

# 제11회 「2023 빅콘테스트」 데이터 분석 계획서

\* 해당란에 ☒ 표시

참가분야	<input type="checkbox"/> 생성형AI 분야 <input type="checkbox"/> 데이터신기술 분야 <input type="checkbox"/> 정형데이터 분석 분야 <input checked="" type="checkbox"/> 비정형데이터 분석 분야 <input type="checkbox"/> 빅데이터플랫폼 활용 분야		
세부리그	<input type="checkbox"/> 어드밴스드 리그 <input type="checkbox"/> 스타터 리그 <small>*정형데이터 분석분야에 한함(선택)</small>		
<small>*해당시 체크</small>	<input type="checkbox"/> 지정주제 리그 <input type="checkbox"/> 자유주제 리그 <small>*빅데이터플랫폼 활용분야에 한함(선택)</small>		
개인/팀여부	<input type="checkbox"/> 개인 <input checked="" type="checkbox"/> 팀(총 4명)	개인/팀명	E1I3S4T3F1J4
지도교사명	<small>*스타터 리그에 한함(선택)</small>		
대표ID	rhkdduf627@naver.com		

※ 5장 내외로 목차는 준수하여 자유롭게 작성

분석 주제명	음성 추출과 객체 추적을 활용한 선수 체력 지표 개발 및 콘텐츠 화
분석 배경 및 목적	<p><b>I. 기존 씨름에 대한 조사 및 고찰</b></p> <p>씨름은 한국 고유의 스포츠로 두 사람이 살바나 바지 허리춤을 잡고 힘을 겨루어 상대방을 넘어뜨리는 경기이다. 특히 강호동, 이만기와 같은 유명 스포츠 스타가 등장하면서 한때 한국 내에서 크게 인기를 끌었던 스포츠였지만 2010년 이후 한국 씨름의 인기는 하락세를 보였다. 최근 공중파에서 씨름 예능을 진행하며 씨름의 인기가 다시 올라가는 듯 보였지만, 이 역시 과거의 씨름 인지도로 이어지지는 못했다. 그 결과 현재는 예전과 달리 인기있는 종목으로 인식되지 않고 있으며, 사람들의 관심이 없어짐에 따라 씨름이라는 한국 전통의 스포츠가 위기에 놓여있는 상황이다. 따라서 본 팀은 씨름이 위기에 처하게 된 원인을 여러 논문과, 뉴스 기사, 전문가의 의견을 토대로 조사, 분석하여 한국 씨름의 하락 원인을 다음과 같이 선정하였다.</p> <p>1. 긴 준비 시간</p> <p>다른 스포츠와 다르게 씨름은 본격적인 경기 전 살바를 잡고 있는 긴 준비 시간이 존재한다. 실제 경기는 약 10초 가량 이루어지지만 살바를 잡는 시간은 평균적으로 60초 가량 소요된다. 긴 준비 시간에 비해 실 콘텐츠인 경기 시간은 매우 짧기 때문에 관람하는 사람은 지루함을 느끼게 된다. 이러한 이유로</p>

	<p>준비시간에도 씨름만의 색다른 콘텐츠가 필요하다.</p> <p>2. 단조로운 경기</p> <p>씨름에는 대중이 생각하는 것보다 더욱 화려하고 다양한 기술들이 존재한다. 그러나 실제 경기 영상과 논문에 의하면, 현재 씨름 경기력에 가장 영향을 끼치는 것은 화려한 기술이 아닌 선수의 몸무게이다. 그로 인해 선수들이 펼치는 기술과는 별개로 몸무게에 따라 경기의 승패가 결정되는 단조로운 경기가 주로 펼쳐진다. 이러한 경기는 씨름에 대한 흥미를 떨어뜨리는 요인으로 작용하고, 이는 시청자로 하여금 콘텐츠가 부족하다고 느껴지도록 한다. 이에 따라 기존의 단순히 몸무게로 이기는 씨름과 버티는 씨름이 아닌 더욱 다양하고 변칙적인 전략을 활용하는 경기가 필요하다.</p> <p>3. 순수한 콘텐츠의 부족</p> <p>국내에서 인기를 끌고 있는 축구, 야구와 같은 스포츠의 경우, 경기 전에 선수에 대한 소개, 경기 전략 등 다양한 콘텐츠를 제공한다. 또한 교체시간, 대기시간의 경우 짧은 하이라이트 및 이벤트로 관람객들이 지루하지 않게 한다. 하지만 씨름은 이러한 콘텐츠가 부족하다고 판단한다.</p> <p>결과적으로, 한국 씨름 하락의 원인은 콘텐츠의 부재라고 할 수 있다. 이에 씨름 영상 데이터 분석을 통해 다양한 콘텐츠 생성과 더 나아가 코치진과 선수에게 더욱 다양한 전략을 펼칠 수 있도록 도움을 주고자 한다.</p> <p><b>II. 체력 지표를 통한 경기 분석과 콘텐츠 생성</b></p> <p>씨름 영상 데이터 분석을 통해 '선수들의 체력 지표'를 설정하고자 한다. 해당 지표를 통해 선수들과 코치진에게는 다양한 전략을 활용해 원활한 경기 진행을 할 수 있도록 돕고, 시청자들에게는 콘텐츠 생성을 통해 씨름의 흥미를 높여 시청자 유입을 늘리고자 한다. 궁극적으로 한국 씨름의 부흥과 대중의 많은 관심이라는 목표를 달성하고자 한다.</p>
<p><b>분석 내용 요약</b></p>	<ol style="list-style-type: none"> <li>1. 데이터 수집: 여러 씨름 영상과 씨름 아카이브에서 데이터를 수집</li> <li>2. 전처리: 핵심 영상 추출과 분석에 필요한 음성 추출을 진행</li> <li>3. Object Tracking: 선수의 움직인 정도를 파악</li> <li>4. Speech Extraction: 해설진 음성을 통해 기술 사용 횟수를 추출</li> <li>5. 데이터 분석: 추출한 특징을 통해 체력 기준 산출과 검증을 진행</li> <li>6. 콘텐츠 생성: 체력 지표를 통해 콘텐츠와 전략을 제공</li> </ol>

## I. 분석을 위한 추가 데이터 수집

기본으로 주어진 데이터로는 체력 지표를 생성하기 위한 분석에 있어 한계가 존재한다. 이에 제공 데이터 외에 살바 TV에서 제공한 추가적인 영상 데이터를 활용하였다. 최종적으로 사용한 데이터는 다음과 같다.

1	위더스 제약 2023 민속씨름 보은장사 씨름대회] 금강장사 결정전
2	위더스 제약 2023 민속씨름 보은장사 씨름대회] 단체전 결승 경기 및 백두장사 결정전(8강~장사 결정전)
3	위더스 제약 2023 민속씨름 보은장사 씨름대회] 한라장사 결정전(8강~장사결정전) 및 단체전 예선-결승 선발 잔여경기
4	위더스 제약 2023 민속씨름 보은장사 씨름대회] 태백장사 결정전 및 금강급 2조 예선전

해당 영상에서 체급별 장사 결정전 70여개의 경기를 수집하였다. 각 경기에서 경기 시간, 살바 시간, 휴식 시간, 누적 판수 등 총 8개의 라벨링을 진행하였다. 추가적으로 씨름 아카이브 센터에서 선수의 키와 몸무게를 수집하여 체력 지표 모델 생성을 위한 특징을 구축하였다.

최종적으로 수집된 데이터에서 실제 분석에 사용해야 하는 '금강장사 결정전' 영상 데이터를 Test 데이터로 설정하였고, 그 외 영상을 train 영상으로 설정하였다.

## II. 데이터 전처리 - 영상 추출, 음성 추출

수집된 영상 데이터에서 학습 및 분석에 활용되는 부분만 추출하는 작업을 진행하였다. 각 판마다 살바를 잡는 순간의 5초 전부터 승부의 결과가 표시되는 순간을 기준으로 영상을 추출하였고, 이를 바탕으로 Object Tracking을 수행하였다. 추가적으로 기술 사용 횟수 분석을 위해 영상에서 해설진의 음성을 추출하였다.

## III. Object Tracking - 움직임 정도 추출

선수들의 체력에 영향을 가장 많이 미치는 요인 중 하나가 '경기가 얼마나 격하게 진행되었는지'라고 판단하였다. 영상 데이터에서 선수들의 격함 정도를 파악하고자 Object tracking 기술을 활용하여 선수들의 이동거리를 측정하였다. 이를 위해 2가지의 주요 Task(Key Point Detection, Object Tracking)를 수행하였다.

### 1. Key Point Detection

Tracking의 초기 위치를 지정하기 위한 방법으로, Key Point Detection 모델을 활용하였다. 두 선수가 붙어 있을 경우 정확한 Key Point Detection이 제대로 이루어지지 않아 두 선수가 살바를 잡기 직전에 Key Point Detection 모델을 적용하였다. RTMPose 모델을 활용하여 각 관절의 위치를 파악하고, 해당 위치를 기준으로 Tracking을 진행하였다.

## 2. Object Tracking

Object tracking 모델로 'CoTracker'를 사용하였다. Key Point Detection을 통해 뽑은 관절의 위치를 CoTracker 모델에 전달하면, CoTracker는 해당 위치를 초기 Point 좌표로 설정한다. 이후 **Point가 이동한 거리를 측정하여 선수의 움직임 정도를 추출**하였다.

다만 데이터의 특성으로 인해 이동 거리 측정에 있어 문제점이 존재한다. 경기 영상을 분석해보면, 경기 도중 화면이 Zoom-in 되어 영상을 송출하게 된다. 또한 선수가 씨름판 가외 또는 밖으로 나갈 경우 Zoom-out이 이루어진다. 주어진 경기 영상을 그대로 분석할 경우 Zoom-in/Zoom-out 여부에 따라 선수의 움직임 정도가 일정한 비율의 값을 가지지 못하게 된다. 이는 거리 계산에 있어 정확한 계산이 될 수 없는 치명적인 문제점이다.

이러한 문제를 해결하고자 확대 비율에 따라 거리값을 조정하는 과정을 거쳐 같은 비율을 가지도록 하였다. 이를 위해 경기 영상에 존재하는 전광판의 길이 차이를 측정하여 Zoom 이전과 이후의 비율을 계산하였다. 계산 결과에 따라 Zoom-in된 경우에는 측정된 길이에 0.6784를 곱해 같은 비율의 움직인 거리를 가지도록 조정하여 확대 비율에 따른 차이를 제거하였다. Zoom-out 되었을 경우 원래의 값과 같은 비율을 가지므로 추가적인 처리는 진행하지 않았다.

위와 같은 과정을 통해 Key Point Detection Object tracking을 활용하여 해당 경기의 선수의 움직임 정도를 추출하였다.

## IV. Speech extraction – 기술 횟수 추출

선수들의 체력에 영향을 미치는 또 다른 주 요인은 '선수가 기술을 얼마나 많이 사용하였는가'로 판단하였다. 주어진 분석 데이터 영상 내에서 선수들의 움직임을 통해 기술 사용 횟수를 추출하려 하였으나, 카메라 각도로 인해 정확한 기술 사용 횟수를 추출함에 있어 많은 어려움이 존재하였다. 따라서 직접 영상을 통해 기술 횟수를 분석하여 기록하는 것보다 전문가의 의견을 통해 기술 횟수를 파악하는 것이 더 높은 정확성을 가질 것이라 판단했다. 해당 영상에서는 해설진이 경기 해설을 위해 선수들이 어떤 기술을 사용하고 있는지 해설한다. 따라서 경기 중에 해설진의 음성을 통해 경기에 사용되는 기술과 기술 횟수를 추출하고, 이를 체력 측정에 활용하고자 한다. 이는 다음과 같은 절차를 통해 이루어진다.

### 1. 음성 분리

주어진 영상 데이터에는, 해설진과 장내 아나운서의 목소리가 존재한다. 이로 인해 2명의 목소리가 동시에 송출되는 경우 제대로 된 음성 인식이 이루어지지 않아 정확한 기술명이 추출되지 않았다. 이를 해결하기 위해 음성 분리 모델 'Mossformer'를 사용하여 해설진과 장내 아나운서의 목소리를 분리하는 과정을

거치고, 이후 해설진의 목소리만을 사용하여 기술명과 사용 횟수를 추출하였다.

## 2. STT (Speech To Text)

해설진의 목소리로부터 기술명과 사용 횟수를 추출하기 위해 음성을 텍스트로 변환하는 과정을 수행하였다. STT 모델 'Whisper'를 사용해 텍스트로 변환하여 분석이 용이하도록 하였다.

## 3. 기술 횟수 추출

해당 텍스트로부터 기술명과 사용 횟수를 추출하였다. 기술 종류는 대한씨름 협회 Youtube 채널을 참고하였다. 또한 기술명이 제대로 인식되지 않을 경우 추가적인 전처리를 진행하였다.

앞선 과정들을 통해 해당 경기에 얼마나 많은 기술이 사용되었는지를 파악하여, 경기의 격함 정도를 분석하였다.

## V. 데이터 분석 - 체력 감소 설정

'체급에 따른 선수들의 체력에는 큰 차이가 없다'라는 논문 내용에 기반하여 모든 선수의 체력을 100으로 설정하였다. 이후 경기가 흘러가는 양상을 파악하여, 이를 토대로 체력이 감소하는 형태를 결정하고자 한다. 경기에 따른 체력 감소를 설정하기 위해 다음과 같은 분석을 진행하였다.

먼저 선수마다 경기 기록을 누적하여 데이터를 생성하였다. y로 누적 판수, x로는 누적 경기 시간, 살바 시간, 이동 거리 등 총 8개의 특징을 설정하였다.

Y	누적 판수
X	키, 몸무게, 살바시간, 경기시간, 휴식시간, 승리여부, 기술횟수, 이동거리

예를 들어 A라는 선수의 첫 번째 경기의 y값은 1이며 해당 경기의 기록들로 X를 이룬다. 두 번째 경기의 y값은 2이며 이전 판수와 해당 경기 기록의 합산이 X를 구성하게 된다. 이와 같은 방법으로 70여개의 경기를 분석하여 선수들에 대한 데이터를 생성하였다.

해당 데이터를 이용해 다음과 같은 회귀식을 설정한다.

$$(누적판수) = a_1 * (누적 경기시간) + \dots + a_n * (이동거리)$$

실분석에 사용하는 분석 영상을 제외한 영상 데이터를 학습에 사용하여 회귀계수를 구한다.

분석 영상에서는 매 판의 기록을 X로 두고 회귀식에 대입하여 체력의 감소 정도를 정한다. 한 경기마다 기본 체력 감소를 10으로 설정한다. 해당 값은 추후 실험을 통해 시청자들이 가장 흥미를 느낄 수 있을만한 수치로 변경할 예정이다. 예를 들어 A라는 선수의 첫 번째 경기의 기록을 입력했을 때 y의 값이 1.4가

	<p>나온다면, 1경기를 치룬 선수가 수치적으로 봤을 때는 1.4만큼의 경기를 치렀다고 판단할 수 있다. 이는 본래의 경기보다 0.4만큼의 경기를 더 치렀다고 판단하여 총 14만큼의 체력을 감소한다. 이후 2번째 경기에서 y의 값이 2.3이 나온다면 13의 체력을 감소하게 된다. 반대로 3번째 판에 y의 값이 2.9가 나온다면 다른 경기에 비해 체력 감소가 더 적었다고 판단할 수 있기에, 9의 체력을 감소한다.</p> <p>위에서 설정한 회귀식을 바탕으로 매 경기마다 선수들의 체력을 감소시킨다.</p>
<b>분석결과 활용 및 시사점</b>	<p>본 팀의 데이터 분석을 기반으로 한 선수 체력 지표 생성은 다음과 같은 시사점을 지닌다.</p> <p>1. 콘텐츠 생성</p> <p>해당 분석 과정을 통해 선수 개인마다 체력이라는 새로운 지표를 생성하고, 이를 통해 시청자들에게 새로운 콘텐츠를 제공할 수 있다.</p> <p>첫 번째로 시청자들이 보고 있는 경기 화면 사이드바에 선수들의 체력 바를 보여줌으로써 새로운 볼거리를 제공할 수 있다. 체력 지표를 통해 시청자들이 경기의 승부를 예측하는 즐거움과, 체력이 낮은 선수가 승리했을 때의 역전승과 같은 기존에는 느낄 수 없었던 새로운 유형의 재미를 제공할 수 있다.</p> <p>두 번째로 준비 시간에도 선수들의 이력과 함께 체력을 보여줌으로써 새로운 콘텐츠를 제공한다. 기존 축구와 야구와 같은 형태의 UI를 보여주고, 해설진들 또한 체력이라는 지표를 통해 준비 시간에 더 다양한 해설을 진행할 수 있을 것이라 예상된다.</p> <p>2. 다양한 경기 전략</p> <p>선수들과 코치진에게도 객관적인 정보를 제공함으로써 더 다양한 전략과 경기 운영을 펼칠 수 있다. 체력이 부족한 선수를 만났을 때는 장기전을 펼치거나 기수를 더욱 과감하게 시도하는 등 다양하게 경기를 풀어 나갈 수 있다. 나아가 선수들과 코치진이 다양한 전략을 사용함으로써 기존에 몸무게에 의해 결정되던 단조로운 경기에서 여러 경기를 보여주면서 시청자들에 더 재미를 줄 것으로 예상된다.</p> <p>3. 씨름의 부흥</p> <p>새로운 콘텐츠를 시청자들에게 제공함으로써 기존 단조로운 씨름에서 벗어나 새로운 시청자들을 유입할 수 있을 것으로 예상된다. 씨름에 대한 관심도가 높아지고, 이에 따른 씨름 스타 발굴, 게임과 같은 형태의 경기 운영을 통해 한국 씨름이 부흥할 것으로 기대한다.</p>

※ 제출자료는 평가에 반영 예정