용하는 선형 구조를 갖기 때문에, 일중 가격 변동에 포함된 고빈도 정보를 활용하지 못하며 장기 기간에 걸친 다중 스케일 지속성 구조를 반영하기 어렵다는 한계가 있다.

이를 보완하기 위해 Andersen et al.[9]은 고빈도 수익률로 구성된 RV가 변동성의 보다 풍부한 정보를 담고 있음을 보였다. 이에 기반하여 Corsi[1]는 일·주·월 세 스케일의 과거 RV를 선형 결합하는 HAR-RV 모형을 제안함으로써, 다중 시간 스케일에서 나타나는 변동성의 이질적 자기상관 구조를 간결하게 반영하였다. HAR-RV는 해석 가능성과 예측력을 동시에 갖춰 이후 RV 예측 문헌의 기준 모형으로 널리 사용되어 왔다. 그러나 HAR-RV 역시 선형 가산 구조에 머물러 있어, 피처 간 비선형 상호작용이나 고차원 입력 변수를 활용하는 데 구조적 한계를 가진다.

기계학습 기반 흐름은 이러한 선형성의 한계를 극복하려는 시도에서 출발한다. Audrino and Knaus[3]는 HAR 구조에 최소 절대값 수축 및 선택 연산자(LASSO, Least Absolute Shrinkage and Selection Operator) 변수 선택을 결합하여 고차원 설명 변수 환경에서의 변동성 예측 가능성을 탐색하였고, Christensen et al.[4]은 랜덤 포레스트·그래디언트 부스팅 등 비선형 기계학습 모형이 변동성 예측에서 경쟁력 있는 성능을 보일 수 있음을 보고하였다. 이들 연구는 고차원 피처와 비선형 관계의 활용 가능성을 제시하였으나, 트리 기반 모형은 기본적으로

시계열 순서 정보를 명시적으로 모형화하지 않으며, 예측 기간이 달라질 때 어떤 모형이 상대적으로 우수한지를 체계적으로 비교한 연구는 상대적으로 부족하다.

딥러닝 기반 접근, 특히 장단기 기억 신경망(LSTM, Long Short-Term Memory) 계열 모형은 시계열 패턴을 데이터로부터 자동 학습할 수 있다는 점에서 주목받아왔다. 그러나 금융 시계열은 표본 수 대비 신호 대 잡음 비율이 낮아 과적합이 발생하기 쉽고, 모형 구조와 하이퍼파라미터에 따라 성능의 변동성이 크다는 실용적 한계가 보고되고 있다. Gu, Kelly, and Xiu[2]는 기계학습 방법이 자산 가격 예측에서 유의미한 성과를 보일 수 있음을 보였으나, 이들 역시 단일 예측 기간과 단일 자산군을 중심으로 분석을 수행하였다.

VRP관련 연구도 RV 예측과 밀접하게 연결된다. Carr and Wu[5]는 옵션 포트폴리오를 활용해 위험 중립 기대분산을 근사하고 RV와의 차이를 VRP로 정의하는 방법을 제시하였다. Bollerslev, Tauchen, and Zhou[6]와 Bekaert and Hoerova[10], Drechsler and Yaron[11]은 VRP가 변동성 위험의 가격화를 나타내는 지표임을 실증하였으며, 이후 연구는 예측력 분석[12][13], 시장 간 전이[14], 자산군별 이질성[15]-[17]으로 확장되었다.

다중 자산(Cross-asset)·다중 기간 분석의 필요성을 살펴보면, 기존 RV 예측 연구는 대부분 특정 자산 중 주로 주가지수를 대상으로 단일 기간의 예측 성능을 비교하는 구조를 취한다[14][15][17]. 주식·채권·원자재를 함께 다루는 연구는 각 자산별 고빈도 데이터나 옵션 자료를 전제로 하는 경우가 많아, 데이터 접근성 측면에서 재현이 어렵다. 또한 예측 기간이 단기(1거래일)에서 장기(252거래일)로 확장될 때 선형·트라이·신경망 모형의 상대적 우위가 어떻게 달라지는지를 동일한 피처와 검증 절차 아래에서 체계적으로 비교한 연구는 제한적이다.

결국 기존 연구들은 각각 통계적 해석 가능성, 비선형 관계 포착, 시계열 패턴 학습 측면에서 장점을 가지나, 다중 자산·다중 예측 기간 환경에서 일관된 성능을 보장하지 못하며, 동일한 실험 조건 아래에서 모형 간 성능 차이를 정량적으로 비교한 증거가 부족하다. 본 연구는 이러한 공백을 메우기 위

해, 일별 공개 데이터로부터 체계적으로 설계한 35개 피처와 WEns를 활용하여 주식·채권·원자재에 걸친 cross-asset 환경에서 11개 자산·8개 예측 기간을 포괄하는 통합 RV 예측 프레임워크를 제시한다.

III. 데이터 및 방법론

3.1 자산 구성

본 연구는 SPY, QQQ, IWM, EFA, EEM, TLT, IEF, AGG, GLD, SLV, USO의 총 11개 ETF를 분석 대상으로 사용한다.

이들 자산은 각 주식, 채권, 원자재를 대표하는 대형 ETF로서, 2010년부터 2024년까지의 분석 기간 동안 일별 OHLCV 자료가 대부분 결측 없이 연속적으로 제공된다. 또한 거래량과 시가총액 규모 측면에서 상대적으로 높은 유동성을 유지하고 있어, 자산군 간 비교가 가능한 패널 데이터 구성을 위해 적합한 표본으로 판단된다.

OHLCV는 각 거래일 t 에 대해 다음의 다섯 가지 변수를 포함한다: 시가(Open), 고가(High), 저가(Low), 종가(Close), 거래량(Volume)을 의미한다.

예측 기간 집합 $h \in \{1, 5, 22, 60, 90, 120, 180, 252\}$ 는 각각 일간·주간·월간·분기·반기·연간 수준을 포괄하도록 설계하여 단기부터 장기까지의 예측 구조 변화를 체계적으로 비교할 수 있다. 데이터는 Yahoo Finance API를 통해 수집하였으며, 총 거래일 수는 약 3,780일이다.

표본 시작 시점을 2010년으로 설정한 이유는 다음과 같다. 첫째, 11개 ETF와 변동성 지표인 VIX 등 주요 지표의 공통 가용 구간을 안정적으로 확보하기 위해서다. 둘째, 2008~2009년 금융위기 직후 구간은 VIX 급등과 유동성 위축이 동시에 나타난 극단적 비정상 구간이므로, 이를 학습 표본에 포함할 경우 특정 위기 국면에 과적합될 가능성이 있다.

본 연구에서 "안정적으로 존재하는 자산"이란 2010년 1월 4일 이후 Yahoo Finance에서 시가, 고가, 저가, 종가, 거래량의 5개 항목이 연속 거래일 기준으로 대부분 결측 없이 제공되며, 단기 결측이 존재하더라도 1~2거래일 수준에 그치는 자산을 의미한다. 별도의 거래량 하한이나 유동성 컷오프를 적용

하지는 않았으나, 선택된 11개 ETF는 각 자산군에서 대표성을 가진 대형 ETF이므로 유동성 문제는 실질적으로 크지 않다. 분석 대상 자산의 구성은 표 1에 제시한다.

표 1. 자산 구성
Table 1. Assets group

Group	Assets
Equity	SPY, QQQ, IWM, EFA, EEM
Bond	TLT, IEF, AGG
Commodity	GLD, SLV, USO

3.2 데이터 전처리

결측치는 전방 채움(Forward fill)으로 처리하였다. 이는 미래 정보 누출을 방지하면서 시계열 연속성을 유지하고 표본 손실을 최소화하기 위한 선택이다. 선형 보간은 미래 시점의 정보를 사용할 수 있으므로 시계열 예측에서 적절하지 않을 수 있으나, 전방 채움은 과거 관측치만을 이용한다. 또한 결측행을 그대로 삭제할 경우 5일, 22일 등 다중 기간 기반 피쳐 계산의 안정성이 저하될 수 있다. 전방 채움 적용 시 표본 수는 41,261개였고, 이를 제거하면 38,489개로 감소하였다.

전방 채움의 구조적 편향을 명시한다. 결측일에 전방 채움을 적용하면 당일 시가·고가·저가·종가가 모두 직전일 종가와 같아지므로, 해당일의 일중 범위 $\log(H/L)$ 가 0이 된다. 이는 Garman-Klass, Parkinson, RogersSatchell 등 range 기반 추정량을 확정적으로 하향 편향시킨다. 그러나 본 연구에서 결측은 거래일 불일치에 의한 1~2일 수준의 산발적 결측에 한정되며, 22거래일 롤링 기간 기준으로 단일 결측일이 차지하는 비중은 최대 4.5%이므로 전반적 영향은 제한적인 것으로 판단된다.

추가 강건성 검증 결과, 전방 채움을 제거한 경우 롤링 R^2 는 0.782에서 0.795로, 중위 R^2 는 0.123에서 0.182로 소폭 상승하였다. 이는 전방 채움이 성능을 인위적으로 부풀리지 않았음을 의미하며, 오히려 보수적인 방향으로 작용했음을 시사한다. 따라서 본 연구의 핵심 결과는 결측 처리 방식에 과도하게 의존하지 않는다.

3.3 데이터 정의

본 연구는 변동성 관련 변수를 모두 분산 단위로 통일하여 정의한다. 먼저 RV는 미래 h 거래일 동안의 수익률 제곱합을 연율화한 값으로 정의하며, 본 연구의 직접 예측 대상이다. IV는 변동성 지수를 제공한 값으로 근사하고, VRP는 IV와 RV의 차이로 정의한다. VRP는 예측된 RV로부터 사후적으로 산출되는 파생 지표이다. 이는 VRP를 직접 타깃으로 설정할 경우 타깃 정의식($IV - RV$)에 포함된 IV 정보와 VIX 계열 피쳐 간 중복이 발생할 수 있어, 예측 대상과 설명 변수의 역할을 분리하기 위한 설계이다.

본 연구의 예측 기간 h 는 모두 거래일 기준으로 정의하며, $h \in \{1, 5, 22, 60, 90, 120, 180, 252\}$ 로 설정한다. RV는 자산 i 의 시점 t 이후 향후 h 거래일 구간 $[t+1, t+h]$ 에서 실현된 수익률 제곱의 평균을 연율화하여 구성한다.

본 연구에서 사용하는 일별 가격 및 거래량 자료는 각 거래일의 시가, 고가, 저가, 종가, 거래량으로 구성된다. 종가는 로그수익률과 HAR 기반 RV 계산에 사용된다. 반면 시가·고가·저가·종가 정보는 종가만으로는 반영하기 어려운 하루 중 가격 움직임을 포함하므로, range 기반 변동성 추정량을 계산하는데 활용된다. 이러한 추정량은 일중 가격 범위를 반영한다는 점에서 종가 기반 변동성 측정보다 더 많은 정보를 담을 수 있으며, 이론적으로도 더 높은 효율성을 가진다. 또한 거래량은 유동성 관련 피쳐를 구성하는 기초 변수로 사용된다.

선행연구에서는 5분 수익률을 이용하여 일별 RV를 구성하거나[9], 고빈도 RV와 점프 분리 기법을 활용한 기계학습 예측 모형을 제시한 바 있다[4]. 그러나 고빈도 데이터는 구독 비용이 높고, 채권 및 원자재 등 비주식 자산에 대해서는 장기간 자료를 확보하기 어렵다는 실무적 제약이 존재한다. 이에 본 연구는 Yahoo Finance를 통해 무료로 수집 가능한 일별 가격 및 거래량 자료를 사용하면서도, range 기반 추정량을 활용하여 고빈도 데이터가 제공하는 정보 효율을 일부 근사하고자 하였다. 이러한 접근은 데이터 접근성과 재현 가능성 측면에서

장점을 갖는다. 다만 일별 가격 및 거래량 자료 기반 접근은 일중 점프와 연속 변동을 분리할 수 없다는 한계를 가지며, 이에 대한 논의는 결론에서 다룬다. 또한 일별 수익률은 로그수익률로 정의한다.

식 (1)에서 P_t 는 yfinance에서 제공하는 수정 종가(Adjusted close)를 의미 한다. 수정 종가는 배당과 주식분할 효과를 반영하므로, 종가 기반 수익률 계산에 적합하다. 반면 range 기반 변동성 추정량 계산에는 수정되지 않은 원시 시가·고가·저가·종가 자료를 사용한다.

$$r_t = \ln(P_t/P_{t-1}) \quad (1)$$

이 경우 배당 또는 분배로 인한 가격 점프가 range 추정량에 영향을 줄 수 있으므로, 그 영향을 별도로 점검하였다. 분석 결과, 11개 ETF 전체에서 배당락 후보일(시가 갭 0.5% 초과)은 전체 거래일의 약 15.9%를 차지하였고, 해당 일자의 일중 범위 $\log(H/L)$ 는 정상일 대비 평균 1.27배 정도 높게 나타났다. 그러나 22거래일 롤링 기간에서 단일 관측치가 차지하는 비중은 약 4.5%에 불과하므로, 배당락에 따른 점프가 기간 기반 range 추정량에 미치는 전반적 영향은 제한적인 것으로 판단된다.

3.4 데이터 구성

본 연구의 데이터는 여러 자산을 여러 시점에 결

쳐 함께 관측한 날짜 $\times$ 자산 구조로 구성된다. 각 날짜 t 에서 피쳐는 $t-1$ 까지의 정보로 계산하고, 타깃은 $[t+1, t+h]$ 구간에서 계산된 전방 실현분산이다. 따라서 각 기간 h 에 대해 표본의 끝부분 h 일은 타깃을 계산할 수 없으므로 학습과 평가에서 제외된다. 또한 예측 기간이 길어질수록 타깃 기간이 서로 중첩될 가능성이 커지므로, 학습-테스트 분할 경계와 교차검증 폴드 경계에는 예측 기간 h 에 해당하는 정화 구간(Purge gap)을 두어 검증 누출 가능성을 완화하였다[18]. 이러한 구성은 동일한 피쳐 세트와 동일한 검증 절차 아래에서 모델 간 성능 차이를 비교하기 위한 기본 단위를 제공한다.

3.5 피쳐 구성

피쳐는 (i) HAR 기반 과거 시차 변수, (ii) 일별 가격 range 기반 변동성 추정량(HF proxy), (iii) IV 기간구조 및 VRP 통계(IV surface), (iv) 거래량 기반 대안 데이터로 구성된 다중 그룹을 사용하며, 총 35개 피쳐를 활용한다. IV surface 그룹에는 IV_VRP와 IV_VRP_ma22가 포함되는데, 이는 과거 시점에서 관측된 VRP(IV - 과거RV)로서 미래 RV와 독립적이다. 과거 VRP는 시장의 변동성 위험 가격화 수준을 반영하므로, 미래 RV 예측을 위한 피쳐로 활용하는 것은 타깃 정보 누출 없이 유의미한 설명력을 추가할 수 있다.

표 2. WEns Ridge 피쳐 목록

Talbe 2. List of WEns Ridge features

Category	Variables	Description
Realized volatility	LogRV_lag(w), LogRV_Std(w), RV_Mom(w)	Lagged logRV; rolling std; momentum [w=1,5,10,22 (lag) / 5,22 (Std·Mom)]
Range estimators	Parkinson(w), GarmanKlass_22, RogersSatchell_22	Range-based vol estimators (H-L / OHLC) [w=5,22 (Pk only)]
Returns & intraday	Ret_lag1, Ret_abs_lag1, Range_Close_Ratio, Overnight_Vol, Overnight_Ret	Prev-day return / abs return; (H-L)/C; overnight vol·return
VIX levels & dynamics	IV_VIX(9D/1M/3M), IV_VIX_chg, IV_VIX_ma5, IV_VIX_std5	log VIX levels (9D/1M/3M); daily change; 5d MA; 5d rolling std
VIX term structure	IV_VIX_TermSlope, IV_VIX_ShortSlope	VIX3M - VIX; VIX - VIX9D
Variance risk premium	IV_VRP, IV_VRP_ma22	VIX ² /100 - RV_SPY; 22d MA
Market context	SPY_LogRV, Corr_SPY	SPY log-RV; 22d rolling corr with SPY
Alternative data	Amihud, Vol_Ratio, Vol_Surprise, PV_Corr, Order_Imb, Kyle_Lambda	Illiquidity; vol ratio; vol z-score; price-vol corr; order imbalance; price impact

각 그룹의 피쳐 수는 Base 12개 + HF Proxy 7개 + IV Surface 10개 + Alternative 6개로 구성된다. 초기 설계에서는 GARCH(1,1) 기반 일별·주간 조건부 변동성 추정치인 GARCH_Daily, GARCH_Weekly를 포함한 37개 피쳐를 검토하였으나, 실험 결과 이 두 피쳐가 Ridge 내부 검증에서 정규화 강도 선택을 왜곡하여 22거래일 폴링 R^2 를 약 0.006 하락시키는 것이 확인되어 최종 피쳐 세트에서 제외하였다. 모델 간 비교에서는 동일한 35개 피쳐 세트를 적용하여 비교의 공정성을 확보하였다. 또한 피쳐 간 스케일 차이를 줄이기 위해 표준화를 적용하였으며, 스케일링 파라미터는 학습 구간에서만 추정하고 이를 테스트 구간에 적용함으로써 정보 누출이 발생하지 않도록 하였다. 피쳐의 목록은 표 2에 제시한다.

3.6 학습 과정

시계열 데이터의 시간적 선후 관계를 엄격히 보존하기 위해 전체 표본을 날짜순으로 정렬한 뒤, 전 반부 80%를 모델 학습에, 후반부 20%를 성과 평가에 배정한다. 전방 타깃은 $[t+1, t+h]$ 구간을 사용하므로, 인접 관측치 간 타깃 기간 중첩이 필연적으로 존재한다. 이를 방지하기 위해 학습-테스트 분할 경계에 $\text{purge gap} = h$ 를 적용한다.

하이퍼파라미터 튜닝은 테스트 세트 정보 없이 수행하기 위해 학습 세트 내부를 다시 80/20으로 분할한 내부 검증(Inner holdout) 구조를 사용한다. 이 내부 분할에도 동일하게 $\text{purge gap} = h$ 를 적용하여 타량간 h 에 대해 내부 검증을 독립적으로 수행하여 기간별 최적 하이퍼파라미터를 선택하는 방식을 채택하였다.

주식·채권·원자재 자산군은 변동성 수준과 VIX 민감도가 이질적이므로, 각 자산군에 대해 하이퍼파라미터 탐색과 최종 모델 피팅을 자산군별로 독립적으로 수행한다. 이를 통해 자산군 간 이질성을 반영하면서도 동일한 피쳐·분할 절차를 유지하여 비교의 공정성을 확보하였다.

3.7 모델 구성

본 연구는 해석 용이성과 일반화 성능을 함께 보

기 위해 선형·트리·신경망 모델을 포함한다. 주요 비교군은 표 3과 같다.

표 3. 모델 그룹

Table 3. Model group

Group	Model
Benchmark	HAR-3
Linear	Ridge, Lasso, ElasticNet
Tree	Random Forest, XGBoost, LightGBM
Neural network	MLP, BiLSTM-A (Bi-LSTM + Temporal Attention)
Ensemble	WEEns (Ridge + XGBoost)

3.8 Weighted Ensemble 구성

본 연구는 Ridge와 XGBoost를 결합한 WEEns 모델을 제안하며, 최종 예측값은 다음 식 (2)와 같이 산출된다.

$$\hat{y}_{WEEns} = p_w \hat{y}_{Ridge} + (1 - p_w) \hat{y}_{XGBoost} \quad (2)$$

식 (2)에서 Ridge는 다중공선성에 상대적으로 강하고 선형적 구조를 안정적으로 포착하는 장점이 있으며, XGBoost는 비선형 상호작용을 보정하는 역할을 수행한다. 가중치 p_w 는 고정 값이 아니라 예측 기간별 내부 검증 세트(Inner holdout)의 R^2 최대화를 통해 선택된다. 구체적으로 $p_w \in \{0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ 중 내부 검증 성능이 가장 높은 값을 채택하며, 이 과정에서 테스트 세트 정보는 일절 사용하지 않는다. 예를 들어 22거래일 기간에서는 $p_w = 0.8$ 이 선택되었고, 60거래일에서는 $p_w = 0.5$ 로 기간에 따라 최적 가중치가 달라진다. 이러한 기간별 adaptive 가중치 설계는 단일 고정 가중치 대비 과적합 위험을 줄이면서 각 기간의 예측 구조를 보다 잘 반영할 수 있다.

3.9 하이퍼파라미터

Ridge는 다중공선성이 큰 피쳐, 특히 range 계열에서 안정적인 추정을 위해 사용한다. 본 연구는 자산군 간 이질성을 반영하기 위해 자산군별로 내부

검증을 통해 독립적으로 α 를 선택한다. 또한 Ridge의 규제 강도에 대한 민감도 분석으로 $\alpha \in [0.001, 5000]$ 범위에서 성능이 큰 폭으로 흔들리지 않음을 확인하였다.

ElasticNet은 L1과 L2를 결합하여 확장 피쳐 환경에서 과적합을 줄이기 위한 모델로 포함되며, 22거래일 기간에서 성능 개선이 보고되었다. 최적 파라미터는 $\alpha = 0.01$, $l1_ratio = 0.1$ 로, Ridge 성격 90%와 LASSO 성격 10%를 혼합한 형태이다.

BiLSTM-A(Bi-LSTM + Temporal Attention)는 35개 전체 피쳐를 입력으로 사용하며, 과거 약 1개월에 해당하는 22거래일을 입력 시퀀스 길이로 설정하였다. 이는 22거래일 월간 예측과 입력 기간을 맞춰 단기 패턴 포착에 집중하기 위한 선택이다. Temporal Attention은 가산적(Additive) 방식으로 시퀀스 내 각 시점에 중요도 가중치를 부여하여, 예측에 더 유용한 시점의 정보를 강조한다. 은닉층 크기 {32, 64}와 드롭아웃 {0.1, 0.3}은 내부 검증으로 탐색하며, 자산군별로 독립 선택한다. 재현성을 위해 단일 시드 0을 고정한다.

Random Forest / XGBoost / LightGBM은 35개 피쳐 기반 비교군으로 포함되어, 기간별 성능을 비교한다. 구체적인 하이퍼파라미터 탐색 범위는 표 4에 제시한다. Random Forest / XGBoost / LightGBM은 35개 피쳐 기반 비교군으로 포함되어, 기간별 성능을 비교한다.

표 4. 모델별 하이퍼파라미터 탐색 범위

Table 4. Hyperparameter search space by model

Model	Hyperparameter	SearchSpace
Ridge	α	{0.01,0.1,1,10,100,1000}
XGBoost	n_estimators	{100,300}
	max_depth	{3,5}
	learning_rate	{0.05,0.1}
BiLSTM-A	hiddenunits	{32,64}
	dropout	{0.1,0.3}
WEEns	pw	{0.4,0.5,0.6,0.7,0.8,0.9}

3.10 평가 지표

모델 순위 비교는 폴링 R^2 를 기준으로 수행하며, 중위 R^2 와 RMSE를 보조 지표로 함께 보고한다. 모델 간 차이의 통계적 유의성은 Diebold-Mariano 검정

을 통해 평가하였다 [19]. 예측 타깃이 일정 기간 $[t+1, t+h]$ 의 평균으로 구성되어 인접한 날짜 간에 타깃 기간이 필연적으로 겹치므로, 예측오차 간에는 최대 $h-1$ 차의 자기상관이 존재할 수 있다. 이에 따라 DM 검정에서는 기간 h 를 Newey-West HAC 대역 폭으로 사용하여 중첩 구조에 의한 분산 과소추정을 보정하였다 [20]. 또한 120거래일 이상의 장기 기간에서는 점근 근사의 한계를 보완하기 위해 블록 크기 h , 반복 횟수 5,000회의 블록 부트스트랩을 보조 검정으로 사용하여 유의성을 재확인하였다 [21].

3.11 실험 개요

본 연구는 11개 ETF 패널 자료를 대상으로, 8개 예측 기간에서 3.3절에서 정의한 RV 직접 예측 + VRP 파생 산출의 2단계 구조에 따라 전방 실현분산을 예측하며, 주요 기호와 정의는 표 5와 같다.

예측 시점의 IV 대응치인 IV_t^2 는 이미 관측된 값이므로, 모델 간 비교와 선택은 1단계 RV 예측 성능 비교를 기준으로 일관되게 수행할 수 있다.

표 5. 기호 및 정의

Table 5. Symbol and definition

Symbol	Definition
$RV_{i,t}(h)$	Forward RV of asset i , h trading days ahead (annualized)
IV_t^2	Implied variance proxy, $(VIX_t/100)^2$
$VRP_{i,t}(h)$	Variance risk premium, $IV_t^2 - RV_{i,t}(h)$
h	Forecast horizon (trading days), $h \in \{1,5,22,60,90,120,180,252\}$
i	Asset index ($i=1, \dots, 11$; 11 ETFs)
t	Time index (trading day)
r_t	Daily log return, $\ln(P_t/P_{t-1})$
R-squared	pooled and median R-squared

3.12 공정 비교 설계

모델 간 성능 비교의 공정성을 확보하기 위해, 3.5~3.7절에 기술한 데이터 분할, 피쳐 구성, 하이퍼파라미터 튜닝 절차를 모든 모델에 동일하게 적용

하였다. NaiveRV, GARCH, IV-only 베이스라인 모델은 피쳐 구성이 상이하나 동일한 데이터 분할과 평가 절차를 공유하며, IV-only 모델은 Ridge와 동일한 α 탐색 그리드를 사용한다. 무작위성이 존재하는 모델(BiLSTM-A, Random Forest, XGBoost, LightGBM)은 단일 시드(seed = 0)를 고정하여 재현 가능성을 확보하였다.

IV. 실험 결과

4.1 기간별 모델 R^2

기간별로 가장 높은 풀링 R^2 를 기록한 모델은 다음과 같다. 중위 R^2 기준 비교는 4.3절을 참조한다. 1, 5, 22, 180, 252거래일 구간에서는 WEns가, 60, 90, 120거래일 구간에서는 Ridge가 최고 성능을 보였다. 8개 예측 기간 중 6개에서 WEns 또는 Ridge가 1위를 차지하였다.

1~22거래일 구간에서는 WEns(22d: 0.8044)와 Ridge(22d: 0.8026)가 상위 1~2위를 기록하였다. 60~120거래일 구간에서는 Ridge가 0.757~0.788 범위에서 1위를 유지하였으며, 180~252거래일 구간에서도 WEns가 0.717~0.759로 상위 2위 이내를 유지하였다.

4.2 WEns 성능

22거래일 예측에서 WEns는 가중치 $pw=0.8$ 로 Ridge와 XGBoost를 결합하고 35개 피처를 자산군별로 독립 적용하며 내부 검증 기반으로 가중치를 선택

하는 구조로, 풀링 $R^2 = 0.8044$, Median $R^2 = 0.2469$, RMSE = 0.4591을 기록하였다. Ridge의 풀링 R^2 는 0.8026, ElasticNet은 0.8009로 세 모델 모두 근소한 차이를 보였으며, 모델 간 차이는 통계적으로 유의하지 않은 수준이다. 반면 BiLSTM-A와 MLP는 풀링 R^2 가 각각 0.5361과 0.5644로, 35개 전체 피쳐 환경에서 과적합 문제로 크게 열위를 보였다.

그림 2에서 60거래일 이상으로 기간이 확장될 경우 중위 R^2 가 음수 영역으로 진입하는 현상이 관찰된다. 90d 중위 R^2 는 WEns 기준 -0.074, 252d에서는 -0.849로, 자산별 R^2 분포의 중앙값이 음수임을 의미한다.

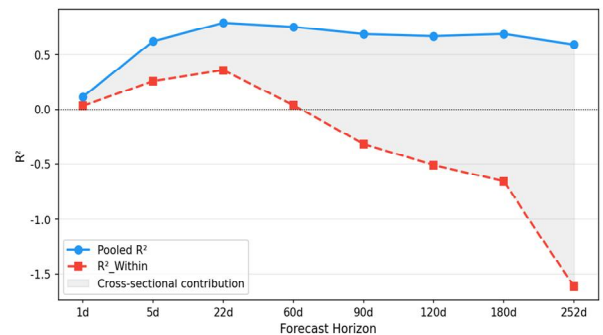


그림 2. 기간별 WEns Pooled R^2 vs R^2 _Within
Fig. 2. WEns Pooled R^2 vs R^2 _Within by horizon

패널 R^2 분해 관점에서, 풀링 R^2 는 어느 자산이 더 변동성이 높은지를 나타내는 자산 간 횡단면 변동과, 같은 자산의 변동성이 시간에 따라 어떻게 달라지는지를 나타내는 시계열 방향 변동을 모두 포함한다. 전체 모델·기간별 풀링 R^2 비교는 그림 3에 제시한다.

Model	1d	5d	22d	60d	90d	120d	180d	252d
WEns	0.118	0.631	0.804	0.788	0.756	0.757	0.759	0.718
Ridge	0.117	0.627	0.803	0.799	0.786	0.775	0.744	0.692
ENet	0.117	0.627	0.801	0.774	0.747	0.747	0.720	0.675
LASSO	0.117	0.622	0.797	0.770	0.749	0.744	0.722	0.683
RF	0.116	0.607	0.777	0.774	0.761	0.753	0.734	0.693
XGBoost	0.111	0.616	0.777	0.725	0.687	0.675	0.719	0.651
LightGBM	0.104	0.610	0.770	0.740	0.697	0.672	0.679	0.641
HAR-3	0.110	0.582	0.766	0.786	0.783	0.770	0.739	0.691
MLP	0.075	0.479	0.564	0.498	0.542	0.527	0.515	0.497
BiLSTM-A	0.072	0.371	0.536	0.638	0.444	0.597	0.561	0.553

그림 3. 전체 모델×기간 Pooled R^2 히트맵
Fig. 3. Model × horizon Pooled R^2 heatmap

높은 풀링 R^2 수치를 시계열 예측력으로 오해하지 않도록, 자산별 평균을 제거한 후 계산한 R^2_{Within} 을 표 6에 함께 제시한다.

표 6. WEns Pooled R^2 vs R^2_{Within} 비교
Table 6. WEns Pooled R^2 vs R^2_{Within} decomposition

day	Pooled R^2	R^2_{Within}	P-W
1d	0.1176	0.0337	+0.084
5d	0.6311	0.2689	+0.362
22d	0.8044	0.3831	+0.421
60d	0.7877	0.1664	+0.621
90d	0.7558	-0.074	+0.830
120d	0.7571	-0.166	+0.923
180d	0.7592	-0.245	+1.004
252d	0.7178	-0.849	+1.567

4.3 중위 R^2

90일 이상 기간에서 중위 R^2 가 음수로 전환되는 현상은 자산군 간 이질성에서 기인한다. 표 6에서 확인되듯, 채권(TLT 0.123, IEF 0.152, AGG 0.159)과 SLV(0.003)의 22d R^2 가 주식(SPY 0.569, QQQ 0.454) 대비 현저히 낮아, 단일 풀링 모델의 중위 R^2 를 음수 방향으로 끌어내린다.

22거래일 기간 WEns의 자산별 R^2 를 표 7에 제시한다. 주식 자산군에서는 SPY(0.569), QQQ(0.454), EFA(0.415)가 상대적으로 높은 예측력을 보였으며, 원자재 중 USO(0.381)도 양호한 수준을 나타냈다. 반면 채권(TLT 0.123, IEF 0.152, AGG 0.159)과 SLV(0.003)는 낮은 R^2 를 보였다.

표 7. 22거래일 WEns 자산별 R^2
Table 7. 22day WEns Per-Asset R^2

Asset	Class	R^2
SPY	Equity	0.569
QQQ	Equity	0.454
EFA	Equity	0.415
EEM	Equity	0.274
IWM	Equity	0.185
AGG	Bond	0.159
IEF	Bond	0.152
TLT	Bond	0.123
USO	Commodity	0.381
GLD	Commodity	0.093
SLV	Commodity	0.003

4.4 다중 모델 공정 비교

동일한 35개 피처, 동일한 자산군별 내부 검증 절차, 동일한 학습-테스트 분할을 적용하여 모든 모델을 공정하게 비교한 결과를 22거래일 기간 기준으로 표 8에 제시한다.

표 8. 22거래일 기간 모델별 성능 비교
Table 8. 22day model performance comparison

Model	Pooled R^2	Median R^2	RMSE
HAR-3	0.7659	0.1702	0.5023
Ridge	0.8026	0.2583	0.4612
ENet	0.8009	0.2619	0.4632
LASSO	0.7973	0.2420	0.4674
RF	0.7775	0.1659	0.4896
XGBoost	0.7773	0.1378	0.4899
LightGBM	0.7697	0.1255	0.4982
BiLSTM-A	0.5361	-0.5286	0.7003
MLP	0.5644	-0.5295	0.6852
WEns	0.8044	0.2469	0.4591

22d 풀링 R^2 기준으로 WEns(0.8044) > Ridge(0.8026) > ENet(0.8009) > LASSO(0.7973) 순으로 규제 선형 계열이 상위 4위를 차지하였으며, 트리 기반 모델인 RF(0.7775), XGBoost(0.7773), LightGBM(0.7697)이 5~7위를 기록하였다. BiLSTM-A(0.5361)와 MLP(0.5644)는 최하위로, 1위 대비 각각 0.268p, 0.244p 낮았다. 이는 자산군별 독립 피팅 구조에서 표본 크기가 제한적인 반면 모델 파라미터 수가 많아 과적합이 발생하기 쉬운 점, BiLSTM-A의 입력 시퀀스 길이인 22거래일이 60거래일 이상 장기 예측에 충분한 시계열 맥락을 제공하지 못하는 점이 복합적으로 작용한 결과로 해석된다. 한편 Median R^2 기준으로는 ElasticNet이 0.2619, Ridge가 0.2583으로 WEns의 0.2469보다 소폭 높게 나타났으며, 모델 선택 기준에 따라 최적 모델이 달라질 수 있음을 보여준다.

4.5 피처 기여

피처 그룹이 실제로 예측력에 기여하는지 확인하기 위해 Ridge 기반 서브셋 실험을 수행하였으며, 결과는 표 9에 제시한다. HAR+Range의 8피처가 풀

링 R^2 0.7899, Median R^2 0.2127로 HAR의 3피처 대비 각각 +0.033, +0.020의 성능 향상을 보였다. Full의 35피처는 0.8026으로 HAR+Range 대비 +0.013의 추가 개선에 그쳐, 피처 수 증가에 따른 한계 효용 감소가 확인되었다.

HAR 기본 3개 피처의 풀링 R^2 가 0.757인 상태에서 출발하여, HAR+Range의 8피처 조합이 0.790으로 상승하고, 35개 전체 피처를 사용한 Full 모델이 0.8026에 이르렀다. Range의 5피처만으로도 0.775에 달해 단기 변동성 패턴 포착에 효과적임을 보였으며, IV Surface의 10피처 단독으로는 0.210에 그쳐 과거 RV 정보 없이는 예측력이 크게 제한됨을 확인하였다. GARCH 계열 피처(GARCH_Daily, GARCH_Weekly)는 초기 검토에서 변수 중요도 상위권에 위치하였으나, Ridge의 내부 검증 기반 정규화 강도 선택을 왜곡하는 부작용이 확인되어 최종 35개 피처 세트에서 제외하였다.

표 9. Ridge 22거래일 피처 그룹 서브셋 실험
Table 9. Ridge 22day feature group subset experiment

Ridge_22day	Feature	Pooled R^2	Median R^2
HAR	3	0.7573	0.1931
Rangeonly	5	0.7749	0.0918
IVSurfaceonly	10	0.2101	-0.7032
HAR+Range	8	0.7899	0.2127
Full	35	0.8026	0.2583

단일 방법의 편향을 줄이기 위해 Ridge 정규화 계수, XGBoost gain, permutation importance의 3개 방법에서 도출한 변수 중요도의 평균 순위를 합의 순위로 산출하였다. 22거래일 기간에서는 RogersSatchell_22, LogRV_lag1, Parkinson_5, GarmanKlass_22, Range_Close_Ratio가 상위권을 차지하였다. HF Proxy(range 기반) 피처가 상위권에 반복적으로 포함되었으며, IV Surface 피처인 IV_VRP_ma22, IV_VIX도 중요도 상위 10위 내에 포함되었다.

이 결과만 보면 HF Proxy가 중요도 순위를 주도하는 것으로 보일 수 있으나, 표 10의 그룹별 제거 실험을 함께 고려하면 해석은 달라진다.

표 10. WEEns 22거래일 기준 피처 그룹 제거 실험
Table 10. WEEns 22day feature group removal experiment

Removed group	Feature	Full R^2	w/o Group	ΔR^2
IVSurface	10	0.8036	0.7872	+0.0164
HFProxy	7	0.8036	0.7883	+0.0153
Alternative	6	0.8036	0.8020	+0.0016

WEEns 기준으로, 풀링 R^2 가 0.8036인 상태에서 IV Surface 변수군의 증분 기여도는 +0.0164로 가장 컸고, HF Proxy는 +0.0153, Alternative Data는 +0.0016이었다.

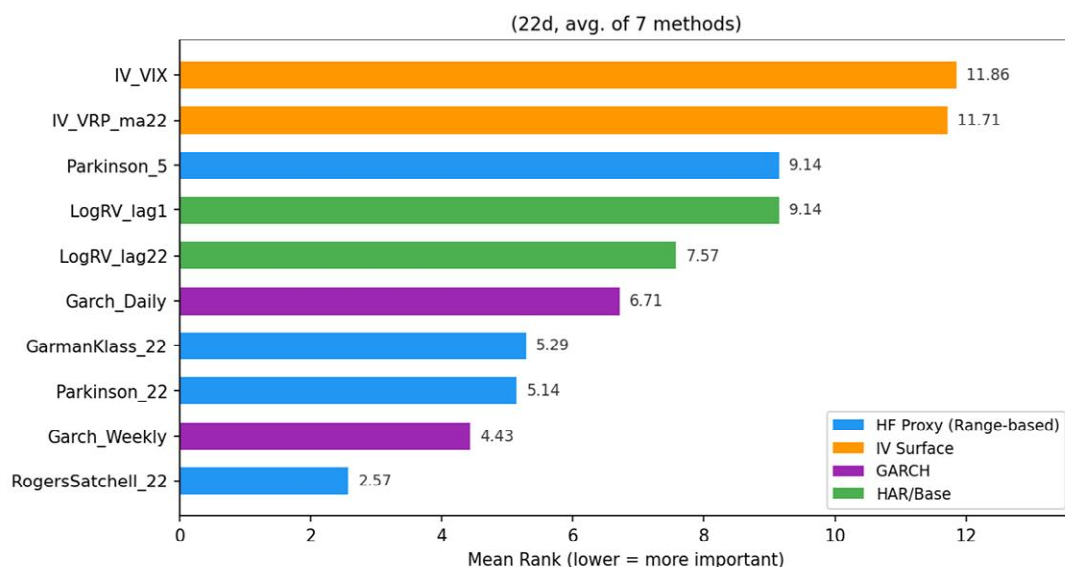


그림 4. 22거래일 피처 중요도 합의 순위 Top-10
Fig. 4. 22day feature importance consensus Top-10

즉, 순위 관점에서는 HF Proxy가 반복적으로 상위권에 등장하지만, 그룹 단위 제거 실험에서는 IV Surface($\Delta R^2=+0.0164$)가 HF Proxy(+0.0153)보다 0.0011 큰 기여를 보였다. 이는 자산의 과거 실현 변동성 정보와 옵션시장이 반영하는 미래 기대 정보가 서로 다른 경로로 예측력에 기여할 수 있음을 보여준다. 이러한 순위와 제거 실험 간 차이는 피처들 사이의 정보 중복에서 비롯되며, 22d 기준 합의 순위 상위 10개는 그림 4에 제시한다.

HF Proxy 피처인 RogersSatchell_22, Parkinson_22, GarmanKlass_22는 서로 비슷한 정보를 담고 있어 개별 순위는 높게 나타나지만, 셋을 한꺼번에 제거해도 나머지 피처가 상당 부분 대체하므로 성능 하락이 상대적으로 작다. 반면 IV Surface는 개별 순위가 낮더라도 HF Proxy가 담지 못하는 옵션 시장 정보를 독립적으로 제공하기 때문에, 해당 그룹을 제거하면 성능이 더 크게 떨어진다. 따라서 두 분석 결과는 모순이 아니라 서로 다른 측면에서 피처 기여를 보여주는 것으로 해석해야 한다.

추가로 선형 모델(Ridge)과 트리 기반 모델(XGBoost, Random Forest) 간 변수 중요도 순위의 합의를 Kendall의 W로 점검하였다. 모든 예측 기간인 1~252d에서 p값이 0.001 미만으로 나타나, 서로 다른 구조의 모델들이 중요하게 보는 변수 순위가 통계적으로 유의하게 일치함을 확인하였다. Kendall의 W는 1일 0.666에서 22일 0.822까지 상승한 뒤 장기 구간에서 완만하게 하락하는 패턴을 보였으며,

22일 기간에서 $W=0.822$ 로 가장 높은 합의를 보였다. 이는 중기 예측 기간에서 모델 구조와 무관하게 공통적으로 유효한 정보 변수가 존재함을 시사한다. 예측 기간별 세부 결과는 표 11에 제시한다.

표 11. 예측 기간별 Kendall's W
Table 11. Kendall's W by Forecast Horizon

Horizon	KendallW	p-value
1d	0.666	<0.001
5d	0.758	<0.001
22d	0.822	<0.001
60d	0.807	<0.001
90d	0.808	<0.001
120d	0.774	<0.001
180d	0.781	<0.001
252d	0.766	<0.001

다만 이러한 결과는 예측상 중요도를 보여주는 것이며, 개별 피처의 인과적 효과를 직접 의미하는 것은 아니다.

4.6 고변동성 구간 분석

그림 5는 2010~2024년 VIX 수준을 시계열로 나타낸 것으로, 분위수 구간(하위 50%, 50~90%, 상위 10%)을 음영으로 표시하였다. COVID-19(2020.03), 유럽 재정위기(2011.08) 등 급등 구간이 상위 10% 국면에 해당한다. 기간별 중요도를 비교하면 핵심 변수군은 예측 기간에 따라 이동하는 패턴을 보였다.

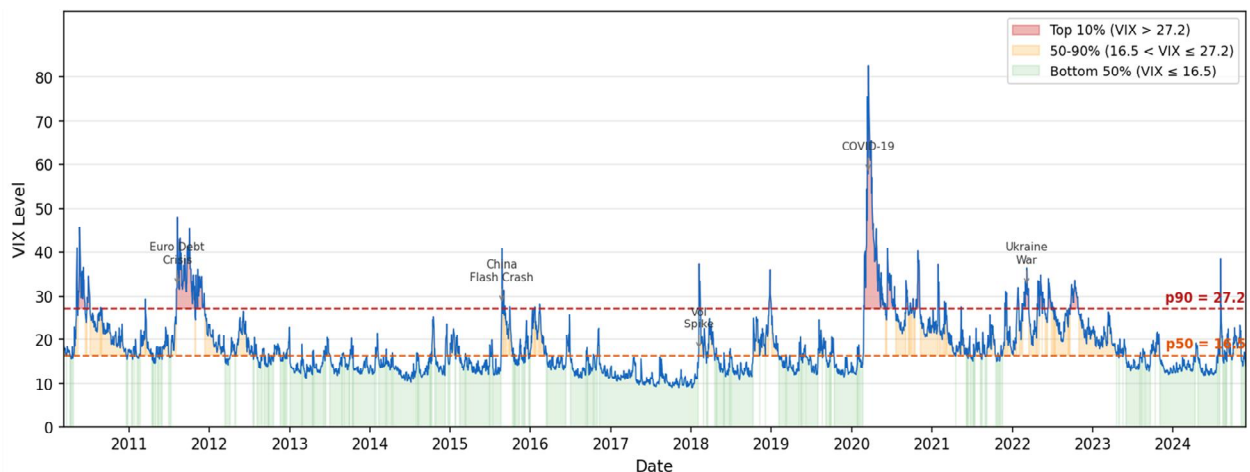


그림 5. VIX 시계열 및 분위수 국면 구분 (2010 - 2024)

Fig. 5. VIX time series with quantile regime shading (2010 - 2024)

1일 기간에서는 Range_Close_Ratio와 RogersSatchell_22가 상위권에 위치하였고, Parkinson_5와 LogRV_lag1이 뒤를 이었다. 5~22일 기간에서는 RogersSatchell_22가 1위로 부상하며 22일 기간에서도 선두를 유지하였고, GarmanKlass_22, Parkinson_5, Parkinson_22 등 22일 윈도우 HF 변수들의 중요도가 상승하였다. 60일 이상의 중장기 구간에서는 RogersSatchell_22, GarmanKlass_22가 계속 상위권을 유지하는 가운데, IV_VRP_ma22의 중요성이 상승하여 장기 예측에서 옵션시장 내재 정보의 역할이 커짐을 확인하였다.

표 12에서는 VIX 분위수 기준 국면 분석 결과, HAR-3 대비 WEns의 개선 폭은 VIX 수준에 따라 달라졌음을 확인할 수 있다.

표 12. 22거래일 VIX 분위수 구간별 Pooled R^2
Table 12. 22day Pooled R^2 by VIX Regime

VIX	HAR-3	WEns	ΔR^2
VIX bottom 50%	0.739	0.781	+0.042
VIX 50~90%	0.698	0.722	+0.024
VIX top 10%	0.602	0.765	+0.163

VIX 하위 50% 구간에서는 R^2 가 0.739에서 0.781로, VIX 50~90% 구간에서는 0.698에서 0.722로 상승하였다. 반면 VIX 상위 10% 구간에서는 0.602에서 0.765로 개선 폭이 가장 크게 나타났다. VIX 상위 10% 구간에서 WEns의 개선폭(+0.163)은 하위 50% 구간(+0.042)과 50~90% 구간(+0.024)보다 각각 3.9배, 6.8배 크다.

4.7 베이스라인과의 비교

본 연구의 RV 예측 성능을 기존 문헌에서 제안된 대표적 방법론들과 비교하기 위해, 세 가지 베이스라인 모델을 자산군별 독립 피팅, 기간별 내부 검증, 동일 학습-테스트 경계를 포함한 동일한 데이터 분할-평가 절차 하에 구현하였다.

첫째, [3]의 접근에서 착안한 단순 과거 실현분산 예측(NaiveRV)으로, h거래일 이동평균 과거 실현변동성을 예측치로 사용한다. 둘째, GARCH(1,1) 기반 예측으로 표준 GARCH 다단계 예측 절차

[7][8]에 따라 각 자산의 학습 세트에서 모수를 추정 후, 테스트 기간 동안 매 시점의 실제 수익률을 반영하여 조건부 분산을 순차적으로 갱신한다. 이후 h일 후까지의 단계별 기대 분산을 평균하여 h-기간 예측치로 변환한다. 이 과정에서 테스트 세트 정보는 모수 추정에 사용되지 않는다. 셋째, IV-only 모델로 [5]의 개념에 따라 VIX 계열 10개 피쳐만을 사용한 Ridge 회귀를 동일한 내부 검증 절차로 피팅한다.

22거래일 기준 폴링 R^2 결과는 다음과 같다: NaiveRV = 0.7139, GARCH(1,1) = 0.6895, IV-only = 0.2094. 제안 모델 WEns는 폴링 R^2 0.8044로 모든 베이스라인을 상회하며, 소표본 보정 DM 검정 [22] 결과 WEns vs NaiveRV의 DM 통계량은 -8.698, WEns vs GARCH는 -12.136, WEns vs IV-only는 -39.456으로 모두 $p < 0.001$ 수준에서 통계적으로 유의한 우위가 확인되었다. 전체 예측 기간에 걸친 DM 검정 결과(BH 다중비교 보정 포함)는 표 13에 제시한다.

표 13. WEns vs 베이스라인 DM 검정 결과 (BH 다중비교 보정)

Table 13. Diebold - Mariano test results: WEns vs. baselines (BH multiple comparison adjustment)

Horizon	HAR-3	NaiveRV	GARCH	IV-only
1d	2.525	32.321	21.037	16.596
5d	8.493	21.738	11.290	38.664
22d	5.395	8.698	12.136	39.456
60d	0.159	0.101	10.983	28.160
90d	-1.885	-0.121	7.643	22.816
120d	-0.717	1.409	8.191	20.683
180d	1.092	3.499	7.284	17.977
252d	0.654	3.570	6.957	12.915

IV-only 모델의 RV 예측 성능은 $R^2 = 0.2094$ 로, 이는 VIX 계열 피쳐가 11개 ETF 전체에 동일하게 적용되어 자산 간 횡단면 변동성 차이를 구분하지 못한 데 기인한다. 이는 자산별 이질성이 큰 다중 자산 패널에서 공통 IV 지표만으로는 RV 예측력이 제한됨을 시사하며, 자산군별 과거 실현변동성 정보와의 결합이 필수적임을 보여준다. GARCH(1,1)의 성능은 $R^2 = 0.6895$ 로 NaiveRV보다 낮게 나타났는데, 이는 단기 GARCH 분산을 h-기간 평균으로 변

환하는 과정에서 발생하는 장기 평균 회귀(Mean reversion) 편향과, 극단적 변동성 국면에서의 모수 불안정성이 복합적으로 작용한 결과로 해석된다.

4.8 확장 피쳐 세트의 기여

[3]의 LASSO-HAR 접근을 직접 모사하기 위해 HAR 3개 피쳐인 LogRV_lag1, LogRV_lag5, LogRV_lag22만 사용한 LASSO를 동일한 학습·테스트 분할 하에 구현하여 비교하였으며, 이하 이 모델을 LASSO-HAR라 한다. 표 14에 기간별 Pooled R² 비교를 제시한다.

표 14. LASSO-HAR vs LASSO 기간별 Pooled R²
Table 14. LASSO-HAR vs LASSO Pooled R² by horizon

	LASSO-HAR	HAR-3	LASSO(35)	WEns	$\Delta(35-LH)$
1d	0.1160	0.1097	0.1173	0.1176	+0.001
5d	0.5669	0.5819	0.6216	0.6311	+0.055
22d	0.7292	0.7659	0.7973	0.8044	+0.068
60d	0.7534	0.7860	0.7705	0.7877	+0.017
90d	0.6931	0.7835	0.7491	0.7558	+0.056
120d	0.6073	0.7700	0.7442	0.7571	+0.137
180d	0.6150	0.7390	0.7219	0.7592	+0.107
252d	0.6070	0.6906	0.6826	0.7178	+0.076

LASSO(35피쳐)는 8개 기간 전체에서 LASSO-HAR(3피쳐)를 상회하며, 성능 격차는 +0.001(1d)에서 +0.137(120d) 범위에 분포한다. 22d 기준 격차는 +0.068이며, 90d 이상 장기 구간에서 격차가 0.056~0.137로 확대된다. 한편 LASSO-HAR(22d: 0.729)는 Ridge 기반 HAR-3(22d: 0.766)보다 1d 구간을 제외한 전 기간에서 낮은 성능을 보였다.

4.9 강건성 검증

WEns 양상블에 대해 수행한 다중 분할 안정성, alpha 민감도, 5-fold 시계열 교차검증, 데이터 누출 테스트, permutation importance, DM 검정 결과는 전반적으로 일관된 패턴을 보였다. Commodity IV 대체(GVZ/OVX) 실험에서는 VIX를 GLD에는 GVZ, USO에는 OVX로 교체하였을 때 전체 풀링 R²가

0.782에서 0.774로 0.008p 하락하였다. 또한 잔차 진단에서 비정규성, 자기상관, ARCH 효과가 관찰되었으나, HAC 기반 검정과 블록 부트스트랩을 통해 통계적 비교의 왜곡 가능성을 완화하였다.

한편 본 연구는 30일 만기 옵션을 기반으로 산출된 VIX를 모든 예측 기간에 동일하게 적용하였으므로, h가 30일을 초과하는 경우 IV의 만기와 예측 기간 사이에 불일치가 존재한다. 예를 들어 252거래일, 즉 1년 예측 기간에 30일 기준 VIX를 사용하면 장기 변동성 기대를 충분히 반영하지 못할 수 있다. 각 기간에 적합한 만기의 IV를 활용하는 방향, 예를 들어 60거래일 기간에는 VIX3M을 사용하는 접근은 향후 연구 과제로 남긴다.

4.10 실험 결과 요약

이상의 결과를 종합하면 다음과 같다. 첫째, WEns는 22d 기준 Pooled R² 0.8044로 8개 예측 기간 중 5개에서 최고 성능을 기록하였다. 둘째, 1~22d에서는 HAR+Range(8피쳐) 조합이 Pooled R² 0.7899를 달성하며 전체 피쳐(0.8026) 대비 0.013p 차이에 그쳤고, 60d 이상에서는 IV_VRP_ma22가 피쳐 중요도 Top-5에 진입하여 장기 예측에서 IV 정보의 기여가 커짐을 확인하였다. 셋째, 22d 기준 WEns(0.8044), Ridge(0.8026), ENet(0.8009), LASSO(0.7973)가 1~4위를 차지한 반면, BiLSTM-A(0.5361)와 MLP(0.5644)는 최하위로 1위 대비 0.268p, 0.244p 낮았다. 넷째, WEns는 NaiveRV·GARCH(1,1)·IV-only 대비 22d Pooled R²를 각각 +0.090p · +0.115p · +0.595p 상회하며(p < 0.001) 통계적으로 유의한 우위를 보였다. 다섯째, LASSO(35피쳐)는 LASSO-HAR(3피쳐) 대비 8개 기간 전체에서 우수하며, 격차는 22d +0.068p, 120d +0.137p이다.

V. 결 론

본 연구는 누구나 접근할 수 있는 일별 가격·거래량 자료와 변동성 지표만으로, 주식·채권·원자재를 포괄하는 다중 자산 실현분산(RV) 예측 프레임워크를 제안하였다.

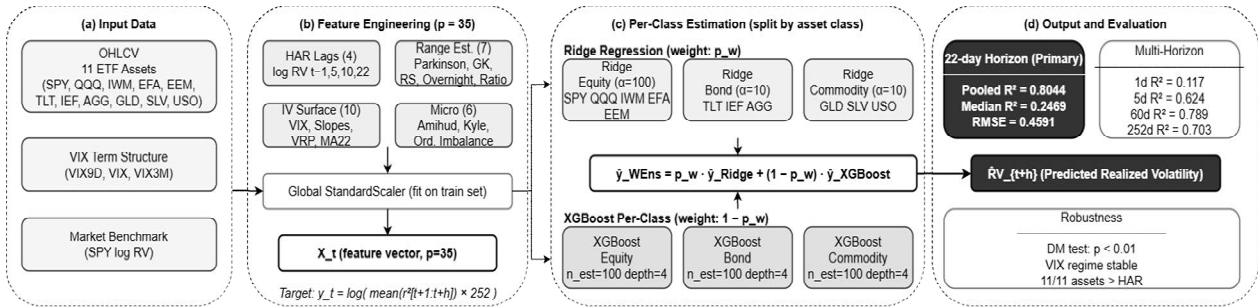


그림 6. WEns 모델의 전체 흐름
Fig. 6. Overall workflow of the WEns model

2010년부터 2024년까지의 11개 ETF 자료를 바탕으로, 미래 RV를 직접 예측하고 그 결과를 IV와 결합해 VRP를 도출하는 2단계 구조를 사용하였다. 본 연구의 전체 모델 흐름은 그림 6에 제시한다.

실증 결과, 1~252거래일 전 구간에서 Ridge와 WEns가 일관되게 1~2위를 유지하였다. 이는 특정 예측 기간에만 유효한 기법이 아니라, 단기부터 연간에 이르기까지 안정적으로 작동하는 구조임을 뜻한다. 특히 22일 예측에서 WEns는 Ridge와 XGBoost를 결합해 두 모델 단독 성능보다 높은 0.8044를 기록하였으며, 이는 선형 모델에 비선형 모델의 보정이 더해질 때 소폭이나마 성능이 개선됨을 보여준다. 또한 전방 채움 제거, VIX 만기 점검, 자산별 변동성 지표 교체 실험에서 풀링 R^2 가 0.774~0.782 범위로 유지되어, 보고된 성능이 특정 전처리나 지표 선택에 좌우되지 않는 견고한 결과임을 확인하였다.

변수 중요도 분석에서는 Ridge, XGBoost, permutation 세 가지 방법이 22일 예측에서 Kendall's $W = 0.822$ 의 강한 합의를 보였다. 이는 중요하게 평가된 변수들이 특정 모델에만 치우친 결과가 아니라, 데이터에 공통으로 존재하는 신호임을 뒷받침한다. 다만 개별 변수 순위와 그룹 단위 기여는 일치하지 않았다. HF Proxy 변수들은 개별 중요도는 높았으나 서로 비슷한 정보를 공유하여 일부를 제거해도 남은 변수가 이를 대신할 수 있었다. 반면 IV Surface 변수들은 옵션시장이 반영하는 미래 기대 정보를 고유하게 담고 있어 대신할 변수가 없어 제거 시 성능 하락이 더 크게 나타났다. 따라서 예측 시스템을 설계할 때는 개별 변수의 순위보다 각

그룹이 다른 변수로 대체하기 어려운 정보를 얼마나 담고 있는지를 함께 보아야 한다는 점을 확인할 수 있다.

본 연구에는 몇 가지 한계가 있다. 첫째, 일별 시가·고가·저가·종가 자료로는 하루 안의 급격한 가격 점프와 완만한 변동을 구분할 수 없으며, 이는 고빈도 자료에서만 가능한 분석이다. 둘째, 모든 예측 기간에 30일 만기 VIX를 동일하게 사용하여 1개월을 넘는 장기 예측에서는 VIX의 기간과 예측 기간이 맞지 않는다. 셋째, 자산군 간 차이는 별도로 학습으로 반영하였으나 시장 국면 전환을 모델이 스스로 감지하지는 못하며, 고변동성 구간의 성능 개선도 사후적으로 구간을 나눈 결과일 뿐이다. 넷째, 예측 정확도만 평가하였을 뿐 예측값이 실제 투자 수익에 기여하는지는 검증하지 않았다.

이러한 한계를 보완하기 위해 향후 연구에서는 예측된 RV를 활용한 변동성 타이밍 전략의 수익성 평가, 자산군별 특성에 적응하는 모델, 시장 국면 변화에 따라 가중치를 자동 조정하는 동적 앙상블, 그리고 자산별 옵션 정보를 직접 활용한 내재변동성 측정으로 확장할 수 있을 것이다.

References

- [1] F. Corsi, "A simple approximate long-memory model of realized volatility", *Journal of Financial Econometrics*, Vol. 7, No. 2, pp. 174-196, Mar. 2009. <https://doi.org/10.1093/jfincnec/nbp001>.
- [2] S. Gu, B. Kelly, and D. Xiu, "Empirical asset pricing via machine learning", *The Review of*

- Financial Studies, Vol. 33, No. 5, pp. 2223-2273, May 2020. <https://doi.org/10.1093/rfs/hhaa009>.
- [3] F. Audrino and S. Knaus, "Lassoing the HAR model: A model selection perspective on realized volatility dynamics", Journal of Financial Econometrics, Vol. 14, No. 2, pp. 392-427, 2016. <https://doi.org/10.1093/jjfinec/nbv015>.
- [4] K. Christensen, M. Siggaard, and B. Veliyev, "Volatility forecasting with machine learning and intraday commonality", Journal of Financial Econometrics, Vol. 22, No. 1, pp. 1-35, Jan. 2024. <https://doi.org/10.1093/jjfinec/nbac020>.
- [5] P. Carr and L. Wu, "Variance risk premia", The Review of Financial Studies, Vol. 22, No. 3, pp. 1311-1341, Mar. 2009. <https://doi.org/10.1093/rfs/hhn038>.
- [6] T. Bollerslev, G. Tauchen, and H. Zhou, "Expected stock returns and variance risk premia", The Review of Financial Studies, Vol. 22, No. 11, pp. 4463-4492, Nov. 2009. <https://doi.org/10.1093/rfs/hhp008>.
- [7] R. F. Engle, "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation", Econometrica, Vol. 50, No. 4, pp. 987-1007, Jul. 1982. <https://doi.org/10.2307/1912773>.
- [8] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity", Journal of Econometrics, Vol. 31, No. 3, pp. 307-327, Apr. 1986. [https://doi.org/10.1016/0304-4076\(86\)90063-1](https://doi.org/10.1016/0304-4076(86)90063-1).
- [9] T. G. Andersen, T. Bollerslev, F. X. Diebold, and P. Labys, "Modeling and forecasting realized volatility", Econometrica, Vol. 71, No. 2, pp. 579-625, Mar. 2003. <https://doi.org/10.1111/1468-0262.00418>.
- [10] G. Bekaert and M. Hoerova, "The VIX, the variance premium and stock market volatility", Journal of Econometrics, Vol. 183, No. 2, pp. 181-192, Dec. 2014. <https://doi.org/10.1016/j.jeconm.2014.05.008>.
- [11] I. Drechsler and A. Yaron, "What's vol got to do with it", The Review of Financial Studies, Vol. 24, No. 1, pp. 1-45, Jan. 2011. <https://doi.org/10.1093/rfs/hhq085>.
- [12] B. Zhang, Y. Chen, and Y. Lu, "Nonparametric estimation of variance risk premiums", Journal of Financial Economics, Vol. 142, No. 1, pp. 312-328, Oct. 2021. <https://doi.org/10.1016/j.jfineco.2021.05.008>.
- [13] P. Carr, L. Wu, and Z. Zhang, "Using machine learning to predict realized variance", arXiv preprint, arXiv:1911.03307, Nov. 2019. <https://doi.org/10.48550/arXiv.1911.03307>.
- [14] M. Hattori, I. Shim, and Y. Sugihara, "Cross-stock market spillovers through variance risk premiums and equity flows", Journal of International Money and Finance, Vol. 119, Art. no. 102480, Dec. 2021. <https://doi.org/10.1016/j.jimonfin.2021.102480>.
- [15] J. R. H. Ornelas and R. B. Mauad, "Volatility risk premia and future commodity returns", Journal of International Money and Finance, Vol. 96, pp. 341-360, Sep. 2019. <https://doi.org/10.1016/j.jimonfin.2017.07.008>.
- [16] T. Dierckx, M. Davis, and W. Schoutens, "Machine learning for the FX variance risk premium", Quantitative Finance, Vol. 22, No. 1, pp. 129-143, Jan. 2022. <https://doi.org/10.1080/14697688.2021.1983063>.
- [17] S. Heston and V. Todorov, "Exploring the variance risk premium across assets", AEA Conference Paper, 2024.
- [18] M. L. de Prado, "Advances in Financial Machine Learning", Wiley, Feb. 2018.
- [19] F. X. Diebold and R. S. Mariano, "Comparing predictive accuracy", Journal of Business & Economic Statistics, Vol. 13, No. 3, pp. 253-263, Jul. 1995. <https://doi.org/10.1080/07350015.1995.10524599>.
- [20] W. K. Newey and K. D. West, "A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix",

Econometrica, Vol. 55, No. 3, pp. 703-708, May 1987. <https://doi.org/10.2307/1913610>.

[21] H. R. Künsch, "The jackknife and the bootstrap for general stationary observations", The Annals of Statistics, Vol. 17, No. 3, pp. 1217-1241, Sep. 1989. <https://doi.org/10.1214/aos/1176347265>.

[22] D. I. Harvey, S. J. Leybourne, and P. Newbold, "Testing the equality of prediction mean squared errors", International Journal of Forecasting, Vol. 13, No. 2, pp. 281-291, Jun. 1997. [https://doi.org/10.1016/S0169-2070\(96\)00719-4](https://doi.org/10.1016/S0169-2070(96)00719-4).

김 건 우 (Gun-Woo Kim)



2006년 12월 : 호주뉴캐슬대학교
컴퓨터공학과(공학사)

2007년 9월 : 호주뉴캐슬대학교
정보공학과(공학석사)

2017년 8월 : 한양대학교
컴퓨터공학과(공학박사)

2021년 9월 ~ 현재 :

경상국립대학교 컴퓨터공학과 부교수

관심분야 : 인공지능, 시멘틱 헬스케어, 데이터마이닝

저자소개

박 상 훈 (Sang Hoon Park)



2023년 2월 : 경상국립대학교
컴퓨터공학과(공학사)

2024년 9월 ~ 현재 :

경상국립대학교 AI융합공학과
석사과정

관심분야 : 인공지능,

거대언어모델, XAI, 금융모델링

성 아 영 (Ayeong Seong)



2024년 2월 : 경상국립대학교
컴퓨터공학부(공학사)

2026년 2월 : 성균관대학교
인공지능융합학과(공학석사)

2026년 3월 ~ 현재 :

경상국립대학교 RISE사업단
연구원

관심분야 : 인공지능, 멀티모달, 컴퓨터 비전, 자연어
처리

하 유 진 (Yu Jin Ha)



2023년 2월 : 경상국립대학교
컴퓨터공학부(공학사)

2026년 2월 : 경상국립대학교
AI융합공학부(공학석사)

2026년 3월 ~ 현재 :

경상국립대학교 RISE사업단
연구원

관심분야 : 인공지능, 멀티모달, 컴퓨터 비전, 헬스케어