

SEMD 분해 및 LSTM 모델을 활용한 제조업 업황실적 예측 모델 구축 및 영향 요인 분석

김대영¹⁾

요약

최근 코로나-19 팬데믹을 거치며 제조업 분야의 업황에 내포된 불안정성이 상당 수준 증가하는 현상이 확인되었다. 이에 본 연구에서는 해당 불안정성을 개선하기 위한 정책 수립 등에 직간접적으로 기여하고자, 제조업 업황 실적을 한 분기 이상 선제적으로 예측할 수 있는 딥러닝 모델을 구축하는 한편, 제조업 업황실적에 시차를 두고 영향을 미치는 요인들의 한계 기여 경향 등을 파악하고자 하였다. 본 연구는 SEMD 분해를 기반으로 한 신호 추출 및 NMIFS-HHO 알고리즘을 기반으로 한 주요 변수 추출을 활용하여 4개월 후 제조업 업황 예측에 유의미하게 기여하는 것으로 파악되는 설명변수 집합을 도출, 이를 기반으로 LSTM을 전제로 한 순환신경망을 학습시킴으로써 제조업 업황실적을 선제적으로 예측할 수 있는 딥러닝 모델을 구축하였다. 연구 결과, 본 모델은 매우 적은 오차 아래 정밀한 제조업 업황실적 예측을 전개할 수 있음이 확인되며, 이때 SHAP 분석을 기반으로 변수별 한계기여도를 확인한 결과 자기계열상 약 3~4년 주기의 장기신호, 경기종합지수, 정부대출금리 등이 예측상 주요 설명변수로서 기능함이 확인되었다.

주요용어 : 제조업, 업황실적, 순환신경망, SEMD 분해, NMIFS-HHO

1. 서론

최근 코로나-19 팬데믹을 거치며 제조업 분야의 업황 내 불안정성이 급증하는 현상이 대두되었다. 일례로, 한국은행 경제통계시스템(ECOS)의 기업경기실사지수 통계 가운데 제조업 분야의 월별 업황 실적 자료 의거 시, 2010년 1월부터 2019년 12월까지의 제조업 업황실적 분산은 92.95였던 반면, 코로나-19 팬데믹이 시작된 2020년부터 현재 2023년 11월까지의 월별 제조업 업황실적 자료에 대한 분산은 약 316.36로 도출되었다. 이는 근 3년간 제조업 업황실적의 분산이 지난 9년의 분산에 비하여 약 4배 가량 폭증하였음을 내포하며, 따라서 현재 제조업계의 안정적 수입 및 경영활동은 코로나-19 이전 시기에 비하여 상당 부분 저해되었다고 판단되는 바이다. 또한, 제조업 분야에서의 인력 규모 감소 또한 최근 큰 문제로서 대두되고 있다. 통계청의 경제활동인구조사 및 2023년 지역별 고용조사 분석결과, 15세 이상 인구 규모는 일정하게 유지된 반면, 경제활동인구는 28,096(천명)(2022년 1월)에서 29,375(천명)(2023년 11월)로 증가함으로써 전체 시장의 측면에서는 경제활동 참가율의 증가가 발생하였음이 확인

1) 서울시 서대문구 연세로 50, 연세대학교 응용통계학과, 학사과정.
E-mail: lgtlsafg@yonsei.ac.kr

되었으나, 제조업 분야에서는 산업중분류상 전자부품, 컴퓨터, 영상, 음향 및 통신장비 제조업, 산업소분류상 기타금속 가공제품 제조업·전자 부품 제조업·직물제품 제조업·봉제의복 제조업 모두에서 전년 대비 취업자 감소가 발생함은 물론, 감소규모 또한 약 15,000명에서 22,000명에 달하였음이 확인되었다. 본 연구는 해당 문제의 심각성, 그리고 제조업이 시장에서 갖는 중요성을 고려하여 제조업 업황 불안정성을 축소하기 위한 노력이 전개될 필요성이 존재한다고 생각하였으며, 나아가 업황 실적에 영향을 미치는 요인들의 추출 및 요인별 한계기여도를 파악할 필요성이 또한 존재한다고 판단하였다.²⁾ 본 연구는 이에 선제적으로 월별 제조업 업황실적을 정밀하게 예측할 수 있는 모델을 구축, 이를 기반으로 한 영향 요인 분석을 전개함으로써 제조업 업황의 불안정성을 사전에 축소할 수 있는 정책적 대응을 제안하는 것을 주요 목적으로서 설정하였다. 구체적으로 본 연구는 통계적·경험적 모드 분해를 통해 제조업 업황실적 시계열 내 존재하는 고유 모드 함수들을 추출, 이를 주기에 따라 단기, 중기, 장기 신호로서 분류하여 분석, 이와 함께 NMIFS-HHO 알고리즘을 기반으로 한 변수선택을 전개함으로써 주요 설명변수 집합을 도출, 이를 전제로 LSTM 기반 순환신경망을 학습시켜 최종적으로 예측 모델을 구축하였으며, 해당 모델에 대한 XAI 분석으로서 샵플리 값 분석을 전개하여 변수별 중요도 및 한계기여 방향을 파악하여 정책 제안을 전개하고자 하였다.

2. 이론적 배경 및 선행연구 검토

2.1 선행연구 검토

본 연구에서는 제조업의 업황 실적 변화에 대한 선제적 대응을 취하고자, 제조업의 매출과 산업발달 전반에 영향을 미치는 것으로 나타난 주요 요인들을 파악한 후, 해당 요인들이 업황실적에 선제적 영향 요소로 기능할 수 있음을 고려하여 연구의 설명변수로서 설정 및 활용하고자 하였다. 이에 관련 선행연구 분석 결과, 다음과 같은 영향 요인들이 제조업 매출에 영향을 미치는 것으로 나타났다. 먼저 전수민·이기성(2019)은 중분류상 제조업 매출액 증가율을 종속변수로 둔 패널 회귀분석 모형 활용하여 노동소득분배율, 노동생산성(부가가치/취업자 수), 자본집약도 등이 제조업 매출에 유의한 음의 방향으로의 영향을 미칠 수 있음을 확인하였으며, 실질GDP증가율과 연구개발집중도, 차입금평균이자율(시중금리), 전기 매출액 증가율 등이 매출성장성에 유의하게 기여함을 제시하였다. 다음으로 김바우 등(2021)은 코로나-19 시기 경제변동을 고려하여 균형가격 결정모형과 한국은행 2019년 산업연관표 데이터를 활용하여 원유 및 천연가스, 철광석, 비금속광물의 중량당 원자재 단가 변화가 산업별 생산비에 미치는 영향을 확인하였으며, 이때 제조업계의 경우 각 원자재의 2017년부터 2021년 3분기 단가 변화

2) 2022년 기준 실질 GDP 대비 제조업 GDP는 약 27%에 달하며, 절대적 제조업 총생산 수준은 525,452.5십억원에 달한다.

분에 따라 약 2.8%, 0.62%, 0.04%의 생산비 인상효과가 발생함이 확인되었다. 한편, 이동주(2020)는 확률적 프린티어 생산함수 모형을 전제로 중요소생산성상 기술진보가 증가할수록 기계 산업의 수출 경쟁력이 증가함과 기술진보와 규모효과의 증가는 전자 산업의 수출 경쟁력 제고 영향을 미침을 확인 및 제시하였으며, 주연화 등(2015)은 산업의 생산성, 마크업 해외중간재 등 제조업계 수출 성과를 결정하는 산업 내 요인을 확인하고자 고정 및 확률효과모형을 동시 적용 및 분석함으로써 산업의 자본집약도 및 해외중간재 비중이 유의한 영향 요인으로서 고려될 수 있음을 확인 및 제시하였다. 나아가 오혜진 등(2020)은 Global Value Chain(GVC) 전방 참여 여부가 금융위기 이후 제조업계의 TSI와 수출액, 양측 모두에서 양의 방향으로 유의하게 영향을 미쳤음을 제시하였으며, 권평오(2013)는 확률효과모형을 활용하여 해외직접투자 증가율, 전기 국내설비 투자율 등 요소가 제조업의 수출증가율 변화에 통계적으로 유의한 영향을 미침을 확인하였으며 유희경(2005)은 거시 경영환경요인에 집중하여 제조업계의 매출 영업이익과 관련하여서는 경상수지, 노동소득분배율, 그리고 도시근로자가구 평균소비성향이, 경상이익률에는 경제성장률, 어음부도율 등이 통계적으로 유의한 영향을 미치는 요인으로서 고려될 수 있음을 제시하였다. 또한, 권선희·이기성(2014)은 고정효과 회귀모형을 활용하여 2000년~2012년 교역액 상위 11개국 대상 환율변화가 제조업체의 수익률에 유의한 영향을 미침을 확인하였으며, 김종훈·남현정(2023)은 2019년 사업체패널조사 및 분산최대화회전(varimax) 주성분 분석을 전제로 한 요인분석 및 다중회귀분석을 활용함으로써 제조업체와 ICT 사이 관계규명을 시도, 그 결과 ICT 투자는 ICT 스마트화(융합)에 긍정적 영향을 미치며, 이에 따른 ICT 스마트화의 증진은 제조업체의 원가 절감, 품질 향상, 납기 단축, 생산성 향상에 통계적으로 유의한 양의 방향으로의 영향을 미치는 것을 확인하였다.

선행연구 검토 결과, 제조업의 업황 및 발달에는 크게 노동생산성, 자본집약도, 연구개발집중도, 시중금리, 원자재 단가 변화, 중요소생산성상 기술진보와 규모효과의 증가, GVC 전방참여 여부, 해외직접투자·국내설비투자율, 경상수지, 도시근로자가구의 평균소비성향, 환율변화, ICT 스마트화 등 요소가 중요한 역할을 수행하는 것으로 확인되었다. 본 연구에서는 해당 요소들을 활용, 이에 직간접적으로 대응될 수 있는 데이터들을 종합하여 설명변수 집합을 가정함으로써 유의한 분석을 전개하고자 하였다.

2.2 분석 방법론의 이론적 배경-SEMD 분해

본 연구에서는 계절성, 주기성 등의 개별 영향이 중첩됨에 따라 일반적인 단변량 시계열 자료가 형성되는 현상을 고려하여 제조업 월별 업황실적 시계열에 대한 자료 분석을 전개하였다. 이에 따라 본 연구에서는 고속 푸리에 변환, 변동 모드 분해법 등과 같은 시계열 자료 분해법 가운데 비정상(non-stationary) 시계열 자료에도 광범위하게 적용할 수 있는 동시에 유의한 신호 분해를 수행할 수 있는 경험적 모드 분해, 즉 Empirical Mode Decomposition(EMD), 그중에서도 Statistical EMD(SEMD)를 진행함으로써 시계열 분석 및 개별 고유 모드 함수(Intrinsic Mode Function(IMF))에 따른 설명변수 생성을 완료하였다. 여기서 EMD는 Huang et al.(1971)이 제시한 분해법으로,

비선형적, 비정상 시계열 등 모든 종류의 시계열에 대해 국소적 최댓값과 최솟값들을 전제로 한 보간법(interpolation)을 원시계열의 지점마다 실시하여 연결, 이를 바탕으로 원시계열에 대한 일종의 천장과 바닥이라 할 수 있는 띠(envelope)를 생성한다. 이때 EMD는 형성된 띠의 평균치를 원시계열에서 제하는 과정을 반복, 해당 과정을 통해 도출되는 자료가 두 가지 조건(극대 및 극솟값의 수와 zero-crossing의 개수 차이가 1 이하, 상방 및 하방 띠의 평균치가 0)을 만족하는 경우, 이를 고유 모드 함수(IMF)로 정의하고 해당 IMF를 원자료에서 제한 다음 앞선 과정 전반을 반복하는 방식의 알고리즘을 지속적으로 전개함으로써 원시계열을 다수의 IMF와 이로 분류되지 않는 잔차(residual) 시계열로 분해한다. 이를 통해 EMD는 원시계열 자료에 존재하는 단기적·장기적 주기와 계절성 전반을 추출할 수 있다는 장점을 갖게 된다.

이때 EMD는 띠 형성 과정에서 극대, 극솟값에 대한 보간법을 활용한다는 점에서 오버슈팅 또는 언더슈팅, 즉 분해되는 원시계열의 흐름 또는 진동의 위치에 비해 과다하게 높거나 낮은 띠가 발생할 가능성이 높다는 단점을 갖는다. 이를 해결하기 위해 Kim et al.(2012)은 전통적인 EMD가 갖는 한계를 극복하는 대안으로서 Statistical EMD(SEMD)를 상정하였다. SEMD는 기존의 띠 형성 과정에서의 보간법을 가우시안 커널 등을 활용한 평활법(smoothing)으로 전환하는 방식으로 변경함으로써 오버슈팅 등 극적인 변동 발생을 제한할 수 있게 된다(알고리즘 1).

알고리즘 1: < SEMD(Statistical EMD) Process >

- 1) 원 시계열 $x(t)$ 에서 국소적 최댓값과 최솟값의 집합 Z 추출
 - 2) Gaussian kernel 기반 상부 envelope U_λ , 하부 envelope L_λ 형성
(Kernel $\sim N(0,1)$, λ : 평탄화모수)

$$ex) U_\lambda(t) = \frac{1}{m\lambda} \sum_{i=1}^m z_{(i)} K\left(\frac{(t-t_i)}{\lambda}\right) \quad (m: \text{샘플 수}, z_{(i)}: Z \text{ 내 } i\text{번째 극대값})$$

$$(t_i: z_i^{(1)} \text{ 대응 시점}, K(\cdot): \text{kernel}, \lambda \text{는 } k\text{겹 교차검증으로 최적화})$$
 - 3) 국소적 평균치 도출에 따른 시계열 생성: $m_\lambda = \frac{1}{2}(U_\lambda + L_\lambda)$
 - 4) 후보군 $h_{1,\lambda} = x(t) - m_\lambda$ 도출, 후보군에 대해 1 ~ 3) 과정 IMF 조건 만족시까지 반복
 - 5) 4)에 따른 최종 $h_{1,\lambda} \equiv IMF_1$ 로 설정, $r_\lambda = x(t) - IMF_1$ 도출
 - 6) 1)부터 5) 반복을 통한 k 개 IMF를 도출, 나머지 자료 잔차계열로 분류.
-

이때 SEMD는 평탄화 모수 λ 에 대한 최적화 방안으로서 k -겹 교차검증을 활용하며, 이를 통해 분해 알고리즘 진행 시 특정 데이터 범위, 즉 시계열상 특정 시기에 국한된 커널 기반 평활화가 진행되지 않도록 하여 더욱 탄력적인 IMF 추출이 가능하도록 한다. 본 연구에서는 SEMD를 실현하고자 Kim et al.(2012)이 상정한 R 패키지인 “EMD” package 내 `semd()` 함수를 활용하여 분석을 전개, 이를 통해 제조업 업황실적 시계열의 움직임 내 존재하는 규칙적 신호 등 정보를 취득하고자 하였다.

3. 연구 방법론

3.1 분석 데이터

<표 3.1> 분석 데이터셋

변수 대분류	변수 중분류	변수 종류	데이터 출처
반응변수	제조업 업황실적	월별 제조업 업황실적 (매출액 기준)	ECOS
설명변수	경제·재정 변수	환율(원/달러)	ECOS
		정부대출금리	ECOS
		한은기준금리	ECOS
		통합재정수지	KOSIS
		경제심리지수	ECOS
		경제활동동인구	ECOS
		경제활동참가율	ECOS
	경기 변수	동행종합지수	ECOS
		선행종합지수	ECOS
		후행종합지수	ECOS
		생산자물가지수	ECOS
		설비투자지수	ECOS
		해외직접투자 금액	ECOS
	원자재 변수	월평균 두바이유 추이(\$/배럴)	KOSIS
		월평균 브랜트유 추이(\$/배럴)	KOSIS
		월평균 WTI 추이(\$/배럴)	KOSIS
		월별 원유 도입단가(\$/배럴)	KOSIS
		월별 원유 도입물량 (백만배럴)	KOSIS
	자기계열 변수	업황실적 IMF1	ECOS
		업황실적 IMF2	ECOS
		...	...
		업황실적 IMFk	ECOS

본 연구에서는 앞서 제조업 업황에 영향을 미치는 것으로 나타난 노동생산성, 자본 집약도, 연구개발집중도, 시중금리, 원자재 단가 변화, 중요소생산성상 기술진보와 규모효과의 증가, GVC 전방참여 여부, 해외직접투자·국내설비투자율, 경상수지, 도시근로자가구의 평균소비성향, 환율변화, ICT 스마트화 등에 대응될 수 있는 설명변수들을 가정, 이를 기반으로 한 데이터 수집을 진행하였다(표 3.1). 이때 본 연구에서는 대응되는 설명변수를 크게 경제·재정 변수, 경기 변수, 원자재 변수, 그리고 자기계열 변수로 분류, 이때 세부 변수로는 총 19개 변수를 상정하여 데이터셋을 구축하였다. 먼저 환

율변화, 시중금리, 경상수지 등에 대응될 수 있는 변수로는 월별 평균 환율(원/달러), 정부대출금리, 한은기준금리, 통합재정수지, 경제심리지수, 경제활동인구. 그리고 경제활동참가율을 상정하였다. 다음으로 선행연구에서 확인된 도시근로자가구의 평균소비성향, 노동생산성, 국내설비투자율 등에 대응될 수 있는 설명변수로는 월별 동행종합지수, 생산자물가지수, 설비투자지수. 해외직접투자 금액 등을 상정하였다. 한편 원자재 단가 변화에 대응될 수 있는 변수로는 두바이유, 서부텍사스유 등 원유의 월별 가격과 도입단가, 도입물량 등을 상정하였으며, 본 연구에서는 추가적으로 자기계열 상 중첩 신호가 향후 본계열에 미치는 영향력이 유의할 수 있음을 가정함으로써 반응변수 기준 4개월 전 IMF들을 설명변수로서 또한 상정하였다(각 데이터는 월별 자료로서 수집됨). 이때 본 연구에서는 선제적 제조업 업황 실적 예측을 주요 목적으로 한다는 점과 분기(3개월)별 정책 대응을 추구한다는 점에서 탄력적 행정을 위한 약 1개월의 처리 기간을 전제로 반응변수와 설명변수 사이 4개월 시차를 설정, 이에 따라 해당 시차를 고려함으로써 설명변수의 분석범위로 2009년 9월부터 2023년 7월 데이터가 설정되었으며, 반응변수 데이터로는 2010년 1월부터 2023년 11월 매출액 기준 제조업 업황실적 기업경기실사지수(BSI) 데이터가 설정되었다.

3.1.1 설명변수

본 연구에서는 선행연구에서 확인된 요인 가운데 환율변화, 시중금리, 경상수지 등에 대응될 수 있는 변수로는 경제·재정 설명변수(월별 평균 환율(원/달러), 정부대출금리, 한은기준금리, 통합재정수지, 경제심리지수, 경제활동인구. 그리고 경제활동참가율)를 상정, 도시근로자가구의 평균소비성향, 노동생산성, 국내설비투자율 등에 대응될 수 있는 설명변수로는 경기 설명변수(월별 동행·선행·후행 종합지수, 생산자물가지수, 설비투자지수. 해외직접투자 금액)를, 원자재 단가 변화에 대응되는 변수로는 원자재 변수(12개월 기준 월별 두바이유, 브렌트유, 서부텍사스유(WTI) 가격, 배럴당 도입단가, 도입물량)를 상정하였다. 나아가 본 연구는 추가적으로 자기계열 상 중첩 신호가 향후 본계열에 미치는 영향력이 유의할 수 있음을 가정함으로써 제조업 업황실적 시계열의 SEMD 분해를 통해 발생하는 과거 IMF들을 설명변수로서 함께 상정하였다. 이때 본 연구에서는 선제적 예측 가능성을 고려하여 반응변수와 4개월 시차를 두어 설명변수 집합을 추출하였으며, 모든 데이터는 월별 데이터로서 ECOS 및 KOSIS에서 수집되었다.

3.1.2 반응변수

본 연구에서는 제조업계의 업황을 4개월 선제적으로 예측하는 모델을 구축, 이를 기반으로 정밀한 변동폭 예측 및 안정화 방안 도모를 추구한다는 점에서 연구의 반응변수로서 ECOS에서 제공하는 월별 제조업 매출액 기준 업황 실적 BSI 데이터를 상정 및 활용하였다. 이때 반응변수의 분석범위는 설명변수와 시차를 고려하여 2011년 1월부터 2023년 11월까지를 분석 기간으로 설정하였다.

3.2. NMIFS-HHO 알고리즘 변수선택 기반 차원축소

본 연구에서는 구축된 설명변수 데이터셋 가운데 반응변수인 월별 제조업 업황 실적과 실질적 영향관계를 유지하고 있는 주요 설명변수들을 추출, 이를 통해 분석의 복잡도를 줄이고 데이터셋 전반의 차원을 축소하고자 NMIFS-HHO 알고리즘을 활용하였다. NMIFS-HHO 알고리즘은 상호정보량을 기반으로 사전에 설정한 개수의 주요 변수를 순차적으로 추출하는 NMIFS 필터링, 그리고 주어진 설명변수 조합 가운데 가장 높은 유의성을 갖는 변수 집합을 탐색 및 반환하는 Harris Hawk Optimization (HHO) wrapping을 연결하는 일종의 하이브리드 변수선택 방법으로, 변수 중요도를 기반으로 차원을 먼저 축소하는 동시에 개별 변수의 자체의 중요도가 낮더라도 교호작용 효과에 따른 영향력을 고려하여 최종적 변수 집합 반환을 시도한다는 점에서 그 의의 및 성능이 대두된다. NMIFS-HHO 알고리즘의 구체적 구조는 다음과 같다(알고리즘 2, 3). 먼저 NMIFS 알고리즘의 경우 FIM, MIFS-U, mRMR 등과 마찬가지로 설명변수와 반응변수 사이 상호정보량인 $I(X; Y)$ 를 최대화하는 동시에 중복성을 최소화하는 변수를 순차 추출하는 방식을 따른다. 여기서 상호정보량이란 어떠한 변수가 다른 변수의 불확실성을 설명하는 크기를 엔트로피(entropy: H)를 활용하여 정량화하는 방식으로, 변수 간 비선형적 연관을 포착하는 것은 물론, 반응변수에 대한 설명변수의 설명력을 비교 분석할 수 있도록 한다는 점에서도 그 의의가 크다.

$$\begin{aligned} \text{상호정보량 (MI)} I(X; Y) &= \sum_{x \in X, y \in Y} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right) \\ &= \sum_{x \in X, y \in Y} p(x, y) \log p(x, y) - \sum_{x \in X} p(x) \log p(x) - \sum_{y \in Y} p(y) \log p(y) = H(X) + H(Y) - H(X, Y) \end{aligned}$$

NMIFS 알고리즘은 개별 설명변수와 반응변수 사이의 상호정보량을 추출, 이에 따라 상호정보량을 최대화하는 첫 번째 설명변수를 선택하여 초기 변수선택 집합의 원소로서 포함하고 본래 설명변수 집합에서 이를 제외한다. 이후 NMIFS 방법론은 중복성(redundancy)을 고려하여 나머지 설명변수 집합에 대한 Greedy Search를 진행한다. 구체적으로 NMIFS 방법론은 변수 간 중복성을 변수선택에 반영하기 위한 통계량으로서 후보 설명변수와 변수선택 집합 원소 간 normalized MI(NI)를 각각 도출한 후, 현재 단계에서 선택된 변수들의 집합인 S 의 원소 개수인 $|S|$ 를 기반으로 반응변수와 후보 설명변수 사이 상호정보량에서 $NI / |S|$ (adaptive redundancy term)를 제외한 통계량 G 를 도출하고 이를 최대화하는 중요 변수를 변수선택 집합에 추가하는 과정을 사전에 정해진 횟수만큼 반복 시행하게 된다. 이를 통해 NMIFS 방법론은 mRMR, FIM, JMI 등 기존에 존재하는 상호정보량 기반 필터링에 비해 상대적으로 적은 연산량(비용) 아래 중복 영향의 적응적 제거를 실현한다는 점에서 높은 효율성을 확보함이 확인되며, 이에 효과적 변수선택 수단으로 평가되고 있다 (Estevez et al., 2009).

알고리즘 2: *NMIFS algorithm (Estevez et al., 2009)*

Step1) 초기 F (설명변수 집합), 초기 S (변수선택 집합) = $\{\emptyset\}$

Step2) for $\forall f_i \in \text{set } F$, $I(f_i; Y)$ (Y : 반응변수), $\text{find } \hat{f}_i = \text{argmax}_{f_i} I(f_i; Y)$,
 $\text{update } F \leftarrow F \setminus \{\hat{f}_i\}$, $S \leftarrow \{\hat{f}_i\}$

Step3) Until $|S| = k$:

for $\forall f_i \in F$, value G 도출.

$$NI(f_i; f_s) = I(f_i; f_s) / \min\{H(f_i), H(f_s)\}, G = I(f_i; Y) - \frac{1}{|S|} \sum_{f_i \in S} NI(f_i; f_s)$$

$\text{find } \hat{f}_i = \text{argmax}_{f_i} G$, $\text{update } F \leftarrow F \setminus \{\hat{f}_i\}$, $S \leftarrow \{\hat{f}_i\}$

이때 NMIFS 방법론은 필터링의 하위 범주라는 점에서, 사전에 설정한 변수 개수까지 순차적으로 변수를 선택함에 따라 교호작용 시 유의한 영향력이 발생하는 개별변수의 중요성을 간과할 수 있다는 한계를 갖는다. 본 연구에서는 해당 한계를 보완하고자 때 지능(Swarm intelligence)을 기반으로 한 wrapper를 연결하고자 하였으며, 이에 변수 선택된 집합 S 에 HHO-algorithm을 적용 및 활용하였다.

Harris Hawk Optimization(HHO) 방법론은 Heidari et al.(2019)에 의해 개발된 메타-휴리스틱 최적화 알고리즘으로, 매목 수리과 파라부테오속 해리스 매가 먹잇감을 사냥하는 방식을 재현하여 변수조합 지형을 탐색하는 알고리즘을 말한다. HHO 알고리즘은 기타 wrapper 방식과 유사하게 여러 탐색자의 존재를 가정하여 각 존재(해리스 매)의 위치를 변수의 선택 여부(미선택:0, 선택:1)와 대응시키며, 유동적 타깃 움직임 가정하고 최적 변수조합 후보군을 탐색하는 탐색(exploration) 단계와 탐색된 후보군보다 개선된 변수조합의 존재성을 확인하는 네 가지 활용(exploitation) 단계(soft besiege, hard besiege, soft besiege with progressive dive, hard besiege with progressive dive)를 활용하여 변수 지형 전반을 균형 있게 탐색하여 최적 변수조합 집합을 반환한다. 이때 HHO 알고리즘은 전체 탐색 공간에 대한 균등분포를 시작점으로 미분으로부터 자유로운(gradient-free) 최적화를 진행하며 Levy flight 함수를 활용함으로써 국소적 최적치로부터의 탈출이 용이하다는 장점을 가지며, 앞선 두 단계(탐색 단계, 활용 단계)의 균형을 달성하기에 유리하다는 장점을 갖는다(Hamed and Mohamed, 2023). 본 연구에서는 Chantar et al.(2021)이 제시한 Binary-HHO 알고리즘을 활용 및 응용하여 HHO 알고리즘을 전개하였으며, 구체적으로는 동일한 논리 구조를 사용하되 세 가지 개선사항(변수별 상한 또는 하한의 서로 다른 단위에 따른 영향최소화를 위해 정규화를 통한 상한 및 하한값 조정, 각 탐색 단계별 위치 이항변환 시 Thaher and Arman (2020)의 접근을 사용함으로써 전달 함수 가운데 시그모이드 함수를 활용하여 변수선택 여부 반환, 이항변환 시 모든 변수가 선택되지 않을 경우 행벡터가 0 벡터로 귀결되는 경우를 방지하기 위하여 랜덤하게 한 개 변수를 선택)을 적용함으로써 최적 변수조합 추출을 시도하였다(알고리즘 3).

알고리즘 3:

Binary HHO- algorithm

$$\begin{pmatrix} X_{11} & X_{12} & \dots & X_{1k} \\ X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \dots & \dots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nk} \end{pmatrix} (X_i : i \text{ 번째 Hawk } (i \in [n]), X_j : j \text{ 번째 특성 (설명변수)} (j \in [k]))$$

(STEP 1) Hawk 개수 (= n) 및 알고리즘 반복횟수 (= T) 설정, $t = 1$

(STEP 2) 개별 항에 대하여 베르누이 분포 기반으로 초기 HHO 행렬 생성 ($\in \{0, 1\}$)

(STEP 3)

while ($t < T$):

개별 Hawk의 위치 ($X_{ij} \in \{0, 1\}$)에 따른 n 개 변수집합별 성능 확인 ($F(X_i)$)

성능을 최대로 하는 변수집합(행)을 X_{prey} 로 설정.

foreach X_i :

draw $E_0 \sim Uniform(-1, 1)$, $r_5 \sim Uniform(0, 1)$, $r \sim Uniform(0, 1)$

set $E = 2E_0(1 - \frac{t}{T})$, $J = 2(1 - r_5)$

if $|E| \geq 1$:

draw $r_1, r_2, r_3, r_4, p \sim Uniform(0, 1)$

if $p \geq 0.5$:

$X_{i, rand} \equiv$ HHO 행렬에서 랜덤하게 1개 Hawk(행) 선택

$X_i(t+1) = X_{i, rand}(t) - r_1 |X_{i, rand}(t) - 2r_2 X_i(t)|$

$X_i(t+1)$ 의 각 요소 sigmoid 함수 기반($\lambda = 0.5$ 초과 시 1, 이하 시 0) 이항변환

else:

$X_n(t) = \sum_{i=1}^n X_i(t)$, $Ub(Lb)$: 설명변수별 상한(하한) 벡터 정규화 값

$X_i(t+1) = (X_{prey}(t) - X_n(t)) - r_3(Lb + r_4(Ub - Lb))$

$X_i(t+1)$ 의 각 요소 sigmoid 함수 기반 이항변환

elif $0.5 \leq |E| < 1$:

if $r \geq 0.5$: (Soft Beseige)

$\Delta X_i(t) = X_{prey}(t) - X_i(t)$

$X_i(t+1) = \Delta X_i(t) - E |J \times X_{prey}(t) - X_i(t)|$

$X_i(t+1)$ 의 각 요소 sigmoid 함수로 이항변환

else: (Soft Beseige with progressive dive)

$Y = X_{prey}(t) - E |J \times X_{prey}(t) - X_i(t)|$

$\beta = 1.5, u, v \sim Uniform(0, 1)$

$\sigma = \Gamma(1 + \beta) \sin(\pi \beta / 2) / (\Gamma((1 + \beta) / 2) \beta 2^{((\beta - 1) / 2)})^{1/\beta}$

$LF(levy flight pattern) = 0.01u \sigma / |v|^{1/\beta}$

$Z = Y + S \times LF$; $S \sim runif(0, 1)$ with size $(k \times 1)$, F : 성능 평가 함수

$X_i(t+1) = \begin{cases} Y & (F(Y) < F(X_i(t))) \\ Z & (F(Z) < F(X_i(t))) \end{cases}$

$X_i(t+1)$ 의 각 요소 sigmoid 함수로 이항변환

```

else (|E| < 0.5):

    if r ≥ 0.5: (Hard Beseige)
        ΔXi(t) = Xprey(t) - Xi(t)
        Xi(t+1) = Xprey(t) - E|ΔXi(t)|
        Xi(t+1)의 각 요소 sigmoid 함수로 이항변환

    else: (Hard Beseige with progressive dive)
        Y' = Xprey(t) - E|J × Xprey(t) - Xn(t)|
        β = 1.5, u, v ~ Uniform(0,1)
        σ = Γ(1+β) sin(πβ/2) / (Γ((1+β)/2) β 2((β-1)/2))1/β
        LF(levy flight pattern) = 0.01uσ/|v|1/β
        Z' = Y' + S × LF; S ~ runif(0,1) with size (k × 1), F: 성능 평가 함수
        Xi(t+1) = { Y' (F(Y') < F(Xi(t)))
                  Z' (F(Z') < F(Xi(t)))
        Xi(t+1)의 각 요소 sigmoid 함수로 이항변환

    t = t + 1

return Xprey

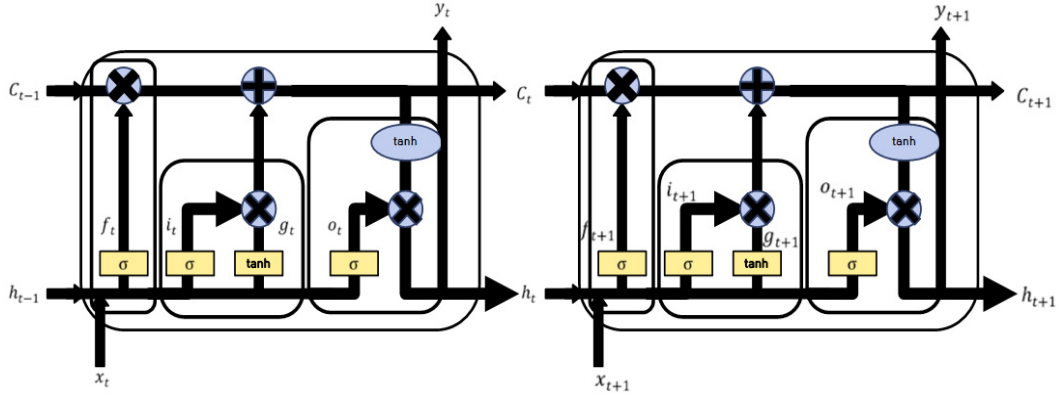
```

본 연구에서는 일차적으로 앞선 데이터셋을 시간 순서를 유지하며 8:2의 비율로 분할, 이를 통해 전자를 학습(훈련) 데이터, 후자를 평가 데이터로 설정하였으며, 이차적으로는 학습 데이터 내 설명변수 집합을 대상으로 75%(k=16개)의 상위 변수를 순차 추출하는 NMIFS 필터링을 적용하였다(NMIFS 내 상호정보량의 계산 및 엔트로피 계산에는 sklearn.feature_selection 패키지 내 mutual_info_regression 함수 및 엔트로피의 성질 $I(X;X) = H(X)$ 사용). 이후 본 연구는 필터링을 통해 구축되는 변수선택 집합을 대상으로 총 n=10개의 해리스 매 및 T=50번의 반복 위치 업데이트를 설정하였으며 이와 함께 초기 HHO 행렬 생성 시 성공확률 p=0.6인 베르누이 분포를 기반으로 초기 변수 후보군을 생성할 것을 가정하여 HHO 알고리즘을 (알고리즘 3)와 같이 실행함으로써 최종 최적 변수조합을 도출하였다. 이때 본 연구에서는 HHO 알고리즘 실행 시 성능 평가 함수 F(x)로서 일반적으로 자주 활용되는 머신러닝 모델 가운데 총 30개의 트리를 상정한 랜덤 포레스트(RandomForest) 모델을 활용하는 동시에 평균 절댓값 오차(MAE)를 평가방법으로 설정하였으며 구체적 모델 실현 수단으로는 Python 내 scikit-learn의 sklearn.ensemble 패키지의 RandomForestRegressor() 함수를 활용하였다.

3.3. LSTM 기반 순환신경망 모델 구축, 학습 및 평가

다음으로, 본 연구는 학습 데이터를 전제로 앞서 차원 축소된 설명변수 집합, 그리고 반응변수 데이터셋을 활용하여 LSTM 레이어 기반 순환신경망(RNN)을 학습시켰다. 여기서 LSTM이란 Long-Short-Term-Memory의 약자로, 순환신경망이 길이가 긴 시

계열 데이터를 처리하면서 발생하는 경사 소멸 문제를 해결하는 수단 가운데 하나로써 제시되는 딥러닝 방법론이다(그림 3.1). LSTM은 (그림 3.1)과 (수식 3.1)과 같이 망각



<그림 3.1> LSTM 레이어의 구조

LSTM layer; f : forget gate, c : longterm state, h : shortterm state, i : input, o : output

$$\begin{aligned} c_t &= f_t \otimes c_{t-1} (\text{previous information}) + i_t \otimes g_t (\text{updated information}) \\ g_t &= \tanh(W_c[h_{t-1}, x_t] + b_c) \quad (x_t: \text{input variable}, W_c: \text{weight}) \end{aligned} \quad (3.1)$$

게이트(forget gate, ‘f’)를 활용하여 과거 시계열 데이터 가운데 향후 예측에 필요한 중요 정보를 장기 기억으로 저장하는 한편, 일시적으로 활용할 시계열 정보는 단기 기억으로서 처리함으로써 다음 반응변수 예측 시 경사 소멸을 효율적으로 방지하는 방식의 신경망 구조를 갖는다. 이를 통해 LSTM은 주가나 텍스트, EEG, ECG 신호 등 다양한 시계열 자료의 해석 또는 예측 수행 시 성능상 높은 역량 및 효율성을 갖게 된다(Sen et al., 2022; Nazari, 2020). 본 연구는 LSTM의 해당 특성을 활용하여 순환 신경망 가중치 업데이트 시 발생할 수 있는 경사 소멸 문제를 효과적으로 개선하는 동시에 향후 제조업 업황실적 시계열에 대한 유의한 선제적 예측을 전개하고자 하였다. 구체적으로, 먼저 본 연구에서는 tensorflow.keras 라이브러리를 바탕으로 순환신경망(RNN) 구조를 구축하였으며 이때 총 5개의 LSTM layer를 설정한 후 각 은닉층별로 10, 30, 50, 30, 10개의 노드를 설정하였다. 다음으로 본 RNN에서는 효율적 학습 및 수렴 가속을 위해 LSTM 은닉층별로 가중치 초기화 방안 중 분산 조정 방식의 초기화를 활용하는 He-Normal 초기화를 사용함으로써 모델 학습 시 경사 소멸 문제를 사전에 일부 방지하였다(He et al., 2015). 이후 구축된 모델의 학습에는 최적화기로서 Adam optimizer(학습률=0.001)를 사용하였고 손실함수와 평가방법으로는 각각 MSE와 MAE를 활용하였다. 한편, 본 연구는 모델의 학습 방식으로는 전체 표본을 모두 활용하여 경사를 업데이트하는 방식인 배치 경사 하강법을 설정하였고 과적합 방지수단의 일환으로서 조기중단 알고리즘(patience=200)을 설정하여 모델의 적합 성능을 개선하는 것을 시도하였다. 이를 통해 본 연구에서는 총 1000번의 epoch를 기반으로 RNN 모델을 앞선 학습 데이터에 대하여 학습시켰으며, 이때 예측의 대상이 되는 반응변수인 제조업 업황실적 시계열과 설명변수 시계열은 학습상 편의를 위해 0과 1 사이로 정규화하여

학습을 진행하였다(오차 판단 시는 역변환을 통해 원시계열 상 평균 절댓값 오차(MAE) 등 확인). 이후 본 연구는 평가 데이터를 기반으로 학습된 최종 모델의 예측 성능 및 실효성을 판단, 이와 함께 해당 결과를 기타 앞서 성능함수에 활용한 랜덤 포레스트 모델의 다양한 변이에 따른 성능과 비교함으로써 제조업 업황 예측 시 SEMD 분해 및 LSTM 모델 활용의 유의성 및 성능에 대한 판단을 전개하였다.

3.4. LSTM 기반 순환신경망(RNN) 모델의 해석 및 정책 제언

본 연구는 이후 LSTM 기반 RNN 모델의 유의성 아래 학습된 모델의 해석을 통한 설명변수별 절대적 영향력의 크기, 그리고 영향별 방향성을 확인하였다. 이때 반응변수에 대한 개별 설명변수들의 영향을 해석하기 위한 수단으로는 변수 간 교호 작용 아래 각 요인별 한계기여도를 추출할 수 있는 SHAP 분석을 활용하였다. SHAP 분석은 Lundberg et al. (2017)이 고안한 개념으로, Shapley(1950)가 제시한 게임이론에서의 새플리(shapley) 값, 즉 게임의 참가자별 한계기여 개념을 딥러닝 및 머신러닝 모델 분석에 도입하여 개별 설명변수가 반응변수의 수치에 기여하는 정도를 확인할 수 있도록 한 eXplainable AI(XAI) 기법을 말한다(Rothman, 2022). 구체적으로, SHAP 분석에서 개별 설명변수의 기여도, 즉 새플리 값은 특성 집합을 연합 N , 연합의 부분집합을 S , 분석 대상이 되는 특성을 i 라고 정의 시 (전체 원소에 대하여 생성한 순열별 한계 기여도 계산을 전제로) 수식 3.2와 같은 수식으로서 정리되며, 이때의 새플리 값은 학습이 완료된 모델을 기반으로 각 샘플별 모델 예측값에 대한 개별 변수의 기여 정도를 상호작용 전제하에 정량적 자료로 제시한다는 점에서, 기존의 머신러닝 또는 딥러닝 모델들이 블랙박스로서 갖던 불투명성을 크게 줄인다는 의의가 존재한다(김현주 등, 2023; 김대영, 2023).

$$\begin{aligned} \text{Shapley Value for feature } \{i\} &\equiv \text{marginal contribution of feature } \{i\} \\ \phi_i(N, v) &= \frac{1}{N!} \sum_{S \subseteq N \setminus \{i\}} |S|!(|N| - |S| - 1)!(v(S \cup \{i\}) - v(S)) \\ (* v : \text{MODEL}). \end{aligned}$$

(3.2)

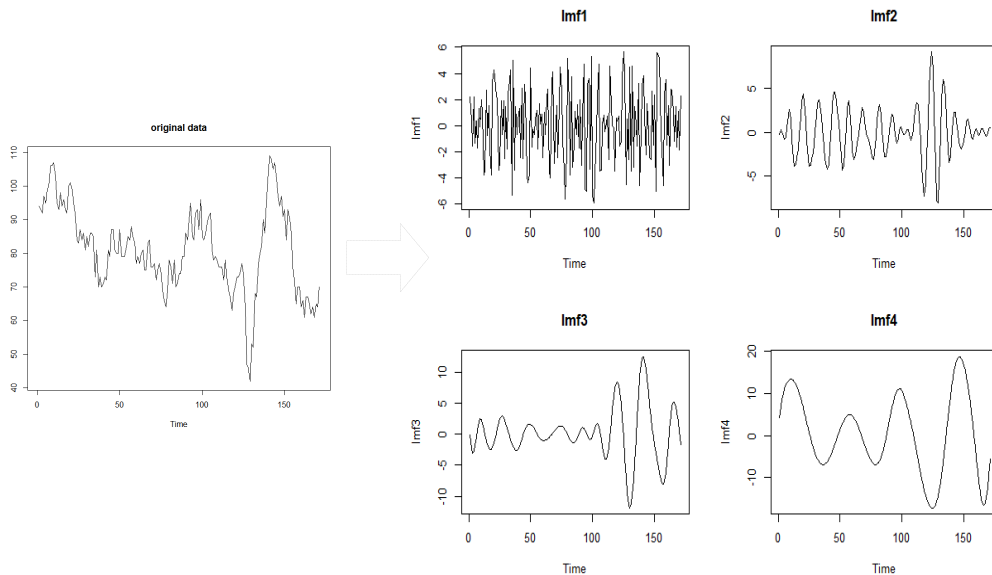
(출처 : Rothman, 2022)

본 연구에서는 SHAP 분석을 전개하기 위한 수단으로서 Python의 shap 라이브러리 내 KernelExplainer 함수를 활용, 해당 함수 적용을 통해 도출되는 개별 변수의 평균적·절대적 변수중요도 및 한계기여 방향성을 시각화 및 분석함으로써 제조업 업황실적에 선제적으로 영향을 미치는 주요 변수 및 해당 변수들과의 연관관계를 파악하였다. 이를 통해 본 연구는 최종적으로 제조업 업황의 급격한 악화 또는 급성장을 예측 및 평가할 수 있는 방안을 도출하는 한편, 분석을 통해 도출되는 변수 간 영향관계에 따른 시사점 등을 파악함으로써, 안정적 제조업 환경 육성을 위한 정책적 접근법을 제시하고자 하였다.

4. 연구 결과

4.1. 제조업 업황 실적 시계열 SEMD 분해 결과

먼저 본 연구에서는 제조업 업황 실적 시계열 흐름상 자기계열 영향을 확인하고자 앞선 데이터 가운데 월별 매출액 기준 제조업 업황 실적 데이터(2009.09.~2023.11.)에 대한 SEMD 분해를 실시하였다. 분해 결과(최적 $\lambda=0.0614$), 총 4개의 IMF 및 잔차계열이 원시계열로부터 분리되었으며, 이때의 각 IMF는 그림 4.1의 결과와 같았다. 먼저 첫 번째 IMF의 경우, -6부터 6 사이에서 매우 짧은 주기의 불규칙한 진폭을 가진 신호가 확인되었다. 다음으로, 두 번째 IMF에서는 약 -5에서 5 사이에서 약 10(개월)의 주기를 갖고 일정한 진폭 아래 진동하는 신호가 확인되었으며, 세 번째 IMF에서는 약 100(개월)을 기준으로 이전까지는 -5와 5 사이에서 약 16의 주기 아래 진동하는 신호가, 이후부터는 -10과 10 사이에서 약 25의 주기 아래 진동하는 신호가 확인되었다. 마지막으로 네 번째 IMF에서는 약 40개월의 주기 아래 -20과 20 사이에서



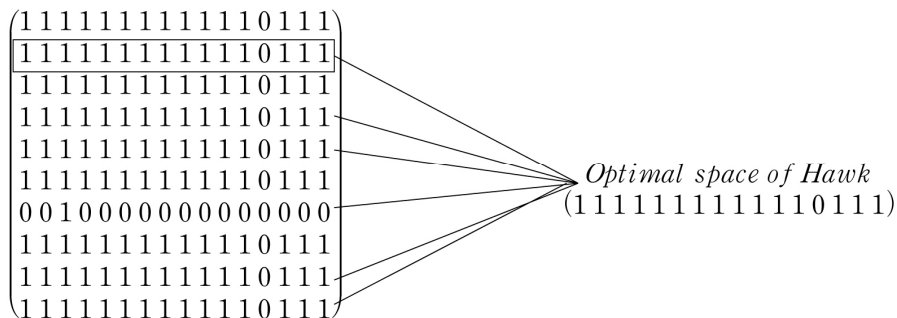
<그림 4.1> SEMD 분해 결과(IMF1~IMF4, 좌측: 원계열 시도표)

진동하는 신호가 확인되었다. 따라서 SEMD 분해 결과, 제조업 업황 실적 시계열 및 수치 변화는 1개월 미만의 초단기적 진동 신호, 약 1년의 주기를 갖는 중장기적 신호, 그리고 약 4년의 주기를 갖는 장기 신호의 결합으로 해석할 수 있음이 확인되었으며, 특히 네 번째 IMF, 즉 약 3~4년의 주기를 갖는 장기 신호에서는 시간의 흐름에 상대적으로 비탄력적인 진폭 변화가 확인된다는 점에서 제조업 업황 흐름에 대한 안정적인 정보 제공이 가능하다는 점이 도출되었다. 이에 본 연구에서는 IMF의 활용이 제조업 업황 예측에 유의한 정보를 제공할 수 있을 것이라는 가정하에 앞서 설계한 연구 과정을 재개하였다.

4.2. NMIFS-HHO 알고리즘 기반 변수선택 결과

다음으로 본 연구에서는 학습 데이터의 설명변수 집합에 대하여 NMIFS-HHO 알고리즘을 기반으로 한 변수선택을 전개하였다. 먼저, 총 필터링 변수 개수를 전체 설명변수 개수의 75%, 즉 16개로 설정한 NMIFS 알고리즘을 전개한 결과 설명변수 가운데 다음과 같은 변수가 주요 변수로서 순차적으로 선택되었다: 월별 동행종합지수(cci), 자기계열상 4개월 전 포착된 3~4년 주기의 장기변동 신호(imf4), 월별 후행종합지수(lgci), 경제활동참가율(eapp), 정부대출금리(inrg), 선행종합지수(lci), 경제심리지수(esi), 생산자물가지수(ppi), 월평균 서부텍사스유가(wti), 자기계열상 두 번째 IMF(imf2), 원유도입단가(opr), 원/달러 환율(er), 원유 도입물량(oq), 브랜트유가(br), 두바이유가(dbi), 설비투자지수(ei). 이를 통해 본 연구는 제조업 업황실적 예측 시, 여러 변수 가운데 경기의 측면에서는 경기종합지수 관련 변수가, 자기계열의 측면에서는 장기변동 신호가, 경제·재정의 측면에서는 정부대출금리, 설비투자지수 및 경제활동규모 관련 변수가, 원자재의 측면에서는 WTI의 가격 변동이 상대적으로 높은 정보를 제공함을 확인하였다. 반면 본 연구는 설명변수 가운데 해외직접투자 금액, 한국은행 기준금리 변동, 그리고 자기 계열상 초단기 신호 등 변수는 (선행연구 검토 시 동일 시점을 기준으로 하는 제조업 업황에 유의미한 영향을 미치는 요인으로 나타나지만) 시차를 둔 제조업 업황실적 예측 진행 시 다른 변수에 비하여 상대적으로 낮은 정보를 제공함을 확인하였다.

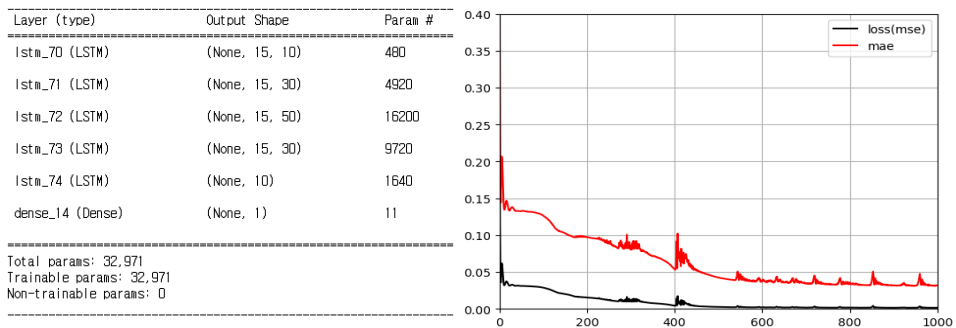
이후 본 연구는 필터링된 16개 변수 집합을 대상으로 HHO 알고리즘을 전개하였다. 본 연구에서는 총 10개의 해리스 매 및 총 50번의 반복을 전제로 한 최적 변수조합 지형 탐색 결과, (그림 4.2)와 같은 HHO 행렬이 도출되었다. 이때 최적 변수조합을 선정한 결과, [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1] 벡터가 반환되었으며, 따라서 앞선 NMIFS 순차 추출 결과에서 원유 도입물량(백만 배럴)(13번째 변수)을 제외한 나머지 15개 변수(cci, imf4, lgci, eapp, inrg, lci, esi, ppi, wti, imf2, opr, er, br, dbi, ei)가 최종적인 최적 변수조합으로서 도출되었다. 이에 본 연구에서는 해당 변수조합을 전제로 한 LSTM 기반 RNN의 학습을 전개, 나아가 정밀한 제조업 업황실적 BSI 변동을 예측을 시도하였다.



<그림 4.2> 50번 반복 탐색을 통한 최종 HHO 행렬

4.3. LSTM 기반 순환신경망 학습 및 평가

본 연구는 NMIFS-HHO 알고리즘을 기반으로 도출된 최적 설명변수 집합을 대상으로 LSTM 기반 순환신경망(RNN) 학습을 전개하였다. 학습 결과 손실함수(MSE)의 그래프 및 평가 지표(MAE)의 그래프는 (그림 4.3)과 같았으며, 이때 MSE 그래프와 MAE 그래프 모두에서 0으로의 유의미한 수렴이 발생함에 따라 본 연구에서는 RNN의 가중치 업데이트가 효과적으로 진행되었음을 확인하였다. 다음으로, 본 연구에서는 학습 데이터에 대한 모델의 적합도를 판단하고자 전체 설명변수 집합을 대상으로 학습한 랜덤 포레스트 모델(RF, 성능 평가 함수와 동일하게 트리 개수 = 30개로 설정), NMIFS 변수선택을 통해 차원 축소된 설명변수 집합을 기반으로 한 RF 모델, NMIFS-HHO 변수선택을 통해 도출된 최적 설명변수 조합을 활용한 RF 모델, 그리고 현재 학습된 NMIFS-HHO 기반 LSTM 순환신경망 모델의 학습 정도를 비교 분석함으로써 모델의 유의성을 확인하였다. 이때 모델별 구체적인 적합도 평가는 MAE, RMSE, 그리고 시계열 시각화를 통해 진행되었다(그림 4.4). 적합도 평가 결과, 네가지

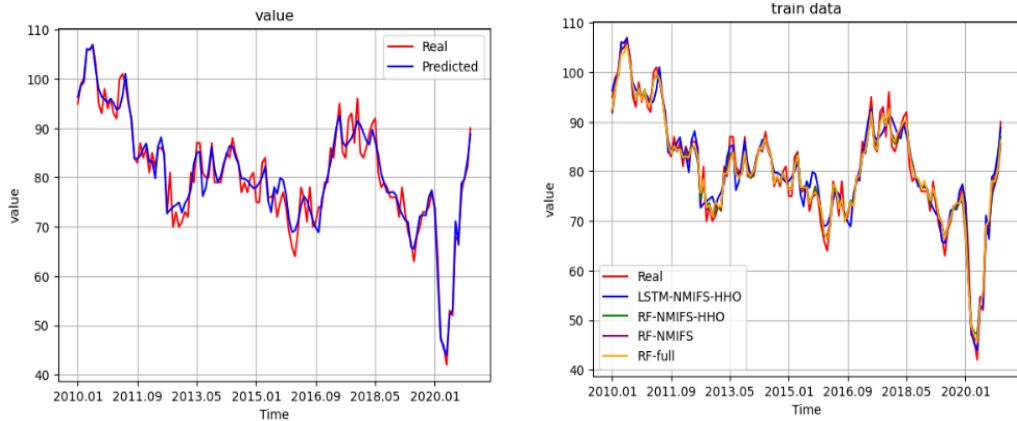


<그림 4.3> 좌측: LSTM 기반 RNN 구조, 우측: 손실함수 및 MAE 곡선

모델 가운데 NMIFS-HHO 기반 RF 모델이 가장 높은 적합도를 가짐이 확인되었으며, 이때 MAE는 1.3840, RMSE는 1.8153임이 확인되었다. 나아가 네 가지 모델 모두에서 1에서 2.2 사이의 평균 절댓값 오차가 발생함에 따라, 모든 모델에서 매우 적은 오차를 갖고 4개월 후 제조업 업황 실적을 설명할 수 있음이 확인되었으며, 이 가운데 RF 모델을 기준으로 변수선택 효과를 비교할 경우, NMIFS에서 NMIFS-HHO로 변수선택이 진행될수록 모델의 적합 정도, 즉 설명력이 개선되는 현상이 확인됨에 따라 상호정보량 기반 필터링 및 떼 지능(Swarm Intelligence) 기반 wrapper를 활용한 변수선택 가운데 NMIFS-HHO 알고리즘의 활용은 제조업 업황실적 BSI 변화 설명 시 유의미한 개선으로 이어질 수 있음이 확인되었다. 반면 예상과는 달리 오차 기준 접근 시 LSTM-기반 모델의 적합도는 네 모델 가운데 가장 낮은 것으로 확인되었는데, 이는 시각화된 시계열 그래프 고려 시 LSTM 기반 모델은 특정 시기의 극댓값 또는 극솟값에 대한 설명력보다는 추세 자체에 대한 설명력이 강한 경향이 다른 모델에 비하여 대두됨에 따른 결과라 해석된다. 이에 본 연구에서는 평가 데이터를 통한 모델 예측력 평가를 기반으로 해당 LSTM 모델의 성능이 강건(robust)한 것인지, 아니면 단순히 설명력이 기타 모델에 비하여 낮은 것인지를 검증하고자 하였다.³⁾

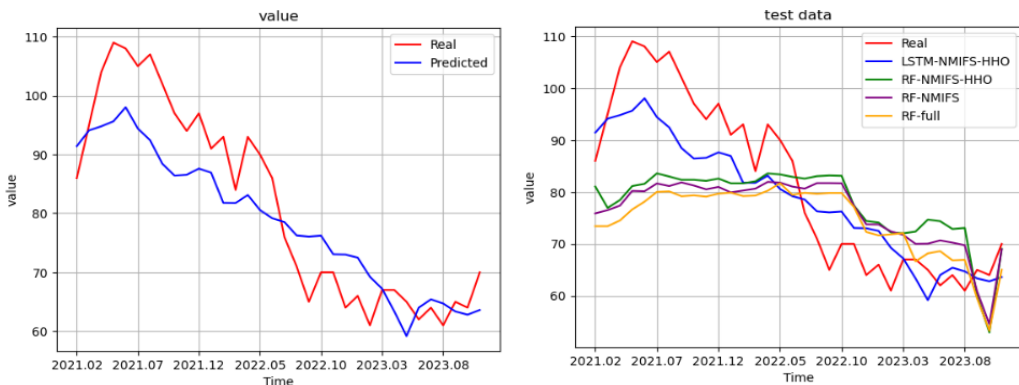
<표 4.1> 학습 데이터에 대한 모델별 적합도(설명력)

Model	RF-Full model	RF-NMIFS	RF-NMIFS-HHO	LSTM-NMIFS-HHO
MAE	1.4576	1.4419	1.3840	2.1189
RMSE	1.8727	1.9095	1.8153	2.6305

<그림 4.4> 좌측: 학습데이터에 대한 LSTM 모델의 적합도 시각화(빨간색: 실제 제조업
업황 실적 시계열), 우측: 학습데이터에 대한 모델별 시계열 적합 시각화

<표 4.2> 평가 데이터에 대한 모델별 예측력

Model	RF-Full model	RF-NMIFS	RF-NMIFS-HHO	LSTM-NMIFS-HHO
MAE	12.4088	12.0304	12.0039	6.6648
RMSE	15.0647	14.1688	13.8755	7.7561

<그림 4.5> 좌측: 평가 데이터에 대한 LSTM 모델의 예측치 시각화, 우측: 평가 데이터에
대한 모델별 시계열 예측치 시각화

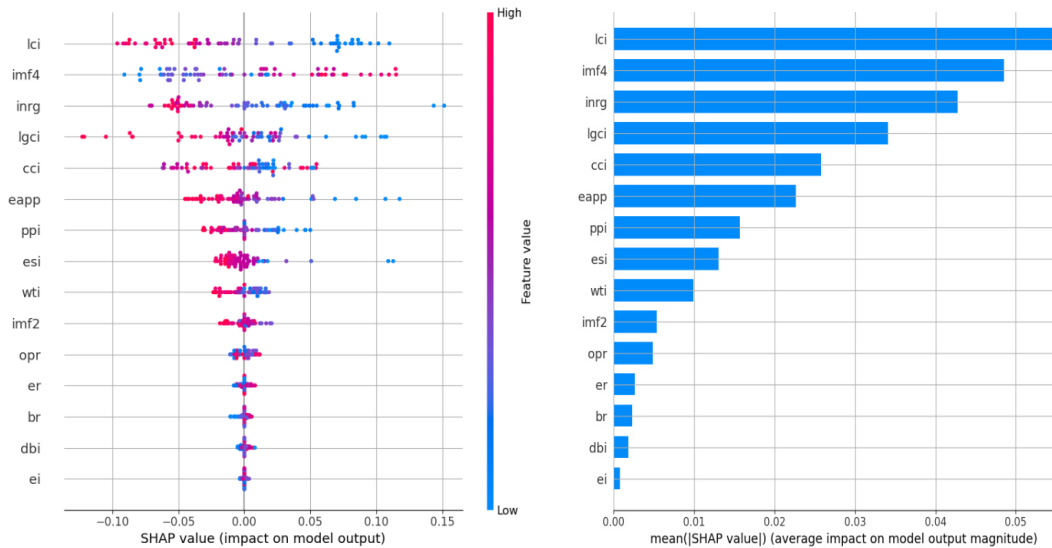
3) 순위와 별개로, 절대적 오차의 수준에서 접근 시, LSTM 기반 모델 또한 1에서 2.2사이의 평균절대값오차(MAE)를 갖는다는 점에서 높은 설명력을 지닌이 확인되었다.

평가 데이터를 전제로 LSTM 기반 순환신경망의 예측력, 즉 예측 성능을 평가한 결과는 표 4.2, 그림 4.5와 같았다. 먼저 오차의 측면에서 LSTM 기반 NMIFS-HHO 모델은 약 6.6648의 평균 절댓값 오차(MAE), 약 7.7561의 RMSE를 갖는 것으로 확인되었으며, 이에 해당 모델은 매우 높은 정확도를 갖고 실제 제조업 업황실적 수준을 4개월 선제적으로 예측할 수 있음이 나타났다. 또한, 해당 모델은 학습 데이터에서의 순위와는 달리 기타 모델(RF-full 모델(모든 설명변수 사용), RF-NMIFS 모델, RF-NMIFS-HHO 모델)과의 비교 분석 시 상대적으로 가장 낮은 오차를 갖고 정밀한 예측을 진행할 수 있음이 확인되었으며, 나아가 시계열의 변동 및 변화폭의 차원에서 타 모델에 비하여 LSTM 기반 모델이 가장 탄력적으로 변화(급락)를 예측할 수 있음이 그림 4.5상 확인되었다. 따라서 본 연구에서는 앞선 논의와 관련하여 LSTM 기반 모델이 기타 모델에 비하여 상대적으로 높은 강건성(robustness)을 확보하고 있음을 확인하였으며, 이를 통해 NMIFS-HHO 변수선택 및 LSTM을 기반으로 한 신경망 차원의 접근이 제조업 업황 실적 예측에 유의미하게 기여함을 검증할 수 있었다⁴⁾.

4.4. LSTM 기반 RNN의 해석 및 변수 간 영향 관계 파악

마지막으로 본 연구는 NMIFS-HHO-LSTM 모델의 유의성 확보 전제하에 4개월 후 제조업 업황 실적 예측상 개별 설명변수들이 미치는 영향력의 크기 및 방향성을 파악, 나아가 제조업 업황의 안정성을 증진할 수 있는 정책적 제안을 함께 전개하고자 하였다. 이에 따라 해당 모델에 대한 SHAP 분석 적용 결과, 그림 4.6과 같은 Shapley 그래프가 도출되었다.⁵⁾ 먼저 변수 중요도의 측면에서 접근 시, NMIFS-HHO 알고리즘을 통해 도출된 최적 변수조합 내 원소 가운데 중요도 상 최상위 및 차상위 변수로는 선행종합지수와 IMF4, 즉 4개월 전 자기계열상 장기 신호가 선정되었다. 이후 순차적으로 월별 정부대출금리, 후행종합지수, 동행종합지수, 경제활동참가율, 생산자물가지수, 경제심리지수, 서부텍사스 유가, IMF2, 그리고 그 외 변수 순으로 반응변수 수치 예측 시 중요도가 높은 것으로 나타났으며, 절댓값 측면에서의 변수중요도에 따른 순차 배열 시 5번째, 7번째, 10번째 변수에서 엘보우 포인트, 즉 기울기가 급격히 변화하는 양상이 확인되었다. 다음으로, 반응변수 예측에 대한 개별 변수별 한계 기여 방향성은 다음과 같았다. 먼저 자기계열상 장기신호(imf4) 및 환율은 4개월 후 제조업 업황 실적에 양의 방향으로의 영향을 미침이 확인되었다. 다음으로 월별 선행종합지수, 후행종합지수, 경제활동참가율, 정부대출금리, 생산자물가지수, WTI의 경우에는 반응변수 예측 시 음의 방향으로의 영향을 미침이 확인되었다. 한편, 그 외 변수의 경우, 특정 방향으로 영향의 방향성을 확정하기에 어려움이 존재함을 확인하였다.

-
- 4) 학습 데이터에 대한 적합도는 절대적인 수치상 기타 모델들과 유사한 반면, 평가 데이터를 기반으로 한 오차와 관련하여서는 약 2배가량의 차이가 존재한다는 점에서 LSTM 모델은 타 모델에 비하여 실제 데이터에 덜 과적합되는 동시에 제조업 업황실적 추세 예측에 대한 탄력성을 유지하고 있음을 확인하였다. 특히 시계열 변동상 기타 모형에서는 2022년 4월부터 2022년 9월 사이 발생한 제조업 업황의 급락을 선제적으로 포착하지 못한 반면, NMIFS-HHO-LSTM 모델은 해당 낙폭 및 추세를 선제적으로 정밀 예측하였음이 확인된다.
- 5) 본 연구에서는 전체 학습 데이터 샘플 가운데 SHAP 연산량 및 비용을 고려하여 전체 학습 데이터 표본에서 랜덤하게 50%의 샘플을 추출한 후, 해당 샘플에 대한 SHAP 분석을 전개함으로써 효율성을 높이고자 하였다.



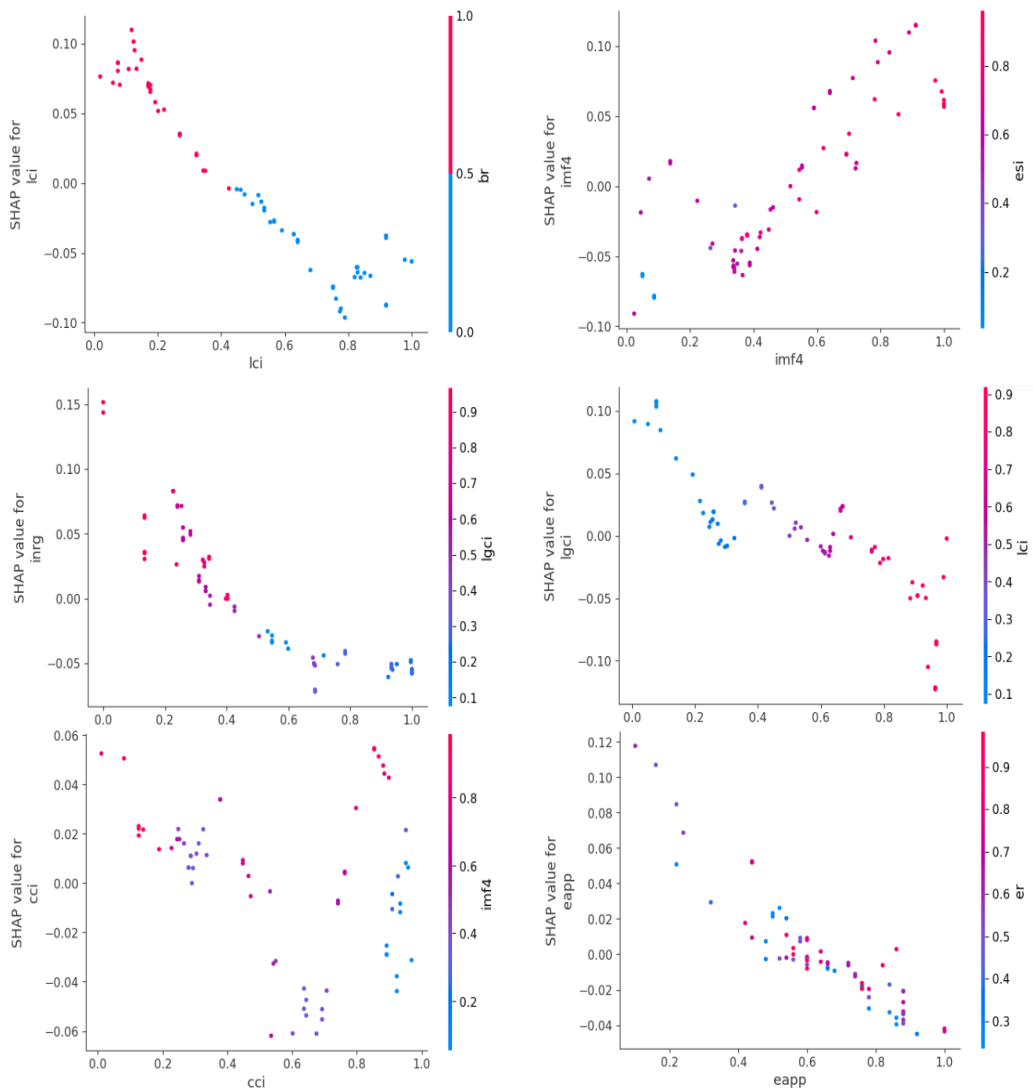
<그림 4.6> 좌측: SHAP 한계기여 그래프, 우측: 새플리 값에 대한 평균 절댓값 기준 변수중요도 순차 나열

이후 본 연구에서는 개별 변수의 한계기여도를 보다 면밀하게 파악하고자 각 설명 변수에 대한 Shapley 한계기여도 그래프를 시각화하였다. 이때 변수중요도의 엘보우 포인트 등을 고려함으로써, 본 연구에서는 전체 변수 가운데 중요도상 상위 40%까지의 변수를 분석·시각화 대상으로 선정하여 SHAP 분석을 전개하였다. 먼저 주요 설명변수 가운데 월별 선행종합지수(lci)는 한국은행 기준금리(br)와 가장 높은 상호작용을 유지하며 반응변수인 4개월 후 제조업 업황실적 BSI에 음의 방향으로의 영향을 미치는 것으로 파악되었으며, 나아가 Shapley 한계 기여 그래프상 선행종합지수가 극단적으로 높거나 극단적으로 낮을 경우, 해당 변수는 4개월 후 제조업 업황실적 예측치 대하여 상대적으로 비탄력적인 영향을 미치는 것으로 확인되었다.

다음으로, 자기 계열상 장기변동 신호(IMF4)의 경우 모든 구간에서 반응변수에 대하여 양의 방향으로의 영향을 미침이 확인되었으며, 이때 장기신호가 안정적 진폭 및 규칙성을 가지고 진동한다는 점을 고려할 경우(그림 4.1) SEMD 분해상 약 3~4년 주기의 장기신호 내 현재 위치를 특정한다면, 다른 변수가 고정되었다는 전제 아래 4개월 후 제조업 업황 실적 규모의 변화 방향성 및 변화 속도를 정밀하게 추정할 수 있음이 확인되었다. 가령 IMF4의 규칙성을 고려할 경우 2024년 1분기는 그림 4.1의 3년 주기상 급격하게 장기변동 신호가 증가하는 시기로 판단된다. 이에 SHAP의 한계기여 방향성을 고려할 경우 2024년 2분기의 제조업 업황 실적은 해당 움직임에 따라 강한 상승 압력을 받을 것으로 예측된다. 따라서 2024년 1분기 경기종합지수, 경제활동참가율, 정부대출금리 등의 변동폭이 낮거나 일정하게 유지된다면, 2024년 2분기 제조업 업황 실적은 현재(2023.12) 수준에 비하여 상대적으로 개선될 것으로 기대할 수 있다.

한편, 정부대출금리 변수의 경우 후행종합지수 변수와 가장 높은 상호작용을 유지하며 반응변수에 음의 방향으로 영향을 미침이 확인되었다. 이때, 정부대출금리는 타

변수와의 교호작용 효과를 고려할 경우, 후행종합지수와 상호작용하는 정도가 높을수록, 반응변수인 4개월 후 제조업 업황실적 수치에 더욱 강한 음의 방향으로의 영향을 미치는 것으로 확인되었다. 마지막으로 주요 상위 40% 설명변수 가운데 후행종합지수 및 경제활동인구 변수는 반응변수 예측치에 대하여 음의 방향으로의 영향을 미치나 동행종합지수 변수의 경우에는 특정 방향으로 영향의 방향성을 확정할 수 없음이 확인되었다. 이때 후행종합지수 변수의 경우, 해당 지수가 매우 높거나 낮은 상태를 제외하면 반응변수 예측치에 대한 해당 변수의 영향력이 상대적으로 비탄력적임이 확인되었으며, 동행종합지수 변수의 경우 IMF4, 즉 SEMD 분해에 따라 도출되는 4개월 전 자기계열상 장기변동 신호와 가장 높게 상호작용하며 제조업 업황실적 예측에 기여함이 확인되었다.



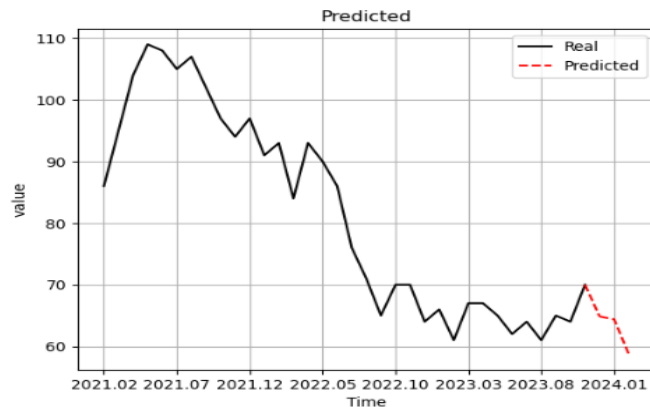
<그림 4.7> 개별 Shapley 한계기여도 그래프 시각화(주요 상위 40% 변수)

이에 따라 본 연구에서는 변수 간 영향 관계에 대한 총 네 가지 핵심 정보를 도출할 수 있었다. 첫째, 제조업 업황 실적 시계열에서 추출되는 약 3~4년 주기의 장기신호는 4개월 후 제조업 업황에 양의 방향으로의 영향을 미치며, 이때 해당 신호의 주기 및 진폭이 규칙적임에 따라 특정 장기 신호상 주기 내 현시기의 위치를 파악할 경우, 향후 제조업 업황실적의 변화 속도 및 크기에 대한 압력을 예측할 수 있음이 확인되었다. 둘째, 경기종합지수 가운데 선행종합지수 및 후행종합지수는 공통적으로 타 설명 변수와의 상호작용 아래 4개월 후 제조업 업황실적 예측치에 대하여 음의 방향으로 한계 기여함이 확인되며, 절대적 수치의 측면에서는 각 경기지수가 전 구간 가운데 극단에 위치하는지의 여부에 따라 제조업 업황실적 BSI 예측치에 대한 영향력이 상대적으로 각각 비탄력적·탄력적일 수 있음이 확인되었다. 셋째, 경제활동참가율 변수는 통념 및 예상과는 달리 그 규모가 증가할수록 오히려 4개월 후 제조업 업황 실적에 음의 방향으로의 영향을 미치는 것으로 나타났다. 이는 시장 내 경제활동인구의 증가 자체는 발생하나, 제조업 영역에서의 경제활동참가인구 유입은 오히려 감소하는, 일종의 직종 변환 효과가 강하게 발생하는 최근 시장 내 현상으로부터 기인되었다고 해석된다. 실제로 통계청 KOSIS 경제활동인구조사 및 2023년 지역별 고용조사 분석결과, 15세 이상 인구 규모는 일정하게 유지된 반면 경제활동인구는 28,096(천명)(2022년 1월)에서 29,375(천명)(2023년 11월)로 증가함으로써 전체 시장의 측면에서는 경제활동참가율의 증가가 발생하였음이 확인되었으나, 제조업 분야에서는 산업중분류상 전자부품, 컴퓨터, 영상, 음향 및 통신장비 제조업, 산업소분류상 기타금속 가공제품 제조업·전자 부품 제조업·직물제품 제조업·봉제의복 제조업 모두에서 전년 대비 취업자 감소가 발생함은 물론, 감소 규모 또한 약 15,000명에서 22,000명에 달한다는 점에서, 제조업 경제활동참가 정도가 상당 수준 축소되었음이 확인되었다. 따라서 최근 경제활동참가율의 증가는 제조업이 아닌 서비스업 등에서의 경제활동인구 증가를 내포한다는 점에서 실질적인 제조업 인구 규모의 축소를 함의하며, 나아가 향후 제조업 업황의 하락과 가시적으로 연결됨에 따라 앞선 SHAP상 영향 관계가 도출되었다고 해석된다. 마지막으로 넷째, SHAP의 변수 중요도상 중요 변수 가운데 월별 정부대출금리는 후행종합지수와의 상호작용에 따라 반응변수에 미치는 영향의 크기가 달라지며, 이때 해당 상호작용의 증가는 반응변수에 대한 더욱 강한 음의 방향으로의 영향으로 이어짐이 확인되었다.

4.5. SHAP 분석을 전제로 한 정책 제언 및 실제 예측치 도출

본 연구에서는 최종적으로 앞선 모든 분석을 종합함으로써 제조업 업황의 불안정성을 개선할 수 있는 대응 방안을 정책적 차원에서 제안하고자 하였다. 특히 본 연구에서는 SEMD 분해 및 NMIFS-HHO 알고리즘을 전제로 한 변수선택이 높은 정보량을 제공함은 물론, 해당 변수선택을 전제로 한 LSTM 기반 순환신경망 모델이 4개월 선제적인 제조업 업황실적 수준을 매우 적은 오차를 갖고 설명 및 예측할 수 있음이 확인됨에 따라, 특정 시기의 제조업 업황 변화를 예측하고자 할 때 해당 모델을 기반으로 한 접근이 유의한 예측치를 제공할 수 있다고 판단하였다. 이에 따라 본 연구에서 상정한 제조업 업황실적 예측 및 대응 프로세스는 다음과 같다. 첫째, 제조업 업황실적 BSI 시

계열에 대한 SEMD 분해($\lambda=0.0614$)를 통해 총 4개의 IMF를 추출한 다음, 초단기 신호와 세 번째 고유 모드 신호를 제외한 나머지 자기계열상 신호(IMF)와 함께 앞선 NMIFS-HHO 설명변수 조합을 LSTM 기반 순환신경망에 투입하여 4개월 후 제조업 업황실적 예측치를 도출한다(실제적 예시: 그림 4.8, 2023년 8월부터 10월 설명변수 데이터를 활용한 2023년 12월~2024년 2월 제조업 업황 실적 추이 예측치: 65 → 64 → 59: 12월부터 2월까지 소폭 감소할 것으로 예측됨).



<그림 4.8> 2023년 12월~2024년 2월 제조업 업황실적 추이 예측

둘째, 이때 예측상 특정 시기에서 업황의 급증 또는 급락이 예측될 경우, SHAP 분석에서 확인된 주요 상위 40% 변수의 추이, 즉 4개월 전 월별 선행종합지수, 자기계열상 장기변동 신호, 정부대출금리, 후행종합지수, 경제활동참가율, 그리고 동행종합지수 시계열의 추이를 평가함으로써 해당 급증 또는 급락 예측에 주요 영향을 미친 변수 움직임을 추출한다. 셋째, 주요 변수 움직임 추출 시, 해당 움직임의 지속성 여부, 나아가 신경망 예측치에 대한 신뢰도를 파악하기 위해, 대응되는 주요 설명변수의 시계열 추이가 갖는 불안정성을 평가하며, 이때 해당 시계열 추세의 불안정성이 낮은 것으로 평가될 경우 앞선 제조업 업황실적 BSI의 급증 또는 급락 예측이 유의하다고 판단함으로써 정책적 대응을 전개한다. 가령, 특정 시기의 제조업 업황 실적이 유의하게 급락할 것으로 예측될 경우, IMF4, 즉 장기변동 신호의 주기상 현재 위치를 파악함으로써 자기계열의 흐름 내 상승 압력 또는 하강 압력의 존재를 파악한 후, 전자의 경우 다른 변수들을 기존 수준의 정도로 유지하려는 정책적 접근을 취하는 반면, 후자의 경우에는 제조업 시장 유입 경제활동인구 증대를 위한 재사회화 프로그램 및 제조업 창업 및 마케팅 지원 프로그램을 확대하는 동시에 제조업계를 대상으로 한 정부대출금리 수준을 인하하는 등의 접근을 전개하는 것이 필요할 것이라 판단된다. 한편 후행 및 선행종합지수의 측면에서는 SHAP 분석상 해당 변수가 4개월 후 제조업 업황 개선에 유의 방향으로의 영향을 미치는 것으로 나타남에 따라, 경기종합지수를 활용한 단기적인 정책 수립 시 현재 후행·선행종합지수의 절대적 수준을 평가한 다음, 해당 지수가 과열되었다고 판단될 경우 각 지수의 세부 지표인 설비투자추계지수, 생산자제품재고지수 등을 검토함으로써 제조업체들의 과도한 단기적 제품 생산 증대 또는 레버리지의 증대

등을 지양·약화하기 위한 초단기적 지원 정책을 전개하는 것이 합리적일 수 있을 것이라고 판단된다. 이를 통해 본 연구에서는 정부 및 기업이 제조업 업황 변화를 4개월 선제적으로 예측 및 대비할 수 있을 것이라 예상하며, 최종적으로는 안정적 매출 및 업황을 유지할 수 있게 될 것이라 기대하는 바이다.

5. 결론 및 제언

본 연구는 제조업계가 현재 마주하고 있는 높은 불안정성 및 인력규모 감소 문제의 심각성을 바탕으로 제조업계 업황을 보다 면밀히 선제적으로 예측함으로써 시장 전반의 안정성을 도모해야 할 필요성에 공감하였다. 이에 본 연구는 제조업계의 안정성에 기여할 수 있는 변수선택 알고리즘 및 예측 모델을 구축, 나아가 주요 관련 영향 변수들의 특성을 분석함으로써 정책적 대응 방안을 제안하고자 하였다. 이에 따라 표 3.1과 같은 데이터를 분석한 결과, 다음과 같은 총 다섯 가지 결론이 도출되었다. 첫째, 제조업 업황실적 시계열에 대한 SEMD 분해는 향후 자기계열 예측에 유의미한 정보를 제공하는 중·장기 신호를 추출하는 데 기여할 수 있다. 구체적으로 SEMD 분해를 활용한 결과, 제조업 업황실적 시계열은 1개월 미만의 주기를 갖는 초단기 신호, 약 20개월의 주기를 갖는 두 개의 중단기 신호, 그리고 약 40개월(3~4년)의 주기를 갖는 장기신호로 분해될 수 있음이 확인되었으며, 특히 약 3~4년 주기의 장기신호는 상대적으로 규칙적인 주기 및 진폭을 갖는다는 점에서 제조업 업황 예측 시 상승 또는 하강 압력에 대한 매우 유의한 정보를 제공할 수 있음이 확인되었다.

둘째, 필터링 및 wrapping을 연결한 하이브리드 변수선택 방식인 NMIFS-HHO 알고리즘은 제조업 업황 실적 예측에 유의미한 정보를 제공하는 변수를 추출하는 데 유의하게 기여할 수 있음이 확인되었다. 구체적으로 본 연구는 베이스라인 모델인 랜덤 포레스트 모델(트리 개수: 30개)을 전제로, 설명변수 집합을 모두 활용하는 경우, NMIFS 필터링만을 활용하는 경우, 그리고 NMIFS-HHO 알고리즘을 활용하는 경우, 총 세 가지 변수선택법 아래 4개월 후 제조업 업황실적 예측 시 발생하는 평균절댓값오차(MAE)를 비교하였다. 그 결과 NMIFS-HHO 알고리즘 사용 시 가장 유의미한 성능 개선이 발생함이 확인되었으며, 이때 제조업 업황 실적 예측 시 상대적으로 많은 정보량을 제공하는 동시에 변수 조합상 반응변수에 높은 영향을 미치는 설명변수로는 월별 동행종합지수, 자기계열상 4개월 전 장기변동 신호, 월별 후행종합지수, 경제활동참가율, 정부대출금리, 선행종합지수, 경제심리지수, 생산자물가지수, 월평균 서부텍사스유가, 자기계열상 두 번째 IMF, 원유도입단가, 원/달러 환율, 브렌트유가, 두바이유가, 그리고 설비투자지수 변수가 존재함이 확인되었다.

셋째, 본 연구는 NMIFS-HHO 알고리즘을 기반으로 한 변수조합을 전제로 LSTM 기반 순환신경망을 학습 시, 높은 설명력 및 예측력을 가진 제조업 업황 실적 예측 모델을 구축할 수 있음을 확인하였다. 구체적으로 본 연구에서 구축한 LSTM 모델은 평가 데이터에 대하여 약 6.6648의 평균절댓값오차를 갖는다는 점에서 4개월 후의 제

조업 업황 실적을 매우 정밀하게 예측할 수 있음을 확인하였으며, 설명력의 측면에서는 학습 데이터에 대하여 2.1189의 평균절댓값오차, 그리고 2.6305의 평균제곱근오차를 갖는다는 점에서 높은 설명력을 갖고 제조업 업황 변화를 설명할 수 있음을 확인하였다.

넷째, 본 연구에서는 SHAP 분석을 통해 LSTM 기반 RNN 모델에서 각 설명변수의 한계기여도 및 한계기여 양상을 파악하고자 하였으며, 그 결과 자기계열 상 장기신호(IMF4)는 반응변수인 4개월 후 제조업 업황 실적에 양의 방향으로의 영향을, 경제활동참가율 변수, 정부대출금리는 반응변수 예측치에 음의 방향으로의 영향을 미침을 확인하였다. 또한, 후행·선행종합지수는 반응변수 예측치에 음의 방향으로의 영향을 미치되, 절대적 수치의 과열 또는 침체 여부에 따라 서로 다른 탄력성을 갖고 한계기여하는 것으로 나타났으며, 새플리 값의 절댓값에 따른 변수 중요도는 월별 선행종합지수, 자기계열상 장기신호, 정부대출금리, 후행종합지수, 동행종합지수, 경제활동참가율, 생산자물가지수, 경제심리지수, 서부텍사스 유가, IMF2, 그 외 변수 순으로 높은 것으로 나타났다. 구체적으로 정부대출금리의 경우에는 월별 후행종합지수와 가장 높은 상호작용을 유지하며, 해당 상호작용의 크기가 클수록 4개월 후 제조업 업황실적에 더욱 강한 음의 방향으로의 영향을 미치는 것으로 확인된 반면, 동행종합지수는 SEMD 분해에 따른 장기신호와 가장 높은 상호작용을 유지하며 제조업 업황실적 BSI 예측에 한계 기여함이 나타났으며, 선행종합지수와 동행종합지수는 각각 양극단을 제외한 구간, 양극단 구간에서 반응변수 예측치에 대하여 상대적으로 높은 탄력성 아래 음의 방향으로의 영향을 미침이 확인되었다. 나아가 경제활동참가율 변수의 경우 가시적으로는 통념과는 달리 SHAP 분석상 4개월 후 제조업 업황실적에 음의 방향으로의 영향을 미치는 것으로 파악되었으며, 해당 현상의 기저 원인으로는 최근 통계청의 경제활동인구조사 및 지역별 고용조사 분석 시 제조업 유입 경제활동인구의 절대적 감소 또는 직종 간 이동(제조업→서비스업)이 상정될 수 있음을 확인하였다.

다섯째, 본 연구는 앞선 분석을 종합함으로써 제조업 업황 실적 예측을 위한 정책적 대응 방안을 제시하였다. 제시한 정책적 프로세스는 다음과 같다. 먼저 제조업 업황 시계열에 대한 SEMD 분해를 통해 총 4개의 IMF를 추출한 다음, 초단기 신호를 제외한 나머지 IMF와 함께 앞선 NMIFS-HHO 설명변수 조합을 LSTM 기반 순환신경망에 투입하여 4개월 후 제조업 업황실적 예측치를 도출한 후, 예측상 특정 시기에서 업황의 급증 또는 급락이 예측될 경우 SHAP 분석에서 확인된 주요 상위 40% 변수의 시계열 추이를 평가함으로써 해당 급증 또는 급락 예측에 영향을 미친 변수 움직임을 추출하는 작업이 필요할 것이라 예상된다. 다음으로, 주요 설명변수의 움직임 추출 시, 해당 변수 움직임의 불안정성을 파악, 이를 통해 추세의 지속가능성 여부를 바탕으로 제조업 업황실적 BSI에 대한 모델의 급락 또는 급증 예측치 신뢰성이 높다고 파악될 경우, 자기계열 내 장기신호의 주기상 현재 위치를 파악함으로써 상승 압력 또는 하강 압력의 존재를 파악한 다음, 전자의 경우 다른 변수들을 기존 수준의 정도로 유지하려는 정책적 접근을 취하는 반면, 후자의 경우에는 제조업 유입 경제활동인구 증대를 위한 재사회화 프로그램 및 제조업 창업 및 마케팅 지원 프로그램을 확대하는 동시에 제조업계를 대상으로 한 정부대출금리 수준을 인하하는 등의 정책을 전개하는 것이

업황 및 시장 안정화에 효과적일 것이라 판단된다.

본 연구는 이에 따라 제조업 업황 실적 변화를 정밀하게 선제적으로 예측할 수 있는 모델을 구축 및 제공하는 한편, 해당 모델을 바탕으로 한 정책적 안정화 프로세스를 제공한다는 점에서 그 의의를 갖는다. 나아가 본 연구는 변수선택 알고리즘 및 SEMD 분해를 통해 반응변수에 영향을 미치는 주요 설명변수의 종류 및 개별 변수의 한계 기여양상을 포착하였다는 점에서도 큰 의의를 갖는다고 할 수 있다.

하지만 본 연구는 제조업 업황지수 예측상 영향을 미칠 수 있는 변수 집합 가정 시 미시적인 영향 요인들을 고려하지 못했다는 점, 원자재 가격 가운데 원유가격 변동만을 고려하였다는 점, 그리고 제조업 시장과 관련한 국제 요인들을 다소 고려하지 못했다는 점에서 한계가 존재한다고 판단된다. 따라서 동일 주제에 대한 후속 연구가 진행될 경우, 철강 가격 변화, 섬유 가격 변화, 주요교역국 금리, 월별 관세 변화 등 추가적 요인들을 고려한 세밀화된 분석이 전개될 필요성이 존재한다고 판단하는 바이다. 또한, 본 연구는 반응변수 시계열에서 추출되는 IMF 자체에 집중한다는 점에서 SEMD 분해상 추세성 등에 대한 정보를 포함할 수 있는 잔차 계열, 나아가 일정 시차 전제 하의 반응변수 원계열 자체를 설명변수 탐색과정에 포함시키지 않았다는 한계, 나아가 모델의 구축 과정상 시차를 4개월로 한정하여 분석을 전개함으로써 다양한 시차를 전제로 한 입체적 접근을 전개하지 못했다는 한계를 갖는다. 이에 후속 연구 등이 진행될 경우, 잔차 계열의 정보에 대한 세부적 분석 및 6개월 또는 그 이상의 중장기적 시차를 갖는 모델 등을 함께 고려한 연구를 수행함으로써 동 주제에 대한 상대적으로 입체적인 연구를 수행할 필요성이 존재한다고 판단하는 바이다.

참고문헌

- 김대영(2023), 베이지안 가법 회귀 트리(BART) 및 심층 신경망을 기반으로 한 선제적 재할용률 분석 및 활용 방안, <제21회 통계청 논문 공모 우수논문집>, 19, 71-103.
- 김바우, 김정현, 강성우(2021), 최근 원자재 가격 상승의 배경과 국내 제조업에 미치는 영향, <월간 KIET 산업경제>, 278, 7-20.
- 김연주, 윤종철, 제상영(2014), 경험적 모드 분해법을 이용한 상대적 구매력 평가설 실증분석, <Journal of the Korean Data Analysis Society>, 16(2), 785-793.
- 김종훈, 남현정(2023), 경쟁강도와 시장수요의 환경불확실성 하에서 ICT투자가 제조업의 비재무적 운영성과에 미치는 영향, <한국생산관리학회지>, 34(3), 391-411.
- 김현주, 박민규, 이지환(2023), SHAP을 활용한 벌크선 메인엔진 연료 소모량 예측연구, <한국항해항만학회지>, 47(4), 182-190.
- 권선희, 이기성(2014), 환율변화에 따른 국내 제조업 수입률 효과 분석, <무역연구>, 10(2), 453-476.
- 권평오(2013), 한국 제조업 부문 해외직접투자의 수출입유발효과에 관한 연구, <통상정보연구>, 15(3), 259-283.
- 박정희, 주혜영(2016), 수출제조업체의 공급업체 개발 지원활동이 상호신뢰 형성과 수출성장에 미치는 영향, <무역연구>, 12(3), 185-204.
- 오혜진, 우서혜, 문희철(2020), GVC 참여가 제조업과 서비스업 수출경쟁력에 미치는 영향 비교분석, <무역학회지>, 45(4), 153-169.
- 이동주(2020), 한국의 생산성 및 효율성이 수출에 미치는 영향 분석: 제조업 11개 부문을 중심으로, <무역상무연구>, 86, 191-218.
- 유희경(2005), 산업별 수익성 비교를 통한 거시적 경영환경의 영향 분석, <경영연구>, 20(2), 73-91.
- 전수민, 이기성(2019), 제조업의 경영성과에 영향을 미치는 요인 분석: 산업별 노동소득 분배율을 중심으로, <무역연구>, 15(2), 213-236.
- 주연화, 강주훈, 박세훈(2015), 한국 제조업의 수출결정요인분석: 고정효과모형과 확률 효과모형, <국제지역학회>, 19(2), 43-59.
- 통계청, 2023년 상반기 지역별고용조사 취업자의 산업 및 직업별 특성:
https://kostat.go.kr/board.es?mid=a10301030300&bid=211&act=view&list_no=427575
- 통계청, KOSIS, 국제유가 도입현황, <재할용가능자원가격조사>, 한국환경공단:
https://kosis.kr/statHtml/statHtml.do?orgId=392&tblId=DT_AA123&vw_cd=MT_ZTITLE&list_id=&scrId=&seqNo=&lang_mode=ko&obj_var_id=&itm_id=&conn_path=E1&docId=0268336543&markType=S&itmNm=%EC%A0%84%EA%B5%AD
- 통계청, KOSIS, 경제활동별 GDP 및 GNI(원계열, 명목), <국민계정>, 한국은행:
https://kosis.kr/statHtml/statHtml.do?orgId=301&tblId=DT_200Y005&vw_cd=MT_ZTITLE&list_id=Q_301009_001_002&seqNo=&lang_mode=ko&language=kor&o

bj_var_id=&itm_id=&conn_path=MT_ZTITLE

통계청, KOSIS, 행정구역(시도)/성별 경제활동인구, <경제활동인구조사>, 통계청:

https://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1DA7014S&vw_cd=MT_ZTITLE&list_id=&scrId=&seqNo=&lang_mode=ko&obj_var_id=&itm_id=&conn_path=E1&docId=02983&markType=S&itmNm=%EC%A0%84%EA%B5%AD

한국은행, 한국은행경제통계시스템(ECOS), 1.3.1., 한국은행 기준금리 및 여수신금리

한국은행, 한국은행경제통계시스템(ECOS), 1.5.4.1., 통합재정수지

한국은행, 한국은행경제통계시스템(ECOS), 3.1.1.1., 주요국 통화의 대원화환율

한국은행, 한국은행경제통계시스템(ECOS), 3.4.1. 해외직접투자(신고기준)

한국은행, 한국은행경제통계시스템(ECOS), 4.1.1., 생산자물가지수

한국은행, 한국은행경제통계시스템(ECOS), 6.1.1.3., 기업경기실사지수(매출액가중 실적)

한국은행, 한국은행경제통계시스템(ECOS), 8.1.2., 경기종합지수

한국은행, 한국은행경제통계시스템(ECOS), 8.1.3., 설비투자지수

한국은행, 한국은행경제통계시스템(ECOS), 8.6.2., 경제활동인구

Cover T.M. and Thomas J.A.(2006), *Elements of Information Theory*, WILEY-INTERSCIENCE.

Chantar, H., Thaher, T., Turabieh, H., Mafarja, M. and Sheta, A.(2021), BHHO-TVS: A binary harris hawks optimizer with time-varying scheme for solving data classification problems, *Applied Sciences*, 11(14), 6516.

Dokeroglu, T., Deniz, A. and Kiziloz, H. E.(2021), A robust multiobjective harris' hawks optimization algorithm for the binary classification problem, *Knowledge-Based Systems*, 227, 107219.

Estevez, P.A., Tesmer, M., Perez, C.A. and Zurada, J.M.(2009), Normalized mutual information feature selection, *IEEE Transactions on Neural Networks*, 20(2), 189-201.

Gunning, D., Stefik, M., Choi, J., Miler, T., Stumpf, S. and Yang, G.Z.(2019), XAI-Explainable artificial intelligence, *Science Robotis*, 4(37), 7120.

Hamed, A. and Mohamed, M.F.(2023), A feature selection framework for anxiety disorder analysis using a novel multiview harris hawk optimization algorithm, *Artificial Intelligence in Medicine*, 143, 102605.

He, K., Zhang, X., Ren, S. and Sun, J.(2015), Delving deep into rectifiers: surpassing human-level performance on imagenet classification, *Proceedings of the IEEE international conference on computer vision*, 1026~1034.

Heidari, A.A., Seyedali, M., Faris, H., Aljarah, I., Mafarja, M. and Chen, H.(2019), Harris hawks optimization: Algorithm and applications, *Future Generation Computer Systems*, 97(2019), 849-872.

Hoffman, K., Lees, J. and Zhang, K.(2023), Local change point detection and

- cleaning of EEMD signals, *Circuits, Systems, and Signal Processing*, 42(8), 4669–4690.
- Huang N.E., Shen, Z., Long, S.R., Wu, M.C., Shih, H.H., Zheng, Q., Yen, N.C., Tung, C.C. and Liu, H.H.(1971), The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis, *Mathematical, Physical and Engineering Sciences*, 454(1971), 903–995.
- Kim, D., Kim, K. D. and Oh, H.S.(2012), Extending the scope of empirical mode decomposition using smoothing. *EURASIP Journal on Advances in Signal Processing*, 2012(1), 168.
- Kim, D and Oh, H. S.(2009) EMD: A package for empirical mode decomposition and hilbert spectrum, *The R Journal*, 1, 40–46.
- Lundberg, S. M., Erion, G. G. and Lee, S. I.(2018), Consistent individualized feature attribution for tree ensembles., arXiv:1802.03888.
- Narkhede, M.V., Bartakke, P.P. and Sutaone, M.S.(2022), A review on weight initialization strategies for neural networks, *The Artificial Intelligence Review*, 55(1), 291–322.
- Nazari, N., Mirsalari, S.A., Sinaei, S., Salehi, M.E. and Daneshtalab, M.(2020), Multi-level binarized LSTM in EEG classification for wearable devices, 2020 28th *Euromicro International Conference on Parallel, Distributed and Network-Based Processing*, 175–181.
- Rothman, D. (2022), *Hands-On Explainable AI(XAI) with Python*, DK ROAD BOOKS.
- Sen, J.(2022), Precise stock price prediction for optimized portfolio design using an LSTM model, 2021 19th *OITS International Conference on Information Technology*, 210–215.
- Shannon, C. (1948), A mathematical theory of communication, *Bell System Technical Journal*, 27(3), 379–423.
- Thaher, T. and Arman, N.(2020), Efficient multi-swarm binary harris hawks optimization as a feature selection approach for software fault prediction, 2020 11th *International Conference on Information and Communication Systems*. 249–254.

SEMD-LSTM based Manufacturing Condition BSI Prediction Model Construction with Factor Analysis

Dae-Young Kim⁶⁾

Abstract

With the recent COVID-19 pandemic, uncertainty has been substantially increased regarding the conditions on manufacturing industry. This research focused on directly or indirectly contributing to the establishments of policies which can be linked to reducing the underlying uncertainties among manufacturing industries by constructing a deep-learning based manufacturing BSI prediction model, while extracting information and marginal contributions of factors related to manufacturing condition BSI under certain time lag. This research used factor selection methods based on SEMD decomposition and NMIFS-HHO algorithms to extract a set of highly relevant covariates, while supervising a LSTM layer based RNN model, which led to the construction of a deep learning model capable of predicting manufacturing condition BSIs with high accuracy. Results show that the considered model can return accurate predictions with extremely low errors, with covariates such as long-term signal(IMF) with a 3 to 4 year period, composite index, and government rate being found important in terms of marginal contributions for prediction using SHAP analysis.

Key words : Manufacturing, Condition BSI, Recurrent Neural Networks, SEMD, NMIFS-HHO algorithm

6) First Author and Corresponding Author, Undergraduate Student, Department of Applied Statistics, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul, Korea. E-mail: lgtlsafg@yonsei.ac.kr