SoccerMon 데이터 기반 훈련 부하 및 심리적 상태 데이터를 활용한 부상 예측 모델링

남길우[1], 조현지[2], 김동근[3]
[1]상명대학교 일반대학원 스포츠ICT융합학과 석사과정
[2]상명대학교 일반대학원 스포츠ICT융합학과 석사과정
[3]상명대학교 휴먼지능정보공학전공, 일반대학원 스포츠ICT융합학과, 지능정보기술연구소 교수
skarlfdn99@naver.com, 202431123@sangmyung.kr, dkim@smu.ac.kr

Injury prediction modeling using SoccerMon data-driven training load and psychological state data

Gil-Woo Nam[1], Hyun-Ji Cho[2], Dong-Keun Kim[3]
[1]Dept. of Sports ICT Convergence, Sanmyung University
[2]Dept. of Sports ICT Convergence, Sanmyung University
[3]Dept. of Human-Centered Artificial Intelligence, Department of Sports ICT Convergence, Institute of Intelligence Informatics Technology Sangmyung University, Professor

요 약

본 연구는 공개 데이터셋 SoccerMon 활용하여 부상 예측 시스템을 제안하였다. 엘리트 여자 축구 2개 팀의 2020-2021년 동안 수집된 선수 일일 관측치(주관적 데이터 33,849 객관적 데이터 10,075)를 선수-일 패널로 리샘플링하여 총 36,550개 관측치를 생성하였다. 휴식일·미기록일은 훈련부하(load)=0으로 처리하였고, ACWR의 결측치는 1.0으로 대체하였다. Tim Gabbett의 훈련 부하 이론을 기반으로 ACWR 위험대, 누적 피로, 단조성 위험 등 파생지표를 생성하고 웰니스 지표(수면, 피로, 스트레스 등)와 결합했으며, MI-RF-L1 하이브리드 피처 선택으로 핵심 20개 피처를 선택하였다. 9개의 머신러닝 모델 중 상위 5개를 앙상블로 구성하였으며, 라벨 기반 층화 분할과 정규화·깊이 제한·조기 종료 등 과적합 제어를 적용한 결과, 3/7/14일 예측에서 각각 ROC AUC 0.944/0.942/0.961을 달성하였으며, F2 기준 임계값 최적화를 통해 재현율을 우선하였다. Precision은 3일 0.156, 7일 0.283, 14일 0.314로 상대적으로 낮았고, 이에 따라 False Positive가 True Positive의 약 2.2 - 4.1배 수준이었으나, 현재 모델은 3/7/14일 지평에서 각각 약 60%, 64%, 71%의 부상을 사전에 예측(재현율)하여 현장 스크리닝 도구로서의 가치를 갖는다.

1. 서론

FIFA의 최신 이적시장 집계에 따르면 2025년 여름 국제 이적료 지출이 역대 최고치를 경신하였다.[1] 임금 측면에서도, Deloitte의 2025년 『Annual Review of Football Finance』는 23/24시즌 유럽 클럽의 총 임금이 €131억에 달하고 임금(수익) 비율이 64%였다고 보고하였다. 즉, 구단 재정의 큰 비중이 선수 급여에 투입되고 있는 걸로 확인하였다.[2] 이러한 프로 축구 구단은 막대한 이적료와 연봉을 투자하지만 핵심 선수의 부상은 즉각적인 전력 손실과 순위 하락, 심하면 강등 리스크로 직결된다.[3] 이는 곧 구단의 재정 리스크에 위험 부담이 된다. 그럼에도 부상 예방·예측은 여전히 경험 의존과 사후 대응에 머무는 경우가 많다. 웨어러블·설문 기반 데이터 인프라가 보급되면서 데이터 주도형 위험 예측이 가능해졌고, 그 중 Tim Gabbett의 급성:만성 부하비 Acute:Chronic Workload Ratio(ACWR)는 실무에서 널리 쓰이는 지표로 자리 잡았다.[4] 또한, Tim Gabbett의 최근 연구에서는 ACWR이 1.3-1.5 범위가 최적의 sweet spot임을 재확인하였다. 특히 1.5 이상에서 부상 위험이 증가한다는 보고가 있으며, Gabbett(2020)등 리뷰에서 반복적으로 논의되었다.[5] 그럼에도 기존 연구들은 과도한 피처 의존, 일관된 일반화 성능 부족, 데이터 단위 불일치 등의 문제가 있다. 이를 해결하기 위해 본 연구에서는 적은 수의 피처로 부상을 예측하여 우수한 성능을 모델링하려고 한다.

2. 관련 연구

SoccerGuard는 여성 엘리트 축구에서 pmSys·(StatSports)·WyScout·의무기록을 통합해 부상을 시계열 분류로 예측하는 프레임워크이며, 전처리 AutoML(LogReg, RF, SVC, XGBoost, LSTM) 대시보드로 구성되어 있다. 총 4,050회 실험에서 입력 5세션·출력 5-7세션·이벤트 비율 0.25-0.5가 가장 안정적이었고, SDV 합성으로 극심한 불균형(부상 43건)을 완화하였다.

3. 연구 방법

3.1 데이터 전처리

본 연구에서는 SoccerMon 데이터를 활용하여 주관 데이터 33,849개, 객관 데이터 10,075개 중 훈련 부하 데이터(급성/만성 부하 비율 Acute:Chronic Workload Ratio, ACWR, 급성 훈련 부하 Acute Training Load, ATL(7일), 만성 훈련 부하 Chronic Training Load－CTL28(28일)·CTL42(42일), Daily Load, Monotony, Strain), 웰니스 데이터(Fatigue, Mood, Readiness, Soreness, Stress), 수면 데이터(Sleep Duration, Sleep Quality)를 활용하였으며, 36,550개의 관측치를 선수-일 단위로 리샘플링하였다. 결측치의 경우 웰니스 데이터는 선수별 평균으로, ACWR의 결측치는 1.0으로 대체하였고, 휴식일은 Daily Load=0으로 처리하였다. 부상 라벨링은 각 관측 시점으로부터 3일, 7일, 14일 이내 부상 발생 여부를 이진 변수로 생성하여, 147건의 실제 부상으로부터 각각 310건, 489건, 704건의 positive 케이스를 도출하였다.

3.2 Tim Gabbett 이론 기반 파생 피처

Tim Gabbett의 훈련-부상 예방 이론을 구체화하여 19개의 도메인 지식 기반 파생 피처를 생성하였다. ACWR danger zone은 급성:만성 비율이 1.5 초과 또는 0.8 미만인 경우를 표시하는 이진 변수로, Gabbett의 위험 구간을 직접 구현하였다. ACWR sweet spot은 ACWR이 0.8-1.3 사이인 경우를 나타내는 이진 변수로, 최적 훈련 구간을 포착한다. Chronic load stability는 28일 대비 42일 훈련 부하의 비율(CTL28/CTL42)로 중장기 훈련 안정성을 측정한다. Training monotony risk는 훈련 단조성이 상위 25%에 속하는 경우를 표시하여, 변화 없는 반복 훈련의 위험을 반영한다. Load spike 3d와 Load spike 7d는 각각 3일, 7일 전 대비 훈련량이 표준편차의 2배(3일) 또는 1.5배(7일) 이상 증가한 경우를 나타내며, 급격한 부하 증가의 위험을 포착한다. Cumulative fatigue는 7일간 피로도의 이동평균으로 급성 피로 누적을 정량화하며, Gabbett이 강조한 주간 부하 관리의 핵심 지표이다. Fatigue trend는 7일 전 대비 피로도 변화량으로 피로 축적 속도를 측정한다. Load variability는 7일간 훈련 부하의 표준편차로 부하 변동성을 정량화한다. Recovery index는 준비도(30%), 수면의 질(30%), 근육통 역수(20%), 스트레스 역수(20%)의 가중평균으로 종합적 회복 상태를 나타낸다. Wellness consistency는 4개 웰니스 지표(피로도, 준비도, 근육통, 스트레스)의 일일 표준편차로, 컨디션 변동성이 클수록 부상 위험이 증가한다는 이론을 반영한다. Wellness decline은 웰니스 지표 평균이 5점 미만인 경우를 표시하는 이진 변수로, 전반적 컨디션 저하를 감지한다. Wellness score는 웰니스 지표들의 평균으로 종합적 웰니스 상태를 나타낸다. Composite risk는 ACWR 위험대(30%), 단조성 위험(20%), 3일 부하 급증(20%), 웰니스 저하(30%)를 가중 결합한 종합 위험 지표이다. Position risk factor는 포지션별 부상 위험도를 반영하는 가중치(0.8, 1.0, 1.2)로 시뮬레이션하였다. 시즌 단계별 리스크 지표로는 먼저 Day of season을 시즌 시작일로부터의 경과 일수로 계산하고, 이를 기반으로 Early season risk(시즌 30일 이내), Late season risk(시즌 300일 이후), Mid season(시즌 30-300일)을 이진 변수로 생성하였다. Early season risk는 준비 부족 상태를, Late season risk는 시즌 후반부의 누적 피로와 경기 압박을, Mid season은 시즌 중반의 안정기를 각각 반영한다. 이러한 19개 파생 피처와 14개 원본 피처(훈련 부하 7개, 웰니스 7개) 총 33개 중에서 MI-RF-L1 하이브리드 피처 선택을 통해 각 예측 기간별로 최종 20개를 선정하였다.

3.3 하이브리드 피처 선택

피처 선택은 Mutual Information(MI), Random Forest Feature Importance(RF), L1 정규화 계수를 가중 결합하는 하이브리드 접근법을 사용하였다. 각 타겟 변수(3/7/14일 부상 예측)별로 독립적으로 피처를 선택하여 예측 기간에 최적화된 피처 세트를 구성하였다. 먼저 전체 33개 피처(원본 14개 + 파생 19개)에서 숫자형 피처만 추출하고, 타겟 변수와 식별자를 제외한 후 결측치를 중앙값으로 대체하였다. 각 피처 중요도는 MI의 경우 SelectKBest를 사용하

여 피처와 타겟 간 비선형 상호정보량을 측정하였으며, 단변량 관점에서 각 피처가 타겟 변수에 제공하는 정보량을 정량화한다. RF는 100개 트리, 최대 깊이 5로 제한하여 학습한 후, 불순도 감소량 기반 피처 중요도를 계산하였다. class_weight = balanced로 클래스 불균형을 처리하였다. L1은 정규화 강도 C=0.05의 L1로지스틱 회귀를 학습하여 절대 계수값을 중요도로 사용하였으며, L1 정규화를 사용하여 상관성 높은 피처 중 하나만 선택하는 특성으로 다중공선성을 자연스럽게 처리하였다. 각 방법의 점수를 최댓값으로 나누어 0-1로 정규화한 후 **(수식 1)**을 이용하여 가중 결합하였다. 최종적으로 상위 20개 피처가 선정되었다. 3일 예측에서는 mid_season, late_season_risk, chronic_load_stability, wellness_consistency, cumulative_fatigue가 상위 5개를 차지했으며, 7일 예측에서는 mid_season, wellness_consistency, late_season_risk, mood, cumulative_fatigue가, 14일 예측에서는 mid_season, mood, cumulative_fatigue, wellness_consistency, late_season_risk가 주요 피처로 선정되었다. 특히 mid_season이 모든 예측 기간에서 1위를 차지하여 시즌 중반기의 중요성이 확인되었으며, 예측 기간이 길어질수록 심리적 요인(mood)의 중요도가 상승하는 패턴을 보였다.

$$Score_i = 0.3\widetilde{MI}_i + 0.4\widetilde{RF}_i + 0.3\widetilde{L1}_i$$

(수식 1) 가중결합 최종 점수.

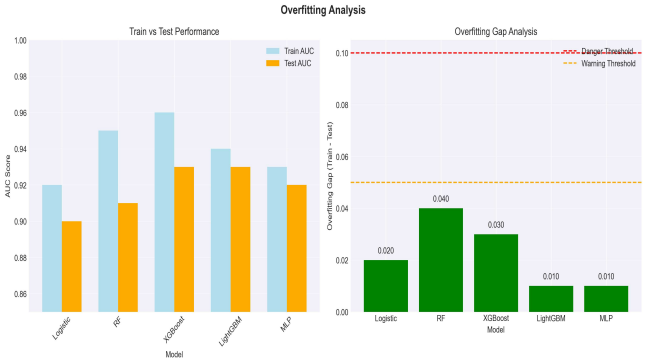
3.4 모델 학습

9개 머신러닝 모델 Logistic Regression balanced / weighted, Random Forest, Extra Trees, LightGBM, XGBoost, Gradient Boosting, Histogram Gradient Boosting(HGB), MLP을 훈련하고, F2-Score 기준 상위 5개를 가중 앙상블로 구성하였다. 각 타깃(3/7/14일)별로 전체 데이터를 80:20 비율로 분할하고, 해당 타깃의 라벨로 stratify하여 Train과 Test의 양성 비율을 유지하였다. 이 설정은 극심한 클래스 불균형(부상률 0.85 - 1.93%)에서 안정적인 모델 학습과 평가를 가능하게 한다. 최종적으로 Train 29,240개로 부상 케이스 3일 248건, 7일 391건, 14일 563건, Test 7,310개로 부상 케이스 3일 62건, 7일 98건, 14일 141건으로 구성되었다. 과적합 제어는 트리 깊이 제한(예: max_depth=3 - 7), 학습률 조정(예: learning_rate=0.03 - 0.05), 정규화(LogReg: L2 C=0.05, 피처선택 단계의 L1(LogReg, penalty='l1', C=0.05), MLP: L2 alpha=0.5), 교차검증(StratifiedKFold)을 적용하였다. 클래스 불균형은 훈련 세트에만 적용되는 샘플링으로 처리했으며, SMOTETomek(Combined) 파이프라인을 사용(내부 SMOTE(sampling_strategy=0.7, k_neighbors=5) + TomekLinks)하였다. 테스트 세트는 원 분포를 유지하였다.

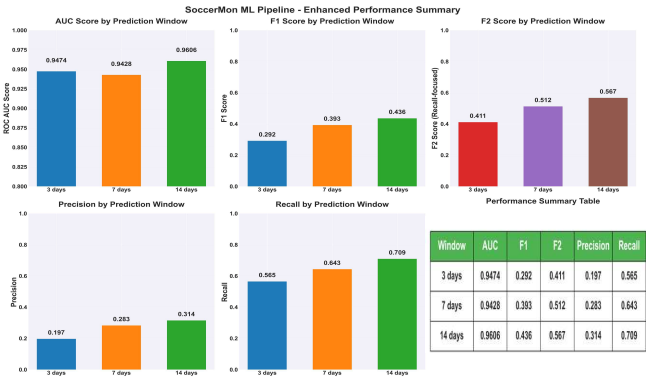
4. 실험 결과

4.1 모델 성능

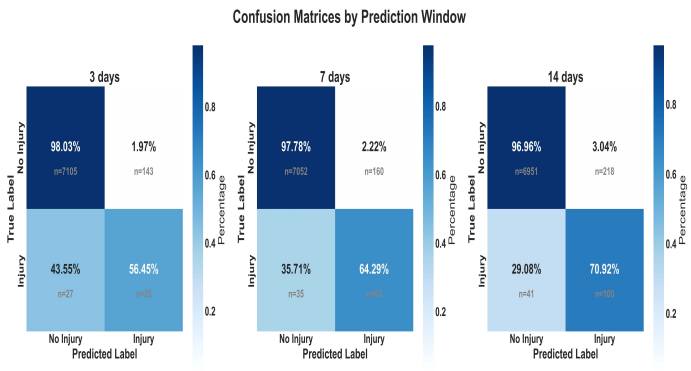
3일 예측에서 HGB가 ROC AUC 0.943(과적합 갭 0.0502), 7일 예측에서 HGB가 0.942(과적합 갭 0.0471), 14일 예측에서 HGB가 0.960(과적합 갭 0.0295)로 최고 성능을 달성하였다. Train - Test 간 AUC 차이(과적합 갭)의 추이는 **(그림 1)**에 제시하였다. F2-Score는 예측 기간이 길수록 향상되어 3일 0.411, 7일 0.512, 14일 0.567을 기록하였다. 예측 기간별 AUC·F2의 비교는 **(그림 2)**에 정리하였다. 다만, Precision은 3일 0.197, 7일 0.283, 14일 0.314로 상대적으로 낮게 나타났다. 양성/음성 분류 양상은 **(그림 3)**에서 확인할 수 있다. 이는 높은 False Positive 비율을 의미하며, 실무 적용 시 추가적인 의사결정 프로세스가 필요함을 시사한다.



(그림 1) 과적합 갭.



(그림 2) 예측 기간별 모델 성능 비교.



(그림 3) 혼동행렬 분석.

4.2 임계값 분석

예측 모델의 실용성을 높이기 위해 F2-score 기준 최적 임계값을 분석하였다. 각 예측 기간별 최적 임계값은 3일 0.396, 7일 0.591, 14일 0.494로 도출되었다. 3일 예측의 경우 낮은 임계값(0.248)을 적용하여 재현율을 극대화하였으며, 이는 단기 부상 위험을 놓치지 않기 위한 보수적 접근이다. 7일 예측에서는 상대적으로 높은 임계값(0.591)이 최적으로 나타났는데, 이는 중기 예측에서 정밀도와 재현율의 균형을 추구한 결과이다. 14일 예측의 임계값(0.494)은 중간 수준으로 설정되어 장기 예측의 안정성을 확보하였다. F2-Score는 예측 기간이 길어질수록 0.381(3일) → 0.512(7일) → 0.567(14일)로 향상되는 추세를 보였다. 이는 장기 예측에서 클래스 불균형이 완화되고(양성 비율 0.85% → 1.34% → 1.93%), 모델이 더 많은 패턴을 학습할 수 있기 때문으로 해석된다.

5. 결론

프로 축구 구단은 더 이상 전통적 경험과 직관에만 의존해서는 안 되며, 스포츠 과학과 데이터 분석의 체계적 활용은 선수 건강 보호와 팀 성과 향상을 동시에 달성할 수 있는 필수 요소이다. 본 연구에서는 Tim Gabbett 이론 기반 19개 파생 피처와 MI-RF-L1 하이브리드 선택하여 최소 피처(20개)로 ROC AUC 0.94+ 성능을 달성하였다. 평균 과적합 갭 0.042는 모델의 우수한 일반화 능력을 시사하며, mid_season, cumulative_fatigue 등 도메인 피처가 상위 중요도를 차지해 이론 기반 접근의 우월성을 입증하였다. 다만 Precision 15-31%로 인한 높은 False Positive(2.2-4.1배)는 독립적 의사결정보다 전문가 판단 보조용 스크리닝 도구로서의 활용이 적절함을 시사한다. 그럼에도 56-71%의 부상을 사전 감지할 수 있어 실무적 가치는 충분하다.

6. 한계 및 향후 연구 방향

본 연구는 단일 데이터셋(SoccerMon) 의존으로 인한 일반화 한계와 극심한 클래스 불균형을 갖는다. 부상 라벨링(3/7/14일)으로 인해 높은 AUC(0.94+)에도 불구하고 낮은 Precision(15-31%)이 나타났다. 또한, 자기보고 기반 웰니스의 응답률이 46.5%로 낮아 결측 대체와 응답 편향의 영향이 잔존할 수 있다는 점에서 한계를 지닌다.

향후 연구에서는 클래스 불균형을 해결하기 위해 대량의 데이터를 이용하고, 리그·성별·시즌이 다른 데이터를 추가해 외부 검증을 수행하고 일반화 성능을 체계적으로 평가할 필요가 있다.

사사

이 연구는 문화체육관광부와 국민체육진흥공단의 지원을 받아 시행한 2025 스포츠산업 융복합대학원 사업의 결과물입니다.

참고문헌

[1] FIFA. Transfer reports hub. (국제 이적 통계 개요), Zurich, Switzerland, FIFA, 2025.

[2] Deloitte Sports Business Group, Annual Review of Football Finance: Europe's Premier, Manchester, Deloitte, 2025.

[3] Hägglund, M., Waldén, M., Magnusson, H., Kristenson, K., Bengtsson, H., & Ekstrand, J., Injuries affect team performance negatively in professional football: An 11-year follow-up of the UEFA Champions League injury study, British Journal of Sports Medicine, 47, 12, 738 - 742, 2013.

[4] Bowen, L., Gross, A. S., Gimpel, M., & Li, F. X., Spikes in acute:chronic workload ratio (ACWR) associated with injury risk in English Premier League players, British Journal of Sports Medicine, 54, 12, 731 - 736, 2020.

[5] Gabbett, T. J., Debunking the myths about training load, injury and performance: empirical evidence, hot topics and recommendations for practitioners, British Journal of Sports Medicine, 54, 1, 58 - 66, 2020.