



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

체육학박사 학위논문

불확실성(uncertainty)을 고려한 스포츠 빅데이터 분석

- Bayesian 추정과 Deep Learning을 활용한
프로야구 심판의 'Ball/Strike' 판정 평가 모델 개발 -

2023년 2월

서울대학교 대학원

체육교육과

황수웅

불확실성(uncertainty)을 고려한 스포츠 빅데이터 분석

- Bayesian 추정과 Deep Learning을 활용한
프로야구 심판의 'Ball/Strike' 판정 평가 모델 개발 -

지도교수 박 일 혁

이 논문을 체육학 박사 학위논문으로 제출함
2022년 12월

서울대학교 대학원
체육교육과
황 수 응

황수응의 박사 학위논문을 인준함
2022년 12월

위 원 장 임 충 훈 (인)

부위원장 박 재 범 (인)

위 원 이 기 봉 (인)

위 원 편 현 응 (인)

위 원 박 일 혁 (인)

국문초록

본 학위논문은 ‘불확실성(uncertainty)을 고려한 스포츠 빅데이터 분석’이라는 큰 주제 아래 두 개의 세부 연구로 구성된다. **Study 1 ‘베이지안 확률 접근을 통한 MLB 심판 판정의 불확실성 분석**’에서는 프로야구 현장에서 수집되는 빅데이터를 활용하여 MLB 심판의 판정오류 발생 확률을 추정하였고, 그 과정에서 k-dependence bayesian network classifier(k-DB)와 로지스틱 회귀분석이 어떤 장단점을 갖는지 논의하였다. **Study 2 ‘인공지능을 활용한 프로야구 심판 판정 평가 모델 개발**’에서는 다양한 조건에서의 실험학습을 통해 판정을 수행하는 인공지능 심판(AI UMPIRE)을 개발하였고, 이를 활용하여 실제 심판의 판정을 평가하는 체계를 제안하였다. 각 세부 연구의 결과를 요약하면 다음과 같다.

[Study 1]

k-DB 모수학습 결과를 바탕으로 추정한 MLB 심판의 판정오류 발생 확률은 ‘Ball-Strike’ 카운트에 따라 큰 차이를 보였다. 2019-2021 MLB 정규시즌 경기에서 3balls-0strike일 때 판정오류 발생확률은 약 19~21%로 볼 카운트 조합 중 가장 크게 나타났다. 반면, 0ball-2strikes에서 판정오류 발생확률은 약 4~8%로 앞선 조건과 비교하여 매우 낮게 나타났다. 정규시즌과 월드시리즈에서의 판정오류 발생 확률을 비교한 결과 월드시리즈에서는 정규시즌과 비교하여 판정오류 발생 확률 자체가 낮게 나타나지만, 홈팀 투수에게 유리한 판정을 보이는 성향은 더 뚜렷하게 나타났다. 또한 0ball-2strikes, 3-2 full count와 같이 삼진에 대한 선택적 유보가 불가능한 상황에서는 정규시즌과 월드시리즈 간에 판정오류 발생 확률에서 큰 차이를 보이지 않았다. 그러나 3balls-0strike와 같이 스트라이크 카운트에 여유 있는 상황에서는 정규시즌보다 월드시리즈에서 더 정확한 판정이 수행되고 있는 것으로 나타났다. 이러한 결과는 심판의 볼 판정은 정립된 규칙 하에 일관되게 이루어지지 않고 경기 상황 또는 경기 중요도에 따라 불확실성이 작용하고 있음을 보여준다.

k-DB분석에서 활용하였던 2021년 MLB 정규시즌 자료를 활용하여 로지스틱 회귀분석을 실시한 결과 비교적 간단한 분류문제에 대해서는 k-DB분석 결과와 1% 미만의 차이를 보였다. 그러나 홈/원정, 볼 카운트, 아웃카운트, 직구/변화구 구분, 투구된 공의 위치와 속도 등을 고려해야 하는 다소 복잡한 문제에 대해서 두 기법은 판정오류 발생 확률 추정치에 다소 큰 차이(3~11%)를 보였다. 이와 같은 결과는 로지스틱 회귀분석에서 통계적 유의성을 나타내지 않는 변수가 포함되었을 때 두드러지게 나타났다. 로지스틱 회귀분석에서 독립변수의 통계적 유의성은 사건 발생의 사례 수, 설명변수들의 조합에 따라 다르게 나타날 수 있다. 따라서 스포츠 데이터 분석에서 통계적 유의성을 나타내지 않는 변수일지라도, 실제로 해당 사건이 스포츠 경기 결과에 영향을 미칠 수 있다는 점이 간과되어서는 안 될 것이다. 이렇듯 Study 1에서 두 기법을 비교한 결과는 해결해야 하는 문제가 어떠한 형태인지에 따라 그 효용성이 달라질 수 있음을 시사한다.

[Study 2]

Study 2의 연구목적은 MLB 심판의 'Ball/Strike' 판정 자료를 바탕으로 심층신경망 학습을 통해 판정을 수행하는 인공지능심판(AI UMPIRE)을 개발하고, 이를 활용하여 심판의 판정을 평가하는 체계를 마련하는 것이다. 최적의 AI UMPIRE를 개발하기 위해 다양한 실험학습을 진행하였으며, 개발된 AI UMPIRE는 test set에 대하여 93.18%의 예측 정확도를 나타냈다. 본 연구에서는 AI UMPIRE의 판정 확률에 기반하여 인공지능 심판이 강한 확신을 갖는 판정임에도 불구하고 실제 심판의 판정은 반대로 선언된 판정들을 선별해내어 경기 상황을 분석하고, 이를 시각화하여 심판 판정의 편향(bias)을 분석하였다.

2019년 MLB 정규시즌에서 4000개 이상의 판정을 수행한 심판 50명의 판정을 AI UMPIRE와 비교한 결과, AI는 강한 확신을 갖고 Ball/Strike 판정을 수행했음에도 실제 심판 판정이 반대되는 판정 [$0.9 < \text{ABS}(\text{AI prob.} - \text{umpiring})$]의 비율은 0.61%~2.8%의 범위를 보였다.

본 연구에서는 '0.9<ABS(AI prob.-umpiring)'에 해당하는 판정들을 시각화하여 패턴을 살펴보고, 해당 판정들이 발생하는 경기 상황을 분석하였다. '0.9<ABS(AI prob.-umpiring)'를 기준으로 심판을 서열화한 뒤 심판을 Top 3와 Bottom 3로 구분하였으며, 분석 결과 Top 3 그룹에서는 AI가 강한 확신을 갖고 있는 판정과 반대되는 판정의 분포가 무선적인 형태를 보였지만, Bottom 3 그룹에서는 특정 위치에서 편향적으로 나타났다(그림 21, 22 참고). 또한 '0.9<ABS(AI prob.-umpiring)'에 해당하는 판정은 0Ball-0Strike 카운트에서 가장 높은 비율을 기록하였다. 반면, 3balls, 또는 2Strikes 상황과 같이 볼 또는 스트라이크 판정에 대한 선택적 여유가 없는 상황에서는 인간의 판정과 AI의 판정이 반대되는 비율이 상대적으로 낮게 나타났다. 즉, 인공지능 심판이 강한 확신을 갖고 있는 판정에 반대되는 판정은 선택을 유보할 수 없는 상황보다, 초구(初球)와 같이 후속 선택지가 많은 상황에서 상대적으로 빈번하게 나타났다. 이와 같은 결과는 프로야구 심판의 Ball/Strike 판정은 인간의 선택적 편향이 작용하고 있음을 실증적으로 보여준다.

본 연구는 스포츠 빅데이터를 기반으로 프로야구 심판 판정의 불확실성을 정량화하여 분석하는 데에 그치지 않고, 인공지능 모델을 개발하여 프로야구 현장에서 실질적으로 활용 가능한 방안을 제시하였다는 점에서 선행연구들과 차별성을 갖는다. 본 연구에서 제안한 AI UMPIRE는 그동안 심판의 고유한 영역으로만 간주되었던 스트라이크존을 평가하기 위한 새로운 패러다임으로써 충분한 역할을 할 수 있을 것으로 기대된다.

주요어 : k-dependence bayesian network classifier, 심층신경망, 인공지능 심판, 스포츠 심판 판정

학 번 : 2018-32629

목 차

I. 서 론

1. 연구 배경	1
2. 연구 내용 및 목표	9

II. Study 1

: 베이지안 확률 접근을 통한 MLB 심판 판정 불확실성 분석

1. 이론적 배경	12
1.1. 빅데이터의 정의와 특성	12
1.2. 체육/스포츠 분야 빅데이터의 특성	13
1.3. 빈도주의 통계와 베이지안 통계의 비교	17
1.4. 체육/스포츠 분야 베이지안 통계의 활용	21
2. 연구 방법	27
2.1. 자료 수집 및 처리 절차	27
2.2. k-Dependence Bayesian Network Classifier	31
2.3. 로지스틱 회귀분석	33
2.4. MLB 판정오류의 과급효과 분석	34
3. 연구 결과	37
3.1. 2019-2021 MLB 정규리그 판정오류 분석	37
3.2. 2017-2021 MLB 월드시리즈 판정오류 분석	46
3.3. k-DB와 Logistic Regression 결과 비교	50
3.4. MLB 판정오류의 과급효과 분석	57
4. Study 1에 대한 논의 및 결론	62

III. Study 2

: AI를 활용한 프로야구 심판 판정 평가 모델 개발

1. 이론적 배경	70
1.1. 프로야구 심판 판정의 평가	70
1.2. 기계학습과 딥러닝	75
1.3. 인공신경망(artificial neural network)	76
1.4. 심층신경망(deep neural network)	80
1.5. 체육/스포츠 분야 기계학습, 딥러닝의 활용	82
2. 연구 방법	88
2.1. 자료 수집 및 처리 절차	88
2.2. ‘Ball/Strike’ 판정 인공지능 심판 개발	91
2.3. AI를 활용한 심판 판정 평가	92
3. 연구 결과	94
3.1. DNN을 활용한 ‘AI UMPIRE’ 개발	94
1) 입력층의 구성과 Optimizer	94
2) 손실 함수와 활성화 함수(activation function)	96
3) Dropout	97
4) ‘AI UMPIRE’(인공지능 심판) 개발	99
3.2. AI를 활용한 심판 판정 평가	103
1) 심판 개인 단위 판정 평가	103
2) 심판 판정 자료 시각화 및 경기 상황 분석	105
4. Study 2에 대한 논의 및 결론	116
4.1 인공지능 심판 개발을 위한 이론적 논의	118
4.2 인공지능 심판의 활용을 위한 실용적 논의	121

IV. 종합 결론

: 불확실성을 고려한 스포츠 빅데이터 분석 125

참고문헌 129

Abstract 137

표 목 차

[표 II-1] 스포츠 빅데이터의 특성과 활용 예시	16
[표 II-2] Study 1 연구 자료	30
[표 II-3] k-DB 활용 판정오류 발생 확률 추정(2021)	40
[표 II-4] k-DB 활용 판정오류 발생 확률 추정(2020)	42
[표 II-5] k-DB 활용 판정오류 발생 확률 추정(2019)	44
[표 II-6] k-DB 활용 판정오류 발생 확률 추정(월드시리즈)	47
[표 II-7] 월드시리즈/정규시즌 판정오류 발생 확률 비교	49
[표 II-8] 판정오류에 대한 로지스틱 회귀분석 결과	51
[표 II-9] k-DB와 로지스틱 회귀분석 결과 비교 1	54
[표 II-10] k-DB와 로지스틱 회귀분석 결과 비교 2	55
[표 II-11] 삼진과 볼넷이 MLB 팀 승리에 미치는 영향	58
[표 II-12] MLB 팀 1승의 경제적 가치 추정	59
[표 II-13] 판정오류의 경제적 파급효과 추정	61
[표 III-1] Study 2 딥러닝을 위한 자료의 구성	90
[표 III-2] 심층신경망 최종 모형의 학습 구조	92
[표 III-3] Optimizer와 Layer 구성에 대한 실험학습 결과	95
[표 III-4] 손실함수와 활성화함수에 대한 실험학습 결과	96
[표 III-5] Dropout에 대한 실험학습 결과	98
[표 III-6] AI UMPIRE 학습 구조	100
[표 III-7] AI를 활용한 심판 판정 평가	104
[표 III-8] AI와 반대되는 판정에 대한 경기 상황 분석(%)	113
[표 III-9] AI와 반대되는 판정에 대한 경기 상황 분석(빈도)	114

그림 목 차

[그림 II-1] Baseball Savant 검색 예시	29
[그림 II-2] k-DB 구조학습 결과(2021)	37
[그림 II-3] k-DB 구조학습 결과(2020)	41
[그림 II-4] k-DB 구조학습 결과(2019)	43
[그림 II-5] k-DB 구조학습 결과(월드시리즈)	46
[그림 II-6] 구종에 따른 판정오류 발생 확률	54
[그림 II-7] plate_x와 out counts에 따른 판정오류 발생 확률	56
[그림 III-1] 인공신경망의 기본 구조	77
[그림 III-2] 퍼셉트론의 구조(Rosenblatt, 1958)	78
[그림 III-3] 다층퍼셉트론 구성 예시(Valero, 2016)	79
[그림 III-4] 심층신경망 구성 예시	80
[그림 III-5] Dropout이 포함된 모형의 구조	97
[그림 III-6] Dropout에 따른 검증 손실과 예측 정확도	99
[그림 III-7] AI UMPIRE 모형 구조 요약	101
[그림 III-8] AI UMPIRE 학습 결과	102
[그림 III-9] Rank 1-3 판정 자료[ABS(prob.-umpiring)<0.1]	107
[그림 III-10] Rank 48-50 판정 자료[ABS(prob.-umpiring)<0.1]	108
[그림 III-11] Rank 1-3 판정 자료[0.4<ABS(prob.-umpiring)<0.6]	109
[그림 III-12] Rank 48-50 판정 자료[0.4<ABS(prob.-umpiring)<0.6]	110
[그림 III-13] Rank 1-3 판정 자료[0.9<ABS(prob.-umpiring)]	111
[그림 III-14] Rank 48-50 판정 자료[0.9<ABS(prob.-umpiring)]	112

I. 서론

1. 연구 배경

“한 마디만 더해도 될까요? 부탁 아닌 부탁을 하고 싶어서요. 지금 저희가 세 경기밖에 안 지났지만 볼 판정의 일관성에 대해 불만이 많거든요. 저희는 안타 하나치기 위해 새벽 3시까지 스윙하고, 안타 하나를 위해 피나는 노력을 하는데, (볼 판정이) 너무 헛갈리는 부분이 많아서 선수 입장 조금만 생각해주셔서 신중하게 잘 봐주셨으면 좋겠습니다.”

- 2020년 5월, KBO리그 경기 종료 직후 A선수의 인터뷰 내용 중-

“*Strikes Out!*” 심판의 세 번째 스트라이크 판정과 함께 경기가 종료된다. 위의 인터뷰는 2020년 5월 한국야구위원회(KBO; Korea Baseball Organization) 정규시즌 경기가 끝난 직후에 이뤄진 한 선수의 인터뷰 내용 일부이다. 이후 해당 경기의 심판위원들은 전원 1군 경기에서 배제되어 퓨처스리그(1군 바로 아래 수준의 2군 리그)로 이동하게 되었다. 프로야구에서 심판 판정의 일관성, 판정 오류(missed call)에 대한 이슈는 선수들의 직관은 물론, 실증적 자료로도 드러나고 있다. Williams(2019)에 따르면 2018년 미국프로야구(MLB; Major League Baseball)에서 ‘Ball/Strike’ 판정오류와 함께 종료된 경기는 총 55경기로, 이는 55경기의 결과는 현재와 다르게 기록될 수도 있었음을 의미한다(황수웅, 박일혁, 2020). 4차 산업혁명, 빅데이터, 인공지능에 대한 관심이 단순한 유행을 넘어서 새로운 패러다임(paradigm)으로 자리 잡고 있는 현대사회의 흐름을 고려한다면 스포츠 경기에서 발생하는 오심(誤審) 역시 더 이상 경기의 일부로 간주될 것이 아닌, 제도적 방안이 필요한 사안이다(Collins, 2010). 축구, 농구 등의 종목에서는 주어진 경기 시간이

모두 흐르면 자연스럽게 경기가 종료되는 반면, 야구경기는 9회에 세 번째 아웃을 선언하는 심판의 판정이 있어야만 경기가 종료된다. 즉, 야구경기는 심판의 볼 판정 하나로 공격기회가 이어질 수도 있고, 반대로 심판의 볼 판정과 함께 바로 종료될 수 있는 구조적 특성을 갖는다. 따라서 프로야구 심판 판정의 공정성과 일관성에 대한 논의는 타 구기종목에 비하여 활발히 이루어지고 있다.

프로야구 심판 판정에 대한 연구는 다양한 관점에서 이루어졌다. Parson 등(2011)의 연구에서는 심판과 투수의 인종 일치 여부에 따라 Strike 판정이 차별적으로 나타나는지 실증적으로 검증하였다. Kim과 King(2014)의 연구에서는 투수의 MLB 올스타전 출전 횟수를 투수의 명성으로 간주하여, 투수의 명성에 따라서 심판의 불공정 판정 발생이 차이를 보이는지 검증하였다. Chen 등(2016)은 선행된 볼 판정 결과가 다음 판정에 영향을 미치는지 통계적으로 검증하여 도박사의 오류(gambler's fallacy) 관점에서 프로야구 심판의 볼 판정을 설명하였다. 또한 황수웅과 박일혁(2020)은 경기 상황과 선수에 대한 정보만으로 투구 판정을 설명하기에는 한계가 있음을 지적하며 구속, 회전수, 구종 등과 같은 투구된 공의 질(quality)을 나타내는 변수를 경기 상황 정보와 함께 고려하여 MLB심판의 판정 오류를 설명하였다.

다양한 선행연구들의 결과를 종합하면, 프로야구 심판의 판정은 투구된 공의 속도, 회전수, 볼 카운트, 구종, 인종, 선수의 명성 등 다양한 요인에 영향을 받고 있다. 즉, 야구에서 볼 판정은 정립된 규칙 하에서 이루어져야 함에도 불구하고, 볼 판정을 선언하는 객체인 심판마다 스트라이크 존(strike zone)을 바라보는 기준에 차이를 보이고 있으며, 실시간으로 달라지는 경기 상황 정보에 따라서 부작위 편향(omission bias), 영향 회피(aversion) 등 인간의 심리가 판정에 작용하는 것으로 밝혀졌다.

프로야구에서 심판의 'Ball/Strike' 판정은 리그 운영 주체로부터 정립

된 규칙 하에 이루어져야 한다. 그러나 실제로 이루어지는 프로야구 심판의 의사결정에서는 개인의 주관이 반영되는 경우가 많기 때문에, 판정은 심판마다 조금씩 차이를 보인다. 관련 선행연구에서는 타자의 키, 타격 자세(batting stance) 등에 따라 심판마다 스트라이크 존을 보는 기준이 다르고, 볼 카운트, 점수 차, 이닝, 선행된 사건, 투구된 공의 질(quality) 등의 경기 상황 정보에 따라 인간의 심리가 다르게 작용하기 때문이라고 보고하고 있다(Parson et al., 2011; Kim, & King, 2014; Green, & Daniels, 2014; Chen et al., 2016). 본 연구에서는 스포츠 상황에서 발생하는 다양한 사건 중 원인과 결과의 관계가 선형관계만으로 정확히 예측이 어려운 현상을 '불확실성(uncertainty)'이라는 속성으로 조작적으로 정의하여 연구를 진행하고자 한다. 즉 프로야구 심판 판정에 영향을 미치는 다양한 요인들을 불확실성 요인으로 간주하여, 동시다발적으로 변화하는 경기 상황과 관련된 정보들을 종합적으로 고려하여 프로야구 심판의 판정을 평가하고자 한다.

현대인들이 일상에서 반복적으로 경험하는 현상은 일정한 규칙 아래 발생하기 때문에 원인과 결과가 비교적 명확하게 드러난다. 예를 들어, 생리학적으로 개인의 음식 섭취량이 증가하였으나 신체활동량이 줄어드는 상황을 가정한다면, 결과적으로 근육량은 감소하고 지방의 양은 증가하게 된다. 또한 경제학의 관점에서 특정 재화의 공급이 일정하게 고정되어 있다고 가정했을 때, 수요가 증가하면 가격은 상승한다. 이렇듯 우리가 일상에서 경험하는 현상들은 비교적 인과관계가 명확히 드러나는 경우가 많다. 이에 반하여, 스포츠 경기에서 발생하는 사건들은 과거에 발생했던 사건이 같은 조건에서 반복되는 것은 매우 드문 일이다. 스포츠 경기에서는 다양한 변수들이 유기적으로 상호작용하여 특정한 하나의 결과로 나타나며, 변수들의 상호작용은 실시간으로 변하는 경기 상황에 따라 또 다른 형태로 나타난다. 따라서 스포츠 상황에서 발생하는 다양한 사

건을 심층적으로 이해하고 정확하게 예측하기 위해서는 ‘불확실성’이라는 스포츠가 갖는 특성을 충분히 고려할 수 있는 연구설계가 뒷받침 되어야 할 것이다. 이와 같은 맥락에서 Gudmundsson과 Horton(2017)은 기존의 스포츠 통계분석이 득점, 평균 등 빈도에 초점을 맞추었다면, 보다 발전된 스포츠 분석에서는 선수들 간의 상호작용과 같은 실시간으로 변화하는 복잡하고 불확실한 문제를 고려할 수 있어야 한다고 보고하였다. 또한 Nalbantis와 Pawlowski(2019)는 금융 및 보험시장의 흐름을 이해하는 데 있어 불확실성을 고려하듯이, 스포츠 상황에서도 특정 사건을 예측하고 설명하기 위해서는 불확실성을 충분히 고려해야 한다고 그 중요성을 강조하였다.

스포츠는 제도화된 규칙 하에서 공정하게 신체적 탁월성을 겨루는 행위로 정의된다(안용규 외, 2011; Parry, 2019). 즉, 스포츠 경기는 본질적으로 ‘공정한 경쟁’을 전제로 한다. 그러나 선행연구들에 따르면 실제로 프로야구 심판의 판정은 다양한 요인에 영향을 받는 ‘불확실성’이라는 속성을 내포하고 있다. 스포츠 경기에서 ‘공정한 경쟁’이라는 전제가 충족되기 위해서는 심판 판정의 불확실성이 최소화되어 일관된 판정이 이루어져야 할 것이다(Huang & Hsu, 2020). 이를 실현하기 위해서는 심판 판정에 영향을 미치는 불확실성 요인에 대한 심층적인 이해와 해당 요인들을 충분히 고려할 수 있는 연구가 필요하다.

프로야구 심판의 볼 판정과 스트라이크 존에 대한 논란은 프로야구가 시작된 이래로 꾸준히 이루어졌다. 그럼에도 불구하고, 2000년대 초반까지는 경기 중 발생하는 다양한 자료를 실시간으로 처리할 수 있는 기술력이 뒷받침되지 못했기 때문에, 투구된 공에 대한 추적을 비롯하여 경기 중 발생하는 불확실한 사건들을 실증적으로 분석하기에는 한계가 있었다. 그러나 2000년대 중반부터 Internet of things (IOT)기술, cloud computing과 같은 정보통신기술의 비약적인 발달로 인해, 대용량의 자료

를 실시간으로 처리할 수 있는 빅데이터 기술이 보편화되면서(Jin et al., 2015) 스포츠 현장에서도 경기 중 발생하는 다양한 형태의 대용량 자료를 실시간으로 처리할 수 있는 환경이 갖추어지게 되었다. 실제로, MLB에서는 2009년 투구추적시스템 (PITCHf/x)이 도입되기 시작하면서 투구된 공을 3차원으로 추적하고 이를 실시간으로 처리하기 시작하였다. 2015년에는 군사 미사일을 추적하는 데 활용되었던 도플러 레이더(doppler radar)기반의 트랙맨 시스템이 도입되면서 투구된 공에 대하여 초당 약 2만 개의 자료를 수집하여 공의 궤적 및 도착점을 실시간으로 추정하는 것이 가능해졌다(Mills, 2017).

빅데이터에 대한 정의는 학자마다 일부 차이를 보이나, 대체로 방대한 양(volume), 실시간으로 이루어지는 자료 처리 속도(velocity), 자료 구성 및 형태의 다양성(variety), 자료 수집 및 처리의 진실성(veracity), 가치(value), 즉 '5Vs'의 특성을 지닌 자료로 합의되고 있다(Shah et al., 2019). 이를 스포츠 상황에 적용해 보면, 스포츠 빅데이터 분석은 스포츠 현장에서 발생하는 다양한 형태의 방대한 양의 자료를 실시간으로 수집 및 처리하여 정확한 결과를 산출하고, 그 결과가 스포츠 현장에서 유의미한 가치를 도출해낼 수 있는 절차로 이해할 수 있을 것이다(최형준, 2020; Morgulev, Azar & Lidor, 2018; Rein & Memmert, 2016; 조정환, 2012). 이렇듯 스포츠 현장에서의 빅데이터의 활용은 정보통신기술의 발달과 함께 대용량 정보의 저장이 가능해지고, 수집된 자료를 실시간으로 처리할 수 있는 환경이 구축되면서 점차 보편화되고 있다(황수웅, 양혜주, 황선환, 2022).

그러나 일부 연구에서는 빅데이터 기술이 보편화되었을지라도 자료의 특성과 상황에 맞는 연구설계와 통계기법에 대한 이론적인 고찰이 충분히 이루어지지 않는다면 그 가치는 제한적일 수밖에 없음을 지적하고 있다(Liu et al., 2016; Becker, King, & McMullen, 2015; Labrinidis &

Jagadish, 2012). 특히, 스포츠 경기와 같이 X라는 설명변수가 투입되었을 때 Y라는 결과가 선형관계만으로는 명확하게 드러나지 않는 불확실성이라는 특성을 갖는 경우에는 더욱 그러할 것이다.

전통적인 체육/스포츠 분야 양적연구에서는 연구자가 관심 있는 현상에 대하여 빈도주의 기반의 통계적 가설검증 절차를 통해 가설의 기각(reject) 여부에 따라 일정한 연구 결론에 닿는다. 그러나 이와 같은 가설검정에서는 설정한 가설의 기각 여부에 따라 이분법적인 결론에 이르게 되기 때문에, 이외의 다양한 가능성을 고려하지 못한다는 한계가 있다. 이와 같은 맥락에서 일부 연구에서는 전통적인 귀무가설 검정(null hypothesis significance testing)은 가설의 기각 여부 이외의 불확실성에 대한 가능성을 배제하고 있음을 지적하며, 체육/스포츠 분야 연구에서 베이저안 통계기법을 활용한 연구설계의 필요성을 강조한 바 있다(Bernards et al., 2017; 김현태, 강민수, 2019).

더불어, 체육/스포츠 분야 연구에서 특정 현상에 영향을 미치는 변수를 탐색하고 그 영향력을 검증하기 위해 선형회귀분석 또는 로지스틱 회귀분석 등의 통계기법을 주로 활용되고 있다. 그러나 해당 분석기법은 설명변수와 종속변수 간의 관계가 선형이 아닌 경우에는 예측력이 크게 저하된다는 한계점이 있다. 다차원 비선형함수로 구성된 예측함수의 적용을 고려해 볼 수 있겠지만, 차원이 증가할수록 필요한 학습데이터의 수가 기하급수적으로 증가하기 때문에 이 또한 현실적으로 어려움이 따른다. 스포츠 상황에서 발생하는 다양한 사건들은 단순히 'X→Y'의 선형관계를 나타내기보단, 선행되는 원인 사건들이 복잡한 상호작용을 통해 발생하는 경우가 많다. 따라서 본 연구에서 조작적으로 정의한 스포츠에서 발생하는 '불확실성'을 선형관계를 다루는 모형만으로 설명하기에는 한계가 있을 것이다.

선형회귀분석과 로지스틱 회귀분석은 모수(parameter)를 추정하기 위

해 각각 최소제곱법(OLS; ordinary least square)과 최대우도법(MLE; maximum likelihood estimation)을 활용한다. 두 기법은 모수를 하나의 고정된 값으로 가정하고, 산출된 회귀계수 값을 해석하여 독립변수가 종속변수에 미치는 영향을 설명한다. 즉, 선형회귀모형과 로지스틱 회귀모형에서는 직접적으로 영향을 미치는 독립변수 1개 이외의 변수들은 일정하게 통제된 상황으로 가정한다. 이를 스포츠 상황에 적용해 보면, 스포츠 경기 중 발생하는 다양한 사건을 고정된 하나의 값으로 가정하고 결과를 해석해야 하는 상황에 직면한다. 따라서 모수를 하나의 고정된 값으로 가정할 경우 실질적으로 스포츠 현장에서 발생하는 불확실성의 특성을 갖는 사건들을 설명하고 예측하기에는 한계가 있다. 또한 독립변수의 수와 차원이 증가할수록 모형에 대한 해석은 더욱 복잡해질 것이다.

실시간으로 수집되는 방대한 양의 데이터를 활용하여 특정한 결과를 도출해낼지라도, 스포츠가 갖는 ‘불확실성’이라는 특성을 충분히 고려할 수 있는 연구설계가 뒷받침되지 못한다면 그 결과의 활용은 매우 제한적일 것이다(황수웅, 양혜주, 황선환, 2022). 이에 대하여 스포츠 빅데이터를 활용한 최근의 연구들에서는 베이지안 추정(bayesian estimation)과 인공지능(artificial intelligence, AI) 기법을 활용한 연구설계가 주목받고 있다. 베이지안 통계는 조건부 확률의 개념을 반영하는 통계기법으로 모수의 참값을 가정하지 않고 다양한 정보를 확률분포로 반영하여 특정 사건이 발생할 확률을 결과로 제시한다. 즉, 베이지안 통계에서는 모수에 대한 점추정이 아닌 확률분포에 기반하여 불확실성을 설명하고, 새로운 데이터의 활용이 가능해지면 학습을 업데이트하여 상황에 맞는 새로운 결과를 제시한다(Kaplan & Depaoli, 2012; Gelman, & Shalizi, 2013; van de Schoot & Depaoli, 2014). 따라서 베이지안 추정은 스포츠 경기 상황에서 실시간으로 변화하는 다양한 정보를 반영하여 유연한 분석모형을 구축할 수 있다는 장점이 있다(Kirk, 2003).

AI는 인간의 인식, 판단, 추론, 문제해결 능력, 학습기능과 같은 인간의 두뇌 작용을 이해하는 것을 목적으로 하며, 현재 AI 분야에서 널리 사용되는 방법론은 기계학습(machine learning)이다(McCarthy, 2007). 기계학습이란 주어진 자료에 대하여 컴퓨터가 일정한 규칙을 찾고 적합한 모형을 탐색하여 특정한 결과를 예측하는 방법을 의미한다. 기계학습의 목적은 인간의 힘으로는 특정한 규칙을 정의하기 힘든 현상에 대처하기 위해 기계에 데이터를 학습시켜 문제 해결을 자동화하는 것이다(Jordan, & Mitchell, 2015). 딥러닝(deep learning)은 기계학습에서 신경망(neural network)의 계층(layer)을 보다 깊고 복잡하게 발전시킨 형태에 해당한다. 딥러닝은 현재 이미지 인식 및 분류뿐만 아니라 의학, 생물학, 경영학 등 다양한 분야에서 특정 현상을 예측하고 분류하는데 활용되고 있으며, 모든 분야에서 높은 예측력을 보이고 있다(고학수 외, 2021). 딥러닝 모형은 전통적인 빈도주의 통계에서의 분류기법들과 달리 수많은 입력변수와 깊은 선형/비선형구조를 취한다. 따라서 딥러닝 모형에서는 선형 관계뿐만 아니라 다차원 비선형 관계를 고려할 수 있으며, 이를 통해 인간이 일정한 규칙으로 정의하기 어려운 현상들에 대한 예측이 가능하다. 이렇듯, 베이지안 접근과 인공지능 기법은 스포츠 상황에서 발생하는 빅 데이터를 분석하는 데에 다양한 이점이 있다(Torgler, 2020; Santos, Wu & Mengersen 2019; Gudmundsson & Horton, 2017).

상기한 내용과 같이, 베이지안 확률 접근/인공지능 기법을 활용한 빅 데이터 분석에서는 정형화된 자료와 비정형자료 모두를 활용하여 특정 현상에 대한 세분화된 분류와 보다 정확한 예측을 가능하게 한다(Yau & Campbell, 2019). 그러나 이러한 장점에도 불구하고 체육/스포츠 분야 연구에서 그 활용은 심리학, 경영학, 사회학 등의 타 사회과학 분야와 비교하여 매우 미비한 실정이다(Stenling et al., 2015). 베이지안 추정과 인공지능 기반의 모델링은 역사적으로 오래된 이론임에도, 실제로 분석을

수행하기 위한 컴퓨터 기술력은 2010년대가 되어서 보편화되기 시작했다. 따라서 그동안 이루어졌던 대부분의 체육/스포츠 분야의 연구는 전통적인 빈도주의 통계분석에 기반한 연구설계가 반복되었으며, 베이지안 통계와 인공지능 기법과 같은 새로운 패러다임에 대한 이론적 논의는 활발히 이루어지지 못했다(Bernards et al., 2017). 2000년대 중반부터 빅데이터 관련 기술은 비약적으로 발전하였으며, 이에 따라 스포츠 현장에서 발생하는 자료 역시 형태와 용량에서 과거와 큰 차이를 보이고 있다. 즉, 스포츠 현장에서 발생하는 자료의 구성이 달라지고 있는 만큼 연구설계 역시 전통적인 접근에서 나아가 보다 다양한 상황을 고려할 수 있는 베이지안 추정과 인공지능 기법의 활용이 필요하다. 본 연구는 이와 같은 문제의식에서 출발한다. 본 연구의 구체적인 목적과 내용은 다음과 같다.

2. 연구 내용 및 목표

상기한 바와 같이, 불확실성을 고려한 스포츠 빅데이터 분석에서 베이지안 추정과 인공지능 기법의 활용은 다양한 장점을 갖는다. 베이지안 추정에서는 실시간으로 업데이트되는 정보를 반영하여, 모수를 고정된 값이 아닌 상황 변화에 따른 확률로 제시할 수 있기 때문에(Kaplan & Depaoli, 2012; van de Schoot & Depaoli, 2014) 스포츠 상황에서 발생하는 불확실성 요인들을 더욱 체계적으로 고려할 수 있을 것이다. 또한 인공지능 기법은 정형화된 자료는 물론 비정형화된 데이터를 활용하여, 선형관계뿐만 아니라 비선형의 심층적인 관계를 고려한 학습이 가능하기 때문에, 빈도주의에 기반한 분류기법들과 비교하여 예측력이 더 높게 나타나는 것으로 밝혀졌다(Barbee, 2020; Russell & Reale, 2021).

그럼에도 불구하고, 체육/스포츠 분야에서 베이지안 추정과 인공지능 기법의 활용은 매우 미비한 실정이다. 이에 대한 원인은 베이지안 추정과 인공지능 기법을 활용한 체육/스포츠 분야 연구 대부분이 이론적 틀

을 제시하기보단, 단편적인 분석을 통한 결과 제시에 그치고 있기 때문이다. 일부 연구에서 스포츠 빅데이터 분석을 위해 베이지안 추정과 인공지능기법 활용을 강조하고 있으나, 해당 기법 활용의 이점을 스포츠 빅데이터의 관점에서 체계적으로 정리한 연구는 찾기 어려운 실정이다. 본 연구는 스포츠 빅데이터를 활용한 분석을 통해 프로야구 심판의 판정의 불확실성을 정량화하여 분석하고, 인공지능을 활용하여 심판의 평가하는 체계를 제안하는 것을 목적으로 한다. 이를 통해 스포츠 빅데이터가 갖는 특성을 충분히 고려할 수 있는 연구설계를 위한 이론적 틀을 제공하고자 한다. 연구 목적에 따른 구체적인 연구 내용은 다음과 같다.

본 학위논문은 ‘불확실성(uncertainty)을 고려한 스포츠 빅데이터 분석’이라는 큰 주제 아래, 총 두 개의 세부 연구로 구성된다. Study 1 ‘베이지안 확률 접근을 통한 MLB 심판 판정의 불확실성 분석’의 연구목적은 첫째, MLB에서 발생하는 심판의 판정오류 발생 확률을 정량화(quantify)하여 분석하고, 둘째, 스포츠 빅데이터를 활용한 프로야구 심판의 판정오류 발생 확률을 추정함에 있어 k-DB와 로지스틱 회귀분석이 각각 어떤 장단점을 갖는지 실증적인 근거를 통해 논의하는 것이다. 이를 위한 첫 단계로 문헌고찰을 통해 스포츠 빅데이터의 정의와 특성을 개념화하고 스포츠 빅데이터 분석에 있어 불확실성 요인을 고려해야 하는 당위성에 대하여 정리하였다. 그 다음 단계로, 프로야구 심판의 판정오류 설명함에 있어 베이지안 네트워크 분석과 최대우도법기반의 로지스틱 회귀분석을 실시하여 각 기법이 어떤 장단점을 갖는지 논의하였다. 또한 Study 1의 마지막 단계에서는 MLB 심판의 판정오류가 직관적으로 어느 수준의 영향력을 갖는지 분석하기 위해 특정한 상황을 가정했을 때, 판정오류가 유발하는 경제적 파급효과를 추정하였다.

Study 2의 연구목적은 MLB 심판의 ‘Ball/Strike’ 판정 자료를 바탕으로 딥러닝을 수행하여 판정을 수행하는 인공지능심판(AI UMPIRE)을 개발

하고, 이를 바탕으로 인공지능에 기반하여 심판의 판정을 평가하는 체계를 마련하는 것이다. 이를 위해 딥러닝 학습에서 가장 기본적인 모형인 Deep Neural Network(이하 DNN) 모형을 기초로 하여, 스포츠 빅데이터 분석의 관점에서 바람직한 자료의 형태, 용량, layer의 구성, optimizer에 따른 결과 비교 등에 대하여 논의하고자 한다. 나아가 개발된 ‘AI UMPIRE’에 심판 개인의 판정 자료를 반영하여 심판 판정을 정량적으로 평가하는 체계를 제안하고자 한다. 이와 같은 절차를 통해, 프로야구 심판 개인의 판정을 평가함에 있어 인공지능 모델의 개발이 어떤 의의를 갖는지 논의를 확장하고자 한다.

II. Study 1: 베이지안 확률 접근을 통한 MLB 심판 판정의 불확실성 분석

1. 이론적 배경

1.1. 빅데이터의 정의와 특성

‘Big data’라는 용어 사용의 기원은 크게 산업계와 학계의 구분으로 나뉜다. 산업계에서는 1990년대 중반 Silicon Graphics Inc.의 John Mashey(1999)와, Gartner Inc.의 Laney(2001)로 알려져 있으며, 학계에서 공식적으로 출판된 자료 중에서는 Weiss와 Indurkha(1998), Diebold(2000) 등이 그 시초로 알려져 있다(Diebold, 2012). 1990년대 중반에서부터 2000년대 초반까지 빅데이터의 개념은 주로 자료의 저장과 관리의 최적화 관점에서 이루어졌다. John Mashey(1999)는 빅데이터에 대하여 ‘일반적으로 사용되는 소프트웨어가 자료를 일정시간 내에 저장하고 처리할 수 없는 크기의 데이터’라 하였으며, Laney(2001)은 현재 빅데이터의 중요한 속성으로 받아들여지고 있는 정보의 양(volume) 처리 속도(velocity), 자료 형태의 다양성(variety)의 관점에서 빅데이터를 설명하였다. 이후 2010년대에는 빅데이터의 특성에 대하여 대용량의 자료를 저장하고 관리하는 차원에서 나아가 실시간 자료 처리(velocity), 자료 구조의 다양성(variety), 정확성 및 진실성(veracity), 가치(value) 등 보다 다양한 관점에서 빅데이터를 설명하고자 하였다(황수웅, 양혜주, 황선환, 2022).

그러나 최근까지도 빅데이터에 대한 정의는 학자마다 조금씩 차이를 보이고 있다. 이에 대하여 Andrea, Marco, Michele(2019)은 빅데이터가 학계, 산업계 모두에서 트렌드로 자리잡고 널리 활용되고 있음에도 개념과 정의는 여전히 모호한 부분이 많음을 지적하며, 빅데이터의 정의와

개념에 대한 이론적 틀을 제시하고자 하였다. Dijcks(2014)는 빅데이터의 특성으로 volume, velocity, variety, value를 제시하며, 방대한 정보의 양, 빠른 자료 수집 및 처리 속도, 다양한 자료의 형태, 유의미한 가치를 갖는 자료의 특성을 모두 충족하는 자료라 정의하였다.

Shah 등(2019)의 연구에서 역시 빅데이터는 거대하고 복잡하며 정형화된 데이터뿐만 아니라 비정형 자료도 포함되며, 5가지 주요 특성은 정보의 용량(volume), 정보처리 속도(velocity), 다양성(variety), 정확성(veracity), 정보의 가치(value)라 보고하였다. 여기서 ‘용량(volume)’은 처리되고 저장되는 정보의 양(quantity)을 의미하며, ‘속도(velocity)’는 데이터가 축적되고 처리되는 속도를 의미한다. 대표적인 예로 실시간 응용 프로그램, 실시간 센서 기록, 이벤트 로그와 같은 실시간 자료들이 해당된다. 다양성(variety)은 수많은 형식으로 데이터를 제공하는 데이터 소스의 구성을 의미하며(Patel et al., 2020), ‘진실성(veracity)’은 자료의 출처, 데이터의 편향, 모호성, 방법론 및 기술 적용과 관련된 진실성을 의미한다(Rubin & Lukoianova, 2013). ‘가치(value)’는 복잡하고 다양한 방대한 양의 자료가 실시간으로 수집된다 할지라도 빅데이터로써 활용되기 위해서는 우리사회에 유의미한 가치를 내포하고 전달할 수 있어야 함을 의미한다(Zikopoulos & Eaton, 2011). 현재 빅데이터의 특성은 학자마다 일부 의견차이를 보이지만 대체로 정보의 용량(volume), 정보처리 속도(velocity), 복잡성(complexity), 정확성(veracity), 다양성(variety), 정보의 가치(value) 등이 주요한 특성으로 알려져 있다.

1.2. 체육, 스포츠 분야 빅데이터의 특성

체육/스포츠 분야에서도 빅데이터 분석에 대한 정의와 특성에 대한 논의는 다양하게 이루어졌다. Watanabe, Shapiro, Drayer(2021)는 체육/스포츠 분야의 빅데이터 분석의 성장에 대하여 논의하고, 미래 스포츠

매니지먼트 분야에서 빅데이터를 활용하기 위한 방향성에 대하여 정리하였다. 연구진은 다양한 학자들이 정의한 빅데이터를 스포츠의 맥락에서 보다 구체적으로 제시하고자 하였다. 그 예로, 스포츠 경기 상황에서 발생하는 빅데이터는 선수들의 경기력을 나타내는 단순 통계뿐만 아니라, 경기에 참석하는 관중들에 대한 다양한 정보를 활용한 마케팅 전략의 수립, 경기 관람 중 소셜 미디어 사용자의 반응, 특정 스포츠 팀 팬들의 조직 행동 등을 설명하는 데에 활용될 수 있음을 빅데이터의 특성 중 하나인 ‘volume’ 차원에서 제시하였다. 나아가 ‘variety’ 차원에서 디지털 플랫폼 상에서 스포츠를 소비하고자 하는 객체들이 서로 상호작용하는 데이터를 활용하여 사회연결망분석(SNA; social network analysis)을 적용한 연구를 예로 설명하였다. 해당 연구에서는 과거 스포츠 현장에서 데이터의 활용은 대부분 운동선수들의 경기력 향상에 초점이 맞춰 이루어졌으나, 최근에는 스포츠 관람 티켓 판매, 효율적인 스포츠 조직의 운영, 마케팅 전략 수립 등 다양한 분야에서 활용되고 있다고 보고하였다.

Lopez(2020)는 빅데이터 분석은 스포츠 조직 내 개인의 성과를 평가할 수 있는 새로운 방법을 모색하고 있는 운영주체들에게 혁신적인 방안이 될 수 있다고 보고하였다. 조직 내 구성원에 대한 데이터를 수집하고 이를 활용한 평가 체계는 조직 운영 환경을 개선시킬 수 있는 효율적인 방법이 될 수 있음을 강조하였다. 그러나 해당연구에서 빅데이터는 다양한 분야에서 혁명의 원천으로 비춰지고 있지만, 스포츠에서 발생하는 사건의 복잡성과 불확실성을 고려하면 여전히 부족한 부분이 많음을 지적하며 스포츠 분야에서 빅데이터를 활용한 연구가 더욱 발전된 형태로 이루어져야 함을 강조하였다.

국내 체육·스포츠 분야에서도 빅데이터 활용에 대한 논의와 활용은 다양한 관점에서 이루어졌다. 빅데이터 활용과 전망을 논의한 조정환(2012)은 빅데이터는 대용량(volume), 다양화(variety), 실시간에 가까운

속도(velocity)의 특성을 갖춘 데이터라고 정의하였다. 해당 연구에서는 스포츠 분야 빅데이터 활용 예시로 다양한 생체신호를 활용한 경기력 분석, 실시간 모멘텀(momentum)지수를 적용한 특정 시점의 경기 상황 분석, 스포츠 마케팅 분야에서 스포츠 팬의 프로파일링, 인공지능을 활용한 스포츠 팬 수요 예측, 사회적 관계 분석 등의 사례를 소개하였다. 최형준(2020)에 따르면 빅데이터 분석을 자료의 용량(volume), 처리 속도(velocity), 다양성(variety), 정확성(veracity), 가치(value)의 특성을 갖춘 자료를 바탕으로 자료의 정제, 분석, 탐색, 분류 및 시각화 등의 절차를 포함한다. 즉, 스포츠 분야에서 빅데이터 분석은 다양한 형태를 갖는 방대한 양의 자료를 신속하게 수집 및 처리하여 정확한 결과를 산출하고, 그 결과가 스포츠 현장에서 유의미한 가치를 도출해낼 수 있는 절차로 그 의미를 정의할 수 있을 것이다.

한편, Baerg(2017)는 스포츠 디지털 정보격차의 중심에는 선수 개인의 움직임, 플레이 스타일 등과 같이, 선수의 특성을 정량화하기 위한 관찰 시스템이 자리 잡고 있다고 보고하였다. 선수 관찰 시스템은 과거부터 스포츠 현장에서 이루어졌으나, 최근 빅데이터 기술의 발달로 팀 전술 전략 수립을 위한 자료뿐만 아니라 개인의 생체정보와 같이 민감한 정보가 동시에 수집 및 분석되고 있음을 지적하였다. Baerg는 스포츠 선수가 조직의 통제 아래 추적(tracking)과 관찰을 무조건 수용할 필요가 없음을 강조한다. 특히 대부분의 선수들이 빅데이터를 제공 받더라도 그 자료로 무엇을 해야 할지 모르는 인식론적 빈곤의 대상이 될 수 있음을 문제점으로 지적하였다. 따라서 운동선수들은 개인정보보호 차원에서 경기 중 추적의 대상이 되는 것을 거부할 수도 있고, 또는 자신만의 데이터를 생성하는 방법으로 수용할 수도 있는 환경이 갖추어져야 함을 강조하였다. 이러한 스포츠 빅데이터에 대한 또 다른 관점은 아무리 빅데이터 기술이 보편화될지라도 정보의 수집, 저장 및 관리, 개인정보 보호 등과 관련된

윤리적 고찰이 꾸준히 이루어져야 함을 시사한다. 앞선 내용들을 종합하여 빅데이터가 갖는 ‘5Vs’의 특성을 기준으로 스포츠 빅데이터의 적용 사례를 표로 정리하면 다음과 같다.

표 II-1. 스포츠 빅데이터의 특성과 활용 예시(출처: 황수웅, 양혜주, 황선환, 2022 재구성)

특성	설명	체육·스포츠 적용 예시
Volume	방대한 양의 정보, 정보의 용량이 제타 바이트 시대로 진입	스포츠 팬들의 SNS 게시물, 운동참여자의 실시간 생체정보(e.g., 심박수, GPS 데이터), 팀 단위 경기 운영 정보(패스 루트, 선수 움직임 추적 데이터 등), 구기 종목 공의 속도, 궤적, 회전수 등의 경기 상황 정보, 스포츠 관람객의 소비 관련 정보 등
Variety	자료 구성 및 형태의 다양성	
Velocity	스트리밍 서비스와 같은 자료의 실시간 생성 및 처리 속도	체육, 스포츠 상황에서 발생하는 모든 형태의 자료가 실시간으로 수집되고 처리되어 활용되는 사례 (예시: 선수의 심박수 변화에 따른 교체 타이밍 설정, 선수의 이동 거리 및 속도에 따른 체력 수준 파악) App, Wearable-device 등을 활용한 실시간 스트리밍 학교체육수업(Kareem, 2020) 디지털 플랫폼에서 이루어지는 스포츠 소비자들의 실시간 상호작용 자료 등
Veracity	수집된 자료의 진실성 및 분석 결과의 정확성	스포츠 소비자 행동 예측(Fried & Mumcu, 2016) 관람 스포츠 티켓의 가격 설정(Drayer et al., 2012) 소셜 웹 자료를 활용한 생활체육 시설의 수요 예측 및 활성화 방안 제시(김미옥 외, 2017)
Value	분석 결과의 유의미한 가치	

1.3. 빈도주의 통계와 베이저안 통계의 비교

4차 산업혁명, 인공지능, 기계학습, 딥러닝 기반의 빅데이터 분석이 하나의 패러다임으로 자리잡고 있는 현대사회에서는 빅데이터 분석과 관련된 컴퓨팅 시뮬레이션, 베이저안 추정 등이 주요한 연구 주제로 다루어지고 있다. 베이저안 통계는 베이즈 정리(Bayes's theorem)에 기초한 통계분석 방법으로 연구자의 관심이 되는 사건의 발생과 관련된 자료들의 확률분포를 활용하여, 특정 사건의 발생확률을 베이즈 정리에 의하여 산출하는 방법으로 정의된다(Gelman et al., 2013). 베이저안 통계는 Thomas Bayes가 제안한 베이즈 정리의 특수한 경우로, 1763년 처음 소개되어 그 역사가 매우 오래되었다. 그러나 오랜 역사에도 불구하고, 베이저안 통계는 계산이 용이하지 않고 매우 복잡하다는 단점 때문에(e.g., 사전분포 유도를 위한 시뮬레이션, 해가 존재하지 않는 경우 수치적 적분, 사전분포에 따른 수많은 parameter) 빈도주의 통계와 비교하여 그 활용은 상대적으로 미미했다. 그러나 현대사회에서는 과학기술의 발달로 계산에 대한 소프트웨어뿐만 아니라 하드웨어적인 문제 역시 많이 해결되었고, 다양한 통계패키지가 개발되어 예전보다 쉽게 베이저안 통계를 적용할 수 있는 환경이 구축되었다.

통계학은 모집단의 특성을 추론하기 위해 관찰된 자료로부터 미지의 모수가 발생할 확률을 산출하여 일정한 의사결정에 이르는 학문이다. 이때의 확률은 특정 사건 발생의 불확실성을 수학적 언어로 표현한 것이며, 그 해석은 빈도주의 통계와 베이저안 통계에서 차이를 보인다(Bayarri & Berger, 2004). 전통적인 빈도주의 통계에서 확률은 무작위성을 고려하여 같은 조건에서 무수히 많은 반복 시행(N)에 대비하여 특정 사건이 발생하는 상대적 빈도(f)를 하나의 고정된 값으로 표현한 것이다. 이러한 '수학적 확률'은 $Pr(A)=f/N$ 의 수식으로 표현되며, 빈도주의 통계에서 확률은 '객관적 확률'을 의미한다. 반면 베이저안 통계에서 확

률은 보편적 무작위성이 아닌 연구자가 사전에 수집한 정보를 바탕으로 사전분포(prior distribution)를 산정하고, 이를 반영하여 특정 사건이 발생할 불확실성을 사후확률(posterior probability)로 제시한다. 즉, 베이저안 통계에서는 특정 사건이 발생할 확률에 대하여 기존의 수집된 자료를 사전분포의 개념으로 활용하고, 새롭게 수집된 정보에 따라 확률분포를 갱신하여 비교적 유연하게 ‘통계적 확률’을 추정한다(이재길, 2018). 따라서 베이저안 접근에서 확률은 ‘주관적 확률’의 성격을 보인다. 베이저안 추론에서 핵심이 되는 사후확률은 다음과 같은 베이즈 정리를 통해 계산된다(Kruschke, 2014).

$$P(A_i | B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B | A_i)P(A_i)}{P(B)}$$

$P(A_i)$:사건 A_i 의 사전확률, $P(B)$:사건 B의 발생확률

$P(A_i | B)$:사건 B가 주어졌을 때, 사건 A_i 의 사후확률

$P(B | A_i)$:사건 A_i 가 주어졌을 때, 사건 B의 발생확률(우도;likelihood)

베이저안 통계를 간단히 요약하면, 모수에 대한 사전분포와 우도함수를 결합하여 사후분포를 유도하고, 이를 통해 모수의 불확실성을 수치화하는 방법으로 정리할 수 있다. 베이저안 추론은 특정한 사건의 발생이 다른 변수의 정보와 관련이 있다면 과거의 정보가 미래를 예측할 때 유용하게 활용될 수 있다는 접근에서 출발한다(Roos et al., 2005). 즉, 연구자의 관심사건에 해당하는 ‘프로야구 심판의 판정오류 발생 확률’에 대한 기초적인 정보를 수집하고, 이를 사전분포의 형태로 표현한다. 이 사전분포는 불확실성을 갖는 분포로 기존 연구 결과와 같은 실증적 자료 또는 전문가의 경험적 자료에 의해 설정될 수 있다. 다음 단계로 프로야구 심판의 판정오류 발생확률에 대한 새로운 정보(e.g., 경기 상황에 따라 판정오류 발생 확률의 변화)를 수집하고, 이를 반영하여 기존의 사전분포를 갱신한다. 이러한 갱신은 사건 발생에 대한 불확실성을 감소시키고 분포의

중심은 사전분포의 평균에서 새로운 자료의 평균 사이의 값으로 이동하게 된다. 따라서 새로 수집된 자료는 연구자가 기존에 설정한 사건 발생에 대한 사전분포를 보다 정교하게 만드는 역할을 한다. 마지막 단계로, 사전확률과 새로 수집된 자료의 우도(likelihood)를 이용하여 사후확률을 계산하여 심판의 판정오류 발생 확률을 계산한다. 이러한 사후확률은 연구 결론에 대한 근거가 되며, 추후 새로운 연구의 사전확률로 활용될 수 있다.

전통적인 빈도주의 통계 접근에서는 귀무가설(null-hypothesis)에 대한 기각(reject) 여부에 따라 의사결정을 수행하는 반면, 베이저안 접근에서는 연구에서 설정한 가설이 발생할 확률을 직접적으로 추정한다는 점에서 가장 큰 차이를 보인다(김청택, 2019). 빈도주의 통계학의 가설검정에서는 귀무가설(null-hypothesis)과 대립가설(alternative-hypothesis)을 설정하고, 유의수준에 따른 1종 오류와 2종 오류의 기각역을 설정하여, 귀무가설의 기각 여부에 따라 연구 결론에 이르게 된다. 즉 귀무가설(null-hypothesis)이 참(true)이라는 가정 아래, 추정된 모수가 관찰될 확률(p-value)를 계산하고, 이 확률이 유의수준보다 작으면 귀무가설을 기각하고 대립가설을 채택하여 대립가설이 참인 것으로 간주한다. 그러나 귀무가설이 참이 아닐 확률과 대립가설이 참일 확률은 동일하지 않다. 따라서 빈도주의 통계에서 가설검증 결과의 해석은 다양한 상황에 대한 가정이 전제되어야 한다는 한계를 갖는다. 반면, 베이저안 통계학에서는 귀무가설과 대립가설의 인위적인 구분에서 벗어나 연구자가 설정한 연구가설에 대한 사후분포를 추정하여 연구가설이 참일 확률을 직접 계산한다. 따라서 사회과학 연구 방법의 논리에서는 베이저안 통계가 빈도주의 통계보다 더욱 직관적인 결과를 제시할 수 있다(Wagenmakers et al., 2008; Stegmüller, 2013).

빈도주의 통계에서 계산되는 관심사건의 발생 확률은 표본의 무작위성과 동일한 조건의 반복을 가정하여 산출된다. 그러나 이와 같이 특정

사건이 동일한 조건에서 반복되는 상황을 가정하는 것은 실험실이나 공장의 생산 라인과 같이 매우 엄격한 통제가 이루어지는 극히 일부의 경우로 한정된다. 체육/스포츠 상황에서 발생하는 다양한 사건들은 한 가지의 원인이 곧바로 결과로 이어지는 것이 아니라, 다양한 원인들이 서로 복합적인 상호작용을 통해 특정한 결과로 도출된다. 또한 스포츠 상황에서 특정 사건의 발생은 다양한 원인이 일정한 수준에 도달해야만 사건이 발생하는 특징을 보이기도 한다. 예를 들어, 프로야구에서 스트라이크 아웃(strike-out) 또는 볼넷에 의한 출루(base on balls)는 ‘세 번째’ 스트라이크 판정 또는 ‘네 번째’ 볼 판정이 선언되어야만 아웃 또는 출루라는 결과로 이어진다. 따라서 프로야구 심판의 볼 판정오류가 경기결과에 미치는 영향력을 분석하고자 할 때, ‘0볼-0스트라이크’ 상황에서의 판정오류와 ‘3볼-2스트라이크’ 상황에서의 판정오류는 분명히 다른 상황으로 간주되어야 할 것이다. 이와 같은 맥락에서, 실험실 검사와 같이 엄격하게 통제된 상황을 가정한 연구설계는 스포츠가 갖는 특성을 충분히 고려하기에는 한계가 있다. 따라서 최근의 다양한 연구들에서 스포츠가 갖는 불확실성을 충분히 고려하기 위해서는 빅데이터를 활용한 베이지안 연구설계의 필요성을 강조하고 있다(Torgler, 2020; Santos, Wu & Mengersen, 2019; Gudmundsson & Horton, 2017)

베이지안 통계에서는 이용 가능한 정보를 확률분포에 기반하여 활용하고, 새로운 정보가 주어지면 실시간으로 정보를 갱신하여 새로운 확률의 추정이 가능하다는 장점이 있다. 이를 스포츠 맥락으로 확장해보면, 특정 사건의 발생을 원인이 일정하게 통제된 상황으로 가정하는 것이 아니라, 원인이 되는 다양한 사건들을 새로운 확률 추정을 위한 정보로 활용하여 스포츠 경기에서 특정 사건의 발생 확률을 추정할 수 있음을 의미한다. 이와 같은 접근은 스포츠 상황을 보다 폭넓게 반영한 결과를 제시할 수 있을 것이다. 다음 절에서는 체육/스포츠 분야 연구에서 베이지

통계를 적용한 관련 연구들을 고찰하여 스포츠 빅데이터 분석에 있어서 베이지안 통계의 적용이 어떤 의의를 갖는지 살펴보고자 한다.

1.4. 체육/스포츠 분야 연구에서 베이지안 기법의 활용

베이지안 통계기법은 오랜 역사와 다양한 장점에도 불구하고, 이론적 복잡성과 물리적인 계산의 양이 많다는 점에서 실질적인 활용은 활발히 이루어지지 못했다. 그러나 현대사회에서는 과학기술의 발달로 복잡한 계산에 대한 효율적인 대처가 가능한 환경이 구축되어 베이지안 통계는 점차 더욱 주목받고 있다. 베이지안 추정은 실시간으로 변화하는 상황에 따른 유연한 대처가 가능하기 때문에, 스포츠 현장에서 수집 및 처리되는 빅데이터 분석에 적합하다. 이에 따라 체육/스포츠 분야 연구에서 베이지안 통계 기반의 연구설계는 점점 더 주목받고 있으며, 실제로 그 활용은 증가하고 있다(Bernards et al, 2017; Santos, Wu & Mengersen 2019).

Bernards 등(2017)의 연구에서는 스포츠 과학 분야가 오늘날 직면하고 있는 통계적 단점을 조명하고, 관행적으로 이루어졌던 연구설계를 보완할 수 있는 잠재적인 발전 방향을 제공하는 것을 목표로 설정하고 있다. 구체적으로, 해당 연구에서는 빈도주의 통계에서 p-value는 연구가설이 참일 확률과 효과의 크기를 제시해줄 수 없기 때문에, 분석 결과의 중요성에 대하여 언급할 수 있는 근거는 실질적으로 없음을 지적하고 있다. 연구진 이에 대한 대안으로 베이지안 접근의 적용을 강조하며, 스포츠과학 분야의 연구에서 그동안 이루어졌던 관행과 같은 연구설계에서 벗어나 보다 발전된 형태로 패러다임의 전환이 필요함을 역설하였다. Bernards 등(2017)의 연구에서는 큰 틀에서 전통적인 가설검증의 한계점에 대하여 실증적 자료보다는 개념적인 설명에 초점을 두었다면, Gucciardi와 Zyphur(2016)의 연구에서는 이미 게재된 논문의 자료를 활용하여 실증적으로 빈도주의 통계와 베이지안 통계를 비교하였다. 해당

연구에서는 동일한 귀무가설을 반복적으로 검증하는 빈도주의 통계와 달리, 베이지안 추정은 사전정보와 새롭게 업데이트되는 정보를 함께 활용하여 결과를 제공하기 때문에 기존의 정보를 활용하여 사전분포를 설정하는 것이 핵심이라고 보고하고 있다. 이를 구체화하기 위해 mental though behavior의 요인부하량(factor loadings)에 대한 사전정보가 있는 모형과 정보가 없는 모형을 각각 비교하여 스포츠 심리 측정분야의 구조 방정식모형에서 사전정보가 어떤 의미를 갖는지 논의하였다. 이를 통해 연구진은 베이지안 통계는 유연하고 다양한 상황을 고려할 수 있는 확률 기반의 모델링이 가능하기 때문에, 체육/스포츠 분야 연구자들은 이와 같은 장점을 충분히 활용할 수 있는 연구설계를 지속적으로 고찰해야 한다고 보고하였다.

Stenling 등(2015)의 연구에서는 앞선 두 연구보다 직관적인 스포츠 사례를 중심으로 체육/스포츠 분야 연구에서 베이지안 추정의 활용에 대하여 논의하였다. 구체적으로, 해당 연구에서는 ‘Sport Motivation Scale-II’ 척도의 타당성과 신뢰성을 분석하기 위해 최대우도법을 따르는 전통적인 확인적 요인분석(CFA; confirmatory factor analysis)모형과 베이지안 모형을 비교하였다. ‘Sport Motivation Scale-II’ 척도는 자기결정 이론(self-determination theory)을 기반으로 스포츠 참여에 대한 선수들의 행동 규제를 측정하기 위한 척도이다. 해당 연구에서는 최대우도법을 따르는 전통적인 CFA보다 베이지안 요인분석모형에서 ‘Sport Motivation Scale-II’의 구성을 설명하기 위한 적합도가 더 높은 것으로 나타났다. 이와 같은 결과에 대하여 연구진은 베이지안 추정은 계산과 해석에 이점을 갖기 때문에 체육/스포츠 분야의 연구에서 더욱 활발히 활용될 필요가 있음을 보고하였다. 더불어, 연구진은 베이지안 추정에서 사전분포를 설정할 때, 연구자의 주관이 반영될 수 있는 부분을 항상 주의하여 사전분포 설정에 대한 근거를 명확히 제시해야 할 것임을 강조하였다.

Higgs와 Stavness(2021)의 연구에서는 베이지안 다층모형 회귀분석을 통해 COVID-19 기간에도 북미 프로스포츠 경기에서 홈 팀 어드밴티지(advantage)가 과거와 같이 작용하고 있는지에 대하여 분석하였다. 그 결과 COVID-19 기간동안 NHL(National Hockey League)과 NBA에서는 홈 어드밴티지가 과거와 비교하여 줄어든 것으로 나타났고, 반면 MLB와 NFL(National Football League)에서는 큰 변화가 없는 것으로 나타났다. 연구진은 해당연구에서 베이지안 추정을 활용한 이유를 스포츠 맥락과 이론적 맥락으로 구분하여 설명하였다. 먼저 스포츠 관점에서는 ‘스포츠 경기의 상대성’을 중요한 요소로 제시하였다. 스포츠 분석에서 상대성을 고려하지 않을 경우 경기력이 약한 팀은 홈/원정 구분 없이 패할 확률이 높고, 반대로 강팀은 지속적으로 승리 확률이 높게 나타나기 때문에 홈 팀의 이점을 체계적으로 분석할 수 없음을 지적하였다. 연구진은 베이지안 통계기법은 상대 팀의 경기력 수준에 대한 정보를 확률 분포에 기반하여 고려할 수 있기 때문에 스포츠가 갖는 특성을 잘 반영할 수 있음을 강조하였다. 또한 이론적 관점에서는 베이지안 통계는 모수에 대한 추정을 고정된 값이 아닌 확률 분포에 기반하여 제시할 수 있기 때문에, 전통적인 빈도주의 통계와 비교하여 유연하고 직관적인 모형을 구축할 수 있음을 장점으로 보고하였다.

Deshpande와 Wyner(2017)의 연구에서는 현대 프로야구에서 포수의 framing이 경기 결과에 중요한 영향을 미칠 수 있는 요인임을 강조하며, 위계적 베이지안 모델(hierarchical Bayesian Model)을 통해 MLB 포수의 framing을 분석하였다. 구체적으로 해당 연구에서는 PITCHf/x 자료를 활용하여 투수가 투구한 공이 투구추적시스템에서 가상의 스트라이크 존의 경계에 걸친 경우를 연구자료로 설정하여 베이지안 로지스틱 회귀분석을 실시하였다. 연구자의 관심 사건은 타자의 스윙 없이 strike 판정을 받은 공이기 때문에 종속변수의 기준 범주($y=1$)는 ‘called strike’로 판정된 사

건으로 설정하였으며, 스트라이크 존의 경계에 투구되었음에도 'ball'로 판정 받은 자료는 $y=0$ 으로 설정하여 연구를 진행하였다. 해당 연구에서는 사전정보의 확률분포를 추정하기 위해 markov chain monte carlo simulation 기법을 활용하였으며, 사전정보로 설정한 볼 카운트의 변동에 따라 called strike 발생 확률의 변화가 각 포수마다 어떠한 차이를 보이는지 비교하여 포수의 경기력 관점에서 framing을 설명하고자 하였다.

Sottas 등(2007)의 연구에서는 스포츠 도핑에서 비정상적인 testosterone glucuronide/epitestosterone glucuronide(T/E)의 탐지를 위해 베이지안 기법을 활용하였다. 해당 연구에서는, 사전분포의 추정을 위해 세계반도핑기구(WADA; World Anti-Doping Agency:WADA)로부터 수집한 운동선수들의 T/E에 대한 평균과 변동계수를 기본 사전정보로 활용하였다. 연구진은 이와 같은 사전정보를 활용하여 비정상적인 T/E비율을 탐지하기 위해 순차적으로 테스토스테론과 에피테스토스테론을 측정하여, 새로운 정보가 수집될 때마다 과거의 측정 결과를 새로운 모형의 사전분포로 활용하여 비정상적인 T/E 비율에 대한 탐지가 점진적으로 정확하게 이루어질 수 있도록 모형을 설정하였다. 연구진은 forensic toxicology 분야에서 사전정보의 활용은 매우 긴요하게 활용될 수 있기 때문에, 이와 같은 맥락에서 베이지안 추정이 매우 효과적일 수 있음을 강조하였다. 즉, 연구자의 관심이 되는 모집단에 대하여 기존의 정보가 존재한다면, 새로운 모형의 사전분포로 활용할 수 있다는 점에서 WADA에 구축된 기존 정보를 활용할 경우 베이지안을 활용한 연구설계가 도핑 관련 문제에서 매우 유용하게 활용될 수 있음을 보고하였다.

Olmedilla 등(2018)의 연구에서는 스포츠에서 발생하는 부상을 설명함에 있어 대부분의 연구가 심리적 요인은 비중있게 다루어지지 않고 있음을 지적하며, 운동선수의 심리적 요인과 부상 발생 사이의 관계를 확률론적으로 설명하기 위해 베이지안 네트워크 분석(bayesian network

analysis)을 수행하였다. 베이저안 네트워크 분석은 입력된 변수들 간의 조건부 종속성(conditional dependency)의 확률적 관계를 방향성 비순환 그래프(DAG; directed acyclic graph)를 통해 나타내는 확률적 그래픽 모델(probabilistic graphical model)이다. 베이저안 네트워크 분석에서는 특정 사건의 발생에 대하여 관련이 있는 것으로 알려진 변수들 중 어떤 변수가 확률적으로 사건 발생에 깊은 연관성을 나타내는 변수인지를 예측한다. 해당 연구에서는 베이스 네트워크 분석을 통해 운동선수의 심리 요인과 부상사이의 확률 관계를 제시하고자 하였다. 구체적으로, 불안(anxiety), 사회적 지원(social support), 자기효능감(self efficacy) 등의 선수들의 심리를 나타내는 요인과 성별, 부상이력/횟수, 신장, BMI 등 다양한 변수를 통합적으로 고려하여 운동선수의 부상 발생 가능성과 심리 요인의 관계를 확률적으로 제시하였다. 해당 연구에 따르면 스포츠에서 자기 효능감은 심리적 요인들 사이에서 중요한 역할을 할 뿐만 아니라, 신체적, 체형적(somatic) 요인, 부상과 관련한 역학적(epidemiological) 요인과의 매우 긴밀한 관계를 보이고 있다.

Baker 등(2020)의 연구에서는 심리적 지속 모델(PCM; psychological continuum model)을 통해 스포츠 소비자의 특성을 세분화하기 위해 베이저안 추정을 활용하였다. PCM은 스포츠 매니지먼트 분야에서 소비자의 태도가 어떻게 형성되고 변화되는지를 설명하는 데에 중요한 역할을 하는 것으로 알려져있다. 연구진은 PCM에 의한 소비자 세분화 관련 연구에서 Beaton 등(2009)이 제시한 PCM ‘Staging algorithm’이 전통적으로 활용되고 있으나, 일부 과학적인 엄격함이 부족한 부분을 지적하며 Bayesian latent profile analysis(Bayesian LPA)을 통해 스포츠 소비자를 세부적으로 분류하고자 하였다. 연구진은 전통적인 PCM 알고리즘과, k-means 군집분석, Bayesian LPA를 수행하여 각각의 결과를 비교하였으며 그 결과 Bayesian LPA 기법이 가장 명확하게 소비자의 특성을 분

류해내는 것으로 나타났다. 이와 같은 결과에 대하여 연구진은 개인의 심리는 고정된 것이 아니라 시간과 개인이 처한 상황에 따라 지속적으로 변화하기 때문에, 전통적인 빈도주의 기반의 접근 보다는 과거의 경험을 확률적으로 반영할 수 있는 베이지안 접근이 보다 효과적으로 스포츠 소비자의 심리를 설명할 수 있는 것이라 보고하였다.

상기한 선행연구들의 결과를 종합하면, 체육/스포츠 분야 연구에서 베이지안 접근은 프로스포츠 구단 또는 선수의 경기력을 설명하기 위한 분석뿐만 아니라 스포츠 심리, 운동생리학, 스포츠 의학, 공공 스포츠 산업 증진을 위한 스포츠경영 분야 등 다양한 체육/스포츠 분야 연구에서 활용되고 있다. 스포츠 상황에서 발생하는 다양한 사건들은 과거의 사건이 같은 조건에서 반복되는 것이 아니라 실시간으로 변화하는 상황에 따라 다양한 변수가 서로 상호작용을 통해 하나의 특정한 결과로 도출된다. 따라서 이를 분석하기 위해서는 다양한 정보들을 확률적으로 고려할 수 있는 유동적인 모형의 구축이 필요하다. 이러한 관점에서 베이지안 기법은 스포츠 상황을 분석하고자 할 때 다양한 장점이 있으나 그 활용은 활발히 이루어지지 못했다. 특히, 베이지안 기법의 적용을 체육/스포츠 관점에서 이론적으로 논의한 연구는 극히 일부에 그치고 있다. 본 학위논문의 ‘Study 1’에서는 스포츠 빅데이터를 활용하여 프로야구 심판 판정의 불확실성을 정량화하여 평가하고, 그 과정에서 베이지안 접근과 전통적인 빈도기반의 로지스틱 회귀분석 결과를 비교하여 스포츠 데이터분석에 있어서 각각의 장단점에 대하여 논의하고자 한다.

2. 연구 방법

2.1. 자료 수집 및 처리 절차

본 학위논문의 ‘Study 1’에서는 스포츠 빅데이터를 활용하여 프로야구 심판 판정의 불확실성을 정량화하여 평가하고자 한다. 그 과정에서 베이 지안 접근과 전통적인 빈도기반의 로지스틱 회귀분석 결과를 비교하여 스포츠 데이터분석에 있어서 각각의 장단점에 대하여 논의하고자 한다. 이를 위한 첫 단계로 문헌고찰을 통해 스포츠 빅데이터의 정의와 특성을 개념화하고, 스포츠가 갖는 불확실성이라는 고유한 특성을 고려해야 하는 이론적 당위성을 제시하였다. 그 다음 단계에서는 실제 MLB 현장에서 수집되는 스포츠 빅데이터를 활용하여 k-Dependence Bayesian Network Classifier(k-DB) 분석을 실시하여 MLB 심판 판정오류 발생의 인과관계를 확률에 근거하여 설명하였다. 이후 k-DB 분석결과를 최대우도법 기반의 로지스틱 회귀분석 결과와 비교하여 불확실성을 고려한 스포츠 빅 데이터 분석에 있어서 각각의 접근이 어떤 장단점을 갖는지 논의하였다. Study 1의 마지막 단계에서는 MLB 심판의 판정오류가 직관적으로 어느 정도의 영향력을 갖는지 분석하기 위해 삼진과 볼넷의 가치에 근거하여 경제적 파급효과를 추정하였다.

‘Study 1’에서 활용하고자 하는 연구 자료의 구체적인 구성은 다음과 같다. 2019-2021년 MLB 심판의 판정 중 Statcast에서 제공하는 가상의 strike zone을 벗어난 투구 중 ‘Ball’ 또는 ‘Called Strike’ 판정을 받은 투구를 연구 자료로 활용하였다. 종속변수로 설정한 심판의 판정오류는 **“투구된 공이 MLB Statcast 스트라이크 존에서 벗어났음에도 불구하고 스트라이크로 판정된 투구”**로 조작적으로 정의하여 연구를 수행하였다. 즉, 본 연구에서는 ‘투수에게 유리한 판정오류’에 한정하여 연구를 진행하였다. 프로야구 심판의 판정오류는 투수에게 유리한 판정오류

뿐만 아니라 스트라이크가 투구되었으나 볼 판정이 이루어지는 타자에게 유리한 판정오류 역시 존재한다. 그러나 본 연구에서는 투수에 초점을 맞추어 연구를 진행하였음을 밝힌다. 이에 따라, 가상의 스트라이크 존을 벗어났음에도 Strike 판정을 받은 판정오류를 $y=1$ (판정오류), 스트라이크 존을 벗어났기 때문에 Ball 판정을 받은 ‘ $y=0$ (정분류)’로 범주화하였다. 또한 자료분석 대상은 plate_x 좌표 기준 $-1.5\sim 1.5$, plate_z 기준 1~4로 설정하여 스트라이크존의 경계(edge)에 위치한 투구로 한정하였으며, 이와 같은 전처리(preprocessing) 절차는 모든 시즌 자료에 대하여 동일하게 진행하였다.

독립변수는 선행연구 고찰 결과에 따라 프로야구 심판의 판정에 영향을 미치는 요인들 중 실제로 프로야구 현장에서 실시간으로 수집 가능한 자료로 선정하였다. 자료 수집은 MLB에서 Statcast data를 공식적으로 관리하는 ‘baseball savant’¹⁾를 통해 수행하였으며, 자료 검색 예시는 아래의 <그림 II-1>과 같다. 수집된 자료는 심판의 판정 결과, 투구된 공의 구종, 속도, 회전수, 공의 움직임, 홈/원정 팀 구분, 투구된 공의 위치, 볼 카운트, 타자에 따른 스트라이크 존의 높이 등 다양한 자료가 포함되어 있다. 문자로 되어 있는 일부 변수는 숫자를 부여하여 범주형 변수로 변환하였다. fastball 변수의 경우 직구계열과 변화구계열로 구분하였다. 직구 계열에는 ‘2-Seam Fastball’, ‘4-Seam Fastball’, ‘Cutter’, ‘Sinker’ 구종이 포함되었으며, 변화구계열은 ‘Changeup’, ‘Curveball’, ‘Eephus’, ‘Forkball’, ‘Knuckle Ball’, ‘Knuckle Curveball’, ‘Slider’, ‘Split Finger’ 구종으로 분류하였다. 또한 투수의 홈/원정 구분 변수 또한 홈 팀은 ‘Home = 1’, 원정 팀은 ‘Away=0’으로 코딩하였다. 수집된 자료는 Comma-Separated Values(CSV)형식으로 저장하여 보관하였으며, 모든 분석은 Software R version 3.5.2을 활용하였다. 또한 다양한 독립변수

¹⁾ https://baseballsavant.mlb.com/statcast_search

중 결측치가 하나라도 있는 경우는 연구대상에서 제외하였다. 본 연구에서 활용한 연구 자료에 대한 세부적인 설명은 <표 II-2>와 같다.

savant Search Player Name or Team

illustrator Gamefeed Scoreboard Probable Pitchers Search Visuals Statistics Leaderboards

11/11 No Games Scheduled.

Statcast Search

An application that lets you search MLB.com's Statcast database. It includes metrics such as Perceived Velocity (PV), Spin Rate (SR), Exit Velocity (EV), Hit Distance (DST), Launch Angle (LA), Batted Ball Direction (BBD), xBA, xSLG, xwOBA, and Quality of Contact.

- Some of these queries are very complicated and take time to run. If the page doesn't fully load, refresh the page and try again.
- Pitch velocities from 2008-16 are via Pitch F/X, and adjusted to roughly out-of-hand release point. All velocities from 2017 and beyond are Statcast, which are reported out-of-hand.
- For the limited subset of batted balls not tracked directly, estimates are included based on the process described [here](#).
- For CSV documentation click [here](#).
- For a detailed tutorial on using the search click [here](#).

Pitch Type: PA Result: Season Type:

Pitch Result: Gameday Zones: Venue:

Batted Ball Location: Attack Zones: Batted Ball Direction:

Count: Season: Situation:

Player Type: Outs: Opponent:

Pitcher Handedness: Batter Handedness: Quality of Contact:

Game Date >= Game Date <= Month:

Team: Home or Away: Runners On:

Position: IF Alignment: OF Alignment:

Inning: Batted Ball Type: Batters:

Flags: Pitchers:

Metric Range:

Group By: Min # of Total Pitches: Min # of Results:

Min PA: Sort By: Sort Order:

Search Change Included Stats Change Total Pitch Parameters Clear Search

그림 II-1. Baseball Savant 검색 예시

표 II-2. Study 1의 연구 자료

변수		설명
종속 변수	umpiring	2019-2021년 MLB 정규시즌 경기 중 ball 또는 called strike 판정을 받은 투구
독립 변수	pitch type	직구/변화구 구분
	Home	투수의 홈, 원정 구분
	P_right	투수 - 좌/우투 구분
	B_right	타자 - 좌/우투 구분
	plate_x	투구된 공이 홈플레이트를 통과할 때 수평적 위치
	plate_z	투구된 공이 홈플레이트를 통과할 때 수직적 위치
	pfx_x	투구된 공의 수평적 움직임
	pfx_z	투구된 공의 수직적 움직임
	speed	투구된 공의 속도
	spin	투구된 공의 회전수
	sz_height	타자에 따른 스트라이크 존의 높이
	score_diff	홈/어웨이 팀 간 점수 차이
	innings	공이 투구되는 시점의 이닝
	outs	공이 투구되는 시점의 아웃카운트
	strikes	공이 투구되는 시점의 strike 카운트
balls	공이 투구되는 시점의 ball 카운트	

2.2. k-Dependence Bayesian Network Classifier

베이지안 네트워크 분석은 입력된 변수들 간의 조건부 종속성(conditional dependency)의 확률적 관계를 방향성 비순환 그래프(directed acyclic graph, DAG)를 통해 나타내는 확률적 그래픽 모델(probabilistic graphical model)이다(Marcot & Penman, 2019). DAG란 확률변수가 다른 확률변수에 미치는 영향이 다시 되돌아오는 것을 고려하지 않는다는 의미로, 베이지안 네트워크 모형에서 확률 변수 X와 Y의 인과관계를 'X(원인)→Y(결과)'의 형태로 나타내며, 이는 조건부 확률 $Pr(Y|X)$ 로 산정된다. 즉, 베이지안 네트워크 분석은 확률변수들 사이에 존재하는 정성적 관계를 그래프 구조로 표현하고, 각 변수 간의 정량적 관계를 조건부 확률 형태로 제시하는 방법이다.

본 학위논문의 'Study 1'에서는 MLB 심판의 판정오류 발생에 영향을 미치는 변수들 간의 확률적 의존성을 분석하기 위해 베이지안 네트워크 분석을 수행하였다. MLB 심판의 판정오류는 복잡하고 다양한 잠재적인 원인들이 존재하기 때문에 선형관계만으로 확률을 산정하기에는 한계가 있다. 베이지안 네트워크는 이와 같이 특정 사건의 발생에 불확실성이 존재하는 경우 유용한 분석 기법이다. 베이지안 네트워크는 불확실성(uncertainty)과 복잡성(complexity)뿐만 아니라 인과관계(causality)에 대해서도 탐구할 수 있는 유용한 도구로 알려져 있다(Ben-Gal, 2008). 베이지안 네트워크 분석은 연결된 노드들 간의 조건부 확률을 토대로 특정 노드가 주어졌을 때, 종속변수로 설정한 노드의 사건이 발생할 조건부 확률을 체인룰(chain rule)에 의해 산출한다.

베이지안 네트워크가 학습하는 과정은 크게 구조 학습 단계와 모수 학습 단계로 구분된다. 일반적으로 베이지안 네트워크 구조는 학습 방법에 따라 종류가 구분되며, 본 연구에서는 k-dependence Bayesian classifier(k-DB)를 통해 프로야구 심판의 판정오류 발생 확률을 분석하

였다. k-DB방법은 하나의 독립변수가 최대 임의의 k개의 독립변수에 영향을 받는다는 가정을 전제한다(Sahami, 1996). 이때, 예측변수 구조의 순서는 종속변수(class variable)와의 상호 정보(mutual information)가 높은 순으로 결정된다. k-DB의 사후분포는 아래의 식과 같이 나타낼 수 있다(Duan, & Wang, 2017; Park, & Hwang, 2020). 따라서 독립변수들이 서로 독립이라는 가정이 존재하는 가장 일반적인 베이저안 네트워크 모형(i.e., naive bayesian network)과 비교하여 k-DB방법은 다양한 경기 상황변수가 유기적으로 결합하여 하나의 결과로 나타나는 스포츠 빅데이터 분석에 적합할 것으로 판단하였다.

$$p(c|x) \propto \prod_{i=1}^n p(x_i | c, x_{i1}, \dots, x_{ip})$$

본 연구에서 수행한 k-DB는 투입된 모든 독립변수들 간의 의존성을 허용하여 독립변수의 개수만큼 k를 설정하였으며, 과적합을 방지하기 위한 k-fold cross-validation에서의 k는 5회로 설정하였다. 또한 수집한 자료 중 일부 연속형 변수를 k-DB 분석에 적합한 형태로 변환하기 위해 Hartemink(2001)에서 제시한 Information-Preserving Discretisation 알고리즘을 활용하였다. 해당 알고리즘의 핵심 아이디어는 정보손실을 최소화하기 위해 원자료를 수많은 범주로 변환한 뒤, 이를 반복하여 각 자료의 상호 정보의 손실을 최소화하는 범주구분을 찾는 것이다. 따라서 해당 알고리즘은 원자료가 갖는 의존성을 고려하여 자료를 변환하기 때문에, 단순히 정량적으로 자료를 구분(e.g., quantile)하는 것보다는 원자료의 속성을 훨씬 더 잘 반영할 수 있다는 장점이 있다(Scutari & Denis, 2022). 이와 같은 모든 데이터 처리 및 분석하는 과정은 Software R의 bnlearn, bnclassify, Rgraphviz 패키지를 활용하였다.

2.3. 로지스틱 회귀분석

종속변수가 연속형 변수 형태를 취하는 선형회귀분석과 달리 로지스틱 회귀분석은 종속변수가 0 또는 1의 범주를 갖는 이항변수의 형태이다. 이와 같은 로지스틱 회귀분석은 연결함수(link function)에 따라서 모형의 형태가 결정되며, 이항분포를 따르는 범주형 변수에 대한 연결함수는 주로 logit(log odds)함수가 주로 활용된다. 즉, 로지스틱 회귀분석에서는 독립변수로 부터 종속변수의 값을 직접 추정하지 않고 log odds의 개념을 통해 독립변수가 종속변수 범주의 발생에 미치는 영향을 추정한다. 또한 최소자승법(ordinary least square, OLS)를 활용하는 선형회귀분석과 달리, 로지스틱 회귀분석에서는 최대우도법을 통해 모수를 추정하게 된다. 본 연구에서는 $y=1$ (판정오류), $y=0$ (정분류)의 형태를 갖는 심판 판정을 설명하기 위해 로지스틱 회귀분석을 수행하였다.

본 연구에서 수행한 로지스틱 회귀분석 절차는 다음과 같다. 첫째, 선행연구 고찰 결과를 바탕으로 독립변수를 설정하고, 둘째, 설정된 변수를 바탕으로 로지스틱 회귀분석을 실시하여 모형의 수렴 여부와 회귀계수의 유의성을 확인하였다. 다음 단계로 각각 회귀계수를 log odds 개념을 통해 해석하여, 심판의 판정에 영향을 미치는 독립변수의 영향력을 검증하였다. 마지막 단계로 추정된 로지스틱 회귀식을 바탕으로 한계효과분석(marginal effects analysis)를 실시하여, 투입된 독립변수의 변화에 따른 프로야구 심판의 판정오류 발생 확률을 추정하고, 그 결과를 베이지안 네트워크 모형과 비교하였다. 로지스틱 회귀분석에 투입한 변수는 앞선 2.1절에 제시한 <표 II-2>와 같다.

2.4. MLB 판정오류의 파급효과

앞선 절차들을 요약하면, K-DB 분석을 통해 MLB 심판의 판정오류 발생 확률을 경기 상황에 따라 분석하여 판정오류 발생의 인과관계를 설명하고, 그 결과를 로지스틱 회귀분석 결과와 비교하여 불확실성을 고려한 스포츠 빅데이터 분석 연구의 내적 타당도의 관점에서 베이지안 접근의 장단점에 대하여 논의하는 것이다. 그 다음 절차로는, 실제로 MLB 심판의 판정오류가 어떤 파급효과를 유발하는지 계량적으로 분석하여 판정오류의 영향력을 추정하였다.

본 연구에서 조작적으로 정의한 MLB 심판의 판정오류는 “*MLB Statcast k-zone에서 벗어난 투구임에도 불구하고, Strike 판정을 받은 투구*”이다. 이를 실제 야구 경기 상황에 적용해 보면, ‘Xballs-2strikes’ 상황에서 ball이 투구되었으나 strike 판정이 이루어질 경우, 판정오류에 의해 투수가 삼진을 얻게 된 것으로 간주할 수 있을 것이며, 이는 투수에게 유리한 결과를 초래한다. 반대로 타자의 입장에서는 ‘3ball-Xstrike’ 상황에서 ball이 투구되었으나 strike 판정이 이루어질 경우, 판정오류에 의해 타자가 하나의 볼넷을 잃은 것으로 간주할 수 있을 것이다. 본 연구에서는 이러한 가정(Xballs-2strikes, 3balls-Xstrikes) 하에서 각 팀이 경험한 판정오류 누적 빈도 자료를 수집한 뒤, 팀의 승리 빈도에 기반한 삼진과 볼넷의 가치를 반영하여 파급효과를 분석하였다. 이후 MLB에서 1승의 가치를 구단의 수익에 대비하여 추정하여, 판정오류로 인해 각 구단들이 잠재적으로 경험하는 경제적 득실이 어느 수준인지 추정하였다. 구체적인 파급효과분석 절차는 다음과 같다.

첫째, MLB 최근 5시즌(2016-2021시즌, 2020시즌은 COVID-19로 인해 제외)자료를 바탕으로 다중회귀분석을 실시하여 일정한 변수가 통제되었을 때, 팀 단위 삼진과 볼넷의 가치를 1승에 대비하여 추정하였다. 통제 변수(시즌, 리그)는 Humphreys와 Pyun(2017)의 연구를 참고하여 모형에

포함하였다. 팀 단위 볼넷과 삼진을 하나의 회귀모형에 함께 포함시킬 경우 통제변수와의 다중공선성 문제가 발생하여, 승리에 대한 삼진의 가치를 추정하는 회귀식에서는 팀의 타격능력을 통제변수로 설정하였고, 반대로 볼넷의 가치를 추정하는 회귀식에서는 팀의 투수력을 통제변수로 설정하여 총 2회 다중회귀분석을 실시하였다. 삼진과 볼넷의 가치를 추정하기 위한 회귀식은 다음과 같다.

$$\widehat{win} = b_0 + b_1(year_i) + b_2(AL/NL_i) + b_3(OPS+_i) + b_4(SO_i) \quad (1)$$

$$\widehat{win} = b_0 + b_1(year_i) + b_2(AL/NL_i) + b_3(FIP_i) + b_4(BB_i) \quad (2)$$

위 식에서 확인할 수 있듯이, 삼진의 가치를 추정하는 회귀식(1)에서는 팀의 타격 능력을 나타내는 세이버매트릭스 지수 ‘OPS+1’를 통제변수로 투입하여, 시즌, 리그, 팀의 타격능력이 일정하게 고정되어 있을 때, 삼진이 승리에 미치는 영향력을 추정하였다. 또한 볼넷의 가치를 추정하는 회귀식(2)에서는 시즌, 리그 변수와 함께 팀의 투수력을 통제하기 위해 세이버매트릭스 지수 ‘FIP2’를 통제변수로 포함하여, 시즌, 리그, 팀의 투수력이 통제되었을 때, 볼넷이 승리에 미치는 영향을 추정하였다.

둘째, 앞선 회귀식(1), (2)를 통해 추정된 삼진과 볼넷의 가치를 판정 오류로 인해 투수가 얻은 삼진, 타자가 잃은 볼넷의 빈도에 반영하여 시즌/팀별로 ‘판정오류를 고려한 승리횟수’를 추정하였다. 구체적으로 ‘Xballs-2strikes’ 카운트와 ‘3balls-Xstrikes’ 카운트에서 ball이 투구되었

1) OPS+ takes a player's on-base plus slugging percentage and normalizes the number across the entire league., $100*((OBP/league_OBP)+(SLG/league_SLG) -1)$

2) Fielding Independent Pitching: $(13*HR + 3*BB - 2*K)/IP + C$, where C is a constant term that re-centers the league-average FIP to match its average ERA

으나 strike 판정이 이루어진 판정오류의 누적 빈도 자료를 수집 후, 승리에 대한 삼진과 볼넷의 가치를 반영하여 판정오류로 인해 각 팀이 얻거나 잃은 승리 빈도를 추정하였다. 이를 통해 판정오류로 인해 투수가 얻은 삼진(Xballs-2strikes)과 타자가 잃은 볼넷(3balls-Xstrikes)이 한 시즌동안 누적되었을 때, 각 팀의 승리에 어느 정도의 영향을 미치는지 탐색하였다.

셋째, MLB에서 1승의 경제적 가치를 추정하기 위해 매시즌 구단의 수익(revenue)을 종속변수로 설정하고 통제변수와 함께 승리횟수를 독립변수로 설정하여 다중회귀분석을 실시하였다. 보다 정확한 추정을 위해 해당 절차에서 역시 구단의 수익에 영향을 미칠 수 있는 요인(시즌, 리그, 포스트시즌 진출 유무)들을 통제변수로 포함하였다. MLB 1승의 가치를 추정하기 위한 회귀식은 다음과 같다.

$$\widehat{revenue} = b_0 + b_1(year_i) + b_2(AL/NL_i) + b_3(PS_i) + b_4(Win_i) \quad (3)$$

마지막 절차로, 위의 회귀식(3)을 통해 추정된 팀의 수익 대비 1승의 경제적 가치를 앞선 절차를 통해 추정된 ‘판정오류를 고려한 승리횟수’에 반영하여, 판정오류로 인해 각 구단이 경험한 경제적 파급효과를 추정하였다. 이를 통해 MLB 심판의 판정오류가 실제로 프로야구 현장에서 어떤 결과를 초래하는지 보다 직관적인 결과를 제시하였다.

3. 연구 결과

3.1. 2019-2021 MLB Regular Season 판정오류 분석

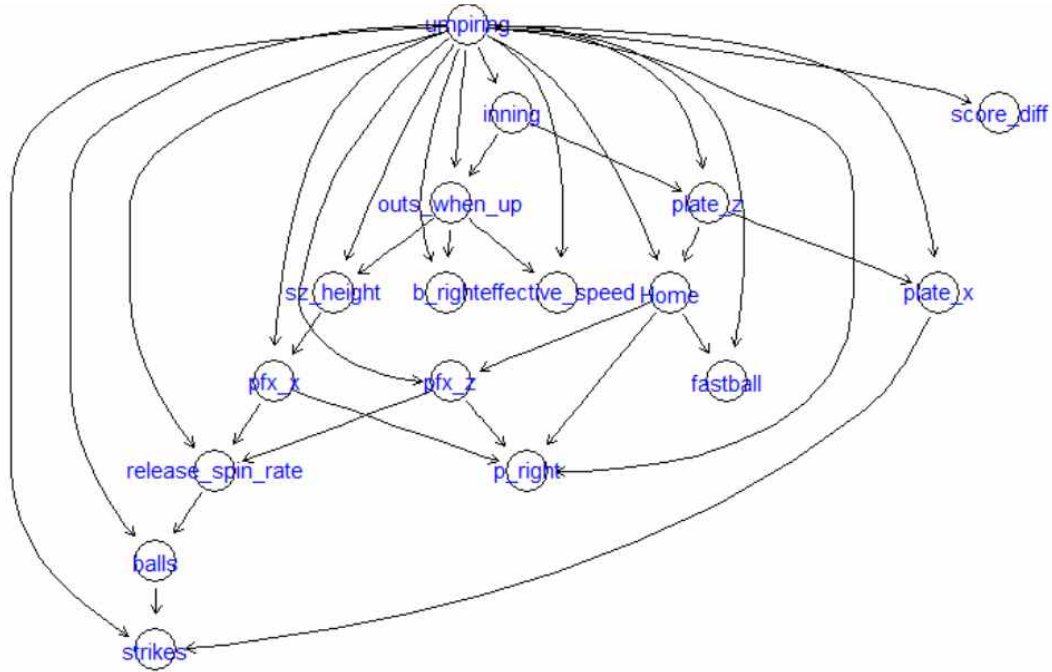


그림 II-2. k-DB 구조학습 결과(2021)

1) 2021 Regular Season

2021년 MLB 정규시즌 데이터로부터 수행한 k-DB의 구조학습 결과는 <그림 II-2>와 같다. 본 연구에서 투입한 변수는 크게 투구된 공의 질 (quality)을 나타내는 변수, 경기 상황을 나타내는 변수, 그리고 홈/원정 구분, 투수와 타자의 좌/우 구분, 타자에 따른 스트라이크 존의 높이 등과 같은 경기 환경을 나타내는 변수로 구분된다. <그림 II-2>에 따르면, 투구된 공의 위치를 나타내는 plate_x와 plate_z는 서로 연결되어 있으며, 또한 투구된 공의 수평적 움직임을 나타내는 pfx_x변수는 p_right변수와 spin 등의 변수로 연결되어 있다. 즉, 투구된 공의 질을 나타내는 변수들은 서로 복잡한 형태로 연결되어 있으며, 이는 경기 상황 또는 환경 변수

로 확장되고 있다. 이와 같은 노드 간의 연결은 변수들 간의 확률적 의존 관계를 의미하며, k-DB에서는 투입된 설명변수들 간의 독립을 가정하지 않고 확률적 의존성을 인정하기 때문에 가능한 구조이다. 반면, score_diff 변수는 다른 변수들과 연결되어 있지 않고 독립적으로 구성되어 있음을 확인할 수 있다. 이는 해당 변수는 다른 변수와 확률적으로 의존적인 관계를 보이지 않음을 의미한다.

<그림 II-2>와 같은 구조학습 결과는 노드들의 확률 값을 기초로 특정 노드가 발생할 확률에 대한 조건부 확률을 계산하는 데 활용된다. 따라서 본 연구에서는 구축한 베이지안 네트워크 모형을 통해 설명변수들의 값이 주어졌을 때, 판정오류 발생의 사후 확률을 산출하여 경기 상황 조합에 따른 판정오류 발생 확률을 추정하였다. 간단한 예시로, 투수의 홈팀 유무에 따라 판정오류가 발생할 확률을 계산하면 다음과 같다.

$$(1) p(ump_error|Home = 1)$$

$$= \frac{p(Home=1|ump_error)p(ump_error)}{p(Home=1)} = \frac{0.5140*0.1191}{0.5057}$$

$$= \mathbf{0.1211}$$

$$(2) p(ump_error|Home = 0)$$

$$= \frac{p(Home=0|ump_error)p(ump_error)}{p(Home=0)} = \frac{0.4859*0.1191}{0.4943}$$

$$= \mathbf{0.1170}$$

판정오류 발생 여부에 대한 사전확률이 0.1191인 상황에서 투수의 홈/원정 구분 변수가 조건부 확률로 포함되었을 때의 판정오류 발생 확률은 홈팀 투수 0.1211, 원정 팀 투수 0.1170으로 나타났다. 본 연구에서 조작

적으로 정의한 판정오류는 투수에게 유리한 판정오류임을 고려한다면, 투수에게 유리한 판정오류는 원정팀의 투수보다 홈팀의 투수에게서 발생 확률이 소폭 높게 나타나는 것을 확인할 수 있다. 이와 같은 결과는 판정오류가 발생했을 때 홈 또는 원정 팀의 투수일 확률, 즉, $p(\text{Home}=0 \text{ or } 1|\text{ump-error})$ 의 조건부 확률 모수를 <그림 II-2>에 제시한 구조로부터 학습하여 추정하였기 때문에 가능한 연산이다.

해당 구조를 바탕으로 추정된 판정오류 발생 확률을 홈/원정, 볼 카운트, 직구/변화구, 좌우투타 구분에 따라 결과를 제시하면 <표 II-3>과 같다. <표 II-3>에 제시한 바와 같이, 아무 조건이 없는(unconditional) 상황에서의 판정오류에 대한 사전확률은 0.1191에 해당한다. 그러나 홈/원정 구분 변수가 조건부 확률로 주어졌을 때는, 앞서 수식으로 제시한 (1), (2) 결과와 같이 홈팀의 투수에게서 판정오류가 발생할 확률이 소폭 높게 나타났다. 볼 카운트에 따른 판정오류 발생 확률 $p(\text{ump-error}|\text{ball-count})$ 은 0ball-0strike 상황에서는 0.15로 나타났으나, 3balls-0strike 상황에서는 0.2160, 0ball-2strikes 상황에서는 0.0445, 3-2 full count 상황에서는 0.09로 볼 카운트에 따른 판정오류 발생 확률은 4%~21%까지 매우 큰 변동을 보이는 것으로 나타났다. 이와 같은 볼 카운트에 홈/원정 구분을 포함하여 판정오류 발생 확률을 추정할 경우 모두 홈팀의 투수에게서 판정오류 발생 확률이 소폭 높게 나타났다. 나아가 홈팀의 투수가 3balls-0strike 상황에서 직구계열의 공을 투구했을 경우에는 판정오류 발생 확률이 0.2238, 변화구 계열을 투구했을 경우에는 0.2096로 직구계열의 투구에서 판정오류 발생 확률이 약 1.5% 높게 나타났다. 또한 같은 조건에서 투타의 좌/우 구분 변수를 포함할 경우 우투수가 우타자를 상대할 경우(0.2313)가 좌투수가 좌타자를 상대하는 경우(0.2137)보다 판정오류 발생 확률이 높게 나타났다.

표 II-3. k-DB 활용 판정오류 발생 확률 추정(2021)

	conditions	Pr(ump_error)
unconditional	p(2021_ump_error)	0.1191
Home/Away	p(ump_error Home)	0.1211
	p(ump_error Away)	0.1170
Ball Counts	p(ump_error 0ball-0strike count)	0.1501
	p(ump_error 3balls-0strike count)	0.2160
	p(ump_error 0ball-2strikes count)	0.0445
	p(ump_error 3-2 full count)	0.0905
Home/Away & Ball Counts	p(ump_error Home, 3balls-0strike)	0.2187
	p(ump_error Away, 3balls-0strike)	0.2125
	p(ump_error Home, 0balls-2strike)	0.0453
	p(ump_error Away, 0balls-2strike)	0.0437
Home, 3-0, fastball	p(ump_error Home, 3balls-0strike, fastball=1)	0.2238
	p(ump_error Home, 3balls-0strike, fastball=0)	0.2096
Left/Right hand	p(ump_error Home, 3balls-0strike, fastball=1, p_right=1, b_right=1)	0.2313
	p(ump_error Home, 3balls-0strike, fastball=1, p_right=0, b_right=0)	0.2137

2) 2020 Regular Season

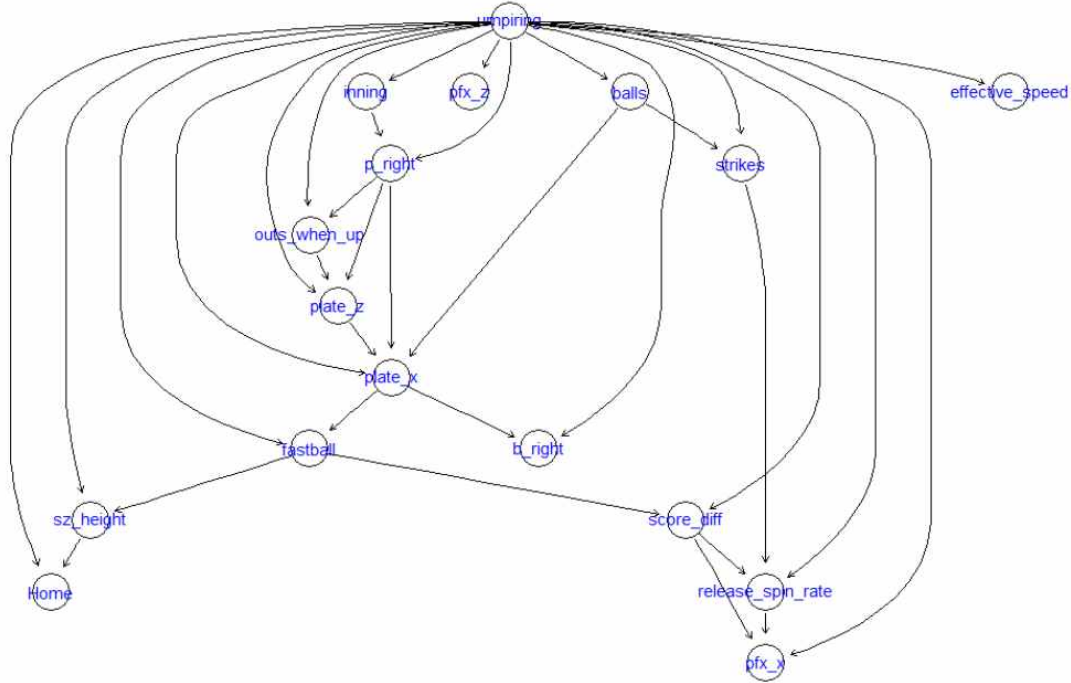


그림 II-3. k-DB 구조학습 결과(2020)

2020년 MLB 정규시즌 데이터로부터 수행한 k-DB 구조학습 결과는 <그림 II-3>와 같다. 2021년 MLB 정규시즌 데이터로부터 k-DB를 수행한 결과와 유사하게 볼 카운트를 나타내는 balls, strikes 변수는 서로 연결되어 있고, plate_z, plate_x 역시 서로 연결되어 있으며, 해당 변수들은 경기 상황 및 환경을 나타내는 또 다른 변수들과 복잡한 확률적 의존성을 보이고 있다. 한편 앞선 2021시즌 분석 자료에서는 score_diff 변수가 독립적으로 떨어져 있는 것으로 나타났으나, 2020년 자료에서는 투구된 공의 속도를 의미하는 effective_speed 변수는 다른 변수들과 확률적 의존성을 나타내지 않았다. 해당 구조학습 결과로부터 추정된 모수를 바탕으로 판정오류 발생 확률을 홈/원정, 볼 카운트, 직구/변화구, 투타 좌우구분을 중점으로 살펴보면 <표 II-4>와 같다.

표 II-4. k-DB 활용 판정오류 발생 확률 추정(2020)

conditions		Pr(ump_error)
unconditional	p(2020_ump_error)	0.1248
Home/Away	p(ump_error Home)	0.1268
	p(ump_error Away)	0.1227
Ball Conunts	p(ump_error no ball-nostrike count)	0.1573
	p(ump_error 3balls-no strike count)	0.1995
	p(ump_error no ball- 2strikes count)	0.0464
	p(ump_error 3-2 full count)	0.1027
Home/Away & Ball Conunts	p(ump_error Home, 3balls-0strike)	0.2025
	p(ump_error Away, 3balls-0strike)	0.1964
	p(ump_error Home, 0balls-2strike)	0.0473
	p(ump_error Away, 0balls-2strike)	0.0456
Home, 3-0, fastball	p(ump_error Home, 3balls-0strike, fastball=1)	0.2068
	p(ump_error Home, 3balls-0strike, fastball=0)	0.1944
Left/Right hand	p(ump_error Home, 3balls-0strike, fastball=1, p_right=1, b_right=1)	0.2039
	p(ump_error Home, 3balls-0strike, fastball=1, p_right=0, b_right=0)	0.2001

2020년 정규시즌 심판의 판정오류에 대한 사전 확률은 12.48%로 2021 시즌 11.91%와 비교하여 소폭 높게 나타났다. 이와 같은 사전 확률에 홈/원정 구분 변수가 포함된 조건부 확률에서는 홈팀 투수의 경우 12.68%, 원정 팀의 투수는 12.27%로 투수에게 유리한 판정 오류 발생 확률은 원정 팀의 투수보다 홈 팀의 투수에게서 소폭 높게 나타났으나, 큰 차이를 보이지는 않았다. 볼 카운트를 고려한 판정오류 발생 확률 $p(\text{ump_error}|\text{ball_count})$ 은 0ball-0strike 상황에서는 약 15%, 3balls-0strike 상황에서는 약 20%, 0ball-2strikes 상황에서는 약 4.5%, 3-2 full count 에서는 약 10%로, 해당 결과 역시 2021년 결과와 유사한 확률을 나타냈다. 또한 홈/원정과 볼 카운트를 함께 고려한 조건부 확률에서 역시 홈 팀의 투수가 유리한 판정을 받을 확률이 소폭 높게 나타났으며, 이와 같

은 확률은 2021년 정규시즌 자료와 유사한 결과를 보였다. 3balls-0strike 상황에서 홈 팀의 투수가 직구계열의 공을 투구했을 때의 판정오류 발생 확률은 20.68%, 같은 조건에서 변화구 계열의 공이 투구되었을 때는 19.44%를 기록하였다. 2021년 자료 분석에서는 같은 조건에서 직구가 투구되었을 때 판정오류 발생 확률은 22.38%, 변화구 계열의 공이 투구되었을 때는 20.96%였음을 고려하면 2020시즌에서는 구종에 따라서 판정오류 발생 확률에 큰 차이를 보이지 않았음을 추론할 수 있다. 또한 우투수가 우타자를 상대하는 상황과 좌투수가 좌타자를 상대하는 상황에서 역시 판정오류 발생 확률은 약 20%로 투수와 타자의 좌/우 구분에 따라서 큰 차이를 보이지 않았다.

3) 2019 Regular Season

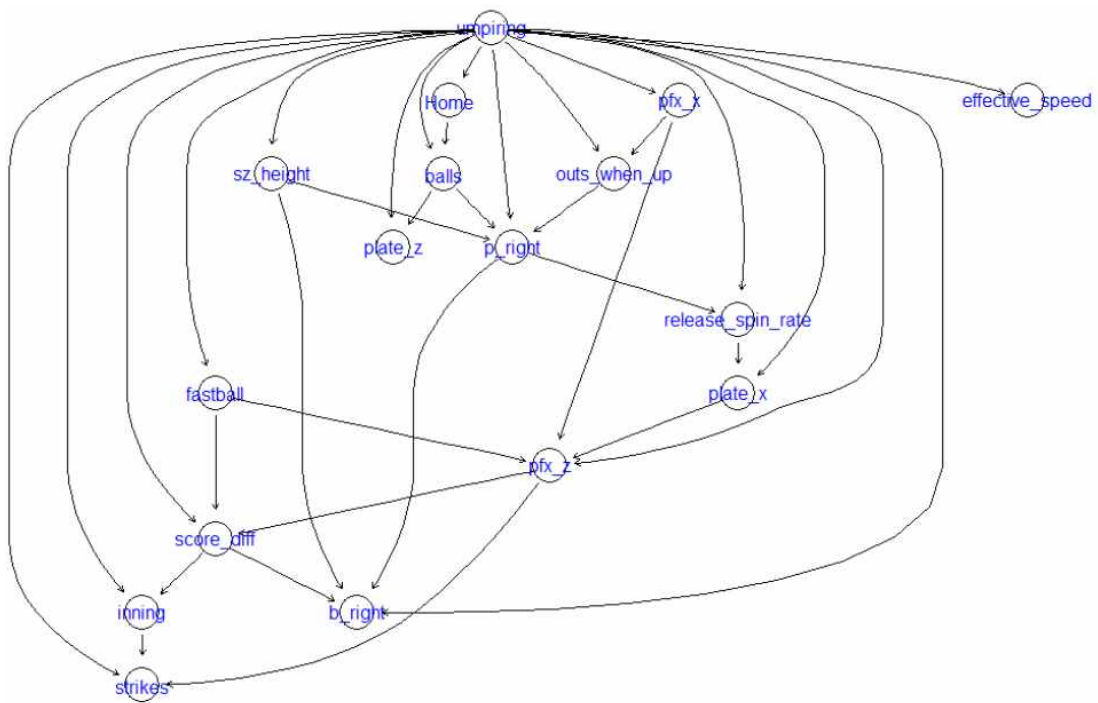


그림 II-4. k-DB 구조학습 결과(2019)

2019년 MLB 정규시즌 데이터로부터 수행한 k-DB 구조학습 결과는 <그림 II-4>와 같다. 앞선 2021, 2020 시즌 자료와 유사하게 투구된 공의 구질을 나타내는 변수들은 서로 연결되어 있으며, 일부 변수들은 아웃 카운트(outs_when_up), 이닝(inning), 점수차(score_diff) 등 경기 상황을 나타내는 변수들과 홈/원정 구분, 투수와 타자의 좌우 구분 등 경기 환경을 나타내는 변수들과 연결되어 있다. 즉, 투구된 공의 속도를 나타내는 effective_speed를 제외하고는 대부분의 변수가 서로 확률적으로 의존적인 복잡한 관계를 보이고 있음을 확인할 수 있다. 해당 구조학습 결과로부터 추정된 모수를 바탕으로 판정오류 발생 확률을 홈/원정, 볼 카운트, 직구/변화구, 투타 좌우구분을 중점으로 살펴보면 <표 II-5>와 같다.

표 II-5. k-DB 활용 판정오류 발생 확률 추정(2019)

conditions		Pr(ump_error)
unconditional	p(2019_ump_error)	0.1489
Home/Away	p(ump_error Home)	0.1519
	p(ump_error Away)	0.1457
Ball Conunts	p(ump_error no ball-no strike count)	0.1977
	p(ump_error 3balls-no strike count)	0.2163
	p(ump_error no ball- 2strikes count)	0.0835
	p(ump_error 3-2 full count)	0.0926
Home/Away & Ball Conunts	p(ump_error Home, 3balls-0 strike)	0.2210
	p(ump_error Away, 3balls-0 strike)	0.2119
	p(ump_error Home, 0ball-2strikes)	0.0852
	p(ump_error Away, 0ball-2strikes)	0.0815
Home, 3-0, fastball	p(ump_error Home, 3balls-0strike, fastball=1)	0.2276
	p(ump_error Home, 3balls-0strike, fastball=0)	0.2080
Left/Right hand	p(ump_error Home, 3balls-0strike, fastball=1, p_right=1, b_right=1)	0.2336
	p(ump_error Home, 3balls-0strike, fastball=1, p_right=0, b_right=0)	0.1972

2019년 정규시즌 심판의 판정오류에 대한 사전 확률은 14.89%로 2020년 12.48%, 2021시즌 11.91%와 비교하여 높게 나타났다. 이와 같은 사전 확률에 홈/원정 구분 변수가 포함된 조건부 확률에서는 홈팀 투수의 경우 12.68%, 원정 팀의 투수는 12.27%로 나타났다. 즉, 2021년, 2020년 자료 분석 결과와 마찬가지로 투수에게 유리한 판정오류 발생 확률은 원정 팀의 투수보다 홈 팀의 투수에게서 소폭 높게 나타났다. 볼 카운트를 고려한 판정오류 발생 확률 $p(\text{ump-error}|\text{ball-count})$ 은 투수의 초구에 해당하는 0ball-0strike 상황에서는 19.77%, 3balls-0strike 상황에서는 21.63%, 0ball-2strikes 상황에서는 8.35%, 3-2 full count 상황에서는 9.26%로, 앞선 2020년, 2021년 자료분석 결과와 비교하여 3-2 full count를 제외한 모든 볼 카운트에서 판정 오류 발생 확률이 높게 나타났다. 또한 홈/원정과 볼 카운트를 함께 고려한 조건부 확률에서는 3balls-0strike에서 홈팀의 투수는 22.1%, 원정 팀의 투수는 21.19%를 기록하였으며, 볼 카운트가 0ball-2strikes로 바뀔 경우 홈 팀의 투수는 8.54%, 원정 팀의 투수는 8.15%로 나타났다. 앞선 자료분석 결과와 같은 맥락에서 투수에게 유리한 판정오류는 홈 팀의 투수에게서 발생할 확률이 높게 나타나고 있으나, 2019시즌에서도 그 차이는 매우 작게 나타나고 있다.

또한, 2020, 2021년 자료 분석에서는 0ball-2strike count에서의 판정오류 발생 확률은 4%대를 기록하였으나, 2019년 자료에서는 8%대를 기록하여 그 차이가 다소 크게 나타났다. 3balls-0strike 상황에서의 판정오류 발생 확률은 2019~2021시즌별 큰 차이가 없음에도, 0ball-2strike count에서는 약 4%의 정도의 차이를 보이는 결과는 주의 깊게 살펴볼 필요가 있을 것이다. 이에 대한 구체적인 논의는 Study 1의 논의 및 결론 파트에 기술하였다.

3.2. 2017-2021 MLB 월드시리즈 판정오류 분석

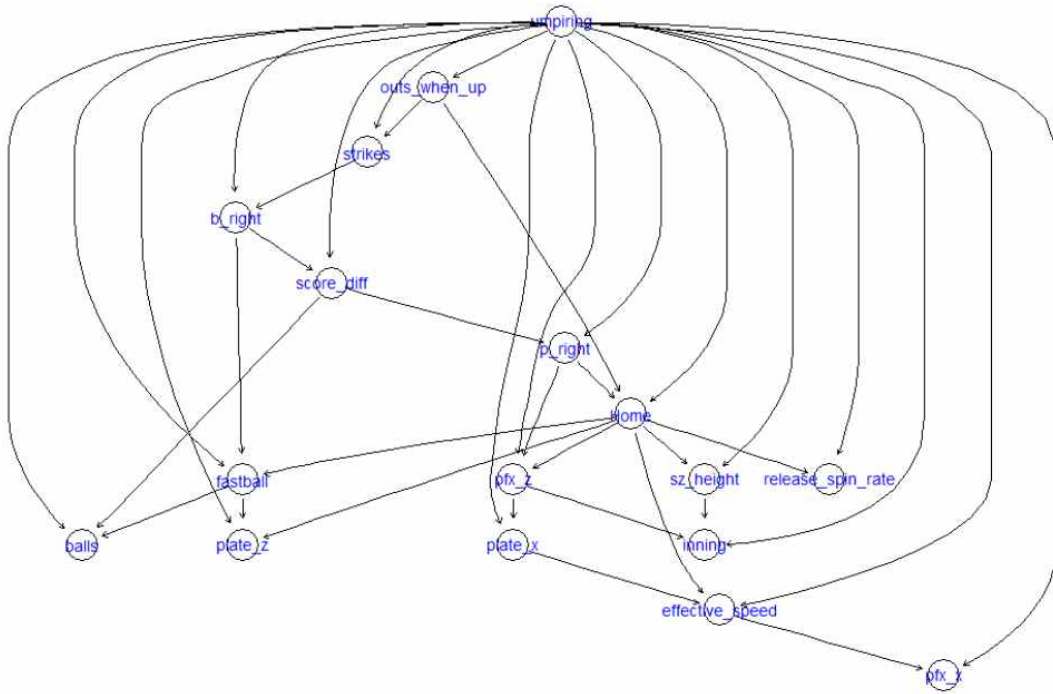


그림 II-5. k-DB 구조학습 결과(월드시리즈)

MLB 심판의 판정오류를 정규시즌과 월드시리즈로 구분하여 비교하기 위해 2017-2021년(최근 5년) 월드시리즈 자료를 대상으로 k-DB를 수행하였다. 구조학습 결과는 <그림 II-5>와 같다. 정규시즌 분석 결과와 유사하게 투구된 공의 질을 나타내는 변수들과 경기 상황 및 환경을 나타내는 변수들은 서로 연결되어 있는 구조를 보였다. 그러나 정규시즌 자료에 대한 구조학습 결과에서는 시즌마다 일부 변수(score_diff, effective_speed)에서 다른 노드들과 연결이 되지 않는 구조를 보였으나, 월드시리즈 자료에서는 모든 변수가 서로 연결된 구조를 보였다. 이와 같은 결과는 월드시리즈에서 발생한 판정오류를 설명하기 위해 투입된 모든 변수가 서로 확률적으로 의존적인 복잡한 관계를 취하고 있음을 의미한다.

표 II-6. k-DB를 통한 판정오류 발생 확률 추정(월드시리즈)

conditions		Pr(ump_error)
unconditional	p(2017-2021 world series_ump_error)	0.1109
Home/Away	p(ump_error Home)	0.1256
	p(ump_error Away)	0.0956
Ball Counts	p(ump_error no ball-no strike)	0.1462
	p(ump_error 3balls-no strike)	0.1487
	p(ump_error no ball-2strikes)	0.0648
	p(ump_error 3-2 full count)	0.0656
Home/Away & Ball Counts	p(ump_error Home, 3balls-0strike)	0.1687
	p(ump_error Away, 3balls-0strike)	0.1257
	p(ump_error Home, 0balls-2strike)	0.0745
	p(ump_error Away, 0balls-2strike)	0.0548
Home, 3-0, fastball	p(ump_error Home, 3balls-0strike, fastball=1)	0.1771
	p(ump_error Home, 3balls-0strike, fastball=0)	0.1249
Left/Right hand	p(ump_error Home, 3balls-0strike, fastball=1, p_right=1, b_right=1)	0.1991
	p(ump_error Home, 3balls-0strike, fastball=1, p_right=0, b_right=0)	0.1376

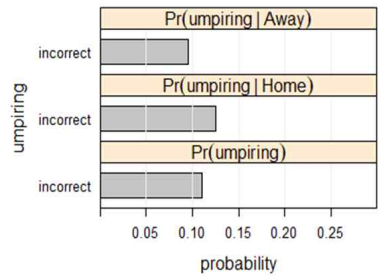
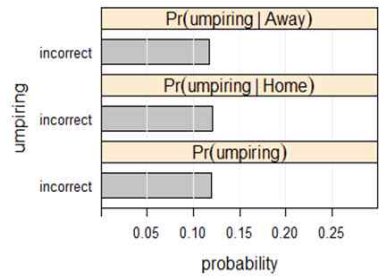
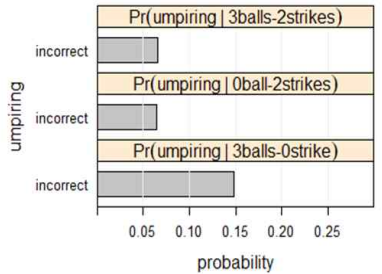
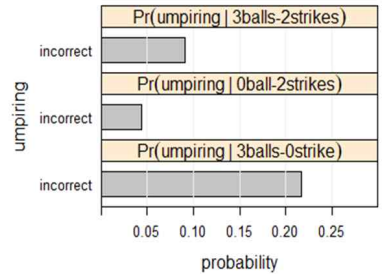
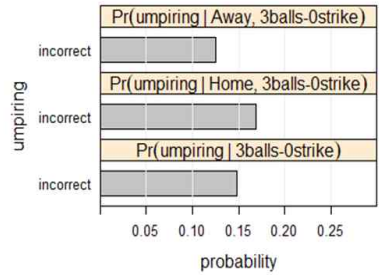
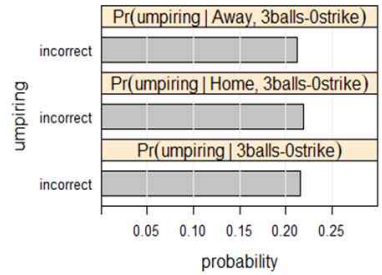
앞선 정규시즌 자료와 같은 조건에서 월드시리즈 판정오류 발생 확률을 정리하면 <표 II-6>과 같다. 2017-2021 월드시리즈에서 판정오류에 대한 사전확률은 11.09%로 2019-2021 정규시즌 자료 분석 결과와 비교하여 가장 낮은 값을 나타냈다. 홈/원정 구분 변수가 포함된 조건부 확률에서는 홈팀 투수의 경우 12.56%, 원정 팀의 투수는 9.56%로 정규시즌 분석 결과와 달리 홈/원정구분에 따라 판정오류 발생 확률에서 다소 차이를 보였다. 볼 카운트를 고려한 조건부 확률에서는 ‘0ball-0strike’ 상황에서 14.62%, ‘3balls-0strike’ 14.78%, ‘0ball-2strikes’ 6.48%,

‘3balls-2strikes’ 6.56%로 나타났다. 같은 조건에서 투수의 홈/원정 구분을 조건부 확률 식에 추가로 포함시킬 경우, 3balls-0strike 상황에서 투수가 홈팀일 경우 판정오류 발생할 확률은 16.87%, 원정 팀의 투수일 경우는 12.57%를 나타냈다. 볼 카운트가 0ball-2strikes인 상황에서는 투수가 홈팀일 경우 판정오류가 발생할 확률은 7.45%, 원정 팀의 투수일 경우는 5.48%를 기록하였다.

언급한 내용에 대한 보다 직관적인 이해를 위해 <표 II-7>에서는 홈/원정과 볼 카운트를 고려한 판정오류 발생 확률을 정규시즌과 월드시리즈로 구분하여 정리하였다. 정규시즌 분석 결과와 마찬가지로 심판의 판정오류 발생 확률은 ‘0ball-2strikes’보다 ‘3balls-0strike’에서 높게 나타나고 있다. 그러나 홈/원정 구분에 따른 판정오류 발생 확률은 정규시즌보다 월드시리즈에서 더 큰 차이를 나타내고 있다. 이와 같은 결과는 홈팀에게 유리한 판정오류는 정규시즌보다 월드시리즈와 같은 특별한 경기에서 더 크게 작용할 수도 있음을 시사한다.

또한, 홈 팀의 투수가 ‘3balls-0strike’ 상황에서 직구계열의 공이 투구되었을 경우와 변화구 계열의 공이 투구되었을 경우 각각의 판정오류 발생 확률은 17.71%, 12.49%를 보였다. 즉, 정규시즌 자료 분석 결과와 마찬가지로, 변화구 계열의 투구보다 직구 계열의 투구에서 판정오류 발생 확률이 높은 것으로 나타났다.

표 II-7. 월드시리즈와 정규시즌 판정오류 발생 확률의 비교

conditions	2017-2021년 월드시리즈	2021 정규시즌
Home/Away	 <p>Prob. = $c(0.0956, 0.1256, 0.1109)$</p>	 <p>Prob. = $c(0.1170, 0.1212, 0.1191)$</p>
Ball counts	 <p>Prob. = $c(0.0656, 0.065, 0.1487)$</p>	 <p>Prob. = $c(0.0905, 0.0445, 0.2160)$</p>
Home/Away +Ball counts	 <p>Prob. = $c(0.1257, 0.1687, 0.1487)$</p>	 <p>Prob. = $c(0.2130, 0.2190, 0.2160)$</p>

3.3. K-DB와 Logistic Regression 결과 비교

앞선 절차에서는 MLB 심판 판정오류 발생의 인과관계를 베이지안 확률 접근에 기반하여 설명하였다. 본 연구의 서론에서도 언급했듯이, 베이지안 접근은 모수를 추정함에 있어 점추정이 아닌 확률분포에 기반하여 사건 발생의 불확실성을 설명하고, 새로운 데이터의 활용이 가능해지면 학습을 업데이트하여 상황에 맞는 새로운 결과를 제시할 수 있다는 장점이 있다(Kaplan & Depaoli, 2012; Gelman, & Shalizi, 2013; van de Schoot & Depaoli, 2014). 이러한 장점을 스포츠 빅데이터에 적용시킬 경우, 스포츠 경기 상황에서 실시간으로 변화하는 다양한 정보를 반영하여 신속하고 유연한 분석모형을 구축할 수 있을 것이다(Kirk, 2003).

본 절에서는 언급한 베이지안 접근의 장단점을 전통적인 빈도기반의 통계분석기법과 비교하기 위해 2021년 MLB 자료를 바탕으로 최대우도법 기반의 로지스틱 회귀분석을 실시하였다. 추정된 로지스틱 회귀식을 바탕으로 MLB 심판의 판정오류 발생 확률을 앞선 2021 정규시즌 k-DB분석 결과와 비교하였다.

1) Logistic Regression 결과

본 연구에서는 로지스틱 회귀분석을 크게 두 차례로 구분하여 실시하였다. 1차 단계에서는 앞선 베이지안 네트워크 분석에서 투입한 변수들을 동일하게 설정하여 독립변수가 판정오류에 미치는 영향력과 통계적 유의성을 탐색하였다. 2차 단계에서는 k-DB분석 결과에서 제시한 조건부 확률식과 같은 조건에서 판정오류 발생 확률을 추정하기 위해 독립변수를 조건부 확률식에 포함한 조건으로 한정하여 분석을 진행하였다. 1차 로지스틱 회귀분석 결과는 <표 II-8>과 같다.

독립변수의 유의성을 확인하기에 앞서 모형적합도 분석을 위해서 수렴(convergence) 여부를 확인한 결과 분리 문제가 없음을 확인하였다.

표 II-8. 판정오류에 대한 로지스틱 회귀분석 결과

Explanatory variable	estimate	exp ^β	z value	p-value
(intercept)	0.581	1.788	2.494	0.013
Home(0 or 1)	0.052	1.053	3.171	0.002**
Innings	0.003	1.003	0.77	0.441
Outs	-0.003	0.997	-0.32	0.749
Strikes	-0.577	0.562	-45.879	< .001***
Balls	0.173	1.189	18.62	< .001***
Plate_x	-0.003	0.997	-0.284	0.776
Plate_z	0.019	1.019	2.086	0.037*
Effective_Speed	-0.019	0.981	-7.863	< .001***
Fastball(0 or 1)	0.240	1.271	8.332	< .001***
pitcher_right(0 or 1)	0.034	1.034	1.608	0.108
batter_right(0 or 1)	0.178	1.195	9.953	< .001***
pxf_x	-0.004	0.996	-0.317	0.751
pxf_z	-0.033	0.967	-1.859	0.063
sz_height	-0.616	0.540	-10.328	< .001***
score_diff	-0.002	0.998	-0.652	0.514
spin	0.0003	1.000	2.037	0.042*

Dependent variable: umpiring_error=1, correct decision=0
Null deviance: 106023 on 145133 degrees of freedom
Residual deviance: 103306 on 145117 degrees of freedom, AIC:103340
Signif. codes: '***' <.001 '**'<.01 '*' <.05

<표 II-8>에 제시된 바와 같이 총 16개의 독립변수 중 유의확률 (p-value)이 .05보다 작게 나타난 변수의 개수는 총 9개로, 그 중 판정오류 발생에 정적인(positive) 영향을 미치는 변수로는 Home($\beta=0.052$, $\exp^\beta=1.053$, $p<.01$), Balls($\beta=0.173$, $\exp^\beta=1.189$, $p<.001$), plate_z($\beta=0.019$, $\exp^\beta=1.019$, $p<.05$), Fastball($\beta=0.240$, $\exp^\beta=1.271$, $p<.001$), batter_right($\beta=0.178$, $\exp^\beta=1.195$, $p<.001$), spin($\beta=0.0003$, $\exp^\beta=1.001$, $p<.05$)로 나타났다. 반대로 판정오류에 부적적인(negative) 영향을 미치는 변수로는 Strikes($\beta=-0.577$, $\exp^\beta=0.562$, $p<.001$), effective_speed($\beta=-0.019$, $\exp^\beta=0.981$, $p<.001$),

sz_height($\beta=-.577$, $\exp^\beta=.562$, $p<.001$)로 나타났다.

제시한 변수들 중 Fastball 변수는 \exp^β 값이 1.271로 심판의 판정오류에 정적인 영향을 미치는 변수 중 가장 큰 영향력을 보였는데, 이는 “다른 조건들이 일정할 때, 직구계열의 투구는 변화구계열의 투구보다 심판의 정분류 확률 대비 판정오류가 발생할 확률($p/1-p$)이 1.271배 높다.”고 해석할 수 있다. 또한 Strikes 변수의 \exp^β 값은 .562로 판정오류에 부적인 영향을 미치는 변수들 중 가장 큰 영향력을 나타냈다. 이는 “다른 조건이 일정할 때, strike 카운트가 1개 증가하면 심판의 정분류 확률 대비 판정오류가 발생할 확률이 0.562배 감소한다”고 해석할 수 있다. 즉, 스트라이크 판정이 추가로 내려질 수 없는 Xballs-2strikes 카운트와 같은 상황에서는 심판의 판정이 보다 정확하게 이뤄지고 있을 것임을 추론할 수 있다. k-DB분석 결과와 마찬가지로 본 연구에서 조작적으로 정의한 판정오류가 ‘strike존을 벗어났음에도, strike판정을 받은 투구’임을 고려한다면, 이는 볼 카운트에 따라 다음 선택지가 존재하는지에 따라서 심판의 판정이 편향적으로 이뤄질 수 있음을 시사하는 결과이다.

2) k-DB와 Logistic Regression 결과 비교

로지스틱 회귀분석 분석에서 회귀계수의 변화는 종속변수의 로그 승산의 변화를 나타내므로, 직관적인 결과 해석에 어려움이 있다. 따라서 대부분의 연구에서는 앞서 제시한 로지스틱 회귀분석 결과에 대한 해석과 같이, \exp^β , 즉 승산비(odds ratio)의 개념을 통해 로지스틱 회귀분석 결과를 해석한다(고길곤, 2017, 황수용 외, 2019). 이와 같은 승산의 변화는 로그 승산에 대한 변화량보다 상대적으로 해석이 용이하지만, 이 또한 k-DB분석 결과에서 제시한 바와 같이 직접적인 확률을 추정하는 것보다는 직관적이지 못하다. 따라서 본 절에서는 k-DB분석 결과에서 제시한 조건부 확률식과 같은 조건에서 판정오류 발생 확률을 추정하기 위

해 독립변수를 조건부 확률식에 포함한 조건으로 한정하여 분석을 진행하였다. 다음 단계로, 추정된 회귀식을 바탕으로 한계효과분석(marginal effects analysis)을 실시하였다. 이를 통해 특정 조건에서의 판정오류 확률을 추정하여 그 결과를 앞선 k-DB 분석 결과와 비교하였다. 이에 대한 결과는 총 두 개의 표로 정리하였다. <표 II-9>는 비교적 간단한 상황에서의 판정오류 발생 확률을 제시하였고, <표 II-10>은 보다 복잡한 상황에서의 판정오류 발생 확률에 대하여 제시하였다. 구체적인 결과는 다음과 같다.

<표 II-9>에 제시된 바와 같이, k-DB와 로지스틱 회귀분석 결과는 대부분의 결과에서 큰 차이를 보이지 않았으며, 그 차이는 1% 내외로 나타났다. 아래의 <그림 II-6>에 제시한 바와 같이, 신뢰구간 역시 $\pm 1\%$ 내외로 나타나 비교적 신뢰할 수 있는 판정오류 확률 추정이 이루어졌음을 확인할 수 있었다. 한편, <표 II-10>에 제시한 바와 같이, 독립변수가 다양하게 포함될 경우 k-DB방법과 logit은 판정오류 발생 확률 차이는 3~11%로 앞선 결과와 다소 차이를 보였다. <표 II-10>의 첫 행에 제시한 ‘p(ump_error|Home, 3balls=0strike, fastball=1, p_right=1, b_right=1, plate_x=1, plate_z=2, ball_speed=90mile)’은 3balls=0strike 상황에서 홈 팀의 우완투수(Home=1, p_right=1)가 우타자를 상대로 오프닝 높이로 바깥쪽으로 90마일의 직구계열의 공을 투구했을 때 판정오류가 발생할 확률을 추정한 것이다. 즉, 앞선 p(ump_error|Home, 3balls=0strike, fastball=1, p_right=1, b_right=1)에서 투구된 공의 수직적, 수평적 위치와 속도가 포함된 것이다. strike와 ball의 구분은 스트라이크 존에서 벗어난 정도가 클수록 명확히 구분되는 ball에 해당하기 때문에, 투구된 공의 위치는 직관적으로 판정오류에 직접적인 영향을 미칠 것으로 예상되는 변수이다. 그럼에도 불구하고, 로지스틱 회귀분석 결과에서 plate_x변수는 통계적 유의성을 나타내지 않았고, 이에 따라 새로운

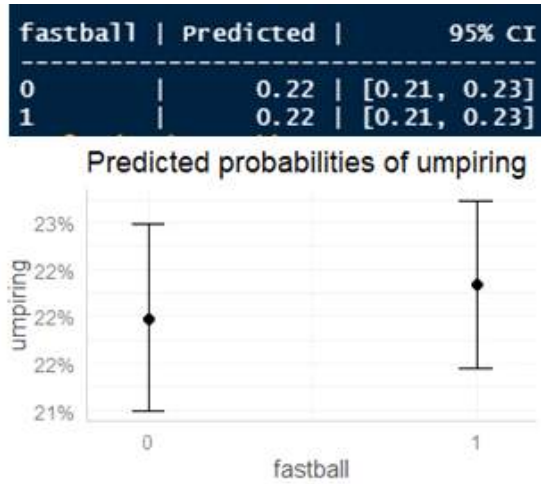


그림 II-6. 구종에 따른 판정오류 발생 확률

표 II-9. k-DB와 로지스틱 회귀분석 결과 비교 1

conditions	Pr(ump_error)		
	KDB	Logit	
unconditional	p(2021_ump_error)		
	0.1191		
Home/Away	p(ump_error Home)	0.1211	0.12
	p(ump_error Away)	0.1172	0.12
Ball Conunt	p(ump_error no ball-no strike)	0.1501	0.14
	p(ump_error 3balls-no strike)	0.2160	0.22
	p(ump_error no ball- 2strikes)	0.0445	0.05
	p(ump_error 3-2 full count)	0.0905	0.08
Home/Away + Ball count	p(ump_error Home, 3balls-0strike)	0.2187	0.22
	p(ump_error Away, 3balls-0strike)	0.2125	0.21
Home/Away + Ball count + Fastball	p(ump_error Home, 3balls-0strike, fastball=1)	0.2238	0.22
	p(ump_error Home, 3balls-0strike, fastball=0)	0.2096	0.22
Home/Away + Ball count + Fastball + pitcher, batter L/R hands	p(ump_error Home, 3balls-0strike, fastball=1, p_right=1, b_right=1)	0.2313	0.24
	p(ump_error Home, 3balls-0strike, fastball=1, p_right=0, b_right=0)	0.2137	0.21

표 II-10. k-DB와 로지스틱 회귀분석 결과 비교 2

conditions		Pr(ump_error)	
		KDB	Logit
unconditional	p(ump_error)	0.1191	
plate_x,z +ball_speed	p(ump_error Home, 3balls-0strike, fastball=1, p_right=1, b_right=1, plate_x=1, plate_z=2, ball_speed=90(mile))	0.3624	0.25
height of strike-zone	p(ump_error Home, 3balls-0strike, fastball=1, p_right=1, b_right=1, plate_x=1, plate_z=2, ball_speed=90(mile), sz_height=1.7)	0.3539	0.26
out counts	p(ump_error Home, 3balls-0strike, fastball=1, p_right=1, b_right=1, plate_x=1, plate_z=2, ball_speed=90(mile), sz_height=1.7, score_diff=0, outs=0)	0.3592	0.28
	p(ump_error Home, 3balls-0strike, fastball=1, p_right=1, b_right=1, plate_x=1, plate_z=2, ball_speed=90(mile), sz_height=1.7, score_diff=0, outs=1)	0.3692	0.28
	p(ump_error Home, 3balls-0strike, fastball=1, p_right=1, b_right=1, plate_x=1, plate_z=2, ball_speed=90(mile), sz_height=1.7, score_diff=0, outs=2)	0.3472	0.28
px_x	p(ump_error Home, 3balls-0strike, fastball=1, p_right=1, b_right=1, plate_x=1, plate_z=2, ball_speed=90(mile), sz_height=1.7, outs=2, px_x=0.5)	0.3102	0.28
	p(ump_error Home, 3balls-0strike, fastball=1, p_right=1, b_right=1, plate_x=1, plate_z=2, ball_speed=90(mile), sz_height=1.7, outs=2, px_x=-0.5)	0.3243	0.28

조건이 추가되었음에도 추정된 판정오류 발생 확률은 큰 변화를 보이지 않았다. 또한 아웃카운트를 고려한 결과에서도 k-DB에서는 34.72~36.92%의 판정오류 발생 확률을 추정하여 아웃카운트에 따라 다소 차이를 보였으나, 로지스틱 회귀분석에서는 모든 아웃카운트에서 고정된 28%의 결과를 나타냈다.

```

> prob_graph
# Predicted probabilities of umpiring

# outs_when_up = 0

plate_x | Predicted | 95% CI
-----|-----|-----
-1.50 | 0.28 | [0.27, 0.30]
-1.00 | 0.28 | [0.27, 0.30]
-0.50 | 0.28 | [0.27, 0.30]
0.00 | 0.28 | [0.27, 0.30]
0.50 | 0.28 | [0.27, 0.30]
1.50 | 0.28 | [0.27, 0.30]

# outs_when_up = 1

plate_x | Predicted | 95% CI
-----|-----|-----
-1.50 | 0.28 | [0.27, 0.30]
-1.00 | 0.28 | [0.27, 0.30]
-0.50 | 0.28 | [0.27, 0.30]
0.00 | 0.28 | [0.27, 0.30]
0.50 | 0.28 | [0.27, 0.30]
1.50 | 0.28 | [0.27, 0.30]

# outs_when_up = 2

plate_x | Predicted | 95% CI
-----|-----|-----
-1.50 | 0.28 | [0.27, 0.30]
-1.00 | 0.28 | [0.27, 0.30]
-0.50 | 0.28 | [0.27, 0.30]
0.00 | 0.28 | [0.27, 0.30]
0.50 | 0.28 | [0.27, 0.30]
1.50 | 0.28 | [0.27, 0.30]

```

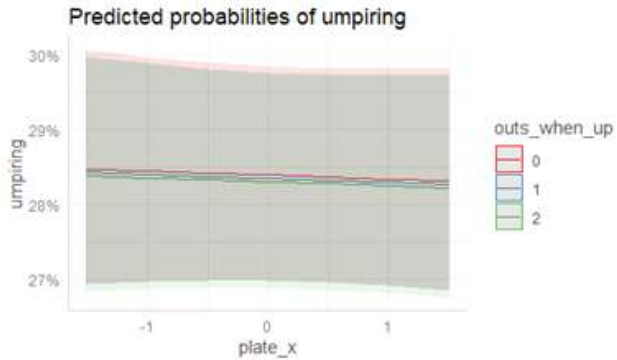


그림 11-7. plate_x와 out counts에 따른 판정오류 발생 확률

위의 그림은 앞서 제시한 내용에 대한 하나의 예시로, 투구된 공의 수평적 움직임을 나타내는 plate_x 변수와 아웃카운트 (out_when_up=0, 1, 2)의 연속적인 변화에 따른 판정오류 발생 확률을 추정한 결과를 시각화하여 보여주고 있다. plate_x와 아웃카운트의 연속적인 변화에도 불구하고, 추정된 판정오류 발생 확률은 28%로 고정되어 있음을 확인할 수 있다. 이를 그래프 형태로 제시한 그림에서는 plate_x의 변화에 따른 판정오류 발생 확률은 실선의 색으로 구분한 아웃카운트에 따라서 변화가 없으며 95% 신뢰구간 역시 27~30%의 비교적 넓은 범위로 나타나고 있다. 이에 대한 구체적인 논의는 ‘4. 논의 및 결론’ 절에 제시하였다.

3.4. MLB 판정오류의 파급효과 분석

앞선 절에서는 k-DB 통해 MLB 판정오류 발생 확률 변화의 인과관계를 설명하고, 그 결과를 로지스틱 회귀분석 결과와 비교하여 스포츠 빅데이터 분석에 있어서 베이지안 접근과 로지스틱 회귀분석의 결과를 비교하였다. 본 절에서는 판정오류가 각 구단들에 어느 수준의 경제적 파급효과를 유발하는지 추정하고자 한다. 구체적으로, 각 팀의 삼진과 볼넷이 승리에 미치는 영향을 추정하고, 이를 판정오류로 인해 ‘투수가 얻은 삼진’, ‘타자가 잃은 볼넷’의 빈도에 반영하여 판정오류로 인해 각 팀이 얻거나 잃은 승리 빈도를 추정한 뒤, 최종적으로 이를 팀의 시즌 수익에 대비하여 경제적 가치를 추정하였다. 구체적인 결과는 다음과 같다.

1) 판정오류를 고려한 승리빈도 추정

판정오류는 투수의 삼진과 타자의 볼넷 개수에 직접적인 영향을 미치기 때문에, 우선적으로 투수의 삼진과 타자의 볼넷이 실질적으로 승리에 어느 정도의 영향을 미치는지 분석하기 위해 다중회귀분석을 실시하였다. 팀 단위 볼넷과 삼진을 하나의 회귀모형에 함께 포함시킬 경우 통제변수와의 다중공선성 문제가 발생하여, 승리에 대한 삼진의 가치를 추정하는 회귀식과 볼넷의 가치를 추정하는 회귀식을 각각 추정하였다.

<표 II-11>의 (1)행에 제시된 바와 같이, 시즌(2016, 2017, 2018, 2019, 2021, 2020시즌은 COVID-19로 제외), 리그(AL/NL), 팀의 타격 능력이 일정하게 고정되어 있을 때, 투수의 삼진이 한 개 증가하면 팀의 승리는 0.046($p < .001$) 증가하는 것으로 나타났다. 또한 (2)행에서는 시즌, 리그, 팀의 투수력이 통제되었을 때, 타자가 얻은 볼넷이 한 개 증가하면 승리는 0.072($p < .001$)만큼 증가하는 것으로 나타났다.

표 II-11. 삼진과 볼넷이 MLB 팀 승리에 미치는 영향

Model	Explanatory variable	Parameter estimate	t-statistic	p-value
(1)	(intercept)	0.0002	2.515	<.05*
	Seasons	-1.075	-2.584	<.05*
	League(AL,NL)	-2.740	-1.875	.063
	Performance(OPS+)	0.834	8.981	<.001***
	SO	0.046	7.413	<.001***
(2)	(intercept)	-60.89	-0.080	.936
	Seasons	0.091	0.240	.810
	League(AL,NL)	3.983	3.057	<.01**
	Performance(FIP)	-18.985	-10.986	<.001***
	BB	0.072	6.185	<.001***

Dependent variable: The Number of Wins for MLB Team *i* in Regular Season *t*
 Control variables: Seasons, League(AL/NL), performance[(1)OPS+, (2)FIP]
 Multiple R-squared of model(1): 0.6325, Multiple R-squared of model(2): 0.6657
 Signif. codes: '***' <.001 '**' <.01 '*' <.05

다음 절차로, ‘Xballs-2strikes’와 ‘3balls-Xstrikes’ 카운트에서 발생한 판정오류 빈도를 각 팀별로 수집하였다. ‘Xballs-2strikes’상황에서 Ball이 투구되었으나 Strike판정이 이루어졌다면 투수는 판정오류로 삼진을 얻게 될 것이며, ‘3balls-Xstrikes’상황에서 Ball이 투구되었으나 Strike가 선언되었다면, 타자가 볼넷을 잃게 될 것이다. 이렇듯 각 팀이 경험한 판정오류 빈도를 수집하여 위의 회귀분석을 통해 추정된 각각의 삼진과 볼넷의 가치에 곱해주어, 한 시즌 동안 누적된 판정오류(‘투수가 얻은 삼진’, ‘타자가 잃은 볼넷’)가 각 팀에게 어느 수준의 승리 증감을 유발하는지 추정하였다.

2) MLB ‘ball/strike’판정오류의 경제적 파급효과

앞선 절차에서는 삼진과 볼넷이 승리에 미치는 영향력을 추정하고, 추정된 계수를 각 팀이 한 시즌동안 실제로 경험한 판정오류 누적빈도에 반영하여 판정오류가 정심(correct decision)으로 이루어졌다고 가정했을 때, 각 팀이 얻게 되는(또는 잃게 되는) 승리 빈도를 추정하였다. 본 절

에서는 ‘판정오류를 고려한 승리 빈도’를 경제적 가치로 환산하기 위해 다중회귀분석을 실시하여 MLB 팀 1승의 경제적 가치를 시즌 별 팀의 수익(revenue)에 기반하여 추정하였다. 그 결과는 아래의 <표 II-12>와 같다.

표 II-12. MLB 팀 1승의 경제적 가치 추정

Explanatory variable	Parameter estimate	t-statistic	p-value
(intercept)	-8311.812	-1.027	.306
Seasons	4.188	1.045	.298
League(AL,NL)	-14.324	-1.038	.301
Postseason appearance	37.796	1.674	.096
Wln	2.174	2.731	<.01**

Dependent variable: Total Revenue(\$million) for MLB Team i in Season t
 Control variables: Seasons, League(AL/NL), Postseason appearance
 Multiple R-squared: 0.2289, Adj. R-squared: 0.2077
 F-statistic: 10.76 on 4 and 145 DF, p-value: <.001
 Signif. codes: ‘***’ <.001 ‘**’<.01 ‘*’ <.05

<표 II-12>에 제시한 바와 같이, 1승의 경제적 가치를 추정하기 위한 회귀식에서는 종속변수를 시즌 별 각 팀의 수익, 독립변수를 승리횟수로 설정하고, 시즌, 리그, 포스트시즌 진출 유무는 통제변수로 포함하였다. 분석 결과 시즌, 리그, 포스트시즌 진출 유무가 통제되었을 때, MLB 구단의 1승은 \$2.174 million의 경제적 가치를 가지는 것으로 나타났다. 본 절에서는 이와 같은 결과를 앞선 절차에서 추정된 ‘판정오류를 고려한 승리 빈도’에 반영하여 최종적으로 각 팀이 판정오류로 인해 잠재적으로 경험할 것으로 예상되는 경제적 득실을 추정하였다.

앞서 추정된 1승의 경제적 가치를 바탕으로, 2021시즌 판정오류로 인해 각 팀이 경험한 경제적 득실을 정리하면 <표 II-13>과 같다. ‘Freq. of SO’, ‘Freq. of BB’, Number of Wins for considering SO+BB(applying coefficients) 행에 제시하였다. <표 II-13>에 따르면 2021년 시즌 기준 MLB 심판의 판정오류로 인해 가장 큰 이득을 본 팀

은 PHI(philadelphia phillies)로 투수가 얻은 삼진의 개수는 101개, 타자가 잃은 볼넷의 개수는 35개로 나타났다. 이와 같은 개수를 추정된 삼진, 볼넷의 가치에 반영하면 2021년 PHI는 판정오류로 인해 약 2.13승을 얻은 것으로 추정할 수 있다. 반면 2021년 시즌 판정오류로 인해 가장 큰 손실을 본 팀은 DET(detroit tigers)로 판정오류로 투수가 얻은 삼진은 50개, 타자가 잃은 볼넷은 52개였다. 이를 추정된 삼진, 볼넷의 가치에 반영하면 2021년 DET는 판정오류로 인해 약 1.44승을 잃은 것으로 추정할 수 있다.

<표 II-13>에서 음영으로 표시한 ‘Estimated economic ripple effects’는 판정오류를 고려한 승리빈도에 앞선 회귀분석 결과 \$2.174 million을 곱해준 결과이다. 즉, 판정오류로 인해 ‘투수가 얻은 삼진’이 ‘타자가 잃은 볼넷’보다 많은 팀의 경우 +(positive)의 경제적 파급효과를 기록하고 있고, 반대의 경우에는 -(negative)의 경제적 파급효과를 기록하고 있다. 그 범위는 \$-3.14m~\$4.62m으로 나타나고 있으며, 2021시즌 CWS(chicago white sox)는 판정오류로 인해 투수가 얻은 삼진의 개수는 59개, 타자가 잃은 볼넷의 개수는 38개로 다소 차이가 있지만, 1승에 대한 삼진과 볼넷의 가치를 반영하여 경제적 득실을 추정할 경우¹⁾ 그 결과는 -0.05로 0에 가깝게 나타났다. 우리나라에서 잘 알려진 MLB 선수, Pittsburgh Pirates 최지만, Sandiego Padres 김하성, Losangles Angels 오타니 쇼헤이(Shohei Ohtani)의 2022시즌 기준 연봉이 각각 \$3.2m, \$6m, \$5.5m임을 고려하면, MLB 팀들은 판정오류로 인해 FA가 아닌 주전급 선수의 1년 연봉 수준의 경제적 득실을 잠재적으로 경험하고 있음을 추론할 수 있다.

1) $[(59*0.046)-(38*0.072)]*2.174=-0.047$

표 11-13. 판정오류의 경제적 파급효과 추정

Team	Freq. of SO (Xballs-2strikes)	Freq. of BB (3balls-Xstrikes)	N of Wins for considering SO and BB	Estimated economic ripple effects (\$million)
PHI	101	35	2.13	4.62
MIL	76	29	1.41	3.06
BOS	87	48	0.55	1.19
CHC	78	43	0.49	1.07
CLE	68	36	0.54	1.17
ATL	62	32	0.55	1.19
LAA	76	42	0.47	1.03
NYM	63	34	0.45	0.98
COL	70	39	0.41	0.90
CIN	86	51	0.28	0.62
KC	58	33	0.29	0.63
NYN	72	44	0.14	0.31
SD	70	44	0.05	0.11
CWS	59	38	-0.02	-0.05
STL	81	55	-0.23	-0.51
MIN	58	39	-0.14	-0.30
OAK	63	45	-0.34	-0.74
ARI	64	46	-0.37	-0.80
TEX	74	55	-0.56	-1.21
TOR	64	48	-0.51	-1.11
MIA	51	40	-0.53	-1.16
WSH	55	43	-0.57	-1.23
BAL	55	44	-0.64	-1.39
PIT	51	43	-0.75	-1.63
SEA	53	46	-0.87	-1.90
TB	53	46	-0.87	-1.90
HOU	70	59	-1.03	-2.23
SF	69	62	-1.29	-2.80
LAD	61	57	-1.30	-2.82
DET	50	52	-1.44	-3.14

Each coefficient of SO and BB for the number of wins is .046, .07
 SO(Xballs-2strikes): Frequency of obtained 'strikeout' by umpiring error(strike calls) under 'Xballs-2stirkes' count.
 BB(3balls-Xstrikes): Frequency of lost 'Base on Balls' by umpiring error(strike calls) under '3balls-Xstirkes' count.

4. Study 1에 대한 논의 및 결론

본 학위논문의 Study 1 ‘베이지안 확률 접근을 통한 MLB 심판 판정의 불확실성 분석’에서는 k-dependence bayesian network classifier(k-DB)를 활용하여 MLB 심판의 판정오류 발생의 인과관계를 설명하였다. k-DB의 구조학습결과를 바탕으로 투입된 변수들 간의 확률적 의존 구조를 탐색하였고, 모수학습을 통해 추정된 확률을 바탕으로 경기 상황의 조합에 따른 조건부 확률을 추정하였다. 다음 절차로, 같은 자료를 활용하여 최대우도법 기반의 로지스틱 회귀분석을 실시하여 추정된 판정오류 발생 확률을 k-DB 분석 결과와 비교하였으며, 이를 통해 스포츠 빅데이터 분석 연구에서 각 기법의 장단점에 대하여 논의하였다. 이러한 절차를 바탕으로 판정오류 발생의 인과관계에 대하여 논의한 후, Study 1의 마지막 단계에서는 MLB에서 판정오류가 어떤 파급효과를 야기하는지 분석하였다. 구체적으로 1승에 대한 삼진과 볼넷의 가치와 팀의 수익을 통해 판정오류가 각 구단에 미치는 잠재적인 경제적 파급효과를 추정하였다.

k-DB의 구조학습 결과 투구된 공의 수직적/수평적 움직임(pfx_x, pfx_z), 회전수 등의 투구된 공의 질(quality)을 나타내는 노드(node) 들은 선(edge)으로 서로 연결되어 확률적 의존성을 나타냈다. 또한 해당 변수들은 투구된 공의 위치를 비롯하여 투수와 타자의 좌/우 구분, ball-strike 카운트, 홈/원정 구분 등 다양한 경기 상황을 나타내는 변수들과 연결되어 서로 복잡한 확률적 의존관계를 나타냈다. k-DB에서는 투입한 독립변수에 대하여 최대 임의의 k개의 변수와 확률적 의존성을 허용한다. 따라서 k-DB구조학습 결과 다양한 변수들이 유기적으로 서로 연결된 구조를 나타냈다. 이와 같은 결과를 통해 본 연구에서 판정오류를 설명하기 위해 투입한 변수들은 서로 확률적으로 의존성이 있는 복잡한 구조를 취하고 있으며, 이는 다양한 원인이 서로 복합적인 상호작용을 통해 하나의 결과를 만들어 내고 있음을 시사한다.

본 연구에서 다룬 MLB 심판의 ball-strike 판정뿐만 아니라, 스포츠 상황에서 발생하는 대부분의 사건은 하나의 원인이 결과로 직결되는 경우보다, 다양한 원인이 서로 복합적인 상호작용을 통해 하나의 결과를 만들어 내는 경우가 많다. 이에 대하여 Gudmundsson과 Horton(2017)의 연구에서는 보다 발전된 스포츠 분석에서는 선수들간의 상호작용과 같은 실시간으로 변화하는 복잡하고 불확실한 문제를 고려할 수 있어야 한다고 보고한 바 있다. 즉, 스포츠 현장에서 수집되는 빅데이터를 기반으로 스포츠 현상을 타당하게 설명하기 위해서는 독립변수와 종속변수 간의 1:1 관계뿐만 아니라, 설명변수들 간의 상호작용 및 비선형 관계를 고려할 수 있는 모형의 적용이 필요할 것이다. 이러한 맥락에서 k-DB는 학습을 통해 변수들 간의 확률적 의존성을 추론해내고 이를 시각화할 수 있기 때문에 기존의 전통적인 접근들과 비교하여 보다 직관적인 결과를 제시해 줄 수 있을 것이다.

MLB 심판의 판정오류에 대한 사전확률은 2021년 11.91%, 2020년 12.48%, 2019년 14.89%로 시기적으로 최근 시즌일수록 판정의 정확성은 높게 나타났다. Williams(2019)의 연구에서 역시 ‘A bad call ratio’로 표현한 판정오류 발생 확률은 2008년부터 2018년까지 시기가 지남에 따라 점차 줄어들고 있는 것으로 보고하였다. 이에 대하여 Williams(2019)는 MLB에서 판정오류에 대한 논란이 지속적으로 부각됨에 따라 판정오류가 줄어들고 있는 추세이긴 하나, 그 비율은 여전히 높다고 보고하였다.

k-DB 모수학습 결과를 바탕으로 추정된 판정오류에 대한 조건부 확률에서는 ‘Ball-Strike’ 카운트에 따라 매우 큰 차이를 나타냈다. 2019-2021 MLB 정규시즌 경기에서 ‘3balls-0strike’일 때 판정오류 발생 확률은 약 19~21%로 볼 카운트 조합 중 가장 크게 나타났다. 반면, ‘0ball-2strikes’에서 판정오류 발생확률은 약 4~8%로 앞선 조건과 비교하여 매우 낮게 나타났다. 즉, 볼 카운트 상에서 추가적인 선택지의 존재

여부에 따라서 심판의 판정은 크게 달라지는 것을 실증적으로 검증할 수 있었다. 같은 맥락에서 Green과 Daniels(2014)의 연구에서 역시, 스트라이크 판정이 2개가 선행된 상황에서는 심판의 스트라이크 존이 줄어들고, 스트라이크 존을 벗어나 볼(ball) 판정이 3개가 선행된 상황에서는 스트라이크 존이 확장되는 것을 실증적으로 검증하였다. 이러한 결과에 대하여 Green과 Daniels(2015)은 프로야구 심판은 경기에 대한 개입을 최소화해야 한다는 ‘영향회피(impact aversion)’ 성향이 이와 같은 결과를 초래할 수 있음을 보고하였다. 이러한 영향회피는 개입하지 않음을 최선으로 간주하는 심리학의 ‘부작위 편향’개념으로 설명될 수 있을 것이다. 이와 관련하여 “엘리트 심판의 심리적 편향, 오심 낚는다(경향신문, 2012.06.25).”의 기사에서는 인간의 부작위 편향은 프로야구 현장에서 오심으로 나타날 수 있음을 지적하며, 심판의 권위보다 정확한 판정의 중요성을 강조한 바 있다(황수웅, 박일혁 2020). Williams(2019)의 연구에서 역시 이러한 결과와 같은 맥락을 보이며, 볼 카운트에 따라 심판의 판정오류 발생 확률은 최대 2배 가까이 차이가 난다고 보고하였다.

정규시즌과 월드시리즈에서의 판정오류 발생 확률을 비교한 결과에서는 월드시리즈에서 판정오류에 대한 사전확률이 더 낮게 나타났다. 이를 통해, MLB심판은 정규시즌보다 월드시리즈에서 더욱 정확한 판정을 수행하고 있음을 추론해볼 수 있을 것이다. 그러나 홈/원정 구분에 따른 판정오류 발생 확률 차이는 정규시즌보다 월드시리즈에서 더 크게 나타났다. 또한 볼 카운트에 따른 판정오류 발생 확률에서는 ‘3balls-0strike’ 상황에서는 약 15%를 기록하였으나, ‘0ball-2strikes’와 ‘3-2 full count’에서는 6%대를 기록하여 정규시즌 자료 분석 결과와 같이 볼 카운트에 따라 심판의 판정오류 발생 확률은 다소 차이를 보였다. 즉, 월드시리즈 경기에서는 정규시즌과 비교하여 판정오류 발생 확률 자체가 낮게 나타나지만, 홈팀 투수에게 유리한 판정을 보이는 성향은 정규시즌보다 더 뚜렷

하게 나타났다. 또한 볼 카운트에 따른 판정오류 발생 확률은 '0ball-2strikes', '3-2 full count'와 같이 삼진에 대한 선택적 유보가 불가능한 상황에서는 정규시즌과 월드시리즈 간에 큰 차이를 보이지 않았으나, '3balls-0strike'와 같이 스트라이크 카운트가 여유 있는 상황에서는 월드시리즈에서의 판정오류 발생 확률이 더 낮은 것으로 나타났다. 이러한 결과를 종합하면, 심판의 볼 판정은 정립된 규칙 하에 일관되게 이루어지지 않고 경기 상황에 따라 불확실성이 크게 작용되고 있음을 시사한다. 이러한 심판의 판정은 경기결과에 충분히 영향을 미칠 수 있기 때문에 보완될 필요가 있다. 이와 같은 맥락에서 Parsons 등(2011)은 심판의 편견과 투수들이 겪는 잠재적인 차별에 대한 가능성은 선수들의 경기력은 물론 임금으로까지 확장될 수 있기 때문에, 프로야구 심판 판정 평가와 관련 연구는 발전된 형태로 지속되어야 한다고 주장하였다.

k-DB분석에서 활용하였던 2021 MLB 정규시즌 자료를 활용하여 최대우도법 기반의 로지스틱 회귀분석을 실시한 결과 $p(\text{ump-error}|\text{Home or Away})$, $p(\text{ump-error}|\text{Ball count})$ 등과 같은 비교적 간단한 분류문제에 대해서는 k-DB분석 결과와 1% 미만의 차이를 보였다. 그러나 홈/원정, 볼 카운트, 아웃카운트, 직구/변화구 구분, 투구된 공의 위치와 속도 등을 고려해야 하는 다소 복잡한 문제에 대해서 두 기법은 판정오류 발생 확률 추정치에 다소 큰 차이(3~11%)를 보였다. 이와 같은 확률 추정치의 차이는 로지스틱 회귀분석 결과에서 통계적으로 유의하지 않은 변수가 포함되었을 때 두드러지게 나타났다. strike와 ball의 구분은 스트라이크 존에서 벗어난 정도가 클수록 명확히 구분되는 ball에 해당하기 때문에, 투구된 공의 위치는 직관적으로 판정오류에 직접적인 영향을 미칠 것으로 예상되는 변수이다. 그럼에도 불구하고, 로지스틱 회귀분석 결과에서는 plate_x변수는 통계적 유의성을 나타내지 않았고, 이에 따라 새로운 조건이 추가되었음에도 추정된 판정오류 발생 확률은 큰 변화를 보이

지 않았다. 또한 아웃카운트를 고려한 결과에서도 마찬가지로, k-DB분석에서는 아웃카운트에 따라 판정오류 발생 확률에 다소 차이를 보였으나, 로지스틱 회귀분석에서는 모든 아웃카운트에서 고정된 28%의 결과를 나타냈다. 즉, 로지스틱 회귀분석에서는 통계적으로 유의하지 않은 변수가 포함된 조건에서는 실제로 독립변수가 변화할지라도 추정된 확률에는 변화가 없었으며, 신뢰구간 역시 비교적 크게 추정되었다.

로지스틱 회귀분석에서 회귀계수는 다른 변수들이 일정하게 고정되어 있는 상황을 가정하여, 특정 변수의 변화가 종속변수에 미치는 영향력을 나타낸다. 이러한 회귀계수는 통계적 유의성 검정 결과, 즉 유의확률(p-value)과 함께 제시된다. 통계적으로 유의한 변수의 회귀계수는 표본의 무작위성과 동일한 조건의 반복을 가정하여 산출된 결과이기 때문에 비교적 안정적인 결과로 받아들일 수 있을 것이다. 즉, Logit에서는 k-DB와 비교하여 상대적으로 엄격한 가정에 의해 정확한 회귀계수를 제시해준다. 따라서 독립변수 하나하나의 세부적인 영향력을 검증할 때에는 k-DB와 비교하여 보다 더 효율적으로 활용될 수 있을 것이다. 그러나 로지스틱 회귀모형에서 통계적으로 유의하지 않은 독립변수의 회귀계수는 큰 의미를 갖지 못하기 때문에 해석에 주의가 필요할 것이다. 로지스틱 회귀분석에서 통계적 유의성은 사건 발생의 사례 수, 설명변수들 간의 조합에 따라 다르게 나타날 수 있다. 특히 불확실성이라는 스포츠 경기의 본질적인 특성을 고려했을 때, 스포츠 현장에서 수집되는 다양한 자료 중 통계적 유의성을 나타내지 못하는 변수일지라도, 그 소수에 해당하는 일부의 사건이 실제 스포츠 현장에서 종속변수 사건 발생에 영향을 미칠 수 있다는 점이 간과되어서는 안 될 것이다. 또한 로지스틱 회귀분석에서는 투입된 독립변수들 간의 상관관계가 존재할 경우 다중공선성에 의해 추정된 계수의 신뢰성에 문제를 야기한다(Senaviratna & Cooray, 2019). 본 연구의 서론에서도 언급했듯이, 스포츠 상황에서 발생하는 다

양한 사건은 하나의 단일 사건이 결과로 이어지기 보다는 다양한 변수가 유기적으로 결합하여 하나의 결과로 나타나는 경우가 많기 때문에(황수웅, 양혜주, 황선환, 2022), 스포츠 빅데이터를 기반으로 선형회귀분석, 로지스틱 회귀분석 등을 수행하고자 한다면 연구자는 독립변수들 간의 상관관계를 어떻게 통제 또는 대처할지에 대한 고민이 필요할 것이다.

반면, k-DB는 하나의 독립변수가 최대 임의의 k개의 독립변수에 영향을 받는다는 가정을 전제로 한다. 또한 구조학습과 모수학습을 통해 추정된 모수, 즉, 연결된 노드들 간의 조건부 확률로부터 연구자가 원하는 다양한 조건을 직접 조합하여 사건 발생 확률을 추정할 수 있기 때문에 로지스틱 회귀분석과 비교하여 비교적 유연한 모형의 구축이 가능하다. 로지스틱 회귀분석과 달리 k-DB에서는 통계적 유의성과 관계없이 데이터 자체가 갖는 구조를 기반으로 노드들의 범주에 따라 변화하는 세부적인 확률 변화를 산출할 수 있다는 장점이 있다. 특히, 사건 발생 확률이 높은 조합을 탐색해 내야 하는 문제, 또는 cut-off를 찾아내야 하는 문제 등에 대해서는 k-DB가 효율적으로 활용될 수 있을 것으로 사료된다. 이와 관련하여 Haghghat 등(2013)의 연구에서는 스포츠 상황에서의 사건을 예측하는 데에 있어 베이지안 네트워크 분석의 장점으로 주어진 자료의 복잡성과 불확실성을 정량화하여 비교적 단순하고 효율적으로 설명할 수 있음을 보고한 바 있다. 또한 Stenling 등(2015)의 연구에서는 베이지안 접근은 계산과 해석에 이점(computational and interpretational advantages)을 갖기 때문에 스포츠 현상을 설명하는 연구에서 더욱 활발히 활용될 수 있음을 보고하였다.

그러나 k-DB 수행에 있어, k-fold cross validation을 위한 k의 숫자가 커지면 커질수록 logit과 비교하여 학습의 시간이 오래 소요된다는 점은 스포츠 현장의 신속성을 고려했을 때 단점이 될 수 있을 것이다. 또한 x 독립변수가 최대 임의의 'k'개의 변수에 영향을 받는다는 가정에서, 연구

자가 설정하는 k 에 따라 구조학습 결과가 달라질 수 있는 점은 모델링의 지나친 유연함이 단점으로 작용할 수 있다. 특히 k -DB에서는 구조학습과 모수학습 결과의 적합성을 평가할 수 있는 지수가 부재하기 때문에, 일관되지 못한 결과가 나타난다면 연구 결과를 다각도에서 재검토하고 해석에 주의를 기울일 필요가 있을 것이다. 이렇듯 본 학위논문의 ‘study 1’에서 활용한 두 기법은 해결해야 하는 문제가 어떠한 형태인지에 따라 그 효용성이 달라질 수 있음을 시사한다. 따라서 연구자들은 스포츠 빅데이터를 분석함에 있어 무조건적인 자료의 대입과 분석이 아닌, 자료가 갖는 특성에 대한 이해는 물론 각 기법들에 대한 충분한 이론적 고찰이 선행되어야 할 것이다.

Study 1의 마지막 단계에서는 ‘3balls-Xstrikes’, ‘Xballs-2srikes’와 같이 특정한 상황을 가정하여, 각 팀이 경험한 판정오류 빈도를 수집하여 경제적 파급효과를 추정하였다. 그 결과 MLB 팀들은 판정오류로 인해 FA가 아닌 주전급 선수의 1년 연봉 수준의 경제적 득실을 잠재적으로 경험하고 있음을 추론할 수 있다. 그러나 판정오류는 본 연구에서 가정한 상황 이외에도 실제로 빈번하게 발생하고 있고, 투구 하나에 경기의 흐름이 크게 좌우될 수 있는 야구 종목의 특성을 고려하면, 실제로 판정오류의 파급효과는 더욱 클 것으로 사료된다. 특히 주자유무, 아웃카운트 등을 종합적으로 고려했을 때의 판정오류의 파급효과는 더욱 커질 것으로 예상해볼 수 있다. Williams(2019)에 따르면 2018년 MLB에서 ‘Ball/Strike’ 판정 오류와 함께 종료된 경기는 총 55 경기에 해당한다. 이는 55경기의 결과는 현재와 다르게 기록될 수도 있었음을 의미한다(황수웅, 박일혁, 2020). 또한 본 연구의 결과에서는 제시하지 않았지만, ‘판정오류를 고려한 승리 추정 빈도’를 단순 합으로 구성하여 각 팀의 시즌 누적 승리 횟수에 반영할 경우 시즌마다 일부 팀의 순위가 바뀌어, 이로 인해 플레이오프(playoff) 진출 팀이 바뀌는 경우도 있는 것으로 나타났다. 이와

같은 결과는 후속연구를 통해 보다 심층적으로 다뤄질 필요가 있다.

본 학위논문의 서론에서도 언급했듯이, 스포츠는 제도화된 규칙 하에서 공정하게 신체적 탁월성을 겨루는 행위로 정의된다(안용규 외, 2011; Parry, 2019). 즉, 스포츠는 본질적으로 ‘공정한 경쟁’을 전제로 한다. 따라서 스포츠의 본질적 가치를 지키기 위한 연구자들의 이론적 논의의 지속될 필요가 있다. 이러한 맥락에서 본 연구는 종속변수가 범주형 변수에 해당하는 사건의 발생 확률을 추정함에 있어 그동안 체육/스포츠 분야에서 활발히 활용되지 못했던 베이지안 접근을 로지스틱 회귀분석 결과와 비교하여 이론적으로 논의하였다는 점에서 의의가 있을 것이다. 본 연구를 통해 체육/스포츠 분야에서 베이지안 추정의 활용이 보다 활발히 이루어질 수 있길 기대한다.

Ⅲ. Study 2

: AI를 활용한 프로야구 심판 판정 평가 모델 개발

1. 이론적 배경

1.1. 프로야구 심판 판정의 평가

프로야구에서 스트라이크 존과 심판 판정에 대한 논란은 프로야구가 시작된 이래로 꾸준히 이루어졌다. 그럼에도 불구하고, 2000년대 초반까지는 프로야구 경기 중 발생하는 다양한 자료를 실시간으로 처리할 수 있는 기술력이 뒷받침 되지 못했기 때문에, 심판의 판정을 실증적으로 분석하기에는 한계가 있었다. 따라서 2000년대 이전의 관련 문헌들에서는 프로야구 심판의 판정에 대하여 실증적 자료에 근거한 논의보다는 역사/철학적, 사회학적 논의가 주를 이루고 있다(McEvoy, 1993; Russell, 1999; Dworkin, & Staudohar, 2002; Russell, 2004). 그러나 2000년대 중반부터 정보통신기술의 비약적인 발달과 함께, 대용량의 자료를 실시간으로 처리할 수 있는 빅데이터 기술이 보편화되면서(Jin et al., 2015) 프로야구에서 역시 심판의 판정을 정량화하여 분석할 수 있는 환경이 마련되기 시작하였다.

Mills(2017)에 따르면 MLB에서 투구추적시스템이(PITCHf/x) 도입된 시점은 2009년이다. 이후 더 발전된 기술인 도플러 레이더 기반의 스탯캐스트(statcast) 시스템이 2015년 MLB에서 자리 잡게 되었다. 학계에서는 해당 시점을 기점으로 투구추적시스템, Statcast등의 기술에 의해 수집되는 자료에 근거하여 프로야구 심판의 판정을 정량화하여 평가하기 시작했다. 즉, 과거에는 기술력의 한계로 실시간 자료수집 조차 불가능했으나, 현대사회에서는 정보의 실시간 수집과 처리에 물리적/기술적 한계

가 적어졌기 때문에, 프로야구 현장에서 역시 ‘빅데이터’ 기반의 분석이 가능해졌다. Parsons 등(2011)의 연구에서는 PITCHf/x 자료를 활용하여 MLB에서 투수에 대한 심판의 인종적/민족적(racial/ethnic) 선호도가 판정에 영향을 미친다는 점을 통계적으로 검증하였다. 해당연구에서는 투수는 심판의 편견(bias)을 미리 예상하고 본인이 유리한 판정을 받을 수 있도록, 또는 심판들이 덜 주관적인 판단을 할 수 있도록 투구를 하는 것으로 나타났다. 연구진은 이러한 심판에 대한 편견과 투수들이 겪는 잠재적인 차별에 대한 가능성은 선수들의 경기력은 물론 임금으로 까지 확장될 수 있기 때문에, 프로야구 심판 판정 평가와 관련한 연구가 지속적으로 이루어져 할 필요가 있다고 보고하였다.

Green과 Daniels(2014)의 연구에서는 MLB 심판의 판정이 공이 투구된 위치 이외의 요소에 영향을 받는지 검증하고자 2009~2011년 MLB 정규 시즌에서 이루어진 판정을 분석하였다. 이 연구에 따르면 심판이 불판정을 내리는 기준이 되는 스트라이크 존은 볼 카운트에 따라 크게 달라지는 것으로 나타났다. 구체적으로, 스트라이크 판정이 2개가 선행된 상황에서는 심판의 스트라이크 존이 줄어들고, 볼(ball) 판정이 3개가 선행된 상황에서는 스트라이크 존이 확장되는 것을 실증적으로 검증하였다. 또한 2-strike 카운트에서 투구가 스트라이크 존의 모서리 부분에 투구되었을 때, 스탠딩 삼진(called strike out)을 선언할 확률은 다른 상황 대비 약 19%나 감소한다고 보고하였다. 연구진은 이와 같은 결과를 통해 MLB 심판의 판정은 공정하게 설정된 규칙에 대비하여 대부분 상황에 따라 영향을 받는 편향된 판정을 내리고 있고, 이는 프로야구에서 경기결과에 충분히 영향을 미칠 수 있기 때문에 보완될 필요가 있다고 보고하였다.

Kim과 King(2014)의 연구에서는 MLB 심판의 판정이 투수의 명성에 따라 차이를 보이는지 검증하고자 자료를 분석하여, matthew effect에

근거하여 심판 판정의 불공정성을 설명하고자 하였다. 이 연구에 따르면 투수의 평판이 'control pitcher'와 같이 리그 내에서 높은 곳에 위치할 때 심판의 strike 판정은 애매한 평판의 투수보다 더 관대하게 이루어지는 것으로 나타났다. 즉, MLB에서 명성이 높은 투수는 심판의 편견에 따라 상대적으로 스트라이크존이 넓게 적용되고 있으며, 이는 투수의 명성에 따라 개인의 경기력이 과대 혹은 과소평가될 수 있음을 보여주는 결과이다. 연구자는 이와 같은 결과에 대해 기존의 선행연구에서는 사회/심리학에 기반하여 심판의 판정을 평가하고자 하였으나, 경제/사회학적 관점에서도 심판의 불공정 판정을 설명할 수 있음을 주요한 의의로 제시하며, 심판의 편향은 불공정 판정으로 이어질 수 있기 때문에 프로야구에서 판정의 공정성 제고를 위한 노력이 지속적으로 이루어져야 한다고 보고하였다.

Chen 등(2016)의 연구에서는 개인의 의사결정에 있어서 도박사의 오류(gambler's fallacy)가 어떻게 작용하는지 실증적으로 검증하기 위해 하나의 사례로 MLB 심판의 판정을 활용하였다. 연구진은 프로야구에서 스트라이크 존은 정립된 규칙 하에서 구성되지만, 스트라이크 존의 경계에 있는 볼에 대해서는 심판의 주관이 반영된 '최적의 판단(best judgment)'이 필요하기 때문에 심판 판정에 대한 실증적인 분석이 필요하다고 보고하고 있다. 해당 연구에서는 PITCHf/x자료를 활용하여 심판의 직전 판정이 스트라이크였을 때, 다음 판정이 볼(ball)로 나타날 가능성이 더 높게 나타나는지 검증하였다. 그 결과 투구된 공이 스트라이크 존의 경계를 통과할지라도, 심판의 직전 판정에 따라 strike 판정을 받을 가능성은 약 0.9-1.3% 포인트 정도 감소한다고 보고하였다.

앞선 연구들에서 심판 판정에 대한 평가는 모두 MLB 사무국에서 제시하고 있는 스트라이크 존에 대한 기준과 투구추적시스템에서 가상으로 설정하고 있는 스트라이크존을 따르고 있다. 그러나 Hunter(2018)의 연

구에서는 심판이 인지하는 스트라이크 존은 심판마다 차이를 보이기 때문에, 심판을 평가함에 있어 스트라이크 존을 기준으로 설정하는 것 보다 일관성(consistency) 차원에서 평가가 이루어져야 한다고 강조하였다. Hunter는 관련 선행연구들 대부분은 MLB사무국에서 제시한 스트라이크 존에 근거하여 심판을 평가하고자 하였지만, 경기 중 심판이 MLB 규칙집(rule book)에서 규정하고 있는 스트라이크 존을 벗어난 볼을 스트라이크로 판정할지라도 판정이 일관되기만 한다면 공정성에 문제가 없음을 강조한다. Hunter에 따르면 실제로 프로야구 선수들은 심판의 판정이 일관되기만 한다면 그 일관성에 부합하는 가상의 새로운 스트라이크 존을 설정하여 시합에 참여한다고 보고하고 있다. Hunter는 심판의 일관성을 평가하는 지수를 총 네 가지로 구분하여 제시하였으며, 해당 지수들을 활용하여 상관분석, 주성분분석 등을 수행하여 심판 판정의 일관성을 평가한 결과 MLB 심판은 대체로 일관된 판정을 하고 있는 것으로 나타났다. 이와 같은 결과를 통해 Hunter는 프로야구 심판 판정에 대한 분석은 프로야구 선수의 경기력을 분석하는 것 이상으로 많은 의의가 있기 때문에 관련 연구들이 발전된 형태로 지속적으로 이루어져야 할 것임을 강조하였다.

Barbee(2020)의 연구에서는 투구된 공에 대한 결과를 예측하는 최적의 통계 모형을 찾기 위해 총 네 가지의 다른 통계기법을 적용하여 분석을 진행하였다. 구체적으로 2015~2018 시즌 MLB에서 이루어진 투구 자료를 활용하여, 로지스틱 회귀분석, 랜덤 포레스트(random forest), 베이지안 로지스틱 회귀분석, 인공신경망(artificial neural network) 분석을 수행하여 어떤 모델이 투구된 공에 대한 결과를 예측하는 데에 가장 적합한지 비교 분석하였다. 이 연구에서는 프로야구에서 스트라이크 존은 운영 주체에서 정의한 일정한 기준에 의해 설정되지만, 스트라이크 존의 높이는 타자의 키에 따라서 다르게 나타나기 때문에, 정량화하여 제시하기에는

한계가 있음을 지적하였다. 분석 결과 예측에 대한 정확도(accuracy)는 인공지능망 모형에서 83.92%로 가장 높게 나타났고, 민감도는 (sensitivity) 베이지안 로지스틱 회귀모형에서 99%, 특이도(specificity)는 인공지능망이 98%로 네 모형 중 가장 높게 나타났다. 해당 연구에서 관심 사건을 스트라이크 아웃으로 정의한 것을 고려했을 때, 가장 중요한 예측지수는 민감도라 할 수 있다. 따라서 해당 결과에 의하면, 베이지안 로지스틱 회귀분석이 투구된 공의 판정을 가장 잘 예측하는 것으로 나타났다. 그러나 연구자는 투구에 대한 예측 모형을 구축하고자 할 때, 학습에 활용한 자료의 수, 관심사건에 대한 균형자료 혹은 불균형자료 사용, 인공지능망에서 layer의 조합 등에 따라 분류 정확도, 민감도, 특이도 등이 달라질 수 있기 때문에 충분한 이론적 고찰을 통해 예측 모형을 구축해야 함을 강조하였다.

다양한 선행연구들의 결과를 종합하면, 프로야구 심판의 판정은 투구된 공의 속도, 회전수, 볼 카운트, 구종, 인종, 선수의 명성 등 매우 다양한 요인에 영향을 받고 있다. 즉, 야구에서 볼 판정은 정립된 규칙 하에서 이루어지고 있음에도 불구하고, 볼 판정을 선언하는 객체인 심판마다 스트라이크 존(strike zone)을 바라보는 기준에 차이를 보이고 있으며, 실시간으로 달라지는 경기 상황 정보에 따라서 부작위 편향(omission bias), 영향 회피(aversion) 등의 인간의 심리가 판정에 영향을 미치고 있는 것으로 밝혀졌다. 이렇듯 프로야구 심판의 판정은 원인과 결과의 관계가 규칙적으로 나타나기 보다는 다양한 변수가 상호작용하여 나타난 결과이기 때문에, 불확실성 요소에 대한 일정한 패턴을 선형관계만으로 설명하기에는 한계가 있다. 따라서 본 연구에서는 인간의 능력으로 일정한 규칙성을 찾기 어려운 문제 해결에 최적화되어 있는 딥러닝 기법을 활용하여 프로야구 심판의 판정을 예측하고 평가하기 위한 체계를 제안하고자 한다.

1.2. 기계학습과 딥러닝

한국정보통신기술협회의 정보통신용어사전에 따르면, 인공지능이란 “컴퓨터로 구현한 지능 또는 이와 관련한 전산학의 연구 분야”이다. 이를 보다 구체화해보면, 인공지능이란 인간의 인식, 판단, 추론, 문제해결 능력의 결과인 자연어 처리, 학습과 같은 인간의 두뇌 작용을 이해하는 학문을 통칭한다(고학수 외, 2021). 현재 인공지능 분야에서 널리 사용되는 방법론은 기계학습(machine learning)이다(McCarthy, 2007). 기계학습이란 주어진 자료에 대하여 컴퓨터가 일정한 규칙을 찾고 이에 대한 특정한 결과를 예측하는 방법을 의미한다(Jordan & Mitchell, 2015). 현대 사회에서 기계학습이 주목받는 이유는 우리 사회에서 발생하고 수집되는 빅데이터는 인간이 직접 일정한 패턴을 찾고 처리할 수 없을 정도의 방대한 양의 데이터이기 때문이다. 즉, 기계학습의 목적은 인간의 힘으로는 일정한 규칙을 정의하기 힘든 현상에 대하여 기계에 데이터를 학습시켜 자동화하는 것이다.

딥러닝(deep learning)은 다양한 기계학습 알고리즘 중 하나로 전통적인 기계학습에서 신경망(neural network)의 계층(layer)을 보다 깊고 복잡하게 발전시킨 형태에 해당한다. 전통적인 기계학습에서 입력변수의 선택은(feature selection) 사람에게 의해 이루어지며, 이는 모형의 적합도/효율성/예측력 등에 영향을 미치기 때문에 인간의 전문 지식이 요구되는 분야로 간주되어 왔다. 그러나 딥러닝에서는 주어진 문제해결을 위해 방대한 양의 자료와 사건에 대한 결과만 주어지면, 비선형으로 구성된 다중 계층이 모형을 스스로 구축하여 필요한 특징을 찾아낸다. 즉, 전통적인 기계학습에서는 복잡한 문제를 풀기 위해서 ‘특징 추출’과 ‘분류’라는 두 단계를 통해 문제를 해결하였다면, 딥러닝에서는 이와 같은 두 단계를 하나로 통합하여 해결하는 자동화 시스템으로 이해할 수 있다. 예를 들어, 프로야구에서 심판의 판정오류가 특정한 상황에 자주 발생한다

고 가정하면, 전통적인 기계학습에서는 투구된 공의 구속, 회전수, 볼 카운트 등 주어진 자료들에 대한 전처리 및 가공을 통해 판정 오류를 예측하는 데 적합한 특징들을 추출한 다음, 이를 학습 데이터로 활용하여 판정오류를 예측하는 체계였다. 그러나 딥러닝에서는 판정오류 발생하는 상황과 관련된 다양한 정보와 주어진 자료가 판정오류인지 아닌지 즉, 지도학습(supervised learning)을 위한 결과 자료만 있으면 자율적으로 학습한 후 일정한 규칙을 찾아내게 된다. 따라서 딥러닝은 주어진 자료의 차원이 아주 크고 복잡한 데이터의 경우에는 전처리 과정을 통해서 손실될 수도 있는 부분을 기계가 자동으로 추출해서 처리할 수 있다는 장점이 있다. 이는 전통적인 기계학습과 딥러닝의 가장 뚜렷한 차이점이라 할 수 있다.

상기한 바와 같이, 딥러닝 모형에서는 은닉층을 깊게 쌓아 입력변수들 간의 선형관계뿐만 아니라 다차원 비선형 관계를 고려할 수 있다. 이를 통해 인간이 일정한 규칙으로 정의하기 어려운 현상들에 대한 예측이 가능하다. 딥러닝 모형은 현재 이미지 분석뿐만 아니라 영상 속 객체인식(e.g., 자율주행, 운동 선수 움직임 인식 등), 음성인식, 번역 등 다양한 분야에서 특정 현상을 예측하고 분류하는데 활용되고 있으며, 모든 분야에서 높은 예측력을 나타내고 있다(고학수 외, 2021). 이와 같은 딥러닝 모형은 인공신경망(ANN; Artificial Neural Network)의 발전된 형태로 두 신경망은 은닉층(hidden layer)의 수에서 차이를 보인다.

1.3. 인공신경망(ANN; Artificial Neural Network)

인공신경망은 뇌의 신경세포망에서 정보처리 체계를 인공적으로 구현한 기계학습 모델로(Jain, Mao & Mohiuddin, 1996; 장병탁, 2017) 주어진 현상에 대한 분류 및 회귀문제를 해결하기 위한 함수를 추정하는 데에 특화된 성능을 보인다(Haykin, 2008). 따라서 체육/스포츠 관련 연구

에서도 인공신경망은 특정한 결과에 대한 예측을 위해 주로 활용되어 왔다(Aslan & Inceoglu, 2007; Young, Holland, & Weckman, 2008). 인공신경망의 기본적인 구조는 아래의 <그림 III-1>과 같다. 아래의 그림에서 원으로 표현된 것을 노드(node), 각각의 노드에 연결선(시냅스)을 포함한 구조를 뉴런(neuron)이라 칭한다. 인공신경망은 주어진 자료로부터 일정한 규칙을 찾는 학습을 한 후 결과를 출력한다. 여기서 학습이란 주어진 현상에 대한 분류와 예측을 위해 최적의 가중치를 추정하기 위한 연산을 반복하는 절차를 의미한다.

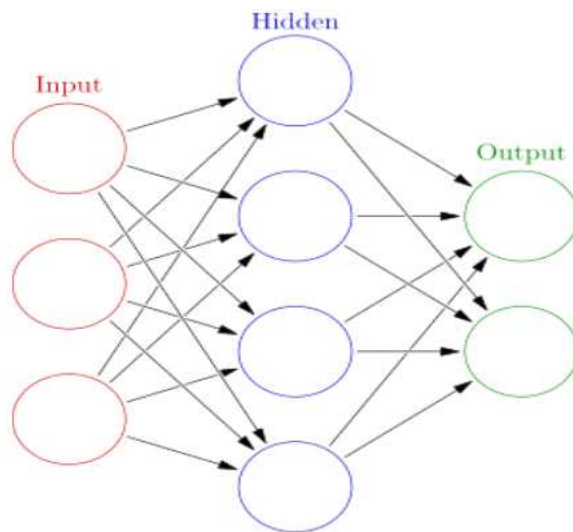


그림 III-1. 인공신경망의 기본 구조

<그림 III-1>에 제시한 인공신경망 구조는 여러 개의 단일 퍼셉트론을 여러 층으로 쌓은 다층퍼셉트론(MLP; Multi Layer Perceptron)으로 구성되어 있다. 퍼셉트론은 생물학적 신경세포의 신경 전달 체계를 수학적으로 모델링한 인공 뉴런이며 그 구조는 아래의 <그림 III-2>와 같다. 단일 퍼셉트론은 n개의 입력뉴런과 한 개의 출력 뉴런을 갖는다. 입력 뉴런의 활성화 값을 x_i 라 하고 시냅스 가중치를 w_i 라 할 때, 출력노드의 활성화 값은 가중치 곱의 총합으로 계산된다. 퍼셉트론은 입력의 총합이 임계값을 초과하면 '1'을 출력하고, 임계값 미만일 경우 '-1'을 출력한다.

즉, 여러 개 입력값을 받은 후 가중치를 곱해 임계값 초과 여부에 따라 1개의 값을 출력하는 개념을 ‘퍼셉트론(perceptron)’이라 하며, 이와 같은 퍼셉트론은 하나의 임계논리유닛(threshold logic unit)으로 구성된 신경망이다(장병탁, 2017; Rosenblatt, 1958).

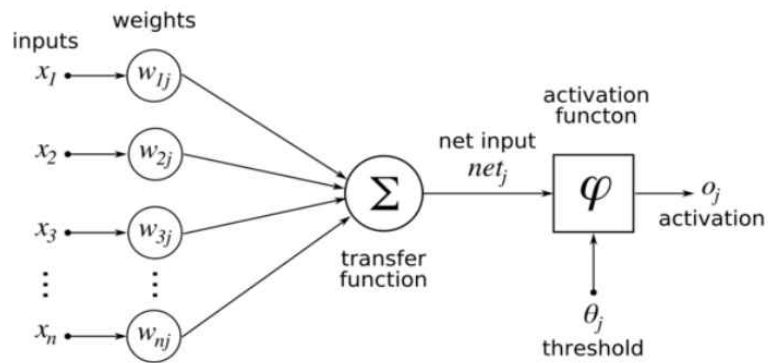


그림 III-2. 퍼셉트론의 구조(Rosenblatt, 1958)

다층퍼셉트론은 앞서 설명한 퍼셉트론을 여러 층으로 쌓은 다층신경망 구조를 의미한다. 다층신경망 모델은 <그림 III-1>과 같이 입력층(input layer), 은닉층(hidden layer), 출력층(output layer)으로 구성되어 있다. 다층퍼셉트론 신경망은 인접한 두 층의 모든 뉴런들 간에는 완전 연결된(fully connected) 구조를 갖으며, 입력층에서 출력층으로 가는 절차가 한 방향으로만 활성화된다(장병탁, 2017).

보다 직관적인 이해를 위해 인공신경망을 적용한 체육/스포츠 분야 선행연구를 통해 다층퍼셉트론 모형을 살펴보면 다음과 같다. Valero(2016)의 연구에서는 <그림 III-3>과 같이 다층퍼셉트론 모형을 구축하여 MLB 경기결과를 예측하고자 하였다. Valero는 입력층에 다양한 세이버매트릭스 지수를 및 경기 상황정보를 투입하여 1개의 은닉층을 갖는 다층퍼셉트론 모형을 구축하였다. 여기서 입력층, 은닉층, 출력층 각각의 뉴런의 수를 I, J, K로 표현하고, 입력유닛, 은닉유닛, 출력유닛의

기호를 i, j, k 라고 가정하자. 세이버매트릭스 지수 $RC(i=1)$ 의 출력 x_1 는 모든 은닉뉴런 j 와의 연결 가중치 W_{j1} 을 통해서 전달된다. 또한 은닉뉴런 j 의 출력은 모든 출력 뉴런 k (Win or Lose)의 입력으로 연결선 w_{jk} 를 통해 전달된다. 즉, k 번째 출력 뉴런의 활성화 값 $f_{win/lose}$ 의 계산과정을 수식으로 표현하면 다음과 같다.

$$f_k = f_{win/lose}(X, W) = \sigma\left(\sum_{j=0}^J W_{jh}^{(2)} \sigma\left(\sum_{i=0}^I W_{ji}^{(1)} x_i\right)\right)$$

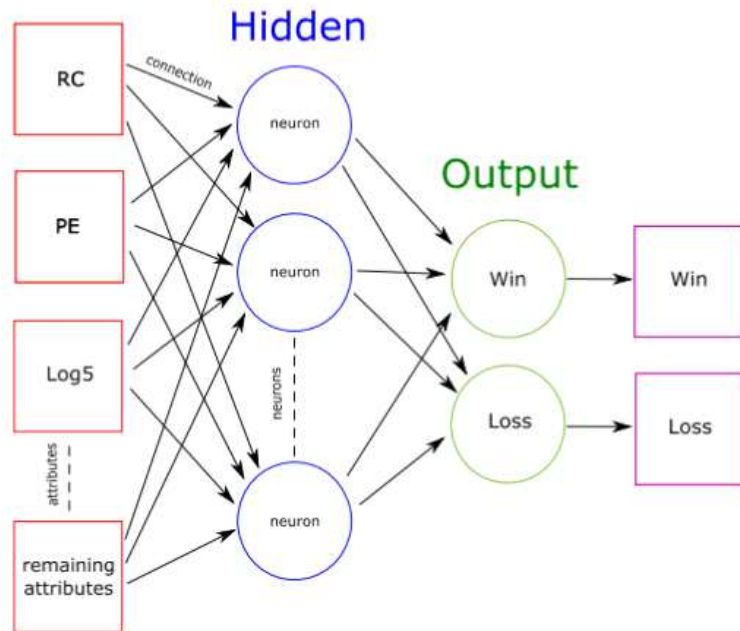


그림 III-3. 다층퍼셉트론 구성 예시(Valero, 2016)

$W^{(1)}$ 은 입력층과 은닉층을 연결하는 가중치 벡터이고 $W^{(2)}$ 는 은닉층과 출력층을 연결하는 가중치 벡터이다. 입력 뉴런에 입력값 벡터 x 가 주어지면 은닉층 뉴런들을 활성화하기 위해 w 를 활용하고, 이는 다시 출력층 뉴런들을 활성화하여 최종적으로 출력값 $f_{win/lose}$ 를 계산한다. 다

층퍼셉트론은 주어진 학습 데이터 집합에 대한 오류를 최소화하도록 가중치 벡터를 반복적으로 수정하여 최적화함으로써 학습한다. $f_{win/lose}$ 는 목표출력 $y_{win/lose}$ 와 비교하여 차이가 계산된다. 이 차이를 교정하기 위해 다층퍼셉트론에서는 가중치들을 변경해가며 최적화 과정을 거치게 된다. 이때 활용되는 기법이 오류역전파 알고리즘에 해당하며, 이에 대한 구체적인 내용은 Hecht-Nielsen(1992)의 연구를 참고하길 바란다.

1.4. 심층신경망(DNN; Deep Neural Network)

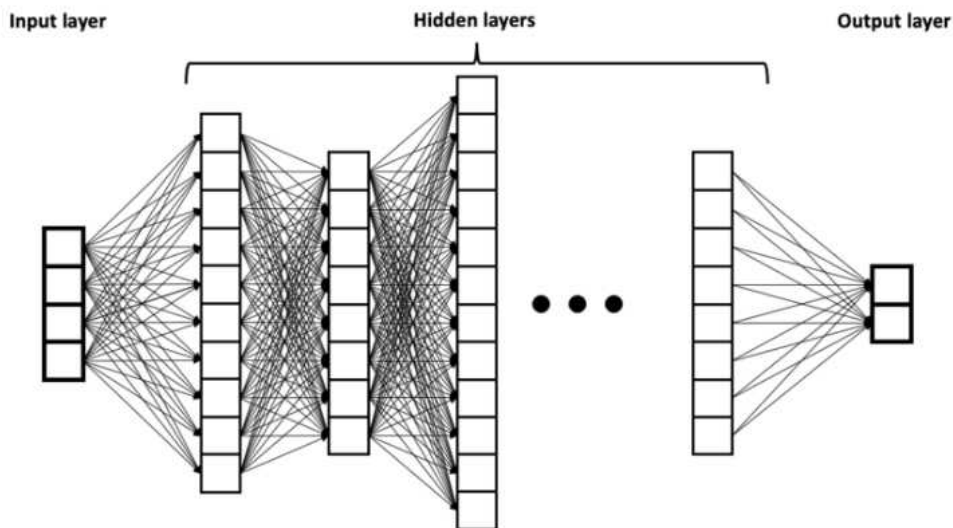


그림 III-4. 심층신경망(deep neural network) 구성 예시

심층신경망(DNN; Deep Neural Network)의 구조는 <그림 III-4>와 같다. DNN은 입력층과 출력층 사이에 여러 개의 은닉층이 있는 인공신경망 모델이다. 다양한 딥러닝 알고리즘(CNN, RNN, DBN 등)의 기본 구조가 되며, 각 딥러닝 알고리즘들은 DNN의 특별한 형태에 해당한다. 기존의 다층퍼셉트론 신경망의 구조가 단순 반복적인 완전연결 구조를 취하는 반면, 심층신경망 구조에서는 다양한 부분 연결 구조와 적층 방식을 통해 주어진 문제의 특성을 고려한다. 앞선 ANN 부분에서 설명했

듯이, input layer의 각 노드(위 그림에서는 네모로 표현)는 입력변수 중 하나의 변수 X_i 의 값이 입력되는 부분이고, 노드를 연결하는 선(시냅스)는 모형의 모수(parameter), 즉, 다음 층으로 연결되는 각각의 가중치를 나타낸다. DNN은 은닉층이 여러 개로 구성되어 있고, 각 층의 뉴런들이 다음 층으로 연결되는 과정이 반복적으로 이루어지면서 출력층에 이르게 되기 때문에, 각각 시냅스 가중치는 현재의 층에서 다음 층으로 가는 방법을 나타낸다.

DNN에 대한 보다 직관적인 이해를 위해 체육/스포츠 분야 연구에서 DNN을 적용한 실제 연구를 살펴보고자 한다. Rahman(2020)의 연구에서는 축구 경기 결과를 예측하기 위해 DNN을 활용하였다. 구체적으로 해당 연구에서는 데이터 발생 시간을 고려하여 분석을 진행하기 위해 DNN 구조를 기본으로 취하는 순환 신경망(recurrent neural network, RNN)모델 중 하나인 장단기 메모리 신경망(long short-term memory, LSTM)기법을 통해 딥러닝을 수행하였다. 기본적인 DNN 알고리즘은 시공간에 대한 순차적인 정보를 고려하지 못한다는 한계가 있으나, RNN에서는 뉴런들의 출력이 다시 입력으로 재배치되는 재귀적인 연결 구조를 갖는다. 특히, LSTM은 장기적 연관성을 효과적으로 포착하고 안정적인 학습이 가능하기 때문에 다양한 순서 데이터 모델링에 널리 사용되고 있다(Greff et al. 2016). Rahman(2020)은 입력변수로 홈/원정 구분, 선수, 포지션, 골의 유형, 카드 유형(경고, 퇴장), 패널티 종류, 교체 등 총 10개의 시간성을 갖는 변수를 입력변수로 설정하였으며 은닉층은 총 두개로 각각 10개의 노드와 5개의 노드로 구성하였다. 목표 출력은 승리한 경기의 경우 $y=1$, otherwise(무 또는 패) $y=0$ 으로 설정하였다. 이를 통해 입력변수에 RNN-LSTM을 적용하여 딥러닝을 수행했을 때 경기결과를 잘 예측하는지 검증하였다. Rahman은 최종적으로 학습된 모형에 2018년 FIFA 월드컵 조별 예선 경기 자료를 투입하여 학습된 DNN 모형이 새로

운 데이터를 잘 예측하는지 검증한 결과 약 63.3%의 예측 정확도를 기록하였다. 저자는 조별 예선보다 16강, 8강, 4강 등 본선에서는 예측력이 낮게 나타났다고 보고하였다. 이와 같은 결과는 스포츠 상황에서 발생하는 불확실성이 예선보다는 본선과 같이 한 경기로 모든 것이 결정되는 경우에는 더욱 크게 나타날 수 있음을 나타내는 결과라 보고하고 있다. 이와 같은 한계점에도 불구하고, Rahman은 체육/스포츠 분야에서 DNN의 활용이 매우 저조함을 지적하며, 해당연구에서 제시한 연구설계를 통해 축구뿐만 아니라 다양한 종목을 위한 분석에서 딥러닝을 활용할 수 있음을 연구의 의의로 제시하고 있다.

상기한 내용을 종합하면, DNN 구조는 CNN, RNN, DBN 등 다양한 딥러닝 모형의 가장 기본적인 구조가 되는 것으로, ANN과의 가장 큰 차이점은 은닉층의 개수에 해당한다. 다수의 은닉층이 있는 신경망 모델을 DNN이라 총칭하며, DNN은 다수의 은닉층에서 깊은 신경망 구조를 취하기 때문에, 보다 복잡한 분류 문제에 있어 정교한 예측이 가능하다는 것이다.

1.5. 체육/스포츠 분야 기계학습, 딥러닝의 활용

Valero(2016)의 연구에서는 프로야구 승부를 예측하기 위하여 세이버 매트릭스 지수를 비롯한 야구 기록들을 입력변수로 설정하여 분류 기반의 데이터 마이닝 기법을 통해 분석을 진행하였다. 구체적으로 해당연구에서는 데이터 마이닝 기법 중 어떤 알고리즘이 프로야구 승부를 잘 예측할 수 있는지 검증하고자, Lazy learning, 인공신경망(Artificial neural network), 의사결정나무(Decision trees), Support vector machines(SVMs) 기법을 적용하여 비교 분석하였다. 해당 연구에서는 WEKA(waikato environment for knowledge analysis)를 통해 feature selection 절차를 수행하여 총 60개의 변수 중 15가지의 입력변수를 설정하였으며, 해당

절차를 통해 프로야구 승부 예측을 위해 중요한 변수는 홈팀 유무, 피타 고리안 기대 승률 등으로 나타났다. 이후 연구자는 총 15개의 입력변수를 각각의 데이터 마이닝 기법에 적용하여 승부를 예측한 결과 SVM의 예측정확도가 가장 높게 나타났다. 그러나 4개의 데이터 마이닝 기법 모두 약 50~60%대의 예측력을 보여 예측 성능이 크게 좋지는 않은 것을 확인할 수 있었다. 연구자는 이와 같은 결과에 대하여 스포츠 경기는 매우 복잡한 다양한 상황이 결합하여 승부가 결정되기 때문에 이러한 스포츠의 특성을 충분히 고려할 수 있는 연구설계가 필요하며, 입력변수들 역시 PITCHf/x 등과 같은 해당 연구에서 포함하지 못한 다양한 형태의 자료가 투입될 경우 승부 예측 정확도가 높아질 것이라 보고하였다.

Park 등(2017)의 연구에서는 딥러닝을 통해 스포츠 선수 평가 모델을 제안하였다. 연구진은 기존의 스포츠 선수 평가 모델은 경기 중 발생한 통계자료에 대한 의존도가 높아 통계로 기록되지 않는 선수의 질적 영향력을 반영할 수 없다는 점을 한계점으로 지적하였다. 이에 대한 대안으로 스포츠 경기 상황이 간접적으로 반영될 수 있는 뉴스기사와 경기 중 발생하는 정량화된 자료를 함께 딥러닝 모형의 입력변수로 투입하여 선수 평가 모델을 제안하였다. 구체적으로 연구진은 프로야구 선수를 특정 기간의 경기력에 따라 긍정적 또는 부정적으로 분류하고, 같은 기간동안 기사화된 자료들에 대하여 사전 분류를 반영하여, 경기력을 나타내는 양적자료와 함께 심층신경망 (Deep neural network, DNN)을 구축하였다. 구축된 DNN 모형은 SVM, Random forest, 다층 퍼셉트론 등의 알고리즘과 예측력 비교가 이루어졌다. 그 결과 DNN을 활용한 모형에서 예측력이 약 70~80%대로 가장 좋은 결과를 나타났다. 연구진은 기사 분류에 대한 정확성이 높아지고, 선수의 경기력을 나타내는 변수의 수가 더 많아지면 DNN모형의 예측력이 더 높게 나타날 수 있을 것이라 보고하고 있다.

Mehrasa 등(2018)은 팀 스포츠에서 선수들 간의 복잡한 상호작용을

분석하여 경기 상황을 예측할 수 있는 모델을 구축하기 위해 딥러닝을 수행하였다. 연구진은 스포츠 분석(sports analytics)에서 선수 또는 팀의 움직임 자료를 활용할 경우, 겉으로 드러나지 않는 특정 팀 또는 선수만의 시공간 활용 패턴을 분류할 수 있으며, 이를 통해 경기 중 발생하는 사건을 예측할 수 있을 것이라고 보고하였다. 구체적으로 해당 연구에서는 팀 또는 선수의 움직임 자료를 분석하여 경기 중 발생하는 패스, 슛 등의 사건을 예측하고, 각 움직임 자료를 통해 팀을 특정화하여 예측하고자 하였다. 이를 위해 1 dimensional convolutional neural network(1d-CNN) 알고리즘을 적용하였으며, 그 결과 아이스하키에서 패스와 볼소유(carry) 동작은 약 80%의 분류 정확률을 보였으며, NBA에서 팀을 분류하는 과제에 대해서는 Phoenix Suns 팀을 제외한 대부분에서 약 90%이상의 분류 정확률을 기록하였다. 연구진은 이와 같은 결과를 바탕으로 딥러닝을 활용한 선수의 움직임 추적 데이터 분석은 팀 단위 스포츠 종목에서 선수 개인의 플레이 특성뿐만 아니라 팀 단위 경기력을 설명하고 예측하는 데에 매우 효과적인 결과를 도출해낼 수 있을 것이라 보고하고 있다.

Kim과 Choi(2020)의 연구에서는 프로야구 영상 자료에 딥러닝을 적용하여 프로야구 자동 해설 생성 인공지능을 제안하였다. 해당 연구에서 제안한 인공지능 모델은 실시간으로 프로야구 경기 영상이 주어지면, 온톨로지(ontology)에 기반한 4가지 딥러닝 모형이 통합적으로 작용하여 누가 어떤 위치에서 무엇을 하고 있는지(who is doing what in which area of the field)에 대한 정보와, 행위에 대한 예상 결과(what results are expected)를 제시한다. 이와 같은 통합된 딥러닝 모형을 구성하는 4가지의 딥러닝 모형은 각각 사건(장면) 분류, 객체(선수) 인식, 수행 동작 인식, 투구 결과 인식(a scene classifier, player detector, motion recognizer, and pitching result recognizer)에 최적화 되어 있는 모형으

로 구성되어 있다. 구체적으로, 사건(장면)분류 알고리즘에서는 프로야구 경기 장면을 13개의 범주로 분류하고, 객체 인식 알고리즘에서는 플레이어의 이미지와 좌표, 레이블을 탐지해 낸다. 동작 인식 알고리즘은 현재의 경기 상황에서 핵심이 되는 포지션을 찾아내고, 투구 결과 인식 알고리즘에서는 현재 투구된 공에 대한 판정 결과를 인식한다. 이를 통해 연구진은 “타자 오늘 2번째 타석입니다.”, “투수 타자를 향해 힘껏 공을 던집니다.”, “타자 친 공, 중견수가 플라이 처리할 수 있을까요?” 등의 자동 해설을 생성하였다. 스포츠 해설은 선수의 행동 묘사뿐만 아니라 주어진 장면에 대해 구체적이고 다양한 정보를 제공해야 한다. 따라서 그동안 스포츠 해설가는 스포츠에 대한 높은 이해와 깊은 배경지식을 갖춘 일부 전문가만이 수행할 수 있는 영역으로 여겨져 왔다. 그러나 본 연구에서 제시하고 있는 자동 해설 인공지능 모델은 스포츠 해설이 단순히 인간만의 영역이 아닐 수 있음을 시사한다. 즉, 이와 같은 연구 결과는 스포츠 분야에서 인공지능의 활용이 우리가 그동안 상상하지 못했던 차원의 다양한 부가가치를 창출해 낼 수 있음을 보여주는 사례라 할 수 있다.

Song과 Marin(2021)의 연구에서는 optimized convolutional neural network(OCNN)을 통해 스포츠 상황에서 발생하는 부상 진단을 위한 인공지능 모델을 제안하였다. 연구진은 스포츠 의학 분야의 데이터는 과거와 비교하여 자료의 용량, 형태에서 큰 변화가 있으며 ANN, DNN 등과 같은 전통적인 알고리즘으로는 부상을 예측하고 진단하기에 한계가 있음을 지적하였다. 이에 따라 연구진은 OCNN기법을 적용하여 부상 진단에 대한 예측 정확도를 전통적인 알고리즘들과 비교하였다. OCNN에서는 신경망을 physical layer, device layer, control layer로 총 세개의 층(layer)으로 설정하였으며, 학습 결과 부상 진단에 대한 정확도는 OCNN 알고리즘이 약 80%로 가장 높게 나타났고, artificial neural network, convolutional neural network, machine learning 기법들은 50-70%의 예측 정확도를 나

타냈다. 이와 같은 결과는 특정 현상에 대한 분류와 정확한 예측에 특화된 기법인 인공지능 기법을 활용할지라도, 데이터의 특성을 충분히 고려하여 그 특성에 적합한 알고리즘을 적용할 필요가 있음을 시사한다.

Joshi 등(2020)의 연구에서는 스포츠 관련 이미지를 통해 종목을 분류하기 위해 'Inception V3'과 신경망 분석을 진행하였다. 연구진은 현대사회에서 콘텐츠 기반의 이미지 분류는 웹상에서 효율적인 검색 결과를 제시하는 데에 매우 중요한 측면임을 강조하며, 주어진 스포츠 이미지의 주변 환경 정보에 따라 스포츠 종목을 분류하기 위한 체계/framework를 제안하였다. 해당 연구에서는 럭비, 테니스, 크리켓, 농구, 배구, 배드민턴 6개의 종목을 구분 범주로 설정하여, 스포츠 이미지만으로 스포츠 종목을 구분해 낼 수 있는 신경망을 구축하였다. 연구진은 이미지의 특징을 추출하는 feature extraction절차에서는 'Inception V3'를 적용하였고, 인공신경망의 활성화함수는 ReLU(rectified linear unit)를 활용하였다. 해당 연구에서는 딥러닝 모형의 분류 정확성 수준을 상대적으로 검증하기 위해 random forest, SVM, KNN과 같은 기계학습기법의 분류 정확도와 비교하였다. 그 결과 인공신경망을 적용한 딥러닝 기법의 분류 정확률이 96.64%로 매우 높게 나타났고 이외의 기계학습 기법들은 90-95%사이를 기록하였다. 특히 신경망에서는 배드민턴과 농구종목에 대한 분류 정확도가 100%로 나타나, 해당 종목들은 이미지를 통해 종목의 특성이 잘 드러나고 있음을 나타냈다. 이와 같은 연구 결과는 이미지가 갖고 있는 정보만으로 스포츠 종목을 효과적으로 분류할 수 있음을 실증적으로 보여주고 있다. 해당 연구에서 제안한 스포츠 종목분류를 위한 이미지 분석에서 나아가 시간과 공간 정보가 포함된 이미지 분석을 수행할 수 있다면, 이는 스포츠 경기 분석의 관점에서 큰 의의가 있을 것이다.

Tirdad 등(2021)의 연구에서는 청소년들이 스포츠 상황에서 경험하는 뇌진탕의 심각성을 언급하며, 스포츠와 관련된 뇌진탕에 대한 대중의 정

서(sentiment)를 식별하기 위해 딥러닝 기법을 활용하였다. 구체적으로 연구진은 스포츠 뇌진탕과 관련된 트위터 게시물들을 수집하여 CNN, LSTM(Long Short-Term Memory) Network, 양방향 LSTM 분석을 수행하였다. 연구진은 트위터 게시물 학습을 통해 구축된 최적의 인공지능 모델을 2018년 FIFA world cup에서 발생한 뇌진탕 관련 게시물 약 8만 2천건에 적용하여 관련한 대중들의 반응이 긍정(positive)/부정(negative)/중도(neutral) 중 어느 범주에 해당하는지 분류하였다. 그 결과 긍정 게시물은 39.13%, 부정 게시물은 26.03%, 중도 게시물은 34.85%에 해당하는 것으로 나타났다. 연구진은 이와 같은 결과를 스포츠 참여자가 경험할 수 있는 뇌진탕에 대하여 사회적 비용을 줄이기 위한 미래 전략 수립의 기초자료로 활용될 수 있음을 보고하고 있다. 스포츠 상황에서 발생하는 비정형 데이터들은 과거에는 물리적인 기술력의 한계로 자료수집 조차 쉽지 않았고, 자료가 수집된다 할지라도 분석할 수 있는 기법은 매우 한정적이었다. 그러나 빅데이터 기술의 발달, 나아가 인공지능 기술의 발전은 정형자료 뿐만 아니라 텍스트와 같은 비정형 자료에 대한 대규모 분석을 가능하게 하였다. 해당 연구는 이와 같은 맥락에서 그 의의가 있을 것이며, 뇌진탕뿐만 아니라 스포츠 상황에서 사회적 비용을 초래하는 다양한 사건들을 소셜 빅데이터, 인공지능 기법을 통해 분석하여 스포츠 위험 요인을 예방하는 차원에 적용된다면 매우 유용하게 활용될 수 있을 것이다.

딥러닝을 위해 개발된 초기의 알고리즘들은 이미지 인식 및 분류가 주를 이루었고, 이후 보다 발전된 방향으로 영상 속 객체 인식 알고리즘들이 개발되었다. 상기한 선행연구고찰 결과들을 살펴보면, 체육/스포츠 자료를 활용하여 딥러닝을 적용한 연구들 역시 이미지 인식, 동영상 속 객체 인식이 다수를 차지하고 있다. 일부 연구에서 스포츠 상황에서 발생하는 자료를 활용하여 딥러닝을 적용하였으나, 스포츠 자료가 갖는 특

성을 충분히 고려한 연구는 매우 일부에 그치고 있다. 특히 Song과 Marin(2021)의 연구결과와 같이, 다양한 딥러닝 기법을 적용했음에도 불구하고 예측력은 알고리즘 간 큰 차이를 보였다. 본 연구의 서론에서 기술한 바와 같이, 실시간으로 방대한 양의 데이터를 활용하여 특정한 결과를 도출해낸다 할지라도, 스포츠 자료가 갖는 특성을 충분히 고려할 수 있는 연구설계와 알고리즘의 적용이 뒷받침되지 못한다면 그 결과의 활용은 매우 제한적일 것이다. 따라서 본 연구에서는 스포츠가 갖는 본질적인 특성을 고려하여 인공지능 학습을 하기 위해 고려해야 하는 요인들을 살펴보고, 실제 MLB 경기기록과 심판 판정 자료를 활용하여 인공지능을 활용하여 심판의 판정을 평가하고자 한다. 본 연구 수행을 위한 연구 방법은 다음과 같다.

2. 연구 방법

2.1. 자료 수집 및 처리 절차

본 학위논문의 ‘Study 2’는 MLB 경기 자료를 활용하여 프로야구 심판의 ‘Ball/Strike’ 판정을 예측하는 인공지능 모델을 개발하는 것을 목적으로 한다. 이를 통해 심판 개인의 ‘Ball/Strike’ 판정이 리그 전체를 대표하는 인공지능 심판(AI UMPIRE)의 판정과 차이를 보이는 판정을 선별해내고, 이를 활용하여 심판 개인의 판정을 평가하는 체계를 구축하고자 한다.

본 연구에서 구축하고자 하는 딥러닝 모형의 목표 출력은 ‘ $y=1(\text{strike})$ ’, ‘ $y=0(\text{ball})$ ’로 설정하였으며, 입력변수의 선정은 수집 가능한 자료 중 선행 연구 고찰 결과와 study 1에서 심판의 판정에 영향을 미치는 요인으로 실증적으로 검증된 요인들을 종합적으로 고려하여 설정하였다. 연구 수행을 위한 자료는 MLB에서 Statcast data를 공식적으로 관리하는

‘baseball savant’와 자체적으로 MLB자료를 관리하고 있는 ‘baseball-reference’ 웹 페이지를 통해 수집하였다. 구체적으로 baseball savant에서는 2019년 MLB에서 타자의 스윙 없이 strike로 판정된 ‘called strike’ 판정 119,155개와 ball 판정 243,579개 총 362,734개의 판정 자료를 수집하였다. 수집된 자료에는 심판의 볼 판정 결과를 비롯하여 투구된 공의 구종, 속도, 공의 움직임, 홈/원정 팀 구분, 투구된 공의 위치, 볼 카운트, 타자에 따른 스트라이크 존의 높이 등 다양한 자료가 포함되어 있다. baseball-reference에서는 ‘Ball/Strike’판정을 수행한 본루심(home plate umpire)의 정보를 ‘bash script’를 통해 scrapping하였으며, 이후 baseball savant에서 수집한 자료와 통합하였다. 자료전처리, 기술 통계, 자료의 시각화 등의 분석 절차는 Software R 4.1.3version을 통해 진행하였으며, 딥러닝은 Python 3.10.6을 jupyter notebook을 통해 진행하였다. 입력변수와 출력변수에 대한 설명은 <표 III-1>과 같다.

표 III-1. Study 2 딥러닝을 위한 자료의 구성

변수		설명
출력층	umpiring	2019년 MLB 정규시즌 경기에서 called strike 또는 ball 판정을 받은 투구
입력층	pitch_type	직구/변화구 구분
	pfx_x	투구된 공의 수평적 움직임
	pfx_z	투구된 공의 수직적 움직임
	plate_x	투구된 공이 홈플레이트를 통과할 때 수평적 위치
	plate_z	투구된 공이 홈플레이트를 통과할 때 수직적 위치
	effective speed	투구된 공의 속도
	spin	투구된 공의 회전수
	sz_top	타자에 따른 스트라이크 존의 상단
	sz_bot	타자에 따른 스트라이크 존의 하단
	release_pos_x	수평적인 공의 투구 위치
	release_pos_z	수직적인 공의 투구 위치
	release_pos_y	포수의 관점에서 투수가 공을 투구하는 위치
	release_speed	투수가 공을 투구하는 시점의 속도
	pitch_number	현재 상대하고 있는 타자를 상대로 투구된 공의 빈도
	innings	공이 투구되는 시점의 inning
	outs	공이 투구되는 시점의 out 카운트
	strikes	공이 투구되는 시점의 strike 카운트
	balls	공이 투구되는 시점의 ball 카운트
	b_right	타자 좌타, 우타 구분
	p_right	투수 좌투, 우투 구분

2.2. 'Ball/Strike' 판정 인공지능 심판(AI UMPIRE) 개발

본 연구에서는 RNN, CNN 등의 딥러닝 알고리즘의 기초가 되는 DNN을 통해 MLB 경기자료를 학습하여 'AI UMPIRE'을 개발하였다. 구체적으로 optimizer, 입력층의 구성, 활성화 함수, dropout의 포함 여부 등의 세부적인 사항을 조정해가며 예측 성능을 비교한 뒤 최적의 모형을 'AI UMPIRE' 선정하였다. 인공지능 학습에 활용한 자료는 무선추출을 통해 60:20:20의 비율로 나누어 60%는 training set으로 활용하였으며, 나머지 40%에 대해서는 20% validation set, 20%는 test set으로 활용하였다. 원자료를 딥러닝 구조에 맞는 형태로 가공하기 위해 아웃라이어의 영향을 최소화하는 'RobustScaler'를 통해 자료를 변환하였고, 학습이 이루어지는 층의 활성화 함수는 'ReLU'로 설정하였다. 출력층의 활성화 함수는 'softmax', 'ReLU', 'Sigmoid' 중 손실함수와의 조합을 고려하여 예측 정확도가 가장 높게 나타난 모델을 활용하였다. optimizer 역시 'rmsprop', 'adam', 'SGD'를 모두를 활용하여 학습 후 예측성능이 가장 높게 나타난 모델을 선택하였다. 마찬가지로, 학습이 이루어지는 층의 구성 역시 32*3, 64*3, 128*3, 256*3 등을 기본으로 dropout을 각 층에 추가/제거해가며 예측 성능이 가장 높게 나타나는 구조를 확인하였으며, 최종 모형에서는 앞선 실험학습 결과를 바탕으로 모델의 일반화 (generalization) 성능을 고려하여 모델을 선정하였다. 본 연구에서 구축한 최종 DNN 모형 구조는 <표 III-2>와 같다.

표 III-2. 심층신경망 최종 모형의 학습 구조

Model	Deep Neural Network
Layers	5 Layers(Units: 256, 128, 64, 32, 1)
Activation functions	ReLU, Sigmoid (last output layer)
Dropout	3 layers(0.5, 0.4, 0.3)
Loss function	Binary crossentropy
Optimizer	SGD
Input data ¹⁾	release_speed, release_pos_x, release_pos_z, release_pos_y, strikes, balls, pfx_x, pfx_z, plate_x, plate_z, outs_when_up, inning, sz_top, sz_bot, effective_speed, release_spin_rate, pitch_number, fastball, b_right, p_right
Output data	2019 MLB Regular season Homeplate umpires' Ball/Strike call (0=Ball, 1=Strike)
Training set	214310 (60%)
Validation set	71435 (20%)
Test set	71435 (20%)

2.3. AI를 활용한 심판 판정 평가

2019년 MLB 정규시즌 경기에서는 총 93명의 본루심(home-plate umpire)이 1인당 최대 5068개에서 최소 729개의 ‘Ball/Strike’ 판정을 수행하였다. 본 연구에서는 93명의 심판 중 볼 판정 개수가 4000개 이상에 해당하는 50명을 선정하여 ‘AI umpire’의 판정의 판정을 기준으로 심판 개인의 판정을 평가하고자 한다. 구체적인 절차는 다음과 같다.

DNN을 통해 개발된 ‘AI UMPIRE’의 Ball/Strike 예측 정확도가 90% 이상의 높은 수준으로 나타난다면, 해당 인공지능 모델은 리그 전체 심판의 판정을 고려하여 판정을 진행하는 가상의 모델로 간주할 수 있을 것이다. 해당 인공지능 모델에서는 각 투구에 대한 분류 확률을 제시해

¹⁾ detailed explanation of each variable is described in Table III-1.

주는 데, 그 확률이 0에 가까우면 'Ball', 1에 가까우면 'Strike'를 의미한다. 즉, 'AI UMPIRE'는 0.90이상 또는 0.1이하의 확률 값을 갖는 투구들에 대해서는 매우 강한 확신을 갖고 'Ball/Strike' 판정을 수행한다. 반면, 0.4~0.6의 확률 값을 갖는 투구는 'AI UMPIRE' 역시 판정에 대한 확신이 부족함을 의미한다. 또한 본 연구에서 수집한 MLB 심판의 판정 자료는 'Strike=1', 'Ball=0'으로 코딩되어 있는 더미변수이다. 따라서 AI UMPIRE가 예측한 확률에 더미변수 1 또는 0을 빼준 차이 값을 구한 뒤, 그 값에 절댓값을 취해주면 ±의 방향성이 없어지기 때문에 값이 크게 나타나는 투구일수록 AI심판과 실제 인간의 판정 결과가 큰 차이를 보인다는 것을 의미한다.

예를 들어, AI 심판이 특정 투구에 대하여 0.1의 예측 확률(AI prob.)을 산출하여 해당 투구를 BALL로 강하게 확신하고 있을 때, 실제 인간의 판정 결과(umpiring)가 Strike(더미변수:1)라면 $ABS(AI\ prob - umpiring)$, 즉, 0.1-1에 대한 절댓값은 0.9로 1에 가까운 값을 나타낸다. 반대로 AI심판이 특정 투구에 대하여 0.9의 확률을 산출하여 Strike라고 강하게 확신하고 있을 때, 실제 인간의 판정 결과가 Ball(더미변수 :0)이라면 $ABS(AI\ prob - umpiring)$ 의 값은 $0.9 - 0$ 에 의해 0.9가 된다. 즉, 앞선 두 예시 모두 $ABS(AI\ prob - umpiring)$ 은 0.9로 같은 값을 갖게 되며, 해당 값은 1에 가까울수록 AI심판과 실제 심판의 판정이 극명하게 대비되는 판정임을 의미한다. 따라서 다른 심판들과 비교하여, 특정 심판의 판정에서 $ABS(AI\ prob - umpiring)$ 가 0.9이상의 값을 갖는 투구의 빈도가 높거나, 또는 0.9이상의 값을 갖는 판정들을 시각화한 자료에서 일정한 패턴이 보인다면, 해당 심판의 판정은 다른 심판들과 비교하여 bias가 있는 판정을 수행하고 있는 것으로 추론할 수 있을 것이다. 본 연구에서는 AI를 활용하여 심판 개인의 판정을 평가하는 절차를 보여주는 '하나의 예시'로 $ABS(AI\ prob - umpiring)$ 가 0.9이상의 값을 갖는 투구의 빈도가 낮은 세

명의 심판(Top 3)과 높은 세 명의 심판(Bottom 3)을 선정하여 판정 자료를 시각화하고 해당 판정들의 경기 상황을 분석하였다. 이를 통해 AI를 심판 판정의 준거로 활용하여 심판 개인의 판정을 평가하고, 나아가 경기 상황에 따라 편향된 판정을 선별해내어 이를 분석하였다. 이와 같이 본 연구에서는 인공지능을 활용하여 프로야구 심판의 판정을 평가하는 체계를 제안하고, 이를 통해 심판 판정의 일관성/공정성에 대하여 논의하였다.

3. 연구 결과

3.1. 인공지능 심판(AI UMPIRE) 모델 개발

본 연구는 인공지능 심판을 개발하는 것뿐만 아니라, 스포츠 빅데이터를 활용한 인공지능 모델 개발에 있어 어떠한 과정과 세부적인 선택 사항의 조정이 최적의 결과를 도출해낼 수 있는지에 대한 부분에도 중점을 두고 있다. 따라서 본 연구에서는 입력층의 구성을 비롯하여 optimizer, 손실 함수와 활성화 함수, Dropout의 포함 여부 등의 세부적인 사항을 조정해 가며 다양한 실험학습을 진행하였고, 개발된 인공지능 모델들의 예측 성능을 비교하여 최적의 모형을 구축하여 ‘AI UMPIRE’를 개발하였다. 이의 구체적인 결과는 다음과 같다.

1) 입력층의 구성과 Optimizer의 선정

입력층의 구성과 Optimizer에 따른 예측 정확도 비교를 위해 손실 함수는 ‘binary crossentropy’, 활성화함수는 ‘ReLU’, ‘sigmoid(출력층)’, epochs=50으로 설정하여 학습을 진행하였다. 구체적인 결과는 <표 III-3>과 같다. training set에 대한 예측 정확도는 adam optimizer와 3*256’조합에서 최대 96.69%를 기록하였으나, test set에서는 오히려 예

측정 정확도가 91.6%로 감소하는 것으로 나타났다. 딥러닝의 목적은 최적화의 관점에서 단순히 주어진 데이터를 잘 설명/예측하는 모델을 개발하는 것이 아니라, 주어진 데이터로부터 새로운 데이터를 잘 예측하는 모형, 즉 일반화(generalization) 성능이 좋은 모델을 개발하는 것이다(Dahl, Sainath, & Hinton, 2013). 따라서 test set에서의 정확도에 집중할 필요가 있다. test set에 대해서는 ‘SGD’ optimizer의 128*3의 입력층을 갖는 모델의 예측 정확도가 93.12%로 가장 높은 값을 기록하였다. 인공지능 모델의 예측 정확도는 optimizer의 변화에 따라 큰 차이는 없었지만, 평균적인 정확도는 SGD optimizer에서 가장 높게 나타났다. 따라서 입력층의 구성과 Optimizer에 대한 다양한 조합 중 예측 정확도가 가장 높게 나타난 layer 128*3, ‘SGD’ optimizer 조합을 바탕으로 손실함수와 출력층의 활성화 함수에 대한 추가적인 실험학습을 진행하였다.

표 III-3. Optimizer와 Layer 구성에 대한 실험학습 결과

	Optimizer	Layers/Units			
		3*32	3*64	3*128	3*256
training set (max. accuracy)	rmsprop	93.37%	93.57%	94.36%	95.53%
	adam	93.41%	93.66%	94.31%	<u>96.69%</u>
	SGD	93.18%	93.23%	93.35%	93.44%
test set	rmsprop	92.97%	92.93%	92.52%	91.95%
	adam	93.15%	93.03%	92.60%	91.60%
	SGD	92.97%	93.03%	<u>93.12%</u>	93.08%

2) 손실 함수(loss function)와 활성화 함수(activation function)

표 III-4. 손실함수와 활성화 함수에 대한 실험학습 결과

loss function	activation function of output layer		
	softmax	ReLU	sigmoid
mean_squared_error	32.89%	92.59%	92.93%
binary_crossentropy	32.89%	67.11%	<u>93.12%</u>
categorical_crossentropy	67.11%	67.11%	67.11%

손실 함수와 활성화 함수의 변화에 따른 예측 정확도 비교를 위해 입력층의 구성은 128*3, optimizer는 'SGD', epochs=50의 조건에서 학습을 진행하였다. 위 표에 제시한 바와 같이, 마지막 출력층의 활성화 함수를 'softmax'로 사용할 경우 모든 투구를 스트라이크로 또는 볼로 예측하여 학습이 제대로 이루어지지 않는 것으로 나타났다. 'ReLU'활성화 함수를 출력층에 활용할 경우 'mean_squared_error'를 손실함수로 활용했을 때 92.59% 예측 정확도를 보였으나, 이외의 손실함수에 대해서는 67.11%로 모든 투구에 대하여 스트라이크로 예측하고 있었다. 마지막으로 'sigmoid'함수를 활용한 모델에서는 'mean_squared_error'에서 92.93%, 'binary_crossentropy'에서 93.12%를 기록하였다. 따라서 손실함수와 출력층의 활성화 함수에 대한 다양한 조합 중 예측 정확도가 가장 높게 나타난 binary_crossentropy-sigmoid 조합을 바탕으로 Dropout에 대한 추가적인 실험학습을 진행하였다.

3) Dropout

딥러닝에서 규제(regularization)는 모델의 과적합(overfitting)을 방지하기 위한 방법이다. 즉, 모델을 더 간단하고 효율적으로 만들어 일반화 성능을 높이는 것을 규제라하며, Dropout은 다양한 규제기법 중 가장 효과적이고 널리 사용되는 방법이다(Dahl, Sainath, & Hinton, 2013). Dropout은 인공지능 모델이 학습하는 동안 무작위로 층의 출력을 0으로 만들어 일부 벡터의 특성을 제외시키는 역할을 함으로써 과적합을 방지한다. 인공지능 모델의 layer는 128*3, 활성화 함수는 입력층에서 'ReLU', 출력층에서 'Sigmoid' 함수를 활용하였으며, optimizer는 'SGD', 손실함수는 'Binary_crossentropy'를 사용하였다. Dropout이 없는 Original 모형에서는 128개로 구성된 layer를 3층으로 그대로 활용하였으며, Dropout이 포함된 모형에서는 <그림 III-5>와 같이 층을 구성하였다. Dropout 포함 유무에 따른 인공지능 모델의 예측 성능 비교 결과는 <표 III-5>, <그림 III-5>와 같다.

```
model128dr = keras.Sequential([
    layers.Dense(128, activation="relu"),
    layers.Dropout(.5),
    layers.Dense(128, activation="relu"),
    layers.Dropout(.4),
    layers.Dense(128, activation="relu"),
    layers.Dense(1, activation="sigmoid")
])

model128dr.compile(
    loss = 'binary_crossentropy',
    optimizer = 'SGD',
    metrics = ['accuracy'])
```

그림 III-5. Dropout이 포함된 모형의 구조

표 III-5. Dropout에 대한 실험학습 결과

	Dropout	Accuracy
epochs=50	No dropout (Original)	93.03%
	2 dropouts (0.5, 0.4)	<u>93.18%</u>
epochs=100	No dropout (Original)	93.08%
	2 dropouts (0.5, 0.4)	<u>93.19%</u>
Layers=128*3 activation functions: 'ReLU', 'Sigmoid (output layer)', optimizer: 'SGD' loss function: 'binary crossentropy'		

<표 III-4>에 제시된 바와 같이, dropout layer가 포함되지 않은 original model의 예측정확도는 epochs=50에서 93.03%, epochs=100에서 93.08%를 기록하였다. 한편, dropout layer가 포함된 모형에서는 epochs=50에서 93.18%, epochs=100에서 93.19%를 기록하였다. 즉, test set에 대한 인공지능 모델의 예측 정확도는 epochs=50과 100 모두에서 dropout이 포함된 모형에서 소폭 높게 나타났다. <그림 III-6>은 epoch=100에 대한 인공지능 모델의 검증 손실(Loss)과 예측 정확도(accuracy)를 나타낸 것이다. Loss에 대해서는 좌측의 original model 그래프에서는 training loss와 validation loss가 겹친 형태로 나타나 거의 일치하고 있으나, 우측의 dropout-regularized model에서는 training loss와 비교하여 validation loss가 낮게 나타나고 있음을 확인할 수 있다. 모델의 accuracy 역시 좌측의 original model에서는 training accuracy와 validation accuracy가 거의 일치하고 있는 반면, 우측의 dropout-regularized model에서는 training accuracy보다 validation accuracy가 뚜렷하게 높게 나타나고 있다. 이러한 결과는 dropout이 인공지능 모델의 과적합을 방지하고 새롭게 주어진 자료에 대한 예측 성능을 의미하는 일반화(generalization) 성능을 높이는 데에 효과적으로 작용하고 있음을 보여주는 결과이다.

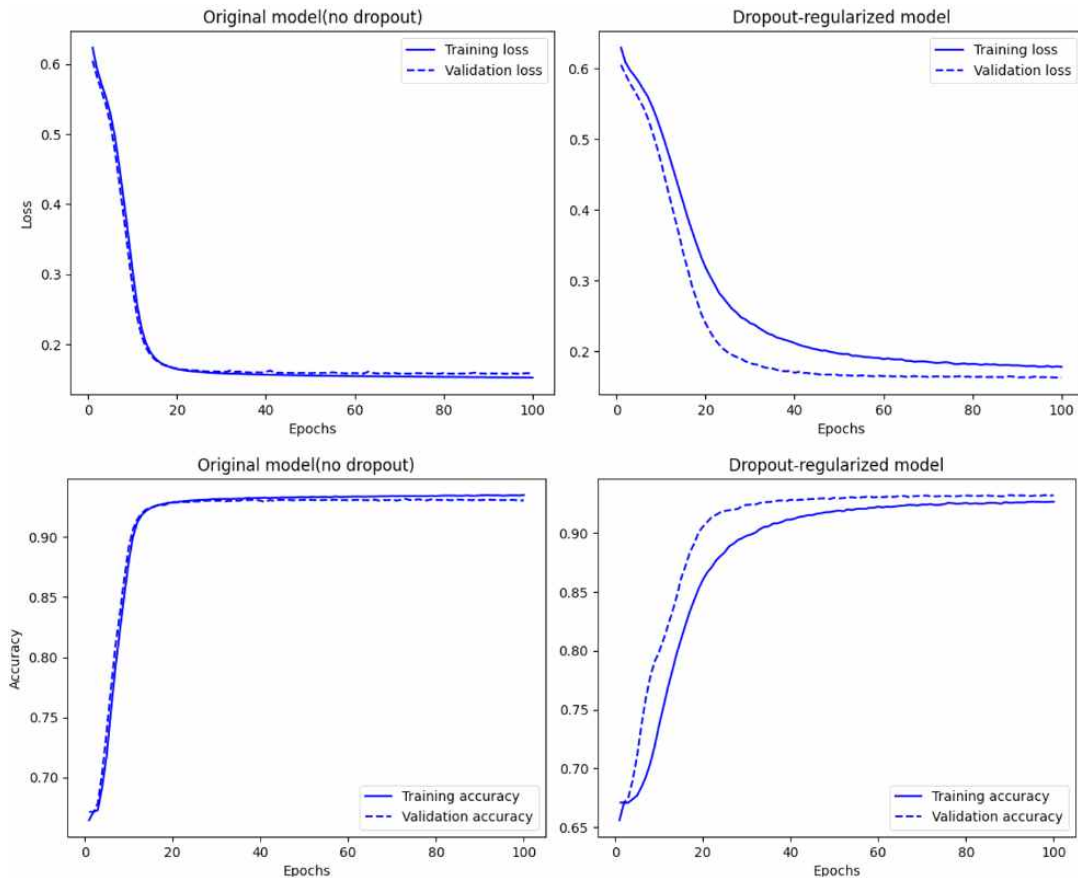


그림 III-6. Dropout에 따른 검증 손실과 예측 정확도

4) 'AI UMPIRE'(인공지능 심판) 개발

앞선 절차에서는 입력층의 구성을 비롯하여 Optimizer, 손실함수, 활성화 함수, Dropout에 대한 세부적인 조절을 통해 스포츠 빅데이터를 활용한 인공지능 학습에 있어, 어떤 구조와 선택적 조합이 높은 예측 성능을 보이는지 살펴보았다. 본 절에서는 앞선 결과와 관련 선행연구 및 교재들을 참고하여 최종모형 'AI UMPIRE(인공지능 심판)'을 제안하였다. 본 연구에서 최종적으로 개발한 인공지능의 구조 및 세부 정보는 <표 III-6>과 같다. 5개의 Hidden layer에 3개의 Dropout을 갖는 구조를 취하며, Dropout에 따른 parameter 개수를 효율적으로 구축하기 위해

표 III-6. AI UMPIRE 학습 구조

Model	Deep Neural Network
Layers	5 Layers(Units: 256, 128, 64, 32, 1)
Activation functions	ReLU, Sigmoid (last output layer)
Dropout	3 layers(0.5, 0.4, 0.3)
Loss function	Binary crossentropy
Optimizer	SGD
Input data ¹⁾	release_speed, release_pos_x, release_pos_z, release_pos_y, strikes, balls, pfx_x, pfx_z, plate_x, plate_z, outs_when_up, inning, sz_top, sz_bot, effective_speed, release_spin_rate, pitch_number, fastball, b_right, p_right
Output data	2019 MLB Regular season Homeplate umpires' Ball/Strike call (0=Ball, 1=Strike)
Training set	214310 (60%)
Validation set	71435 (20%)
Test set	71435 (20%)

output layer에 가까워질수록 각 층을 구성하는 노드의 수는 256, 128, 64, 32와 같이 줄어드는 구조로 설정하였다. 활성화 함수는 ReLU와 Sigmoid(출력층)을 활용하였으며, 본 연구는 ‘Ball/Strike’를 구분하는 이진 분류(binary classification) 문제를 다루고 있기 때문에 손실함수는 ‘Binary crossentropy’를 활용하였다. 또한 optimizer는 ‘SGD’를 활용하였다. 각 층을 구성하는 구조와 순서를 요약한 내용은 <그림 III-7>과 같다.

¹⁾ Detailed explanation of each variable is described in Table III-1.

```

Model: "sequential"
-----
Layer (type)                Output Shape                Param #
-----
dense (Dense)                (None, 256)                 5376
dropout (Dropout)           (None, 256)                 0
dense_1 (Dense)              (None, 128)                 32896
dropout_1 (Dropout)         (None, 128)                 0
dense_2 (Dense)              (None, 64)                  8256
dropout_2 (Dropout)         (None, 64)                  0
dense_3 (Dense)              (None, 32)                  2080
dense_4 (Dense)              (None, 1)                   33
-----
Total params: 48,641
Trainable params: 48,641
Non-trainable params: 0
-----
None

```

그림 III-7. AI UMPIRE 모형 구조 요약

개발된 ‘AI UMPIRE’의 검증 손실과 예측 정확도는 <그림 III-8>과 같다. 해당 인공지능 모델은 Dropout layer가 포함되어 있기 때문에 빨간색 실선으로 표시한 Training loss보다 파란색 점선으로 표시한 Validation loss가 작게 나타났다. 또한 Accuracy 역시 Training accuracy보다 Validation Accuracy에서 높게 나타나 dropout이 인공지능 모델의 일반화에 효과적으로 작용하고 있음을 검증하였다. test set에 대한 정확도는 93.18%, 손실은 .1598로 나타났다.

본 연구에서 제안한 ‘AI UMPIRE’는 2019 MLB에서 이루어진 모든 심판의 ‘Ball/Called Strike’ 판정 자료를 활용하여 개발되었다. 또한 test set을 적용했을 때, 예측 정확도가 93%이상의 높은 성능을 보이고 있다. 이러한 결과는 ‘AI UMPIRE’가 100개의 투구 중 93개 이상은 2019년 실

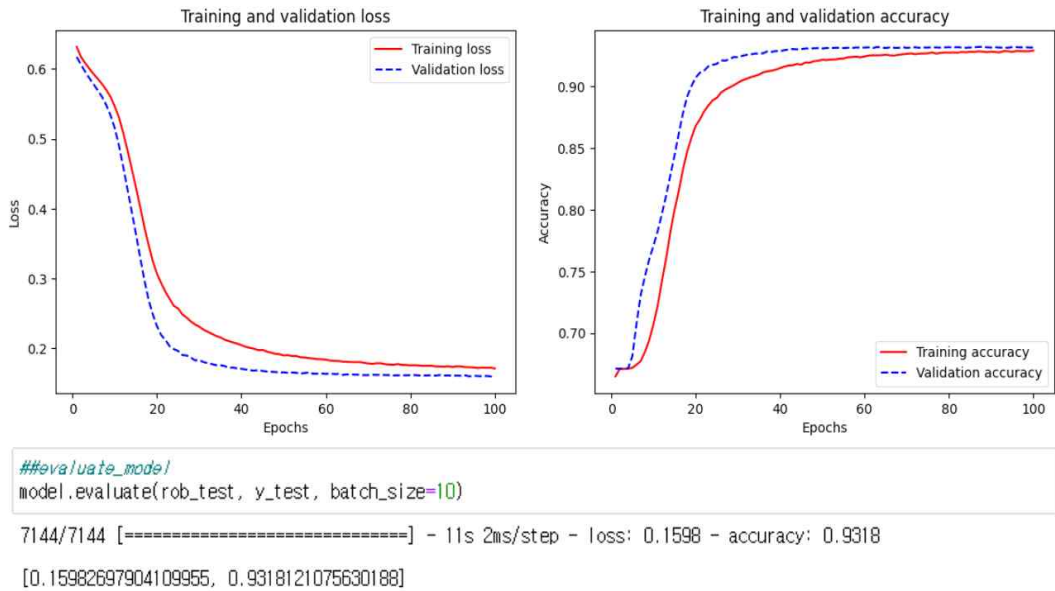


그림 III-8. AI UMPIRE 학습 결과

제 심판 판정과 일치함을 의미한다. 2019년 MLB에서 이루어진 ‘Ball/Strike’ 판정이 30만개가 넘는 것을 감안하면, 실제 심판의 판정과 인공지능 심판의 판정이 일치하는 숫자는 매우 큰 빈도를 보일 것이다. 즉, 본 연구에서 제안한 ‘AI UMPIRE’는 단순히 심판 한 명의 판정을 예측하는 인공지능이 아닌, 2019년 MLB 심판 전체를 대표하는 인공지능 심판으로 간주할 수 있을 것이다. 이와 같은 논리에서, 실제 심판의 판정과 AI UMPIRE의 판정을 비교하여 크게 대비되는 판정을 수행하는 심판 또는 투구를 찾아낼 수 있다면, 이는 심판 개인의 판정을 평가하는데 큰 도움이 될 것이다. 특히 AI UMPIRE가 매우 강하게 확신을 갖고 판정한 공이라면 더욱 그러할 것이다. 본 연구에서는 이러한 인공지능 심판의 ‘Ball/Strike’판정에 대한 확신을 바탕으로 심판 개인의 판정을 평가하였다.

3.2. AI를 활용한 심판 판정 평가

1) 심판 개인 단위 판정 평가

본 연구에서 개발한 인공지능 심판은 'Ball/Strike'판정을 수행할 때, 각 판정에 대한 확신을 의미하는 확률을 제시해준다. 해당 확률은 1에 가까울수록 Strike에 대한 강한 확신을 나타내고, 반대로 0에 가까울수록 Ball에 대한 강한 확신을 나타낸다. 즉, AI UMPIRE는 .9이상 또는 .1이하의 확률을 갖는 투구들에 대해서는 매우 강한 확신을 갖고 판정을 수행했음을 의미한다. 반면 .4~.6의 확률을 갖는 투구는 인공지능 역시 판정에 대한 확신이 부족함을 의미한다. 또한 본 연구에서 수집한 MLB 심판의 판정 자료는 'Strike=1', 'Ball=0'으로 코딩된 더미변수이다. 따라서 AI UMPIRE가 예측한 확률에 더미변수 1 또는 0을 빼준 차이 값을 구한 뒤, 그 값에 절댓값을 취해주면 ±의 방향성이 없어지기 때문에 값이 크게 나타나는 투구일수록 AI심판과 실제 인간의 판정 결과가 큰 차이를 보인다는 것을 의미한다. 이를 수식으로 나타내면 다음과 같다.

$$ABS[AI \text{ prob.} - \text{umpiring}(0,1)]^1$$

본 연구에서는 2019년 Ball 또는 Called Strike판정을 수행한 심판에 대한 자료를 bash-script를 통해 수집하였고, 총 93명 중 투구 판정 횟수가 4000개 이상을 기록한 심판 50명의 자료를 test set으로 구분한 뒤, 'AI UMPIRE'가 해당 test set에 대한 판정을 수행하도록 하였다. 이를 통해 심판 50명의 실제 판정 자료와 AI UMPIRE의 판정 결과를 비교하여, 심판 개인의 판정을 평가하였다. 이에 대한 구체적인 결과는 <표 III-7>과 같다.

1) AI prob: AI UMPIRE가 수행한 'Ball/Strike' 분류 확률을 의미함. 1에 가까울수록 Strike에 대한 강한 확신을 나타냄. 반대로 AI prob.이 0에 가까울수록 Ball에 대한 강한 확신을 나타냄.

표 III-7. AI를 활용한 심판 판정 평가

Rank	freq.	ratio(%)	Rank	freq.	ratio(%)
1	30	0.61%	26	50	1.21%
2	32	0.64%	27	54	1.22%
3	30	0.74%	28	51	1.22%
4	37	0.89%	29	52	1.23%
5	35	0.82%	30	56	1.26%
6	35	0.83%	31	55	1.31%
7	36	0.88%	32	57	1.31%
8	37	0.91%	33	55	1.34%
9	40	0.97%	34	60	1.34%
10	40	1.0%	35	57	1.38%
11	40	1.0%	36	60	1.38%
12	42	1.0%	37	64	1.42%
13	46	1.0%	38	67	1.44%
14	45	1.04%	39	69	1.48%
15	50	1.06%	40	63	1.52%
16	46	1.06%	41	70	1.61%
17	47	1.08%	42	67	1.65%
18	50	1.12%	43	83	1.78%
19	47	1.12%	44	81	1.95%
20	47	1.14%	45	91	1.99%
21	49	1.15%	46	86	2.0%
22	51	1.20%	47	88	2.15%
23	50	1.20%	48	97	2.40%
24	52	1.20%	49	122	2.74%
25	53	1.21%	50	137	2.80%

<표 III-7>은 50명의 심판 판정을 AI가 강한 확신을 갖고 판정을 수행했음에도 불구하고 실제 심판의 판정은 반대로 이루어진 판정의 비율 (the ratio of the cases with the probability difference > 0.9, 이하 ‘ratio of. > 0.9’로 표기)을 기준으로 심판을 서열화하여 정리한 내용이다. 해당 표에서 ‘freq’는 AI UMPIRE가 예측한 확률에 더미변수 1 또는 0을 빼준 차이 값을 구한 뒤, 그 값에 절댓값을 취했을 때, 0.9이상을 기록하는 공의 빈도를 의미한다. 즉, AI가 매우 강한 확신을 갖고 ‘Ball 또는 Strike’ 판정을 수행하였음에도, 실제 심판의 판정은 반대로 이루어진 판정의 빈도를 의미한다. 또한 ‘ratio’는 해당 심판의 Ball/Called Strike에 대한 전체 판정 중 앞선 빈도가 차지하는 비율을 의미한다. 예를 들어, A심판이 2019년 수행한 Ball/Called Strike 판정을 5000개로 가정했을 때, ‘freq’에 해당하는 판정의 빈도가 50개라면, A 심판의 ‘ratio of >.9’는 1%가 된다. 본 연구에서는 AI를 통해 심판을 평가하는 과정을 보다 직관적으로 제시하기 위해 ‘ratio of. > 0.9’를 기준으로 심판을 서열화한 뒤 상위 3명과 하위 3명의 판정 자료를 시각화하여 비교 분석하였다.

2) 심판 판정 자료 시각화 및 경기 상황 분석

앞선 <표 III-7>의 Rank 1-3 심판(top 3)의 자료와 Rank 48-50 심판 (Bottom 3)의 자료를 시각화한 결과는 아래의 <그림 III-9~14>와 같다. <그림 III-9>는 ‘ABS(prob-umpiring)’ 값이 0.1보다 작은 값을 나타내는 판정들에 대한 시각화 자료이다. ‘ABS(prob-umpiring)’이 0.1보다 작다는 것은 AI UMPIRE가 강한 확신을 갖고 판정을 수행한 결과와 실제 심판의 판정 결과가 일치하는 볼에 해당한다. 따라서 <그림 III-9, 10>에서는 Rank top 3 그룹과 bottom 3 그룹 간 패턴에 큰 차이를 보이지 않았다. 또한 ‘ABS(prob-umpiring)’ 값이 .4~.6의 범위에 해당하는 판정들은 AI 또한 판정에 대한 강한 확신을 갖지 못하는 판정에 해당한다.

해당 투구들에 대해서는 top 3 그룹에서는 볼과 스트라이크 판정이 스트라이크 존의 경계를 형성하고 있는 형태로 나타나고 있다. Bottom 3에서도 대부분의 판정이 스트라이크 존의 경계에서 볼 또는 스트라이크 판정이 혼재되고 있음을 볼 수 있다. 그러나 ‘ABS(prob-umpiring)’가 0.9보다 큰 값을 갖는 판정을 시각화한 <그림 III-11, 12>에서는 top 3 그룹과 Bottom 3 그룹 간 다소 뚜렷한 차이를 보였다. top 3 심판에 대한 ‘.9<ABS(prob-umpiring)’의 시각화 자료는 비교적 무선적으로 분포하고 있는 것을 확인할 수 있다. 그러나 bottom 3 심판들에 대한 시각화 자료에서는 한쪽으로 치우쳐진 패턴을 보이고 있다.

심판의 판정이 ‘.9<ABS(prob-umpiring)’에 포함되었다는 것은, AI가 강한 확신을 갖고 판정을 수행하였음에도 실제 심판의 판정은 반대의 판정으로 이루어졌음을 의미한다. 본 연구에서 개발한 AI UMPIRE가 2019년 MLB에서 이뤄진 모든 투수의 Ball 또는 Called Strike 판정 자료를 학습하여 개발되었고, test set에 대한 정확도가 93% 이상임을 감안하면, ‘ABS(prob-umpiring)’가 .9보다 큰 값을 나타내는 판정이 많은 심판은 실제로 판정을 수행함에 있어 특정한 bias가 작용할 수 있음을 시사한다. 따라서 본 연구에서는 ‘.9<ABS(prob-umpiring)’에 해당하는 판정들에 중점을 두고, 해당 판정들을 따로 구분하여 경기 상황에 대한 기술통계 분석을 진행하였다. 그 결과는 <표 III-8, 9>와 같다.

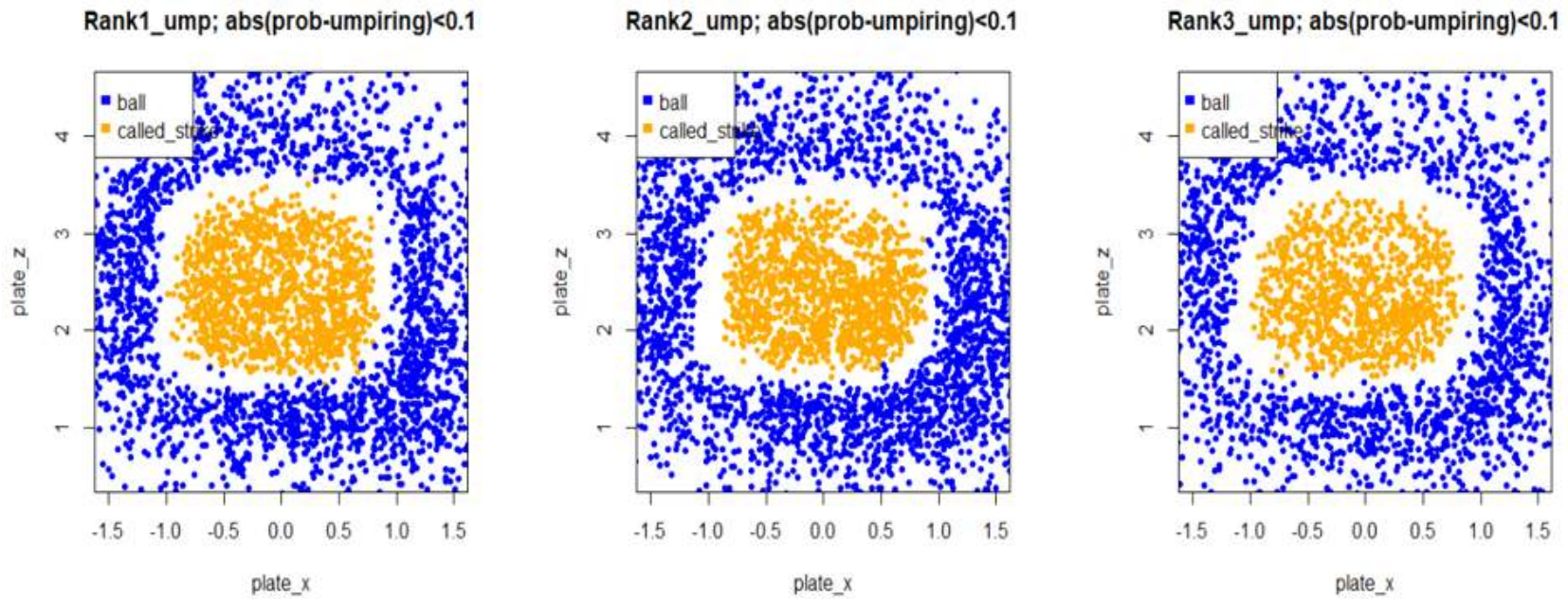
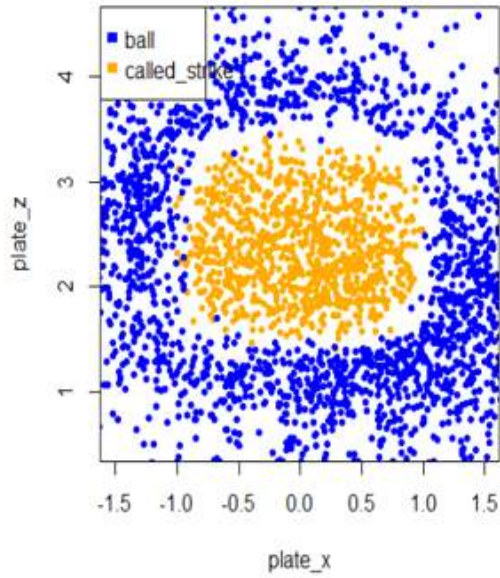
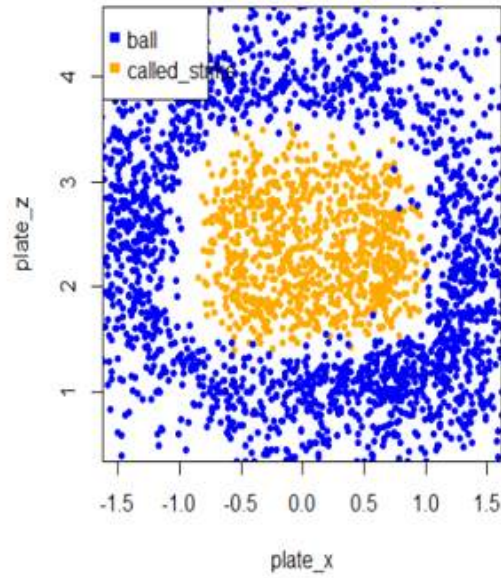


그림 III-9. Rank 1-3 판정 자료[ABS(prob.-umpiring)<0.1]

Rank48_ump; abs(prob-umpiring)<0.1



Rank49_ump; abs(prob-umpiring)<0.1



Rank50_ump; abs(prob-umpiring)<0.1

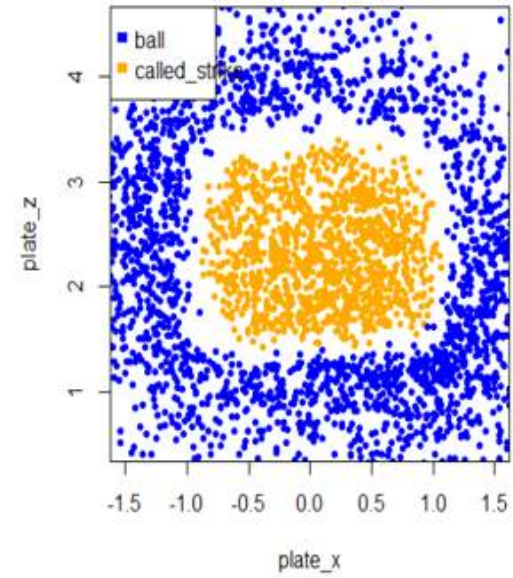


그림 III-10. Rank 48-50 판정 자료[ABS(prob.-umpiring)<0.1]

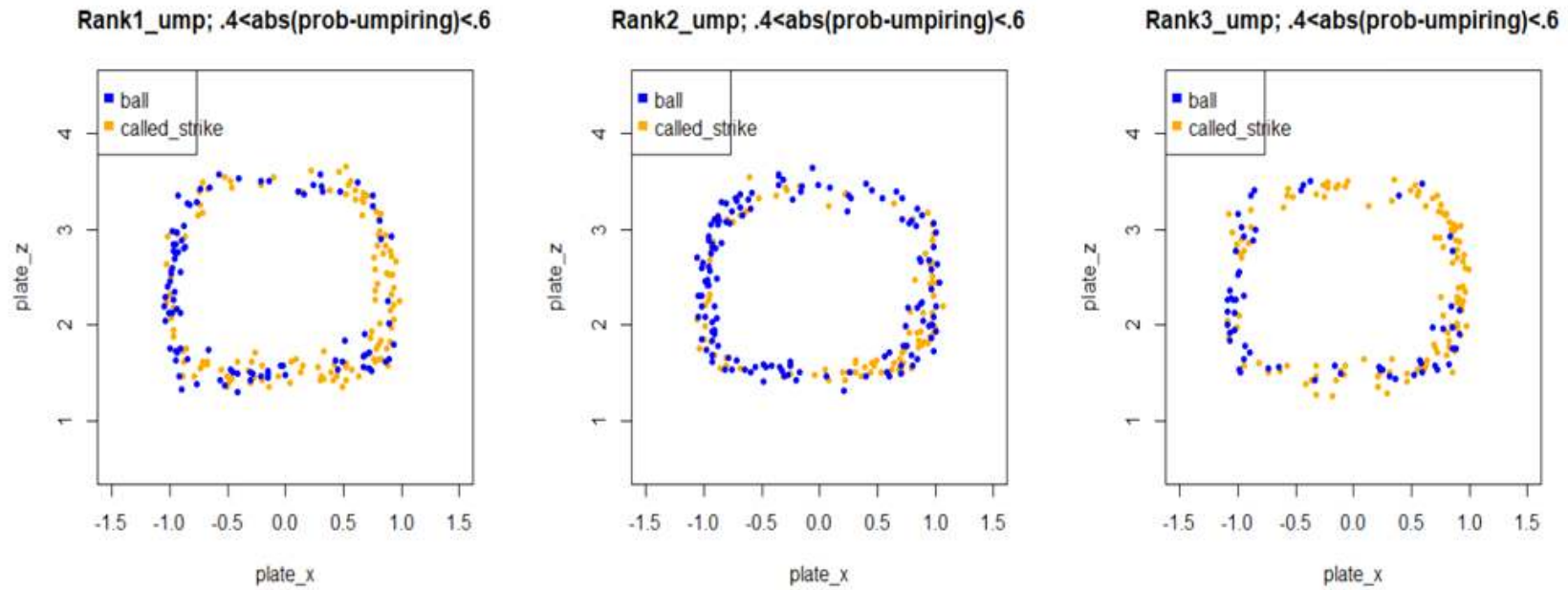


그림 III-11. Rank 1-3 판정 자료[0.4<ABS(prob.-umpiring)<0.6]

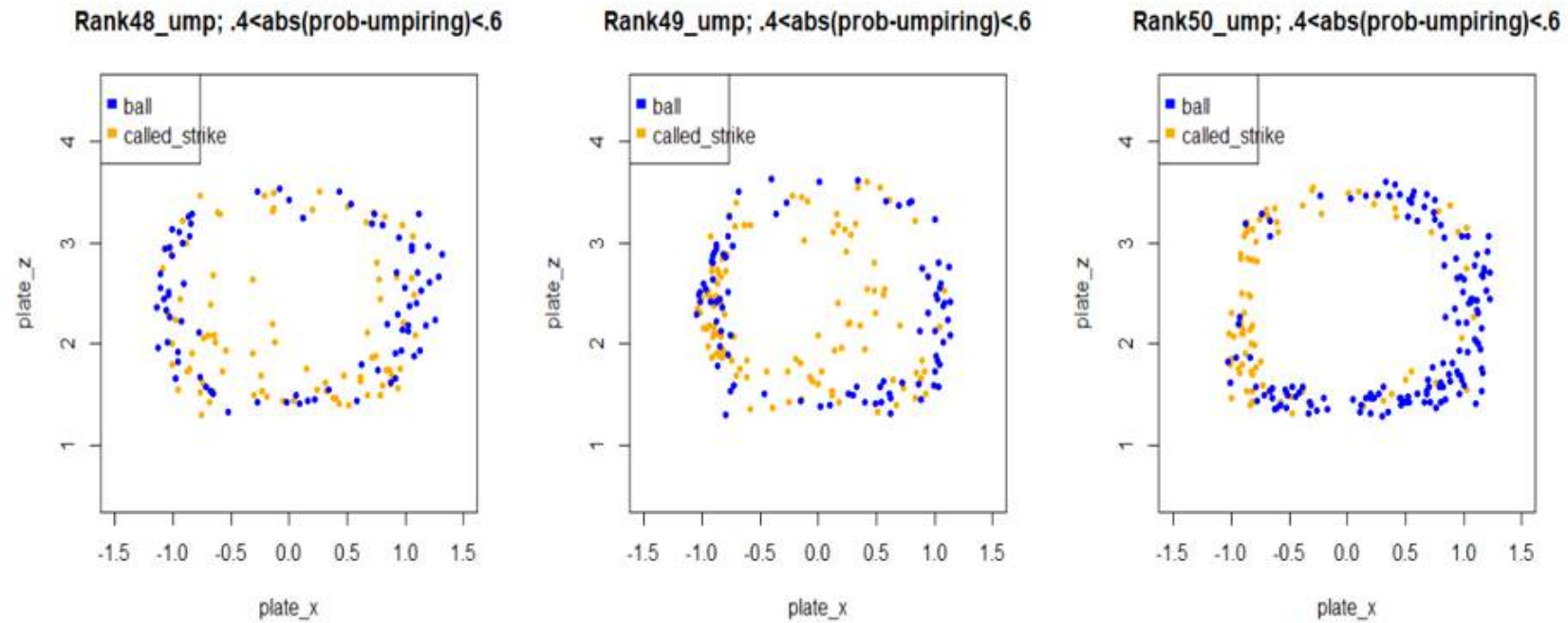


그림 III-12. Rank 48-50 판정 자료[0.4<ABS(prob.-umpiring)<0.6]

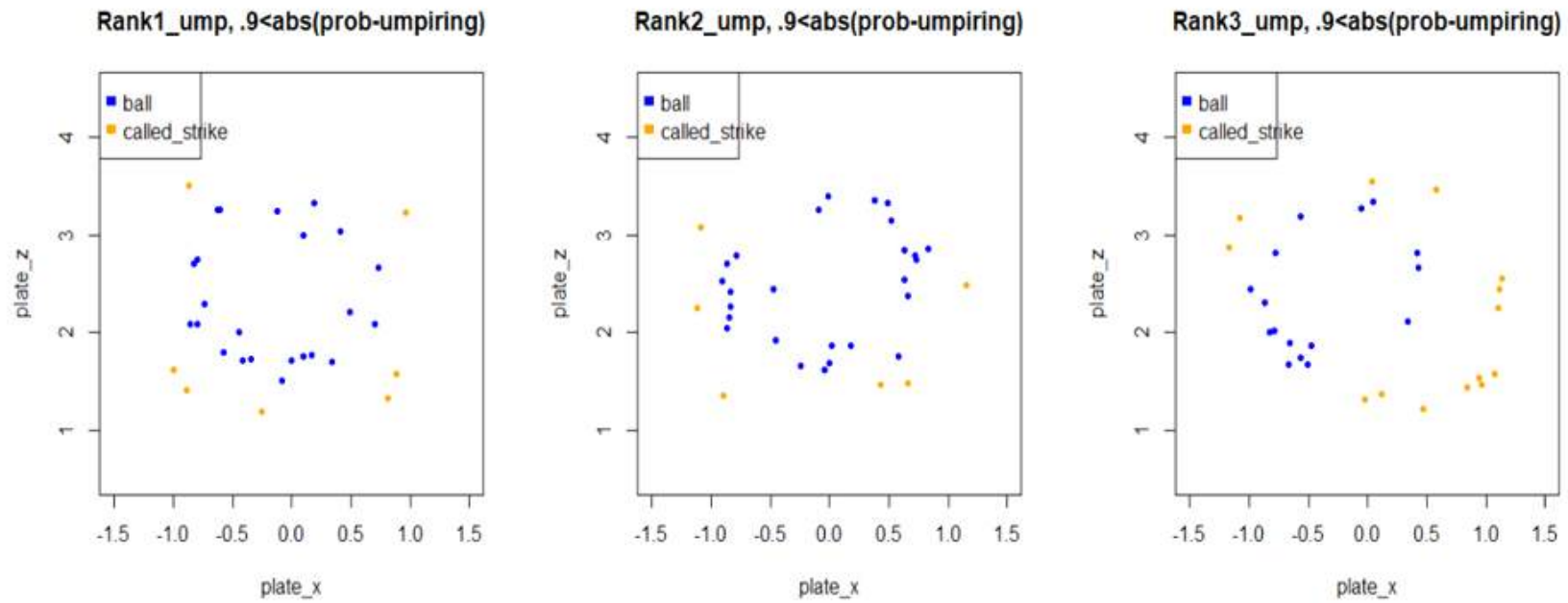


그림 III-13. Rank 1-3 판정 자료[0.9<ABS(prob.-umpiring)]

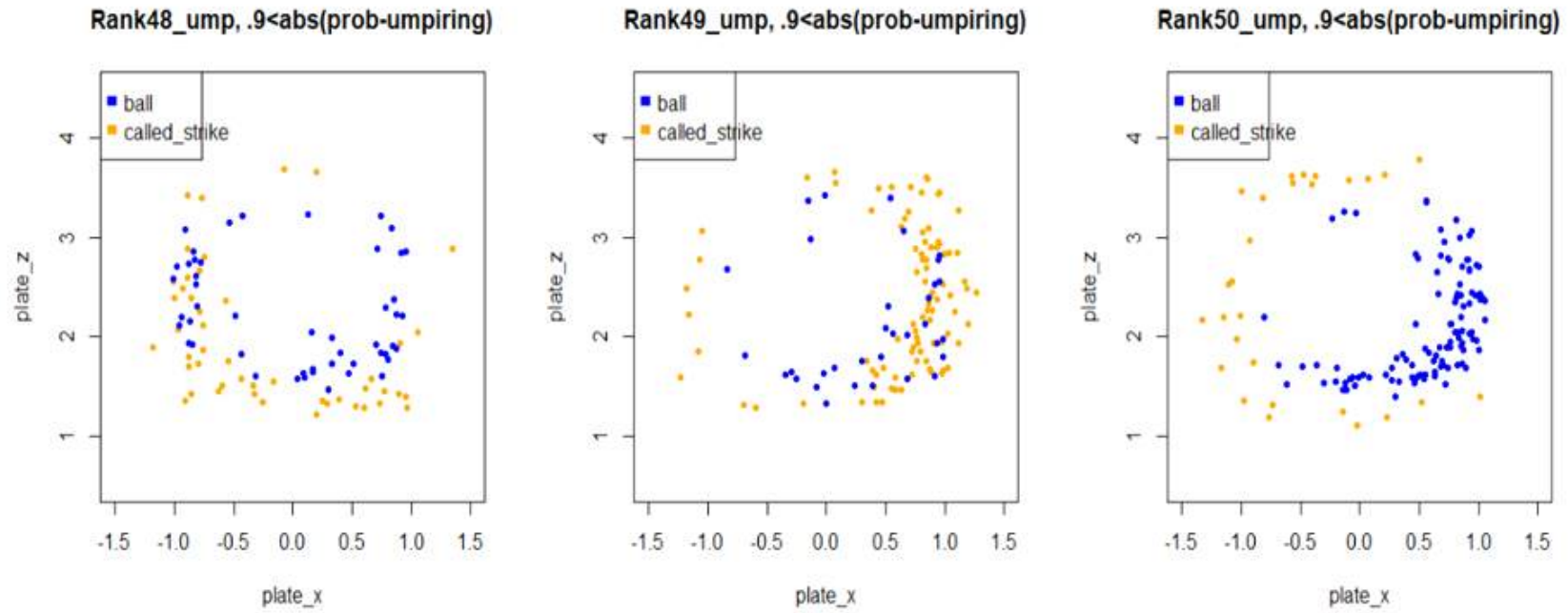


그림 III-14. Rank 48-50 판정 자료[0.9<ABS(prob.-umpiring)]

표 III-8. AI와 반대되는 판정에 대한 경기 상황 분석(%)

Prob. table of 0.9<abs(AI prob - umpiring)		top 3			bottom 3		
		rank1	rank2	rank3	rank48	rank49	ump50
total n		4914	4992	4067	4654	4448	4061
N. of .9<abs(AI prob - umpiring)		30	32	30	137	122	97
%		0.61	0.64	0.74	2.40	2.74	2.80
Ball/Strike	Ball	0.77	0.81	0.53	0.80	0.27	0.33
	Strike	0.23	0.19	0.47	0.20	0.73	0.67
innings	1~3	0.47	0.56	0.30	0.37	0.40	0.30
	4~6	0.27	0.19	0.27	0.33	0.33	0.40
	7~9	0.27	0.25	0.43	0.24	0.26	0.30
	<9	0.00	0.00	0.00	0.06	0.01	0.00
outs	0	0.40	0.44	0.40	0.40	0.39	0.29
	1	0.17	0.38	0.17	0.31	0.22	0.39
	2	0.43	0.19	0.43	0.29	0.39	0.32
pitch_number	1	0.37	0.31	0.43	0.39	0.50	0.40
	2	0.27	0.22	0.33	0.20	0.23	0.23
	3	0.20	0.22	0.10	0.23	0.10	0.24
	4	0.07	0.19	0.07	0.08	0.14	0.08
	5	0.00	0.06	0.07	0.06	0.02	0.03
	<5	0.10	0.00	0.00	0.04	0.02	0.02
Ball/Strike Count	0-0	0.37	0.31	0.43	0.39	0.50	0.40
	0-1	0.10	0.13	0.20	0.12	0.12	0.08
	0-2	0.00	0.03	0.03	0.04	0.02	0.03
	1-0	0.17	0.09	0.13	0.07	0.11	0.14
	1-1	0.13	0.16	0.03	0.13	0.05	0.12
	1-2	0.00	0.06	0.03	0.03	0.01	0.05
	2-0	0.07	0.03	0.03	0.07	0.04	0.08
	2-1	0.03	0.13	0.03	0.04	0.08	0.01
	2-2	0.00	0.00	0.07	0.04	0.00	0.03
	3-0	0.03	0.06	0.00	0.02	0.03	0.02
	3-1	0.00	0.00	0.00	0.03	0.02	0.02
3-2	0.10	0.00	0.00	0.02	0.02	0.00	

표 III-9. AI와 반대되는 판정에 대한 경기 상황 분석(빈도)

Prob. table of 0.9<abs(AI prob - umpiring)		top 3			bottom 3		
		rank1	rank2	rank3	rank48	rank49	rank50
total n		4914	4992	4067	4654	4448	4061
N. of .9<abs(AI prob - umpiring)		30	32	30	137	122	97
%		0.61	0.64	0.74	2.40	2.74	2.80
Ball/Strike	Ball	23	26	16	109	33	32
	Strike	7	6	14	28	89	65
innings	1~3	14	18	9	51	49	29
	4~6	8	6	8	45	40	39
	7~9	8	6	12	39	37	42
	<9	0	0	0	8	1	0
outs	0	12	14	12	55	48	28
	1	5	12	5	42	27	38
	2	13	6	13	40	47	31
pitch_number	1	11	10	13	54	61	39
	2	8	7	10	27	28	22
	3	6	7	3	32	12	23
	4	2	6	2	11	17	8
	5	0	2	2	8	2	3
	<5	3	0	0	5	2	2
Ball/Strike Count	0-0	11	10	13	54	61	39
	0-1	3	4	6	17	15	8
	0-2	0	1	1	5	3	3
	1-0	5	3	4	10	13	14
	1-1	4	5	1	18	6	12
	1-2	0	2	1	4	1	5
	2-0	2	1	1	9	5	8
	2-1	1	4	1	5	10	1
	2-2	0	0	2	5	0	3
	3-0	1	2	0	3	4	2
	3-1	0	0	0	4	2	2
	3-2	3	0	0	3	2	0

위의 <표 III-8, 9>는 Rank 1-3(Top 3), 48-50(Bottom 3) 심판의 '0.9<ABS(prob-umpiring)'에 해당하는 투구들의 경기 상황을 분석한 것이다. <표 III-8>은 전체 판정 대비한 '0.9<ABS(prob-umpiring)'의 비율을 나타낸 표이며, <표 III-9>는 앞선 표에 제시한 비율을 빈도 형태로 나타낸 것이다. <표 III-8, 9>에서 노란색 음영과 함께 굵은 글씨(Bold)로 처리된 값들은 각각의 경기 상황 범주 안에서 가장 높은 비율(빈도)을 차지하는 값들을 의미한다. 먼저 첫 번째 범주 Ball/Strike 구분에서는 Top 3에 해당하는 심판들은 모두 Strike 판정보다 Ball 판정에서 '.9<ABS(prob-umpiring)'에 해당하는 판정이 높은 비율을 차지하는 것으로 나타났다. 반면 Bottom 3에서는 rank 48의 심판은 앞선 결과와 같은 양상을 보였지만, Rank 49와 50에서는 앞선 결과들과는 반대로 Ball 판정보다 Strike 판정에서 '.9<ABS(prob-umpiring)'에 해당하는 판정이 높은 비율을 차지하는 것으로 나타났다. Rank 49 심판은 '0.9<ABS(prob-umpiring)'에 해당하는 122개의 판정 중 89개(73%)의 판정이 Strike에서 나타났으며, Rank 50 심판의 97개 중 65개(67%)의 판정이 Strike에서 나타났다.

경기 초반(inning 1~3), 중반(inning 4~6), 후반(inning 7~9)으로 구분한 이닝에서는 Rank 1, 2, 48, 49의 심판들은 경기 초반에 AI와 반대되는 판정을 수행한 비율이 높게 나타났다. 반면, Rank 3는 경기 후반, Rank 50은 경기 중반에서 '0.9<ABS(prob-umpiring)'에 해당하는 판정의 비율이 높은 것으로 나타났다. 즉, 6명의 심판 중 4명은 경기 초반에 AI와 반대되는 판정을 수행한 비율이 높은 것으로 나타났으나, top 3, bottom 3 집단 간 뚜렷한 차이를 보이지는 않았다. 또한 아웃카운트 역시 두 집단 간 뚜렷한 차이를 나타내지 않았다.

한편 투수가 현재 상대하는 타자에게 투구한 횟수를 의미하는 'pitch_number'에서는 6명의 심판 모두 초구(初球)에서 인공지능 심판의 판정과 반대되는 판정의 비율이 가장 높게 나타났다. 특히 Rank 49 심

판은 '0.9<ABS(prob-umpiring)'에 해당하는 122개의 판정 중 절반을 차지하는 61개의 판정이 모두 '0Ball-0Strike'상황, 즉, 초구에서 나타났다. 볼 카운트에 따른 '0.9<ABS(prob-umpiring)'의 발생 비율 역시 앞선 'pitch_number'의 결과와 같은 맥락에서 0Ball-0Strike 카운트에서 가장 높은 비율을 기록하였다. 반면, 회색 음영으로 표시한 '3balls-XStirkes' 상황이나 'XBalls-2Strikes'상황과 같이 볼 또는 스트라이크 판정에 대한 선택적 여유가 없는 상황에서는 상대적으로 낮은 비율을 나타냈다. 이는 심판의 판정에 있어서 선택을 유보할 수 없는 상황에서는 인공지능 심판이 강한 확신을 보이는 판정에 반대되는 판정의 빈도가 낮고, 비교적 초구(初球)와 같이 후속 선택지가 많은 상황에서는 상대적으로 많이 나타나고 있음을 보여준다. 이는 심판의 Ball/Strike에 대한 판정이 경기 상황에 따라서 일관되지 않을 수 있음을 시사한다.

4. Study 2에 대한 논의 및 결론

초기의 딥러닝 알고리즘들은 이미지 인식 및 분류가 주를 이루었고, 이후 영상 속 객체를 인식하는 알고리즘들이 개발되었다. 따라서 체육/스포츠 분야에서 딥러닝을 수행한 초기 연구들 역시 이미지 인식, 동영상 속 객체 인식이 대부분을 차지하고 있으며(Joshi et al., 2020; Kim & Choi 2020; Mehrasa et al., 2018), 스포츠 상황에서 발생하는 경기 자료를 활용하여 인공지능 모델을 개발하고, 그 과정을 스포츠 빅데이터 활용의 관점에서 논의한 연구는 매우 일부에 그치고 있다(Huang & Li, 2021; Park et al., 2017). 이러한 선행연구들의 대부분이 2010년대 중반 이후를 기점으로 이루어진 것을 고려해보면, 체육/스포츠 분야에서 인공지능의 활용은 역사적으로 매우 초기에 해당한다고 할 수 있다. 따라서 체육/스포츠 분야에서 인공지능 기법의 활용은 아직도 풀지 못한 무수히

많은 과제와 함께 발전하고 있다. 본 연구는 그 풀지 못한 과제의 일부를 해결하는 과정의 일환으로, 스포츠 빅데이터의 특성을 고려한 인공지능 모델 개발에서 고려해야 하는 요인들을 살펴보고, 인공지능 심판(AI UMPIRE)을 개발하여 이를 실제 심판의 판정을 평가하는 데 활용하는 체계를 제안하였다.

본 학위논문의 Study2 ‘인공지능을 활용한 프로야구 심판 판정 평가 모델 개발’은 큰 틀에서 두 가지의 연구목적을 갖는다. 첫째, 스포츠 빅데이터를 활용하여 인공지능 모델을 개발하고자 할 때, 어떤 형태의 입력층의 구조(architecture)와 선택사항에 대한 조합이 최적의 결과를 도출해 낼 수 있는지 실증적 근거를 제공하는 것이다. 둘째, 개발된 ‘AI UMPIRE’를 활용하여 심판 개인의 판정 자료를 정량적으로 평가하는 체계를 마련하는 것이다. 따라서 본 연구에서는 입력층의 구성을 비롯하여 optimizer, 손실함수와 활성화 함수, Dropout의 포함 여부 등의 세부적인 사항을 조정해가며 다양한 실험학습을 진행하였고, 개발된 인공지능 모델들의 예측 성능을 비교하여 최적의 모형을 ‘AI UMPIRE’로 선정하였다. 또한 개발된 AI UMPIRE가 제공하는 판정에 대한 확률(AI Prob.)에 실제 심판의 판정을 대입하여, 인공지능 심판이 강한 확신을 갖고 있음에도 불구하고 반대되는 판정을 선별해내고, 이를 시각화하여 심판 판정의 편향을 분석하였다. 이와 같은 절차에 따라 본 절에서는 연구 결과에 대한 논의를 이론적 측면과 실용적 측면, 두 부분으로 나누어 구성하였다.

4.1. 인공지능 심판 개발을 위한 이론적 논의

본 연구에서는 투구된 공에 대한 Ball/Strike 판정을 수행하는 최적의 인공지능 심판(AI UMPIRE)을 개발하기 위해 다양한 실험학습을 진행하였다. 그 결과 본 연구에서 개발한 ‘AI UMPIRE’의 구조는 5개의 은닉층에 3개의 Dropout을 갖는 구조를 취하며, 각 층을 구성하는 노드는 256, 128, 64, 32로 출력층에 가까워질수록 줄어드는 형태를 취한다. 활성화 함수는 ReLU와 Sigmoid(출력층)로 설정하였으며, 손실함수는 ‘Binary crossentropy’, optimizer는 ‘SGD’를 활용하였다. AI UMPIRE의 test set에 대한 예측정확도는 93.18%를 기록하였으며, 이는 관련 선행연구들과 비교하여 높은 성능에 해당한다(Song & Marin, 2021; Huang & Li, 2021; Park et al., 2017; Valero, 2016).

딥러닝의 목적은 단순히 주어진 데이터를 잘 분류/예측하는 모델을 개발하는 것이 아니라, 주어진 데이터로부터 새로운 데이터를 잘 예측하는 모형, 즉 일반화(generalization) 성능이 좋은 모델을 개발하는 것이다(Dahl, Sainath, & Hinton, 2013). 이러한 과적합과 일반화에 대한 논의는 모델의 복잡도(model complexity)와 긴밀한 관계가 있다(장병탁, 2017). Schmidhuber(2015)는 주어진 문제의 복잡도에 비해 모델의 복잡도가 지나치게 크면 일반적으로 training set에 대한 정확도는 우수하나, 새롭게 주어지는 test set에 대한 일반화 성능은 저하될 수 있음을 보고하였다. 즉, 본 연구에서 수행한 다양한 조건에서의 실험학습 결과는 인공지능 모델이 training set에 대한 규칙을 과도하게 학습할 경우 오히려 새롭게 주어지는 test set에 대한 대처 성능이 저하되며, 간단한 모델이 복잡한 모델보다 과적합될 가능성이 낮음을 보여주고 있다. 이와 같은 결과는 해결하고자 하는 문제에 대한 두 가지 해결방법이 있다면, 더 적은 가정이 필요한 간단한 해결책이 합리적이라는 오컴의 면도날 이론(Ockham’s razor)(Blumer, Ehrenfeucht, Haussler, & Warmuth, 1987)이

스포츠 빅데이터를 활용한 딥러닝에서도 적용되고 있음을 의미한다. 따라서 본 연구에서는 과적합에 대비하고 일반화 성능을 높이기 위해 Dropout 규제를 포함하여, layer의 구성을 출력층으로 갈수록 좁아지는 형태로 구성하여 최종모델을 설정하였다. 그 결과 선행연구들의 결과와 같이 Dropout이 포함된 모형은 그렇지 않은 모형에 비해 test set에 대한 예측 정확도가 높게 나타났다(Jindal, Nokleby, & Chen, 2016; Srivastava et al., 2014; Dahl, Sainath, & Hinton, 2013).

이와 관련하여 스포츠 상황에서 발생하는 부상 진단을 위한 인공지능 모델을 제안한 Song과 Marin(2021)의 연구에서는 다양한 기계학습 기법을 적용했음에도 불구하고 모델의 성능은 알고리즘 간 큰 차이를 보였다. 구체적으로, 부상 진단에 대한 알고리즘의 예측 정확도는 optimized convolutional neural network 알고리즘에서 약 80%로 가장 높게 나타났고, artificial neural network, convolutional neural network에서는 50-70%의 예측 정확도를 나타냈다. 해당연구에서는 복잡한 모형이 무조건적으로 좋은 성능으로 이어지는 것이 아니라, 데이터의 특성에 맞는 알고리즘의 선정과 layer의 구성이 필요하다는 것을 강조하고 있다. Valero(2016)의 연구에서는 다층퍼셉트론 모형을 구축하여 MLB 경기결과를 예측하고자 하였다. 해당연구에서는 앞선 연구와는 반대로 하나의 은닉층으로 구성되는 다층퍼셉트론의 구조를 취하고 있다. 즉, 앞선 Song과 Marin(2021)의 연구와 본 연구의 'AI UMPIRE'와 비교하여 훨씬 간단한 구조를 취하고 있다. Valero(2016)의 연구에서는 승/패를 예측한 모형의 평균적인 예측 정확도가 57.89%로 보고하고 있는데, 목표 출력이 이진분류(binary classification)임을 고려해본다면 57.89%는 예측 모형으로써 큰 의미가 없는 결과이다. 따라서 해당 연구에서는 다층퍼셉트론보다는 본 연구에서 활용한 심층신경망과 같이 보다 더 복잡한 구조를 갖는 모형을 통한 학습이 필요했을 것으로 사료된다.

물론 본 연구에서는 출력층의 분류가 'Ball/Strike'였으나, Song과 Marin(2021)의 연구에서는 부상 진단, Valero(2016) 연구는 승패를 예측하기 위한 모형이었기 때문에 서로 직접적인 비교에는 한계가 있다. 그럼에도 불구하고 앞선 두 연구와 본 연구의 결과로부터 추론해 볼 수 있는 내용은, 복잡한 모형일수록 좋은 성능을 보이는 것이 아니고, 반대로 단순한 모형이라고 좋은 성능을 나타내는 것이 아니라는 것이다. 즉, 데이터의 특성을 충분히 고려하여 그 특성에 적합한 알고리즘의 적용이 필요하고, 과적합과 일반화의 사이에서 최적의 결과를 도출해낼 수 있는 layer의 구성 및 세부사항의 조정이 필요하다는 것이다. 이와 같은 결과는 실시간으로 방대한 양의 데이터를 활용하여 특정한 결과를 도출해낸다 할지라도, 스포츠 빅데이터가 갖는 특성을 충분히 고려할 수 있는 딥러닝 구조가 뒷받침되지 못한다면 그 결과의 활용은 매우 제한적일 수밖에 없음을 실증적으로 보여주고 있다.

Chollet(2021)은 데이터의 특성에 따라 적합한 층의 구성이나 각 층의 유닛 개수가 달라지기 때문에, 이에 대한 문제를 해결하기 위한 정해진 공식은 없다고 보고하고 있다. 즉, 최적의 인공지능 모델을 개발하기 위해서는 적은 수의 layer를 갖는 구조로부터 학습을 시작하여, 점진적인 복잡화 과정을 통해 과적합과 일반화의 최적의 절충점을 찾아내야 한다는 것이다. 이 과정에서 훈련 데이터의 구조와 특성에 대한 이해는 당연히 선행되어야 할 것이다. 이러한 점에 비추어 볼 때, 본 연구에서 제시한 다양한 실험학습 결과는 스포츠 빅데이터를 활용한 인공지능 모델 개발에 있어 최적의 결과를 도출해내는 과정을 체계적으로 정리하였다는 점에서 의의가 있을 것이다.

4.2. 인공지능 심판의 활용에 대한 실용적 논의

본 연구에서 개발한 AI UMPIRE는 test set에 대하여 93.18%의 예측 정확도를 나타냈다. 본 연구에서는 AI UMPIRE의 판정 확률에 기반하여 인공지능 심판이 강한 확신을 갖는 판정임에도 불구하고 실제 심판의 판정은 반대로 선언된 판정들을 선별해내고, 이를 시각화하여 심판 판정의 편향(bias)을 분석하였다.

2019년 MLB 정규시즌에서 Ball/Strike에 대한 판정 빈도가 4000개 이상을 기록한 50명의 판정을 AI UMPIRE의 판정과 비교한 결과, AI는 강한 확신을 갖고 Ball/Strike 판정을 수행했음에도 불구하고 실제 심판의 판정이 반대되는 판정[$0.9 < ABS(AI \text{ prob.} - \text{umpiring})$]비율은 0.61%~2.8%의 범위를 보였다. 이는 특정 심판은 AI가 강한 확신을 갖고 있는 판정과 반대되는 판정이 100개 중 1개 미만으로 나타나지만, 또 다른 심판은 100개 중 약 3개(2.8개)는 AI가 강한 확신을 갖고 있음에도 반대되는 판정을 수행하고 있음을 의미한다. 본 연구에서 활용한 심판 50명의 2019년 정규시즌 누적 판정 빈도가 4061~4992개임을 감안한다면, 각 심판마다 AI와 반대되는 판정을 수행한 빈도는 최소 30개에서 최대 137개에 달한다. 이러한 AI와 대비되는 판정이 무선적으로 나타나지 않고 특정한 경기 상황 또는 Strikezone의 특정 위치에서 지속적으로 발생한다면, 이는 심판의 판정이 어떠한 순간에는 매우 편향적으로 이루어지고 있음을 시사한다. 따라서 본 연구에서는 ' $0.9 < ABS(AI \text{ prob.} - \text{umpiring})$ '에 해당하는 판정을 시각화하여 특수한 패턴을 보이는지에 대한 분석과 함께 해당 판정들이 발생하는 경기 상황을 분석하였다.

<그림 III-13, 14>에 제시한 바와 같이, top 3 그룹에서는 AI가 강한 확신을 갖고 있는 판정과 반대되는 판정이 무선적으로 나타나고 있지만, Bottom 3 그룹에서는 Strikezone의 특정 위치에서 편향적으로 발생하고 있는 것으로 나타났다. 이와 관련하여 Hunter(2018)의 연구에서는 경기

중 심판이 MLB 규칙집(rule book)에서 규정하고 있는 스트라이크 존을 벗어난 볼을 스트라이크로 판정할지라도 판정이 일관되기만 한다면 공정성에 문제가 없다고 주장하고 있다. 그러나 <그림 III-14>에서 확인할 수 있듯이, Rank 50 심판은 우타자의 몸 쪽에 투구된 공은 상대적으로 스트라이크 존에서 멀어 보임에도 일관되게 스트라이크를 선언하였고, 반대로 우타자의 바깥쪽에 투구된 공에 대해서는 스트라이크 존에 가까워 보임에도 일관되게 Ball을 선언하고 있다. 이러한 판정이 단순히 일관되기 때문에 공정한 판정으로 간주될 수 있을지는 논란의 여지가 있다.

‘0.9<ABS(AI prob.-umpiring)’에 해당하는 판정이 발생하는 경기 상황을 분석한 결과 top 3와 bottom 3 모두 초구(初球)에서 인공지능 심판의 판정과 반대되는 판정의 비율이 높게 나타났다. 따라서 AI가 강한 확신을 갖고 있음에도 실제 인간 심판의 판정이 반대되는 판정은 0Ball-0Strike 카운트에서 가장 높은 비율을 기록하였다. 이와 관련하여, Green과 Daniels(2014)의 연구에서는 스트라이크 판정이 2개가 선행된 상황에서는 심판의 스트라이크 존이 줄어들고, 스트라이크 존을 벗어나 볼(ball) 판정이 3개가 선행된 상황에서는 스트라이크 존이 확장되고 있음을 보고한 바 있다. 같은 맥락에서 Williams(2019)의 연구에서는 볼 카운트에 따라 심판의 판정오류 발생 확률은 최대 2배 가까이 차이가 난다고 보고하였다. 이렇듯 선행연구들에서 역시 프로야구 심판의 판정은 볼 카운트에 따라 편향적으로 발생하고 있다고 보고하고 있으며, 이러한 결과는 AI를 활용한 심판 판정의 평가에서 역시 재현되었다. 특히 Rank49 심판은 ‘0.9<ABS(prob-umpiring)’에 해당하는 122개의 판정 중 절반을 차지하는 61개의 판정이 모두 0Ball-0Strike 상황에서 나타났다(표 III-9 참고). 이는 해당 심판의 Ball/Strike 판정이 경기 상황에 따라서 일관되지 않을 수 있음을 다시 한 번 보여주는 결과이다. 또한 Rank 49 심판은 ‘0.9<ABS(prob-umpiring)’에 해당하는 122개의 판정 중 89개(73%)의

판정이 Strike에서 나타났으며, Rank 50 심판의 97개 중 65개(67%)의 판정이 Strike에서 나타났다. 3개의 Strike판정이 하나의 아웃카운트를 만들고, 27개의 아웃이 선언되어야 경기가 종료되는 야구경기의 특성을 고려한다면 투구 판정에 대한 편향이 Ball보다 Strike에서 많이 발생하고 있다면, 이는 심판 개인의 판정을 평가하고자 할 때 주의 깊게 다룰 필요가 있을 것이다.

3balls 카운트, 2Strikes 상황과 같이 볼 또는 스트라이크 판정에 대한 선택적 여유가 없는 상황에서는 인간의 판정과 AI의 판정이 반대되는 비율이 상대적으로 낮게 나타났다. 즉, 인공지능 심판이 강한 확신을 갖고 있는 판정에 반대되는 판정은 선택을 유보할 수 없는 상황보다, 초구(初球)와 같이 후속 선택지가 많은 상황에서 상대적으로 빈번하게 나타나고 있었다. 이러한 결과에 대하여 Green과 Daniels(2015)은 프로야구 심판은 경기에 대한 개입을 최소화해야 한다는 ‘영향회피(impact aversion)’ 성향이 이와 같은 결과를 초래할 수 있음을 보고한 바 있다. 이와 같은 결과는 프로야구 심판의 Ball/Strike 판정은 인간의 선택적 편향이 작용하고 있음을 실증적으로 보여준다.

본 연구에서는 2019년 MLB 정규시즌 모든 심판의 판정 자료를 학습하여 Ball/Strike 판정을 수행하는 인공지능 심판을 개발하는 데에 성공하였고, 해당 인공지능 모델은 강한 확신을 갖고 <그림 III-14>에 제시한 Rank50 심판의 판정에 반대되는 판정을 수행하고 있다. 이렇듯 본 연구는 단순히 과거 현상에 대한 분석이 아닌 스포츠 빅데이터를 기반으로 인공지능 모델을 개발하여 프로야구 현장에서 실질적으로 활용 가능한 방안을 제시하였다는 점에서 선행연구들과 차이를 보인다. 본 연구에서 제안한 AI UMPIRE는 그동안 심판의 고유한 영역으로만 간주되었던 스트라이크존을 평가하기 위한 새로운 패러다임으로써 충분한 역할을 할 수 있을 것이다.

상기한 논의에도 불구하고 본 연구에서 제시한 인공지능 모델 역시 일부 한계가 있다. 첫째, 본 연구에서 활용한 DNN은 합성곱신경망(CNN), 순환신경망(RNN) 등의 기본구조가 되는 모형이다. 따라서 DNN의 심화된 형태인 CNN, RNN 등의 알고리즘을 스포츠 빅데이터에 적용할 경우 본 연구에서 제시한 모델의 성능보다 더 좋은 성능을 나타낼 수도 있을 것이다. 그러나 본 연구에서는 이러한 심화된 구조에 대해서는 검증을 진행하지 못하였다. 둘째, 본 연구에서는 plate_x, plate_z 변수의 범위를 지정하지 않고 모든 판정 자료를 학습 데이터로 활용하였다. 즉, 스트라이크 존의 경계에 있는 투구가 아니라면 볼 또는 스트라이크 판정이 비교적 명확하게 분류될 수 있기 때문에 test set에 대한 예측 정확도 역시 93%로 높은 성능을 기록한 것으로 판단된다. 따라서 스트라이크 존의 경계에 위치한 투구만으로 학습을 진행한다면, 분류 난이도 자체가 높아짐에 따라 모델의 예측 성능은 저하될 수 있을 것이다. 이러한 한계점은 앞서 언급한 CNN, RNN 등을 활용하여 일부분 해결할 수 있을 것으로 사료된다. 이에 대해서는 후속 연구를 통해 추가적인 검증이 필요할 것이다. 셋째, 본 연구에서 개발한 AI UMPIRE의 훈련 데이터의 구성과 예측 성능을 고려해본다면, 다수의 인간심판을 대변하는 인공지능 심판임에는 부정할 여지가 없다. 그러나 다수의 심판을 대변할 수 있는 인공지능 모델일지라도, 주어진 훈련 데이터 자체가 인간의 편향을 반영하고 있기 때문에 AI UMPIRE를 온전히 정확한 판정만을 수행하는 심판으로 간주하기에는 한계가 있다. 이러한 점을 고려하여 본 연구에서는 '0.9 < ABS(prob-umpiring)'와 같이 AI가 강한 확신을 갖고 있는 판정에 집중하여 이에 반대되는 판정들의 경기 상황을 분석하였다. 따라서 이에 대한 해석은 주의가 필요할 것이다.

IV. 종합 결론

: 불확실성을 고려한 스포츠 빅데이터 분석

4차 산업혁명, 빅데이터, 인공지능 등이 새로운 패러다임(paradigm)으로 자리 잡고 있는 현대사회의 흐름을 고려한다면 스포츠 경기에서 발생하는 오심(誤審) 역시 더 이상 경기의 일부로 간주될 것이 아닌, 충분히 기술적으로 통제 가능한 사안이라는 것이 본 연구의 출발점이었다. 선행 연구를 비롯하여 본 연구의 결과를 종합적으로 고려해 본다면, 프로야구 현장에서 심판의 판정은 경기 상황에 따라 불확실성이 크게 작용하고 있다. 이는 “정해진 규칙 하에서 공정하게 신체적 탁월성을 겨루는 행위”로 정의되는 스포츠의 정의와 대치되는 결과이다. 그럼에도 불구하고 그동안의 관련 선행연구에서는 심판의 편향된 판정을 개선할 수 있는 실질적인 대안을 제시하기에는 한계가 있었다. 이에 따라 프로야구 현장에서 스트라이크 존은 심판 고유의 영역이라는 인식이 자리 잡은 채, 심판마다 볼 판정에 차이를 보이는 것이 마치 당연한 것처럼 여겨져 왔다.

그러나 현대사회의 스포츠 현장에서는 수많은 첨단장비가 구축되어 방대한 양의 빅데이터가 실시간으로 수집되고 있으며, 이러한 빅데이터는 인터넷만 가능한 환경에 있다면 누구나 손쉽게 접근할 수 있다. 즉, 심판 판정에 대한 논란을 단순히 심판의 고유 권한이라는 프레임 속에 가둬두기에는 이를 반증할 수 있는 방대한 양의 스포츠 빅데이터가 매 경기 수집되고 있다. 이러한 맥락에서 본 학위논문은 ‘불확실성(uncertainty)을 고려한 스포츠 빅데이터 분석’이라는 큰 주제 아래, 총 두 개의 세부 연구로 이루어졌다. Study 1 ‘베이지안 확률 접근을 통한 MLB 심판 판정의 불확실성 분석’에서는 프로야구 현장에서 수집되는 빅데이터를 활용하여 MLB 심판의 판정오류 발생 확률을 추정하였고, 그 과정에서 k-dependence bayesian network classifier (k-DB)와 로지스틱 회귀분

석이 각각 어떤 장단점을 갖는지 논의하였다. Study 1의 마지막 단계에서는 1승에 대한 삼진과 볼넷의 가치와 팀의 수익을 통해 판정오류가 각 구단에 미치는 잠재적인 경제적 파급효과를 추정하였다. Study 2 ‘인공지능을 활용한 프로야구 심판 판정 평가 모델 개발’에서는 다양한 조건에서의 심층신경망 실험학습을 통해 판정을 수행하는 인공지능 심판(AI UMPIRE)을 개발하였고, 이를 활용하여 실제 심판의 판정을 평가하는 체계를 제안하였다.

Study 1의 연구결과 프로야구 심판 판정의 불확실성은 경기 상황에 따라 크게 작용하고 있는 것으로 나타났다. 특히 ‘0ball-2strikes’, ‘3-2 full count’와 같이 삼진에 대한 선택적 유보가 불가능한 상황보다 초구(初球) 또는 ‘3balls-0strike’와 같이 스트라이크 카운트가 여유 있는 상황에서는 심판 판정의 판정오류발생 확률이 높게 나타났다. 이러한 결과는 심판의 볼 판정은 정립된 규칙 하에 일관되게 이루어지지 않고 경기 상황에 따라 불확실성이 크게 작용하고 있음을 보여준다.

같은 자료를 바탕으로 k-DB와 로지스틱 회귀분석을 실시한 결과 비교적 간단한 분류문제에 대해서는 k-DB분석 결과와 1% 미만의 차이를 보였다. 그러나 홈/원정, 볼 카운트, 아웃카운트, 직구/변화구 구분, 투구된 공의 위치와 속도 등을 고려해야 하는 다소 복잡한 문제에 대해서 두 기법은 판정오류 발생 확률 추정치에 다소 큰 차이(3~11%)를 보였다. 이와 같은 결과는 로지스틱 회귀분석에서 통계적 유의성을 나타내지 않는 변수가 포함되었을 때 두드러지게 나타났다. 이러한 통계적 유의성은 사건 발생의 사례 수, 설명변수들의 조합에 따라 다르게 나타날 수 있다. 따라서 스포츠 데이터 분석에서 통계적 유의성을 나타내지 않는 변수일지라도, 실제로 해당 사건이 스포츠 경기 결과에 영향을 미칠 수 있다는 점이 간과되어서는 안 될 것이다. 이렇듯 Study 1에서 두 기법을 비교한 결과는 해결해야 하는 문제가 어떠한 형태인지에 따라 그 효용성이 달라

질 수 있음을 시사한다.

Study 2에서는 심층신경망 구조를 기반으로 하여, optimizer, 입력층의 구성, 활성화 함수, dropout의 포함 여부 등의 세부적인 사항을 조정해가며 실험학습을 진행하였고, 일반화(generalization) 성능을 고려하여 인공지능심판(AI UMPIRE)를 개발하였다. AI UMPIRE는 test set에 대하여 93.18%를 예측 정확도를 나타냈다. 본 연구에서는 해당 모델을 바탕으로 실제 심판의 판정을 평가하는 체계를 제안하였다.

AI UMPIRE가 강한 확신을 갖고 판정을 수행하였음에도 불구하고, 이와 반대되는 실제 심판의 판정이 특정한 경기 상황 또는 Strikezone의 특정 위치에서 지속적으로 발생한다면 이는 심판의 판정이 어떠한 순간에는 매우 편향적임을 시사한다. 따라서 본 연구에서는 '0.9<ABS(AI prob.-umpiring)'에 해당하는 판정을 시각화하여 특수한 패턴을 보이는지에 대한 분석과 함께 해당 판정들이 발생하는 경기 상황을 분석하였다. AI와 반대되는 판정을 수행한 빈도가 높은 심판(Bottom 3)과 낮은 심판(Top3) 각각 세 명씩을 비교한 결과 Bottom 3의 판정에서 AI와 반대되는 판정이 특정 위치에서 편향적으로 발생하고 있는 것으로 나타났다(그림 III-13, 14 참고). 또한 이러한 AI와 반대되는 판정은 선택을 유보할 수 없는 상황보다, 초구(初球)와 같이 후속 선택지가 많은 상황에서 상대적으로 빈번하게 발생하는 것으로 나타났다. 이와 같은 결과는 본 학위 논문의 Study 1에서 논의한 바와 같이, 프로야구 심판의 Ball/Strike 판정은 인간의 선택적 편향이 작용하고 있음을 시사한다.

본 연구에서 제안한 인공지능 심판의 개발 과정은 실시간으로 방대한 양의 데이터를 활용하여 특정한 결과를 도출해낸다 할지라도, 스포츠 빅데이터가 갖는 특성을 충분히 고려할 수 있는 딥러닝 학습 구조가 뒷받침되지 못한다면 그 결과의 활용은 매우 제한적일 수밖에 없음을 실증적으로 보여주고 있다. 또한 인공지능 심판을 통해 실제 심판의 판정을 평가하는

체계는 그동안 심판을 평가하기 위한 준거가 부재하여 이루어지지 못하였던 심판 판정의 불확실성을 개선시킬 수 있는 실질적인 대안을 제시하였다는 점에서 의의를 갖는다. 이렇듯 Study 2는 단순히 과거 현상에 대한 분석이 아닌 스포츠 빅데이터를 기반으로 인공지능 모델을 개발하여 프로야구 현장에서 실질적으로 활용 가능한 방안을 제시하였다는 점에서 선행 연구들과 차이를 보인다. 본 연구에서 제안한 AI UMPIRE는 그동안 심판의 고유한 영역으로만 간주되었던 스트라이크존을 평가하기 위한 새로운 패러다임으로써 충분한 역할을 할 수 있을 것이다.

스포츠 경기 상황에서는 과거에 발생했던 사건이 같은 조건에서 반복되는 경우는 매우 드물다. 즉, 스포츠 경기에서는 다양한 변수들이 유기적으로 상호작용하여 특정한 하나의 결과로 나타나며, 변수들의 상호작용은 실시간으로 변하는 경기 상황에 따라 또 다른 형태로 나타난다. 따라서 스포츠 상황에서 발생하는 다양한 사건들을 심층적으로 이해하고 예측하기 위해서는 ‘불확실성’이라는 스포츠가 갖는 특성을 충분히 고려할 수 있어야 할 것이다. 이에 대하여 본 연구에서는 베이지안/인공지능 기법의 활용을 제안하였다. 본 학위논문에서는 문헌고찰을 통해 스포츠 빅데이터의 개념을 정립하였고, 실제 프로야구 현장에서 수집되는 빅데이터를 활용하여 프로야구 심판 판정을 정량화하여 분석하고 평가하는 체계를 제안하였다. 본 연구는 그동안 체육/스포츠 분야에서 활발히 논의되지 못하였던 베이지안/인공지능 연구 절차를 체계적으로 제시하였다는 점에서 의의를 갖는다. 본 연구를 발판으로 체육/스포츠분야에서 빅데이터/인공지능 관련 연구가 보다 활발히 이루어질 수 있길 기대한다.

참 고 문 헌

- 고학수, 김용대, 윤성로, 김정훈, 이선구, 박도현, 김시원(2021). *인공지능 원론 : 설명가능성을 중심으로* / 서울: 박영사.
- 김미옥, 고진수, 노승철, 정제훈(2017). 빅데이터 분석을 활용한 기금지원 체육시설 활성화 방안. *디지털융복합연구*, 15(2), 527 - 535.
- 김청택(2019). *사회과학을 위한 고급 통계학 - R을 이용한 분석* / 서울: 학지사.
- 김헌태, 강민수(2019). 체육 및 스포츠 분야에서의 베이지안 통계 분석 방법의 실용적 접근: JASP 프로그램의 활용. *한국체육측정평가학회지*, 21(4), 65-74.
- 안용규, 김동규, 권오륜, 송형석, 김홍식(2011). 스포츠 관련 용어 정의 재정립. 움직임의 철학: *한국체육철학회지*, 19(4), 33-60.
- 이용균(2012. 06. 25). “엘리트 심판의 심리적 편향, 오심 낳는다.” 경향신문, <https://www.khan.co.kr/sports/baseball/article/201206252127185#c2b>
- 이재길(2018). *(R프로그램에 기반한)베이지안 통계학* / 서울: 황소걸음 아카데미.
- 장병탁(2017). *(장교수의) 딥러닝* / 서울: 홍릉과학
- 조정환(2012). 스포츠 빅데이터 활용과 전망. *한국체육측정평가학회지*, 14(3), 1-11.
- 최형준(2020). 국내 스포츠 빅데이터 분석 연구의 현황. *한국체육측정평가학회지*, 22(2), 63-69
- 황수웅, 박일혁(2020). MLB 투구기록을 활용한 프로야구 심판의 ‘Strike’판정 오류 영향 요인 탐색. *한국체육측정평가학회지*, 22(3), 25-38.
- 황수웅, 양혜주, 황선환(2022). 스포츠 빅데이터 활용 이슈와 미래 방향성 고찰. *한국융합과학회지*, 11(5), 53-70.
- Aslan, B. G., & Inceoglu, M. M. (2007). A comparative study on neural network based soccer result prediction. *Paper presented at the Seventh International Conference on Intelligent Systems Design and Applications*.
- Baerg, A. (2017). Big data, sport, and the digital divide: Theorizing how athletes might respond to big data monitoring. *Journal of Sport and Social Issues*, 41(1), 3-20.
- Barbee, J. (2020). *Prediction of Final Pitch Outcome in MLB Games Using Statistical Learning Methods* (Master dissertation, California State University, Long Beach).
- Bayarri, M. J., & Berger, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science*, 19(1), 58-80.
- Baker, B. J., Du, J., Sato, M., & Funk, D. C. (2020). Rethinking segmentation

- within the psychological continuum model using Bayesian analysis. *Sport Management Review*, 23(4), 764-775.
- Beaton, A. A., Funk, D. C., & Alexandris, K. (2009). Operationalizing a theory of participation in physically active leisure. *Journal of Leisure Research*, 41(2), 175-203.
- Becker, D., King, T. D., & McMullen, B. (2015, October). Big data, big data quality problem. In *2015 IEEE International Conference on Big Data (Big Data)* (pp. 2644-2653). IEEE.
- Ben - Gal, I. (2008). Bayesian networks. *Encyclopedia of statistics in quality and reliability*, 1.
- Bernards, J. R., Sato, K., Haff, G. G., & Bazyler, C. D. (2017). Current research and statistical practices in sport science and a need for change. *Sports*, 5(4), 87.
- Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. K. (1987). Occam's razor. *Information processing letters*, 24(6), 377-380.
- Chen, D. L., Moskowitz, T. J., & Shue, K. (2016). Decision making under the gambler's fallacy: Evidence from asylum judges, loan officers, and baseball umpires. *The Quarterly Journal of Economics*, 131(3), 1181-1242.
- Chollet, F. (2021). *Deep learning with Python 2nd Edition*. Simon and Schuster.
- Collins, H. (2010). The philosophy of umpiring and the introduction of decision-aid technology. *Journal of the Philosophy of Sport*, 37(2), 135-146.
- Dahl, G. E., Sainath, T. N., & Hinton, G. E. (2013, May). Improving deep neural networks for LVCSR using rectified linear units and dropout. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 8609-8613). IEEE.
- Deshpande, S. K., & Wyner, A. (2017). A hierarchical Bayesian model of pitch framing. *Journal of Quantitative Analysis in Sports*, 13(3), 95-112.
- Diebold, F.X. (2000), "Big Data Dynamic Factor Models for Macroeconomic Measurement and Forecasting," *Discussion Read to the Eighth World Congress of the Econometric Society, Seattle, August*.
<http://www.ssc.upenn.edu/~fdiebold/papers/paper40/temp-wc.PDF>
- Diebold, F. X. (2012). *On the Origin (s) and Development of the Term 'Big Data'*. Penn Institute for Economic Research, Department of Economics, University of Pennsylvania.

- Drayer, J., Rascher, D. A., & McEvoy, C. D. (2012). An examination of underlying consumer demand and sport pricing using secondary market data. *Sport Management Review*, 15(4), 48-460
- Duan, Z., & Wang, L. (2017). K-dependence Bayesian classifier ensemble. *Entropy*, 19(12), 651.
- Dworkin, J. B., & Staudohar, P. D. (2002). Even umpires sometimes strike out. *NINE: A Journal of Baseball History and Culture*, 10(2), 18-26.
- Fried, G., & Mumcu, C. (2016). *Sport analytics: A data-driven approach to sport business and management*. New York, NY: Taylor & Francis
- Gelman, A., Carlin, John B., Stern, Hal S., Dunson, David B., Vehtari, A., Rubin, Donald B. (2013). *Bayesian Data Analysis (Third ed.)*. Chapman and Hall/CRC. ISBN 978-1-4398-4095-5.
- Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1), 8-38.
- Green, E., & Daniels, D. P. (2014). What does it take to call a strike? Three biases in umpire decision making. In *8th Annual Sloan-MIT Sports Analytics Conference Proceedings*.
- Greff, K., Srivastava, R. K., Koutnik, J., Steunebrink, B. R., & Schmidhuber, J. (2016). LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10), 2222-2232.
- Gucciardi, D. F., & Zyphur, M. J. (2016). Exploratory structural equation modelling and Bayesian estimation. *An introduction to intermediate and advanced statistical analyses for sport and exercise scientists*, 172-194.
- Gudmundsson, J., & Horton, M. (2017). Spatio-temporal analysis of team sports. *ACM Computing Surveys (CSUR)*, 50(2), 1-34.
- Haghighat, M., Rastegari, H., Nourafza, N., Branch, N., & Esfahan, I. (2013). A review of data mining techniques for result prediction in sports. *Advances in Computer Science: an International Journal*, 2(5), 7-12.
- Hartemink, A. J. (2001). *Principled computational methods for the validation discovery of genetic regulatory networks* (Doctoral dissertation, Massachusetts Institute of Technology).
- Haykin, S. (2008). *Neural networks and learning machines (3rd ed.)* New Jersey: Prentice Hall.
- Hecht-Nielsen, R. (1992). Theory of the backpropagation neural network. In *Neural networks for perception* (pp. 65-93). Academic Press.

- Higgs, N., & Stavness, I. (2021). Bayesian analysis of home advantage in North American professional sports before and during COVID-19. *Scientific Reports*, 11(1), 1-11
- Huang, J., & Hsu, H. J. (2020). Approximating strike zone size and shape for baseball umpires under different conditions. *International Journal of Performance Analysis in Sport*, 20(2), 133-149.
- Huang, M. L., & Li, Y. Z. (2021). Use of Machine Learning and Deep Learning to Predict the Outcomes of Major League Baseball Matches. *Applied Sciences*, 11(10), 4499.
- Humphreys, B. R., & Pyun, H. (2017). Monopsony exploitation in professional sport: Evidence from Major League Baseball position players, 2000–2011. *Managerial and Decision Economics*, 38(5), 676-688.
- Hunter, D. J. (2018). New metrics for evaluating home plate umpire consistency and accuracy. *Journal of Quantitative Analysis in Sports*, 14(4), 159-172.
- Jain, A. K., Mao, J., & Mohiuddin, K. M. (1996). Artificial neural networks: A tutorial. *Computer*, 29(3), 31-44.
- Jin, X., Wah, B. W., Cheng, X., & Wang, Y. (2015). Significance and challenges of big data research. *Big Data Research*, 2(2), 59-64.
- Jindal, I., Nokleby, M., & Chen, X. (2016, December). Learning deep networks from noisy labels with dropout regularization. In *2016 IEEE 16th International Conference on Data Mining (ICDM)* (pp. 967-972). IEEE.
- Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255-260.
- Joshi, K., Tripathi, V., Bose, C., & Bhardwaj, C. (2020). Robust sports image classification using InceptionV3 and neural networks. *Procedia Computer Science*, 167, 2374-2381.
- Kaplan, D., & Depaoli, S. (2012). Bayesian structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 650–673). The Guilford Press
- Kareem, M. A. (2020). Application of technology in physical education class. *International Conference on Enhancing Skills in Physical Education and Sport Science.(Edition:1st)*, London: Rubicon Publications, 331-333.
- Kim, B. J., & Choi, Y. S. (2020). Automatic baseball commentary generation using deep learning. In *Proceedings of the 35th Annual ACM*

- Symposium on Applied Computing* (pp. 1056-1065).
- Kim, J. W., & King, B. G. (2014). Seeing stars: Matthew effects and status bias in major league baseball umpiring. *Management Science*, 60(11), 2619-2644.
- Kirk, R. E. (2003). The importance of effect magnitude. *Handbook of research methods in experimental psychology*, 83-105.
- Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Saint Louis: Elsevier Science & Technology, 2014. Web.
- Labrinidis, A., & Jagadish, H. V. (2012). Challenges and opportunities with big data. *Proceedings of the VLDB Endowment*, 5(12), 2032-2033.
- Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. META group research note, 6(70), 1.
- Liu, J., Li, J., Li, W., & Wu, J. (2016). Rethinking big data: A review on the data quality and usage issues. *ISPRS journal of photogrammetry and remote sensing*, 115, 134-142.
- Lopez, M. J. (2020). Bigger data, better questions, and a return to fourth down behavior: an introduction to a special issue on tracking data in the National football League. *Journal of Quantitative Analysis in Sports*, 16(2), 73-79.
- Mashey, J. R. (1999). *Big Data and the Next Wave of {InfraStress} Problems, Solutions, Opportunities*. In 1999 USENIX Annual Technical Conference (USENIX ATC 99).
- Marcot, B. G., & Penman, T. D. (2019). Advances in Bayesian network modelling: Integration of modelling technologies. *Environmental modelling & software*, 111, 386-393.
- McCarthy, J. (2007). From here to human-level AI. *Artificial Intelligence*, 171(18), 1174-1182.
- McEvoy, S. A. (1993). The Umpire Strikes Out: Postema v. National League: Major League Gender Discrimination. *U. Miami Ent. & Sports L. Rev.*, 11, 1.
- Mehrasa, N., Zhong, Y., Tung, F., Bornn, L., & Mori, G. (2018). Deep learning of player trajectory representations for team activity analysis. *In 11th mit sloan sports analytics conference* (Vol. 2, p. 3).
- Mills, B. M. (2017). Technological innovations in monitoring and evaluation: Evidence of performance impacts among Major League Baseball umpires. *Labour Economics*, 46, 189-199.
- Morgulev, E., Azar, O. H., & Lidor, R. (2018). Sports analytics and the

- big-data era. *International Journal of Data Science and Analytics*, 5(4), 213-222.
- Nalbantis, G., & Pawlowski, T. (2019). US demand for European soccer telecasts: A between-country test of the uncertainty of outcome hypothesis. *Journal of Sports Economics*, 20(6), 797-818.
- Olmedilla, A., Rubio, V. J., Fuster-Parra, P., Pujals, C., & García-Mas, A. (2018). A Bayesian approach to sport injuries likelihood: does player's self-efficacy and environmental factors play the main role?. *Frontiers in psychology*, 9, 1174.
- Parry, J. (2019). E-sports are not sports. *Sport, ethics and philosophy*, 13(1), 3-18.
- Park, H., & Hwang, B. S. (2020). The performance of Bayesian network classifiers for predicting discrete data. *The Korean Journal of Applied Statistics*, 33(3), 309-320.
- Park, Y. J., Kim, H. S., Kim, D., Lee, H., Kim, S. B., & Kang, P. (2017). A deep learning-based sports player evaluation model based on game statistics and news articles. *Knowledge-Based Systems*, 138, 15-26.
- Parsons, C. A., Sulaeman, J., Yates, M. C., & Hamermesh, D. S. (2011). Strike three: Discrimination, incentives, and evaluation. *American Economic Review*, 101(4), 1410-35.
- Rahman, M. A. (2020). A deep learning framework for football match prediction. *SN Applied Sciences*, 2(2), 1-12.
- Rein, R., & Memmert, D. (2016). Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *Springer Plus*, 5(1), 1-13.
- Roos, T., Wettig, H., Grünwald, P., Myllymäki, P., & Tirri, H. (2005). On discriminative Bayesian network classifiers and logistic regression. *Machine Learning*, 59(3), 267-296.
- Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6), 386.
- Rubin, V., & Lukoianova, T. (2013). Veracity roadmap: Is big data objective, truthful and credible?. *Advances in Classification Research Online*, 24(1), 4.
- Russell, J. S. (1999). Are rules all an umpire has to work with?. *Journal of the Philosophy of Sport*, 26(1), 27-49.
- Russell, J. S. (2004). Taking umpiring seriously: How philosophy can help umpires make the right calls. *Baseball and Philosophy*, edited by E.

- Bronson*, 87-104.
- Russell, R. L., & Reale, C. (2021). Multivariate uncertainty in deep learning. *IEEE Transactions on Neural Networks and Learning Systems*.
- Sahami, M. (1996, August). Learning Limited Dependence Bayesian Classifiers. In *KDD* (Vol. 96, No. 1, pp. 335-338).
- Santos-Fernandez, E., Wu, P., & Mengersen, K. L. (2019). Bayesian statistics meets sports: a comprehensive review. *Journal of Quantitative Analysis in Sports*, 15(4), 289-312.
- Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, 61, 85-117.
- Scutari, M., & Denis, J. (2022). *Bayesian networks : With examples in R / 2nd ed.*, Chapman and Hall/CRC.
- Senaviratna, N. A. M. R., & Cooray, T. M. J. A. (2019). Diagnosing multicollinearity of logistic regression model. *Asian Journal of Probability and Statistics*, 5(2), 1-9.
- Shah, G., Shah, A., & Shah, M. (2019). Panacea of challenges in real-world application of big data analytics in healthcare sector. *Journal of Data, Information and Management*, 1(3), 107-116.
- Sottas, P. E., Baume, N., Saudan, C., Schweizer, C., Kamber, M., & Saugy, M. (2007). Bayesian detection of abnormal values in longitudinal biomarkers with an application to T/E ratio. *Biostatistics*, 8(2), 285-296.
- Strivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1), 1929-1958.
- Stegmueller, D. (2013). How many countries for multilevel modeling? A comparison of frequentist and Bayesian approaches. *American Journal of Political Science*, 57(3), 748-761.
- Stenling, A., Ivarsson, A., Johnson, U., & Lindwall, M. (2015). Bayesian structural equation modeling in sport and exercise psychology. *Journal of Sport and Exercise Psychology*, 37(4), 410-420.
- Tirdad, K., Cruz, A. D., Sadeghian, A., & Cusimano, M. (2021). A deep neural network approach for sentiment analysis of medically related texts: an analysis of tweets related to concussions in sports. *Brain informatics*, 8(1), 1-17.
- Torgler, B. (2020). Big Data, Artificial Intelligence, and Quantum

- Computing in Sports. *In 21st Century Sports (pp. 153-173). Springer, Cham.*
- Van de Schoot, R., & Depaoli, S. (2014). Bayesian analyses: Where to start and what to report. *The European Health Psychologist, 16*(2), 75-84.
- Wagenmakers, E. J., Lee, M., Lodewyckx, T., & Iverson, G. J. (2008). Bayesian versus frequentist inference. In Bayesian evaluation of informative hypotheses (pp. 181-207). Springer, New York, NY.
- Watanabe, N. M., Shapiro, S., & Drayer, J. (2021). Big data and analytics in sport management. *Journal of Sport Management, 35*(3), 197-202.
- Weiss, S. M., & Indurkha, N. (1998). *Predictive data mining: a practical guide.* Morgan Kaufmann.
- Williams, M. T. (2019). MLB MUST EMBRACE TECHNOLOGY TO FIX POOR UMPIRE PERFORMANCE. In *Phi Kappa Phi Forum* (Vol. 99, No. 3, pp. 22-26). Honor Society of Phi Kappa Phi.
- Yau, C., & Campbell, K. (2019). Bayesian statistical learning for big data biology. *Biophysical reviews, 11*(1), 95-102.
- Young, W. A., Holland, W. S., & Weckman, G. R. (2008). Determining hall of fame status for major league baseball using an artificial neural network. *Journal of Quantitative Analysis in Sports, 4*(4). doi: 10.2202/1559-0410.1131
- Zikopoulos P, Eaton C (2011) *Understanding big data analytics for enterprise class hadoop and streaming data.* McGrawHill, New York, pp 1-166

Abstract

Analysis of the Sports Big Data Considering Uncertainty

-Development of a “Ball/Strike” Umpiring Evaluation Model
Using Bayesian Estimation and Deep Learning-

Soowoong Hwang

Sport Informatics Lab.

Sport Science of Physical Education

The Graduate School

Seoul National University

This doctoral dissertation consists of two sub-research topics under the major theme of “Sports Big Data Analysis Considering Uncertainty”. In Study 1 “A Bayesian approach to quantify the uncertainty of MLB umpires' calls” estimated the probability of MLB umpires' missed Ball-Strike calls using big data collected from MLB games. In this process, we discussed the advantages and disadvantages of the k-dependence Bayesian network classifier (k-DB) and the logistic regression. In Study 2 “Development of the Evaluation Model for MLB umpires' Call Using Artificial Intelligence”, AI UMPIRE judging Ball/Strike was

developed through experimental learning under various conditions. As the next step, a method to evaluate model for MLB umpires' calls was also suggested based on this AI UMPIRE. The detailed results of each study are summarized as follows.

[Study 1]

Firstly, the analysis using the k-DB showed that the probability of MLB umpires' missed calls varies significantly depending on the "Ball-Strike" counts. In the 2019-2021 MLB regular-season, the probability of umpires' missed calls was about 19 to 21% for the 3balls-0strike count, the largest among the ball-strike counts combinations. On the other hand, in the 0ball-2strikes count, the probability of missed calls was about 4% to 8%, which was very low compared to the previous condition. Secondly, comparing the probabilities of umpires' error in the regular season and the World Series, it was found that the total probability of umpires' missed calls was lower in the World Series than the regular season, while the tendency to make decisions in favor of home team pitchers appeared more in the World Series than the regular season. In addition, there was no difference in the probability of missed calls between the regular season and the World Series when the decision leads to strikeout such as 0ball-2strikes and 3-2 full count. However, in the ball counts not immediately leading to strike out, such as 3balls-0strike, more accurate umpiring was performed in the World Series than the regular season. These results imply that the umpires' ball-strike call is not consistently made under the established rules from MLB, and uncertainty plays a critical role depending on the game situation.

Furthermore, the logistic regression was performed using 2021 MLB regular season data used in the k-DB and showed no apparent difference (less than 1%) from the k-DB results for relatively simple classification problems. On the other hand, for rather complex problems that require considerations of home/away, the ball count, the out count, fastball/breaking balls, the pitched location (plate_x, z) and the speed of pitched balls, the two methods showed relatively large differences (3-11%) in the estimated probability of umpires' missed calls. In particular, while statistically insignificant variables do not affect the estimated probability in the logistic regression, they should not be overlooked because in practice such events could affect the result of sports games although their statistical significance is not

captured by the logistic regression due to, e.g., the small number of events or the correlation among variables in the collected data. As such, the result of comparing the two approaches in “study 1” suggests that the strength of each method may differ depending on the form of the problem to be solved.

[Study 2]

Study 2 aims to develop an artificial intelligence umpire (AI UMPIRE) that performs 'ball-strike' calls through deep neural network learning based on MLB umpiring data and to establish a system to evaluate umpires' judgment by using the developed AI UMPIRE. Many experimental learnings considering various conditions (layers, activation function, dropout, etc) were conducted to develop an optimal AI UMPIRE which showed a prediction accuracy of 93.18% for the test set. In this study, the human umpires' calls contradicting to that of the AI UMPIRE with strong confidence (high probability) were analyzed by visualizing them, which would indicate the bias of each umpire.

Comparing the 'ball-strike' calls of 50 umpires judging more than 4000 pitches in the 2019 MLB regular season with AI UMPIRE, the ratio of the human umpire's decisions as opposed to the AI UMPIRE with strong confidence (the ratio of the cases with the probability difference > 0.9) appeared to be in a range of 0.61% to 2.8%. In addition, by classifying the visual patterns of those cases, it was analyzed in which situation such biased decisions are made. Based on the ratio of the biased umpiring (the ratio of the cases with the probability difference > 0.9), the Top 3 and Bottom 3 umpires were selected, and in the top 3 group, the human umpiring contrary to the AI appeared uncorrelated. On the other hand, in the bottom 3 group, biased decisions occurred in the specific area of plate_x, z (Figures 21 and 22).

Also, it was found that this biased umpiring appeared with the highest rate in the 0Ball-0Strike count, while in the deterministic situations such as the 3balls or the 2Strikes, the rate of disagreement between the human umpire and AI was relatively low. In other words, human umpiring opposed to the decision of the AI with strong confidence occurred frequently in situations where there are many follow-up options, such as the first pitch, rather than those where the strikeout decision cannot be postponed. These results empirically show that the MLB umpires' ball/strike decisions are reflecting human bias.

This study has strength compared to previous studies in that it not only quantifies the uncertainty of MLB umpires' decisions based on sports big data but also develops an artificial intelligence model and an evaluation system of the human umpire that can be practically used in the fields of professional baseball. The AI UMPIRE suggested in the current study is expected to play a crucial role in providing a new paradigm for estimating the strike zone, which has only been regarded as a unique role of the human umpire.

keywords : k-dependence bayesian network classifier, deep neural network, AI umpire, sports umpiring.

Student Number : 2018-32629