Data Scientist / Applied ML Scientist Portfolio

Data Science, Data Analysis, Data Engineering, Project Design, Project Management

Kwangmin Kim

2025-08-09

Table of contents

1 NLP를 활용한 Data Governance 시스템 단계적 구축

So-What: 전사 데이터 표준화 체계를 확립해 데이터 품질과 활용도를 향상.

1.1 프로젝트 개요

- 소속: Seegene
- 기간
  - Phase1: 데이터 표준화 시스템 구축 2024.10 ~ 2025.08 (파일럿 완료)
  - Phase2~3: 데이터 거버넌스 시스템 구축 2025.09 ~ 2027.09
- 참여인원: 20명 (Data Scientist, Data Engineer, SW 개발자, BT 개발자, DBA)
- 대내적 의의: 총괄장 수명 프로젝트 및 전사 자동화 시스템 구축의 시발점
  - 실험 자동화, 시약 개발 자동화, 분석 자동화, Data Driven 의사 결정
- 대외적 의의: 글로벌 기술 공유 사업의 시발점
  - Microsoft, Springer Nature, KPMG 등 각 국 정부기관 및 기업 등
- 역할: Technical Lead
  - 표준화 체계, 아키텍처 및 프로세스 구축
  - 1명의 Junior Data Scientist 멘토링: 문제정의, 데이터 분석 역량 강화
  - 19명의 IT/BT 개발자 멘토링: 데이터 거버넌스 70% 이해도 달성

1.2 주요 문제점 및 도전과제

- **문제점**
  - 16개 부서 53개 DB의 83% 메타데이터 불일치로 인한 데이터 활용도 저하 문제
  - 데이터 거버넌스 체계 부재, Data Silo 현상, 데이터 통합 및 검증 체계 부재
- **도전과제**
  - 표준화 체계 구축
    - 문제: 독립적으로 개발된 시스템 및 외주 개발 시스템 통합 불가
    - 해결방안: 표준화 현황 분석 및 표준화 프레임워크 확립
  - 데이터의 품질 평가 자동화
    - 문제: 영문 약어 생성 규칙 구현의 어려움 & 표준화 KPI (품질 평가 지표) 부재
    - 해결방안: 계층적 Rule Engine 설계 & 원칙 기반 평가 지표 개발
  - BT & IT 용어 표준화
    - 문제: (관용어 vs 표준화 원칙) & (SI 단위계 vs 업계 관행)
    - 해결방안: 업무 전문가와의 협업을 통한 사전 구축
  - 조직내 낮은 Data 성숙도 및 교육의 어려움
    - 문제: 임원진과 실무진의 막연한 두려움 & 표준화 지식 부족
    - 해결방안: 추상적 개념 구체화, 동기부여 및 교육

1.3 솔루션 설계 및 전략

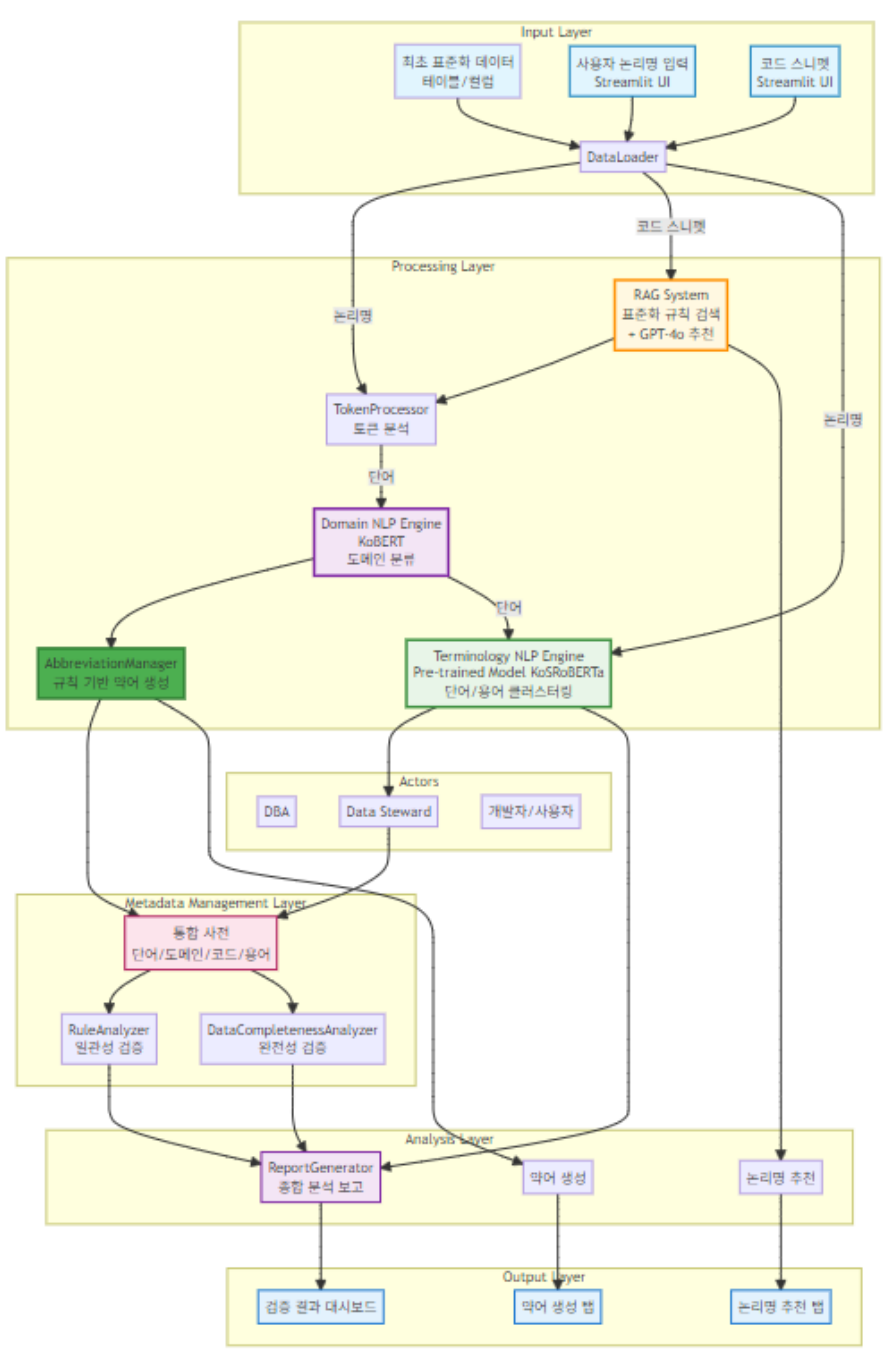

Figure 1: 표준화 프레임워크 아키텍처(단계적 확장·자동화 기반)

- Phase 1: 표준화 프레임워크 구축 (Proof of Concept)
  - 표준화 프래임워크 실증 + 현실적 범위 설정 + 단계적 표준화 진행
  - 표준화 프레임워크 구축 (원칙 + 표준 사전 + 품질 평가)
  - 표준화 프로세스 반자동화 시스템 구축: Hybrid Rule Engine
    - Rule Based 표준화 프로세스 구축
    - Data Driven (딥러닝을 활용) 표준화 프로세스 병목 현상 해결
- Phase 2 ~ 3: 표준화 범위 확장 및 데이터 거버넌스 시스템 구축
  - DB 확장 시 표준화 프로세스 시스템 확장
  - MS Azure Data Factory & Databricks 연동
  - 거버넌스 체계 구축, 표준화 교육 및 캠페인 진행
  - 표준화 모니터링 시스템 구축
  - 자동화 워크 플로우 구축
  - 데이터 활용방안 모색

1.4 기술 스택 및 요구 역량

- 데이터 처리: Python (pandas, NumPy, regex, NLTK, KoNLPy, PyArrow)
- 자연어 처리: PyTorch (Hugging Face Transformers, KoBERT, KoSRoBERTa)
- 머신러닝: scikit-learn (HDBSCAN, PCA, UMAP, silhouette score, 등)
- 시각화 & 모니터링: Streamlit, Matplotlib/Seaborn, Plotly, NetworkX
- RAG: LangChain, OpenAI (GPT-4o), FAISS, Hugging Face, pdfplumber
- 기술적 역량
  - 복잡한 비즈니스 규칙의 알고리즘 설계 능력
  - 딥러닝 모델 설계 및 훈련 데이터 생성 능력
- 업무적 역량
  - 데이터 표준화 프레임워크 구축 (표준화 원칙 및 표준 사전 구축)
  - 도메인 전문가와의 협업 및 요구사항 분석
  - 표준화 정책 수립 및 이행 관리
  - 경영진 및 실무진과의 커뮤니케이션 능력

1.5 결과 및 성과

- 정량적 성과: 데이터 품질 향상
  - 표준화 체계 수립: 0% → 100% (최초 구축)
    - 표준화 원칙, 표준 사전, 표준 데이터 계층 구조, 품질 평가 프로세스 구축
  - 메타데이터 완전성: 29.6% → 100% (80.4% 개선)
  - 메타데이터 일관성: 8.4% → 98.7% (90.3% 개선)
- 효율성 개선
  - 물리명 규칙 검증 시간: 수동 8시간 → 자동 0.73초 (99.27% 단축, 전체 컬럼 기준)
  - 약어 규칙 준수율: 수동 생성 72% → 자동 100% (28% 향상)
  - NLP를 활용한 표준화 프로세스 간소화: 도메인 분류 자동화
- 시스템 구축 성과
  - 표준화 (품질 평가 및 약어생성) 프로그램 핵심 모듈 8개 개발
  - 표준화 세부 규칙 200여개 생성
  - 품질 지표 16개 개발 및 자동 산출 체계 구축
  - NLP(KoBERT)를 통한 도메인 그룹 분류: 용어 중복 방지, 도메인 항목 관리 및 도메인 그룹 관리
    - 예시: "USER_ID", "고객번호", "제품코드", "비밀번호" 등
    - 총 5,632개 훈련 용어, 14개 도메인 그룹, 훈련/테스트 분할 80%/20%
    - 정확도 96.89%, Macro-F1 0.97, '일반단어' 클래스 최저 F1 0.88, 그 외 다수 클래스 0.95 ~1.00
  - NLP(KoSRoBERTa)를 통한 유사 용어 Clustering: 금칙어 관리 및 표준안 관리
    - 예시: "사용자ID", "UserID", "User", "user_id" 등 금칙어 관리
    - Silhouette Score, 0.2752 -> 0.4374 (스키마 정규화로 58.9% 향상, 물리명/논리명 2,214개 기준)
  - RAG(LangChain + FAISS + OpenAI API)을 활용한 표준화 지원
    - 표준화 원칙 정보 검색 및 QnA
    - 코드 스니펫을 입력받아 논리명 추천
    - 데이터 표준화 원칙 (약 200개 Rules) 기반 논리명 추천
    - LLM 논리명을 입력받아 커스텀 약어생성모듈로 물리명 추천
    - 표준화 관련 회의 및 QnA 건 수 감소 (약 70건/주 → 약 4건/주, 94.3% 단축)
- 정성적 성과
  - 표준화 정책 수립, 거버넌스 체계 구축 및 16개 부서 통합 데이터 표준 확립

1.6 기대효과

- 단기 기대효과
  - 데이터 통합 작업 시간 단축 (2시간 → 약 5분, 95.8% 단축)
  - 신규 시스템 구축 시 신속한 표준안 적용 가능
  - 데이터 품질 이슈 사전 예방 체계 확립
- 장기적 비즈니스 가치
  - 부서 간 데이터 공유 활성화, 통합 작업 시간 단축 및 활용도 향상
  - 실험 데이터 모니터링 시스템 구축
  - 글로벌 기술 공유사업 본격화

1.7 추후 과제

- 표준화 범위 확대
- Airflow를 활용한 자동화 워크플로우 구축
- 통합: Big2Core Data Suite(거버넌스 관리) ↔ 인텔리전스 레이어(NLP engines) 간단 API 연계(표준 사전 등록/승인/감사)
- MLOps: FastAPI 서빙, Docker/Helm 배포, Airflow 오케스트레이션, MLflow(실험·모델 레지스트리), GitHub Actions CI/CD, Prometheus/Grafana 모니터링

2 Real-Time PCR 진단 시스템을 위한 지능형 신호 처리

So-What: 불확실한 신호에도 견고한 baseline 처리로 위음성률을 비약적으로 개선.

2.1 프로젝트 개요

- 소속: Seegene
- 기간: 2024.01 ~ 2024.09 (9개월)
- 참여 인원: Data Scientist 3명, Data Engineer 2명, Biologist 2명
- Rule Based 진단 알고리즘을 Data Driven 알고리즘으로의 점진적 개선
  - 기존 rule-based 알고리즘의 한계로 인한 다양한 PCR 신호 패턴 대응 부족
  - 표준화되지 않은 baseline fitting 알고리즘 사용으로 인한 일관성 문제
  - 진단 정확도 향상 및 위양성/위음성 결과 최소화
- 역할: Project Manager & Data Scientist

2.2 주요 문제점 및 도전과제

- 기술적 문제점
  - 신호 노이즈 복잡성: 화학/광학/기계적 반응의 측정 불가능한 노이즈 패턴
  - 알고리즘 분산화: 여러 baseline fitting 알고리즘 병존 및 소통 장애
  - Gray Zone 신호: 시약 성능 및 환경 요인으로 인한 모호한 판독 구간 존재
- 운영적 도전과제
  - 데이터 파이프라인 부재: 체계적인 신호 데이터 수집 및 분석 프로세스 미구축
  - 성능 평가 기준 부재: 객관적인 알고리즘 성능 비교 메트릭 부족
  - 주관적 신호 선별: 1년간 수동으로 특이 신호를 육안 식별하는 비효율적 프로세스
- 제약 조건
  - 호환성 요구: Python에서 C++로의 원활한 포팅을 위한 최소 패키지 사용
  - 이해관계자 다양성: 생물학자, 비전문가 임원 등 다양한 배경의 stakeholder 고려
  - 적은 데이터 포인트: 제한적인 baseline 데이터에서의 robust 알고리즘 필요

2.3 솔루션 설계 및 전략

1. 데이터 파이프라인 구축

- 다양한 PCR 신호 패턴 수집 및 전처리 자동화
- MuDT 전/후 신호 처리 분석 체계 구축
- 성능 평가를 위한 end-to-end 데이터 처리 워크플로우

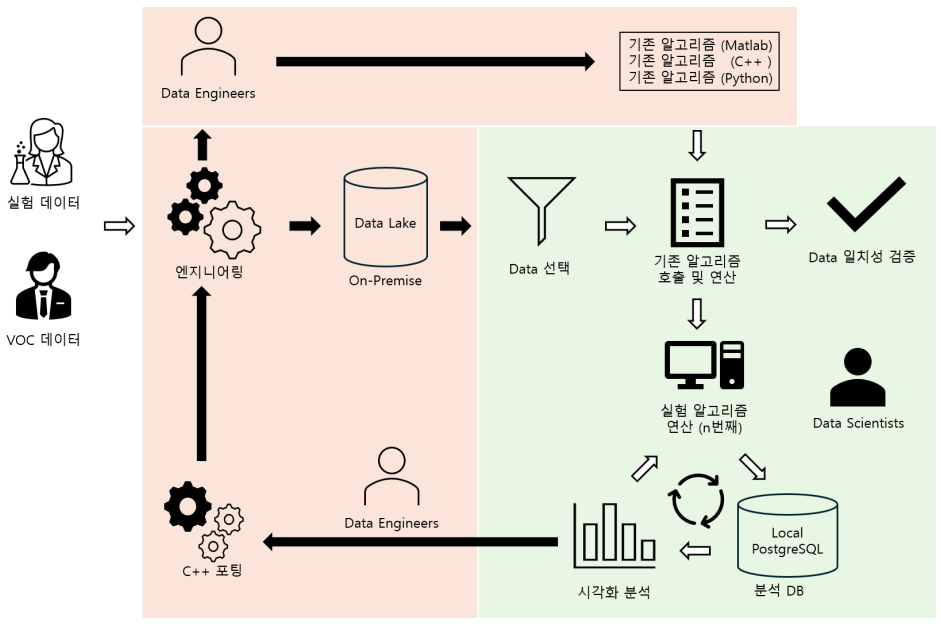

Figure 2: 엔드투엔드 신호 파이프라인(검증·시각화 자동화)

2. 알고리즘 비교 분석

- 1st Panel [After BPN]: normalized **Raw Data**를 보여준다.
- 2nd Panel [CFX]: (대조군1) 타사 기기전용 SW에 내재된 **Black Box** 알고리즘
- 3rd Panel [DSP]: (대조군2) DS팀의 공식적으로 배포된 **Legacy Rule-Based** 알고리즘
- 4th Panel [Auto]: (대조군3) 생물 실험자들이 사용하는 **Legacy Rule-Based** 알고리즘
- 5th Panel [Strep+N]: (실험군1) N+1 번째 [DSP]를 보완용 **Rule-Based** 알고리즘
- 6th Panel [ML]: (실험군2)본인의 특성방정식을 활용한 **data driven ML** 알고리즘
  - Taylor Series에서 함수를 다항식으로 근사할 수 있다는 점에서 착안
  - 다항식 기저 함수를 사용한 선형 회귀로 데이터를 적합하는 방법을 시도
  - 특성 공간 확장을 통해 데이터 내 복잡한 비선형 관계 모델링
  - 적절한 차수 선택과 정규화를 통해 baseline 신호에 적합
  - 로그 정규화 > 기저 함수 > 특성 방정식 > 비용 함수 > 그래디언트 > Momentum > 예측 > 역정규화

3. 시각화 중심 검증 체계: 비전문가를 위한 직관적 성능 평가

- 복수 신호 분석: 6개 알고리즘의 총체적 성능 비교
- 단일 신호 분석: 특이 신호에 대한 세부 성능 평가
- 신호 유형별 분석: 증가/감소/MuDT 특이 신호 패턴별 성능 검증

2.4 기술 스택

- 언어: Python (C++ 포팅 고려), MATLAB (Legacy 알고리즘)
- 라이브러리: NumPy, pandas (최소화 정책)
- 시각화: Matplotlib, Plotly
- 수학적 구현: 특성방정식, 신경망 (without PyTorch, TensorFlow, Keras)

2.5 결과 및 성과

- 알고리즘 성능 검증
  - ML 알고리즘: White noise에 가장 근접한 차감 결과로 최우수 성능 입증
  - 개선된 Rule-based: 기존 대비 특이 신호 처리 능력 향상
  - Black Box 알고리즘: 업계 1위 타사 알고리즘과 성능 비교 완료
  - 위음성률 개선: 0.47% → 0.04% (91.49% 개선)
- 프로세스 개선
  - 개발 프로세스 표준화: 시약 개발 시 일관된 알고리즘 적용 체계 구축
  - 검증 체계: 정성적 평가 방법론을 통한 알고리즘 성능 검증 프레임워크
- 시스템 개선
  - 표준화: 분산된 baseline fitting 알고리즘의 단일화 방향 제시
  - 자동화: 수동적 신호 선별 과정을 체계적 파이프라인으로 대체
  - 시각화 도구: 비전문가도 이해 가능한 직관적 성능 비교

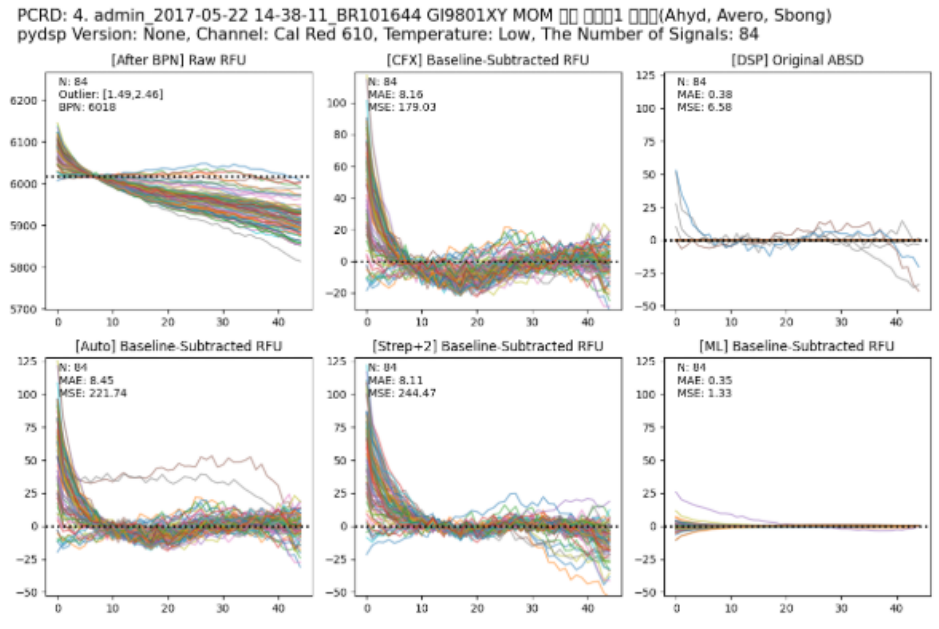

Figure 3: 복수 신호에서 ML이 White noise 근접(최우수) 성능 입증

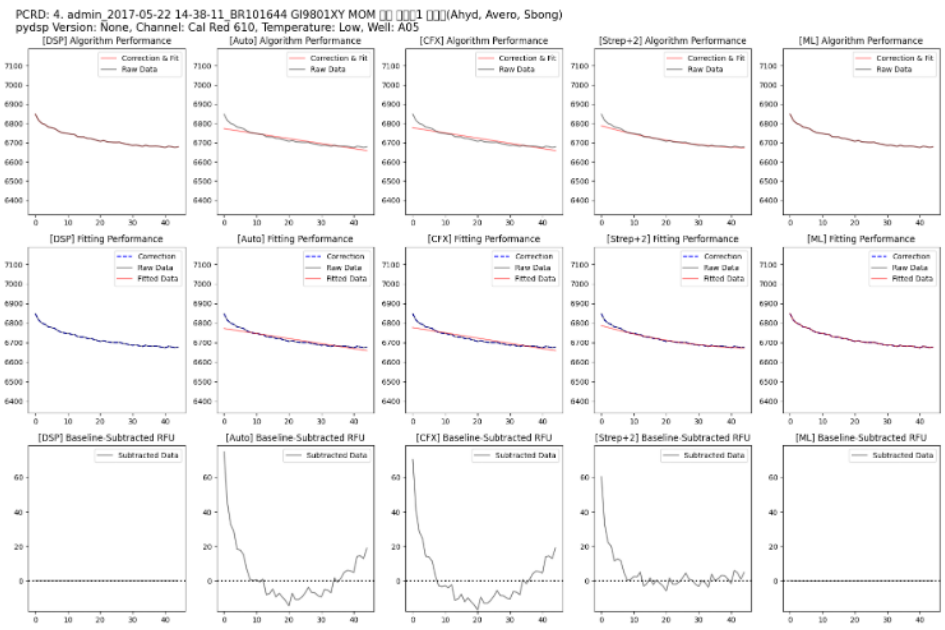

Figure 4: 특이 신호에서 Gray zone 판독 안정성 개선 입증

2.6 기대효과

- 즉시적 효과
  - 진단 정확도 향상: 위양성/위음성 결과 감소로 환자 안전성 제고
  - 개발 효율성: 표준화된 알고리즘으로 시약 개발 시간 단축
  - 품질 일관성: 단일화된 baseline fitting으로 제품 간 성능 편차 최소화
- 장기적 영향
  - 규제 대응 강화: V&V 프로세스 기반의 알고리즘 검증 체계 구축
  - 기술 경쟁력: Data-driven 접근으로 차세대 진단 알고리즘 기술 확보
  - 확장성: 다른 진단 알고리즘 영역으로의 방법론 확산 가능
- 비즈니스 가치
  - 시장 차별화: 업계 최고 수준의 신호 처리 기술 확보
  - 리스크 관리: 진단 오류로 인한 법적/재정적 리스크 감소
  - 혁신 문화: data-driven 의사결정 문화 확산의 기반 마련

3 북미 진출을 위한 진단 알고리즘 안전성 검증 자동화

So-What: 규제 레벨 검증을 자동화해 승인 준비 시간을 대폭 단축.

3.1 프로젝트 개요

- 소속: Seegene
- 기간: 2023.05 ~ 2023.12 (8개월)
- 참여 인원: 데이터 사이언티스트 3명, 데이터 엔지니어 2명, 생물학자 8명, 특허 담당자 3명
- 의료 장비 및 시약 제품의 글로벌 진출 시 각국 정부의 규제 사항 존재
  - 시약의 안정성 검증 & 장비의 안정성 검증
  - **진단 알고리즘의 안정성 검증**
- 북미 진단 시장 진출을 위한 알고리즘 안전성 검증용 통계 분석 문서 작성 반자동화
- 기존 Software Engineering Test보다 더 엄격한 **Advanced Testing** 요구
- 역할: Data Scientist & Project Lead
  - 전체 검증 시스템 설계 및 구현 주도
  - 통계 분석 책임자: Switch Model 기반 검증 방법론 개발
  - Junior Data Scientist 1명 멘토링: 통계 분석, 실험설계 및 리포팅 작성 역량 강화
  - 팀 리더십: 15명 다학제 팀 관리 및 FDA 규제 교육
  - 역할 분배: unit test, integration test, system test, **statistical test**

3.2 솔루션 설계 및 전략

- 알고리즘 안전성을 **통계적으로 입증**하는 시스템 기획
- **Statistical Validation System** 확립을 통한 통계적 분석 입증
- 알고리즘 리스크 정의 및 **정량적 영향도 분석**
- 코드 변화 대응을 위한 **자동화 시스템** 구축
- **SGS 가이던스(EN62304)** 참고
- **FDA General Principles of Software Validation** 문서 기반 시스템 확립
- Structural Testing (코드 기반) & **Statistical Testing** (통계 분석 기반) 병행
- Seegene BT(생명공학)와 IT(정보기술) 부문 협력 체계 구축
- 창의적 Testing Model 기획 및 Statistical Analysis Design 구체화
- Analytical endpoints(비임상, bench): precision/repeatability, reproducibility, linearity, LoD/LoQ, interference/cross-reactivity, stability

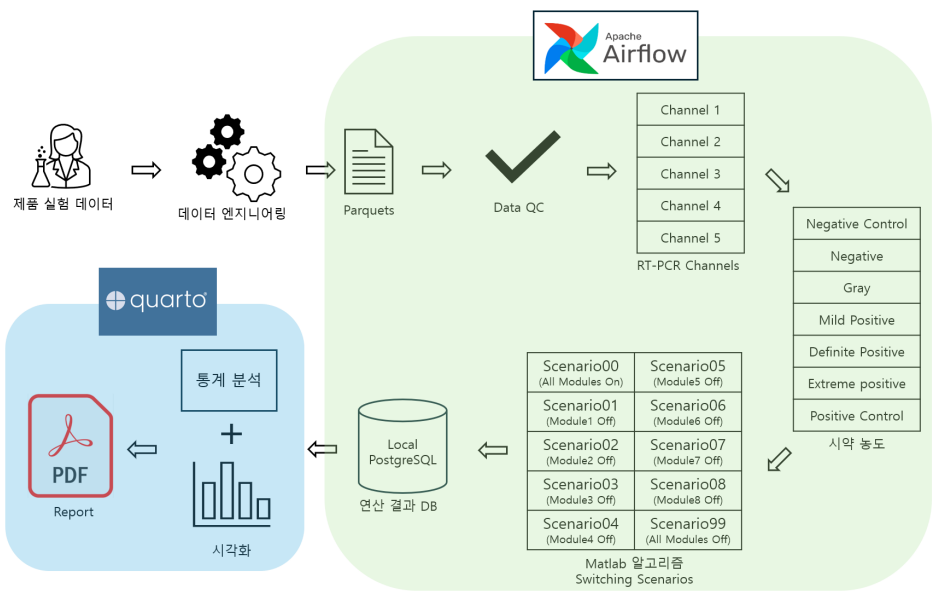

Figure 5: 검증 자동화 파이프라인(레포트 6개월→3주)

3.3 주요 도전과제 및 해결방안

- 문제: BT 부서 생성 데이터 입력 시스템 부재
- 해결: 실험 설계 파일, 의료기기 원시 데이터, 추출 데이터의 **디지털화 시스템** 구축
- 문제: BT 및 Data Science 팀 업무 기술서 부재
- 해결: 부서간 협업을 통한 **업무 문서화** 진행 및 기대 정답 기준 확립
- 5단계 Data QC Process 강화
  - 오타 교정, 결측치 처리, 이상 데이터 처리, 알고리즘 데이터 정합성 1,2차 검정
- 제약 조건
  - 호환성 요구: Python에서 C++로의 원활한 포팅을 위한 최소 패키지 사용
  - 통계 분석 결과 시각화 및 문서화 자동화 시스템 구축

3.4 기술 스택 및 요구 역량

- 규제 지식: FDA Software Validation
- 통계 분석: Statistics (2-Way Repeated Measures ANOVA, McNemar, Breslow-Day, Cochran-Mantel-Haenszel), Analytical Performance Evaluation (CLSI EP05/EP06/EP17), Experimental Design (DoE)
- 프로그래밍: R (Statistical Testing), Python (Engineering), MATLAB (진단 알고리즘)
- 워크플로우: Apache Airflow
- 문서 자동화: Quarto (200페이지 FDA 보고서 자동 생성 시스템)
- 도메인 지식: Biology, Biostatistics, Epidemiology

3.5 결과 및 성과

Table 5.5: P-value Summary of the McNemar Tests for the Negative Concentration

template	concentration	scenario	accuracy	p-value
S	negative	scenario00	100.0	NA
S	negative	scenario01	100.0	NA
S	negative	scenario02	80.6	< 0.001
S	negative	scenario03	100.0	NA
S	negative	scenario04	100.0	NA
S	negative	scenario05	100.0	NA
S	negative	scenario06	100.0	NA
S	negative	scenario07	100.0	NA
S	negative	scenario08	100.0	NA
S	negative	scenario99	80.6	< 0.001

Note:
Inf (infinity), NA (Not Available) and NaN (Not-a-Number) are normal calculation results that occur when positivity or negativity is observed at 100% in the experiments or determined as 100% in the DSP scenarios.
Inf: constant over zero
NaN: zero over zero or Inf over Inf
NA: NaN treated as NA in the caret package

Figure 6: FDA용 통계표 자동 생성(검증 재현성 강화)

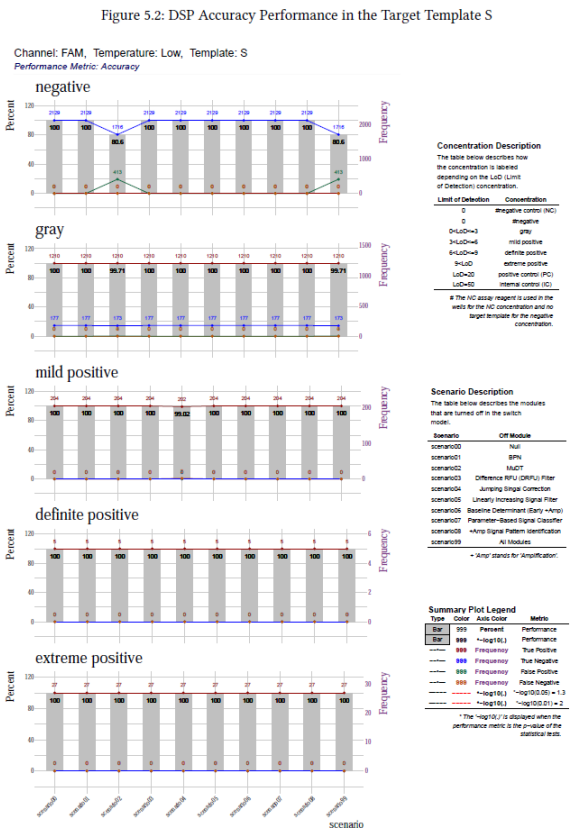

Figure 7: 전반 성능 지표 시각화(이해관계자 커뮤니케이션 향상)

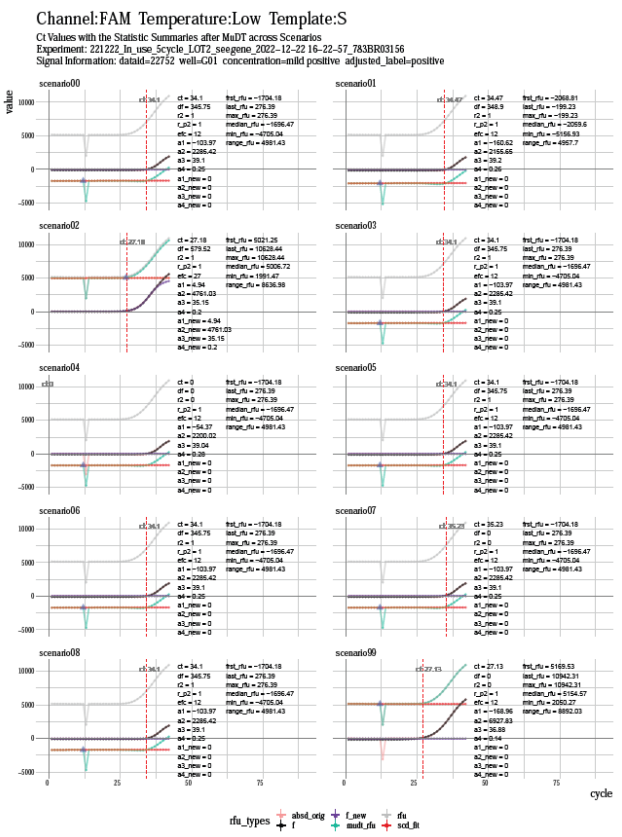

Figure 8: 모듈별 성능 비교(리스크 포커스 분석)

- FDA 제출용 verification & validation report 초안 완성 (FDA 미제출)
- 문서화 시스템: 업무 소통 및 RDB 시스템 구축을 위한 자동화 시스템
- 리스크 관리 통계 분석: 시약/장비 고유 효과 및 교란 요인 위험 관리 분석
- 성능 평가 체계: 사내 최초 알고리즘 및 시약 제품 종합 성능 평가 관리 체계
- 리포팅 자동화: 수동 6개월 → 자동화 3주 (87.5% 시간 단축)
- 99.2%의 시약 + 알고리즘 안전성 통계적 입증
- 혁신성: 사내 고유 Switch Model 기반 모듈별 검증 방법론 개발

3.6 기대효과

- 북미 시장 진출을 위한 FDA 규제 대응 체계 확립
- 알고리즘 안전성에 대한 통계적 증명 체계 구축
- 시약, 장비, 소프트웨어 및 알고리즘 통합 인허가 시스템 구축

4 레거시 Rule-Based 알고리즘의 Data-Driven 전환

So-What: 설명 및 검증 가능한 하이브리드 모델로 레거시의 구조적 해결 방향 제시.

4.1 프로젝트 개요

- 소속: Seegene
- 기간: 2021.10 ~ 2023.04 (1년 6개월)
- 참여 인원: Data Scientist 4명, Data Engineer 2명
- Rule Based 진단 알고리즘을 Data Driven 알고리즘으로의 점진적 개선
  - 복잡한 조건문 기반 신호 보정 시스템의 근본적 한계 해결
  - 10+단계 보정 과정에서 발생하는 systematic bias 및 비선형 상호작용 문제 개선
  - 비효율적인 유지보수 레거시 알고리즘을 데이터 기반 시스템으로 전환
- 프로젝트 구성
  - Phase 1: 레거시 알고리즘 Reverse Engineering (12개월)
  - Phase 2: 알고리즘 개선안 설계 및 제안 (6개월)
- 역할: Data Scientist (Statistical Learning을 활용한 알고리즘 분석 및 개선안 설계)

4.2 주요 문제점 및 도전과제

- 기술적 도전과제
- 알고리즘 문서 부재, 코드 주석 부재 및 가독성이 매우 낮은 변수명의 Legacy MATLAB 코드의 역공학(reverse engineering)의 필요성
  - 유지보수성 저하: 10+단계의 보정과 각 단계별 5+가지 조건으로 수백 가지 실행 경로 생성
  - Systematic Bias 누적: 각 보정 단계에서 발생하는 작은 편향들의 누적으로 최종 결과의 체계적 왜곡
  - 비선형 상호작용: 보정 단계들 간의 복잡한 비선형 상호작용으로 인한 sensitivity analysis 불가능
- 운영적 도전과제
  - 과적합 위험: 특정 장비나 데이터셋에 맞춰진 구체적 조건문들의 새로운 환경에서의 실패 가능성
  - 테스트 어려움: 모든 조건 조합 테스트의 사실상 불가능성과 특정 조건에서만 발생하는 버그 발견의 어려움
  - 예측 불가능성: 유사한 입력 데이터가 미세한 분기점에서 완전히 다른 보정 경로를 따르는 결과 일관성 부족
  - 조직 내 소통 장벽: 통계 비전공자에게 결정론적 규칙의 한계, 편향누적 및 민감도 분석 등의 통계 개념 설명 어려움
  - 기존 워크플로우와의 충돌: 레거시 알고리즘에 익숙한 생물학자들의 새로운 방법론에 대한 저항과 학습 곡선
- 제약 조건
  - C++ 포팅 요구: 최종 목표가 모든 알고리즘을 C++로 포팅하는 것이므로 reverse engineering을 통한 명확한 로직 이해 필수 및 알고리즘 개선안을 low level 프로그래밍으로 구현
  - 확장성 문제: 새로운 노이즈 패턴 발견 시마다 조건문 추가로 인한 복잡도의 기하급수적 증가
  - 팀 문화적 제약: Data Science팀 내에서도 수학/통계적 접근보다 엔지니어링 관점을 우선시하는 문화
  - 통계적 엄밀성보다 실행 및 운영 가능성을 우선하는 개발 철학의 차이

4.3 해결 접근법

1단계: Legacy 시스템 역공학 및 포팅

- Legacy MATLAB 코드 분석
  - 비문서화 알고리즘의 체계적 분석 및 논리적 흐름 해석
    - 주석 부재 및 가독성 낮은 변수명으로 구성된 코드의 의미 추론
    - 50+개 경험적 파라미터들 간의 의존성 분석 및 각 조건문의 의미 추론
    - 10+단계 보정 과정의 수학적/통계적 근거 문서화
- 알고리즘 로직 플로우 명세화
  - 각 보정 단계의 입력/출력 관계 정의
  - 조건분기 구조의 결정 트리 형태 시각화
  - Data Engineer의 C++ 포팅을 위한 상세 기술 문서 작성

2단계: 개선안 설계 및 제안

- 규제 환경 고려사항 분석
  - FDA 규제 대응: 의료기기, 의료시약 및 의료 알고리즘 승인을 위한 알고리즘 설명력(explainability) 요구사항 분석
  - 딥러닝/ML 블랙박스 모델의 설명력 부족으로 인한 규제 리스크 평가
  - 기존 rule-based 접근법의 설명력 확보 명분과 실제 성능 간 trade-off 문제 분석
- Hybrid Modeling 개선안 설계
  - 메카니스틱 모델 기반 접근: 생물학적 메커니즘을 반영한 해석 가능한 모델
  - 로지스틱 시그모이드 일반형을 RT-PCR kinetics의 수학적 표현으로 활용
  - 주요 전처리 함수 3개와 메카니스틱 모델의 합성함수(composite function) 구성

합성 함수 정의

> **Note**
>
> 보안 고지: 규제 목적의 설명력을 위해 개념 및 수식 수준으로만 기술했으며, 소스코드, 세부 알고리즘 로직, 파라미터 규칙 등 민감 정보는 보안상의 이유로 공유하지 않습니다.

$$f(x;\phi_1,\phi_2,\phi_3,\beta) = g_3(g_2(g_1(x,\phi_1),\phi_2),\phi_3) + \text{sigmoid}_g(x;L_{\text{min}},L_{\text{max}},k,x_0)$$

일반형 로지스틱 시그모이드 정의(예시):

$$\text{sigmoid}_g(x;L_{\text{min}},L_{\text{max}},k,x_0) = L_{\text{min}} + \frac{L_{\text{max}}-L_{\text{min}}}{1+e^{-k(x-x_0)}}$$

목적 함수

$$\hat{\theta} = \arg\min_{\phi_1,\phi_2,\phi_3,\beta,\sigma^2} \sum_{i=1}^{n} [y_i - f(x_i;\phi_1,\phi_2,\phi_3,\beta)]^2$$

- 통계적 모델링 통합: 잔차의 확률적 특성을 명시적으로 모델링
  - 결합추정(joint estimation)을 통한 전체 파라미터의 동시 최적화로 systematic bias 방지
  - 합성함수와 실제 데이터 간 잔차의 정규분포 가정 및 white noise 조건 확인

정규분포 가정

$$y_i|x_i,\theta \sim \mathcal{N}(f(x_i;\phi_1,\phi_2,\phi_3,\beta),\sigma^2) = \mathcal{N}(f(x_i;\theta),\sigma^2) = \mathcal{N}(\mu,\sigma^2)$$

$$\text{where } \theta = (\phi_1,\phi_2,\phi_3,L_{\text{min}},L_{\text{max}},k,x_0,\sigma^2)^T$$

규제 환경에서의 설명력(Explainability) 근거

생물학적 해석가능성 (Biological Interpretability)

- 모형-현상 매핑: 합성함수의 각 구성요소(g1/g2/g3, sigmoid)가 신호 전처리·메커니즘과 일대일로 대응되어 파라미터 변화의 의미를 생물학적으로 해석 가능
- RT-PCR kinetics의 수학적 표현을 통해 각 파라미터가 실제 생물학적 프로세스와 직접 연결
- 로지스틱 시그모이드 일반형의 파라미터들($L_{\text{min}}, L_{\text{max}}, k, x_0$)이 각각 물리적 의미를 가짐
- 합성함수의 각 구성요소(g_1, g_2, g_3)가 신호 전처리 단계와 일대일 대응되어 생물학적 메커니즘으로 설명 가능

통계적 검증 가능성 (Statistical Validation)

- 명시적 확률 모델을 통해 불확실성 정량화
- 잔차 분석(ACF, QQ-plot, Ljung-Box test)으로 모델 가정 검증
- 신뢰구간과 가설검정을 통한 성능의 수치적 입증
- 규제 심사에서 요구되는 "통계적으로 유의한 성능"을 객관적으로 제시
- 예시: 정확도 94.5% [95% CI: 93.8–95.2], LoD = 35.2, 선형성 R^2 = 0.992, $p<0.001$

결정론적 재현성 (Deterministic Reproducibility)

- 목적함수가 완전히 명시되어 동일 입력에 대한 동일 출력이 재현적으로 확인됨
- Black-box ML과 달리 모든 계산 과정이 수학적으로 추적 가능
- 알고리즘 변경 시 어떤 부분이 어떻게 바뀌었는지 명확한 documentation

임상적 해석력 (Clinical Interpretability)

- 각 파라미터 변화가 진단 결과에 미치는 영향을 정량적으로 설명
- 민감도 분석을 통해 어떤 요인이 최종 결과에 가장 큰 영향을 미치는지 파악
- 임상의가 결과를 해석하고 의사결정에 활용할 수 있는 명확한 근거 제공

4.4 기술 스택

- 언어: Python (C++ 포팅 고려), MATLAB (Legacy 알고리즘 역공학)
- 라이브러리: NumPy, pandas, pyarrow (최소화 정책)
- 통계적 방법론: mechanistic modeling, composite function optimization, Joint parameter estimation, residual analysis, white noise testing

4.5 성과

- 달성 성과
- Legacy MATLAB 코드의 완전한 reverse engineering 및 문서화 80% 완료
  - Data Engineer팀의 C++ 포팅을 위한 상세 기술 명세서 제공
  - 통계적으로 엄밀한 hybrid modeling 개선안 설계 완료
- 조직적 한계
  - BT(Biotechnology) 부서들의 새로운 방법론에 대한 저항 우려
  - 팀장 차원에서 개선안 도입 거부 결정
  - 기존 워크플로우 유지를 우선하는 조직 문화로 인한 혁신 제약

4.6 Lesson Learned

- 시스템적 사고와 최적화 전략
  - 복잡한 rule-based 시스템의 근본적 한계를 systematic bias와 비선형 상호작용 관점에서 분석
  - 개별 구성요소 최적화가 아닌 전체 시스템의 global optimization 필요성 인식
  - 결합추정을 통한 통합적 접근법이 순차적 최적화보다 우수함을 이론적으로 확립
- 도메인 특화 모델링 역량
  - 메카니스틱 모델과 통계적 방법론을 결합한 hybrid modeling의 실무 적용성 확인
  - FDA 규제 환경에서 설명력과 성능을 동시에 만족하는 방법론 설계 경험
  - 생물학적 현상을 수학적 모델로 정확히 표현하여 도메인 전문가와의 소통 개선
- 레거시 시스템 분석 및 리엔지니어링
  - 문서화되지 않은 복잡한 시스템을 체계적으로 역공학하는 방법론 정립
  - 파라미터와 10+ 보정 단계의 상호의존성을 논리적 플로우로 재구성하는 분석 역량
  - 기존 시스템의 한계를 정량적으로 진단하고 통계적 근거를 바탕으로 개선 방향 제시
- 조직 변화 관리와 기술 도입 전략
  - 기술적 우수성과 조직 수용성 간의 균형점 탐색: BT 부서의 저항과 팀장의 리스크 회피 성향 경험
  - 통계적 개념(편향 누적, 민감도 분석)을 비전공자에게 전달하는 커뮤니케이션 역량의 중요성
  - 기술적 완성도보다 stakeholder buy-in과 점진적 변화 전략이 실행 성공의 핵심 요소

5 진단 장비 QC 프로세스 자동화 및 알고리즘 고도화

So-What: 공정 단축과 비용 절감을 동시에 달성하는 자동화 QC 파이프라인 확립.

5.1 프로젝트 개요

- 소속: Seegene
- 기간: 2020.12 ~ 2021.09 (9개월)
- 참여 인원: 데이터 사이언티스트 1명, Full Stack 개발자 3명, 기계공학자 4명, 특허 담당자 3명
- PCR 진단 시약을 타사 장비 공급업체의 장비에 탑재
- 진단 서비스 결과의 정확도를 위해 **2 Step 장비 QC 프로세스를 통해 장비의 성능 평가**
- 프로젝트의 목적
  1. 부정확한 **QC 알고리즘 개선**
  2. 투입 리소스가 많은 **QC프로세스 과정을 간소화**시켜 현업의 부담을 경감
- **2 Step QC Process**
  - QC Step 1: 자사 시약에 맞게 장비간 **신호 Scale Calibration**
  - QC Step 2: 장비의 성능을 평가하여 **합격/불합격 분류** - 문제점 발생
    - 엑셀을 이용한 **수동검사**, 비효율적인 **데이터 및 장비 추적 관리**
    - 수동 검사 과정에서 신호의 증폭 크기에 따라 **왜곡된 QC 결과** 발생
    - **기계 결함 및 휴먼 에러 구별 불가**
- 역할: Data Scientist & Project Manager
  - 문제정의 및 QC 프로세스 및 알고리즘 개선 방향 제시
  - 데이터 분석 파이프라인 구축
  - 프로젝트 진행 및 추진
  - 프로젝트 결과 보고서 작성 및 특허 출원

Figure 9: 기존 수동 QC 프로세스(병목·왜곡 발생)

5.2 솔루션 설계 및 전략

- Data Engineering: 산재된 **Excel QC data** ETL
- QC Step2의 **장비 성능 평가 지표를 생성**하여 장비 성능 측정 고도화
- **합격/불합격 분류** 뿐만 아니라 **장비 등급을 차등 부여**하여 고객사에 차등 공급
- 시간에 따른 **장비의 성능을 지속적으로 모니터링**하여 장비의 성능 분석
- **QC Process 간소화**
  - QC Step 1 데이터를 통해 QC Step 2 결과를 예측하는 **딥러닝 모델 개발**
  - 예측 결과로 장비성능이 Fail로 확실시 되는 장비에 한해서 QC Step 2 검사 진행
  - Web App 로 분석 결과 및 시각화 Dashboard 제공
  - 실무 담당자가 데이터 업로드 하면 자동으로 분석 결과 제공

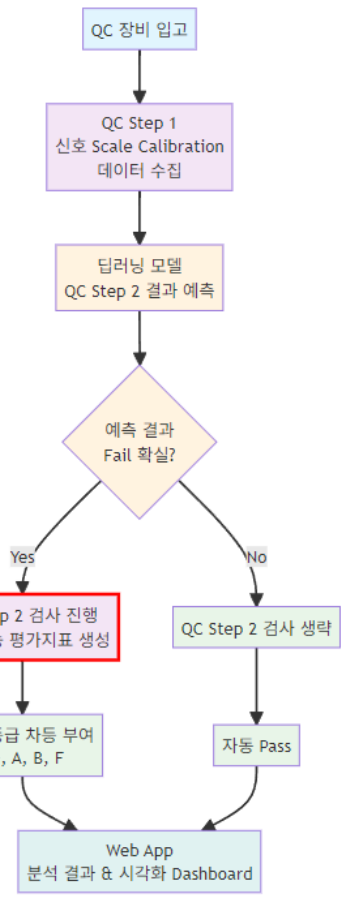

Figure 10: 개선된 자동화 QC 프로세스(QC시간 14배 단축)

5.3 기술 스택 및 요구 역량

- 데이터 엔지니어링: QC Data ETL
- 머신러닝: Clustering (PCA, t-SNE, DBSCAN), Anomaly Detection (Isolation Forest), Outlier Detection (IQR, Z score, 3-Sigma Rule)
- 딥러닝: PyTorch (LSTM), scikit-learn
- 통계/신호처리: SNR, RSS 계산, 시계열 분해 등
- 웹앱 개발: R Shiny (대시보드 및 시각화)
- 도메인 지식: PCR 기술, 의료기기 QC, 통계적 공정관리, 광학 장비 성능 평가

5.4 결과

- PCR기기 2201대를 2552번의 실험해서 만들어진 61,248개의 신호 데이터 확보
- QC Process Step 2 장비 성능 평가 메트릭 생성
  - 신호 증폭 효율성 측정
  - SNR (Signal to Noise Ratio) 측정
  - 기준선 안전성 측정
  - 광학 균일성 측정
  - 장비 온도 균일성 측정
  - 음성 신호 추세 측정
  - 양성 신호 노이즈 측정
  - 시계열 분해 기반 노이즈 측정
  - Outlier 및 Anomaly Data 탐지로 labeling (IQR. Z score, PCA, t-SNE, DBSCAN, 3-Sigma Rule, Isolation Forest)
  - 신호 RSS (Residual Sum of Squares) 측정
- 평가 메트릭 QC 등급 분류: Pass (A+,A,B), Fail (F)

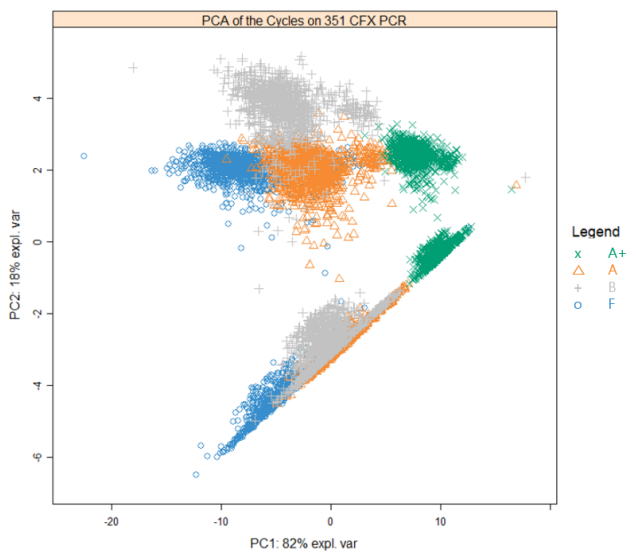

Figure 11: 장비 성능 클러스터링(등급화·공급 최적화 근거)

- LSTM을 활용한 Step 1 데이터를 통한 Step 2 결과 예측 모델 개방
  - 합격/불합격 분류 정확도: 94.5%
  - 장비 성능 등급 분류 정확도: 82.7%

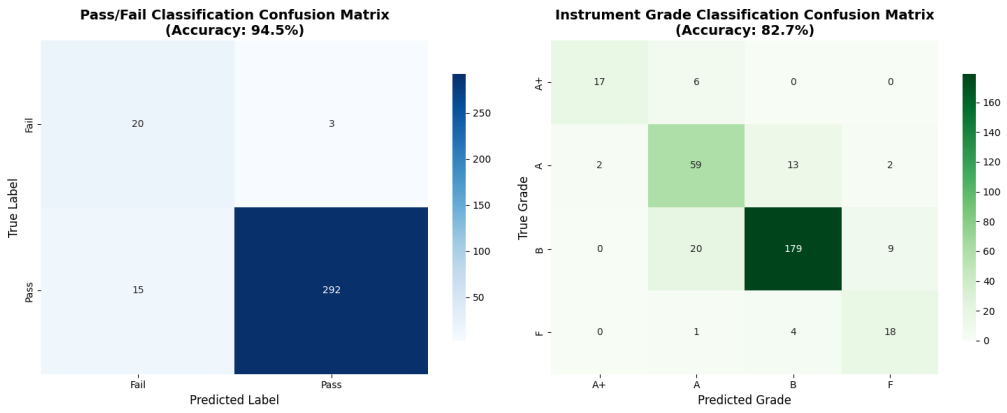

Figure 12: LSTM 혼동행렬(합불 94.5% 정확도)

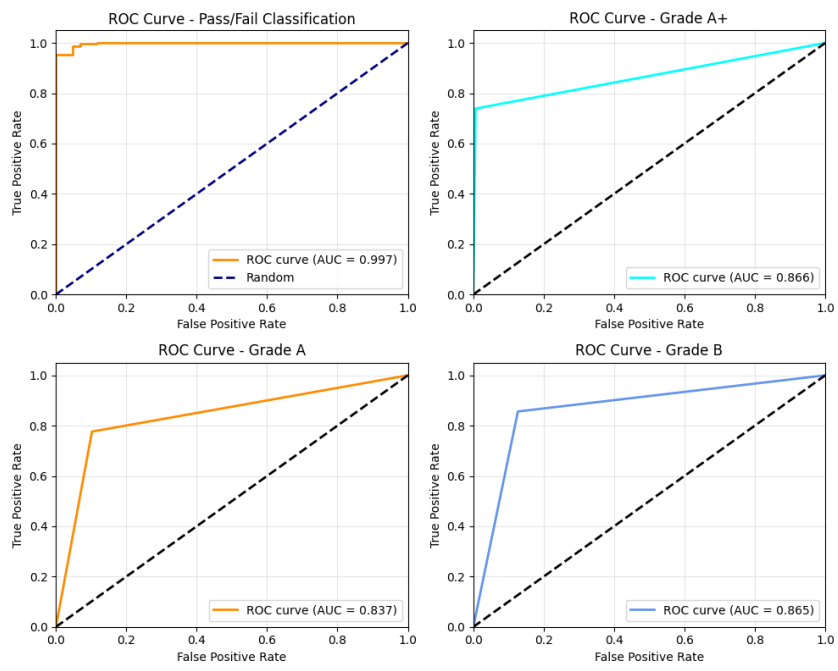

Figure 13: LSTM ROC(등급 분류 82.7% 정확도)

- Web App Dashboard Prototype 개발
  - 실무자가 데이터 업로드 하면 자동으로 분석 결과 제공
  - 시각화 및 데이터 관리 기능 제공
- 총괄장 R&D 부문 우수상 수상 및 2개의 특허 출원

5.5 기대효과

- 편의성 증가: QC시간 약 14배 감소
  - (As-Was: 약 400시간/100대) vs (As-Is: 약 28시간/100대)
- 웹 기반 자동화 플랫폼 제공
  - 연간 비용 약 13배 감소 (QC 시간 및 약 6억원의 비용 감소)
- Mechanical Engineers의 신기술 개발 지원

6 치매 Biomarker 규명: 대사체 통계 분석 및 머신러닝 방법론 비교 연구

So-What: 고차원·소표본 환경에서 임상 타당성 높은 후보를 통계·ML 융합으로 도출.

6.1 프로젝트 개요

- 소속: The Taub Institute, Columbia University Irving Medical Center (CUIMC)
- 2차 세계 대전 후 Baby Boomer 세대의 대규모 치매 발병에 대비한 치매 인자 규명
- 모집단: 장수마을에 거주하는 백인 참여자 (LLFS - Long Life Family Study)
- 기간: 2018.12 ~ 2020.05 (18개월)
- 역할: Research Statistician & Data Scientist
- 참여 인원: 23명 (Epidemiologists, 생물통계학자, 유전통계학자, 신경외과 의사, 생화학자, 임상연구 코디네이터, 데이터 엔지니어)

6.2 주요 문제점 및 도전과제

- 의학적 문제점
  - 알츠하이머병은 증상 발현까지 20년간 진행되어 전임상 단계 생리학 이해 필요
  - 유전적 요인(ex. APOE)이 알츠하이머병에 50% 기여하나 구체적 메커니즘 불명
- 데이터 분석적 도전과제
  - 다중 센터 데이터 수집으로 인한 데이터 품질 불일치
  - 고차원 데이터 (약 3,000개 변수) 대비 작은 표본 크기 (146개 관측치)
  - Mass Spectrometry 데이터의 복잡한 결측치 및 이상치 패턴
  - Multiple testing으로 인한 1종 오류 증가 위험
- 연구 환경적 제약
  - 8개월간 연구소에서 파악하지 못한 강력한 교란자 존재
  - 유전체, 전사체, 단백질체, 대사체 다층 분석 필요성
  - 다학제 팀 간 소통 및 결과 공유 복잡성

6.3 솔루션 설계 및 전략

- 데이터 수집 및 전처리
  - 다중 센터 (New York, Boston, Pittsburgh, Denmark) 혈액 샘플링
  - Mass Spectrometry를 통한 대사체 데이터 디지털화
  - 체계적 데이터 품질 관리 프로세스 구축

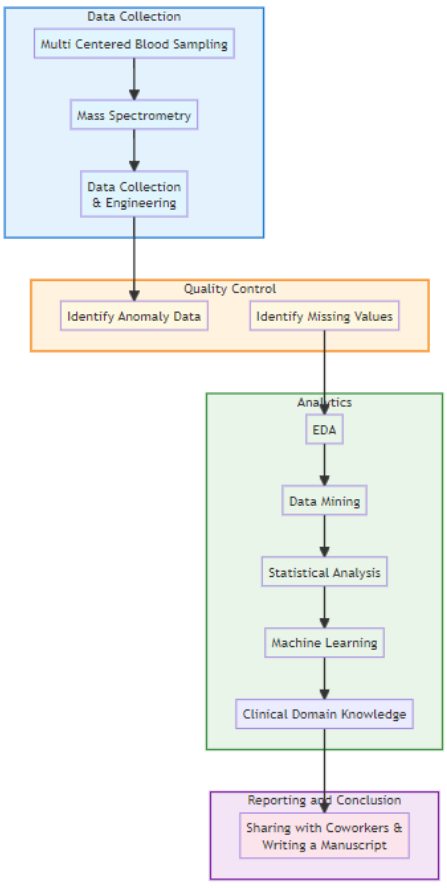

Figure 14: 다층 오믹스 데이터 파이프라인(품질관리·전처리 체계)

- 데이터 품질 관리 (QC) 전략
  - 이상치 및 결측치 식별: 생화학자와의 협업을 통한 실험실 기준 적용
  - Missing Value Analysis: MCAR, MAR, MNAR 분류 및 처리 방안 결정
  - 데이터 포함/제외 기준: rowwise 및 columnwise 결측치 비율 5% 기준 적용
  - 데이터 전처리: Log transformation 및 Standardization
- 3단계 분석 파이프라인
  - 1단계 - EDA & Data Mining & 임상 지식: 패턴 발견 및 교란자 규명
  - 2단계 - Statistical Analysis: 다변량 회귀분석 및 생존분석
  - 3단계 - Machine Learning: 다양한 알고리즘 성능 비교 및 최적 모델 선택

6.4 기술 스택 및 요구 역량

- 프로그래밍: R (Tidyverse, ggplot2, mixOmics, survival, glmnet, etc.)
- EDA: Student t-tests, Wilcoxon Mann-Whitney tests, χ^2 tests, Fisher Exact Tests, ANOVA, Kruskal-Wallis Tests
- 통계분석: Linear/ Logistic regression, Cox Proportional Hazards(PH) regression, Linear Mixed Effect Model, GEE, GWAS, Multiple Comparison Tests
- 머신러닝: Lasso, Ridge regression, Elastic net, Random forests, AdaBoost, Gradient boosting, SVM, PLS, Sparse PLS
- 데이터 마이닝: KNN, PCA, K-means clustering, DBSCAN
- 통계적 검증: Permuted p-values를 통한 multiple Testing correction
- 도메인 지식: 생리학, 생화학, 신경의학, 역학, 생물통계학, 유전학
- 협업 역량: 다학제 팀 (의사, 생물학자, 생물정보학자, 역학자) 소통

6.5 결과

- Sparse PLS 선택 (ML): variable extraction & selection, 해석가능성 확보
- Cox PH Model 선택: 알츠하이머병의 본질적 특성(시간 의존, right censoring, semi-parametric) 반영
- GEE (Stat): 가족 구성원 간 유전적 상관관계 존재, 같은 가족 내 구성원들의 Working correlation이 틀려도 일관된 추정값 산출
- Permutation Test 적용 (Stat): 작은 표본 크기에서 parametric assumption 위반 시 robust한 통계적 추론
- 기존 8개월간 미지의 강력한 교란자를 EDA 및 Data Mining을 통해 3주 만에 규명
- 약 3,000개 대사물질 중 약 60개가 질병과 5% 유의수준에서 유의한 관련성 있음
- 약 60개 중 13개의 대사물질이 질병과 1% 이하의 유의수준에서 유의한 관련성 있음
- GWAS를 통해 유의한 대사물질과 유의한 관련이 있는 유전자 규명
- Sparse Partial Least Squares가 최적 성능의 분류기로 선정 (분류 정확도 84%)

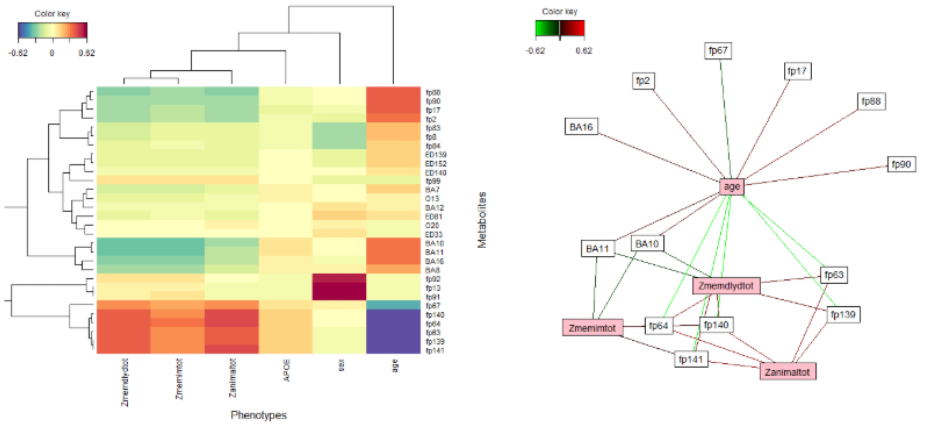

Figure 15: 질환-대사체 상관 구조 탐색(강력 교란자 규명)

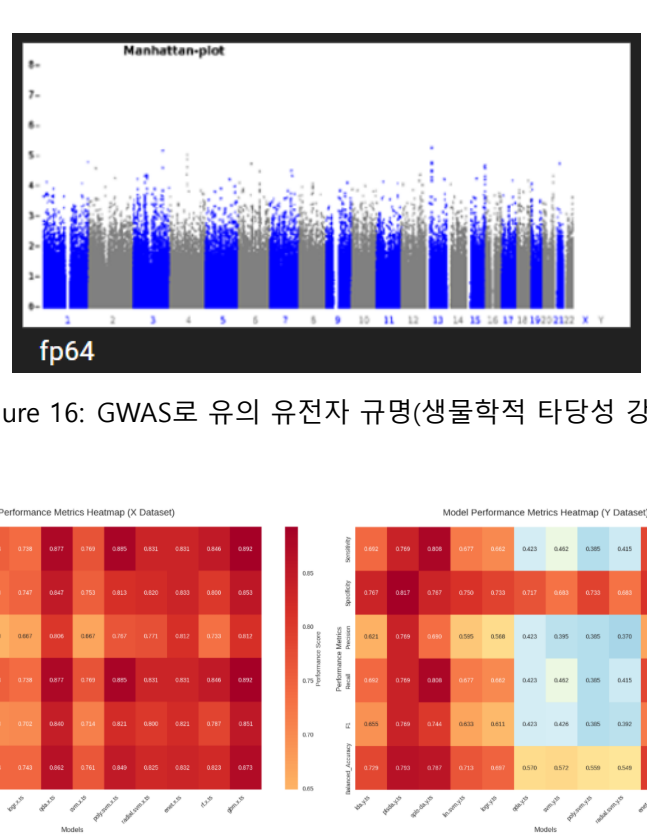

Figure 16: GWAS로 유의 유전자 규명(생물학적 타당성 강화)

Figure 17: ML 분류기 성능 비교(최적 sPLS, 84% 정확도)

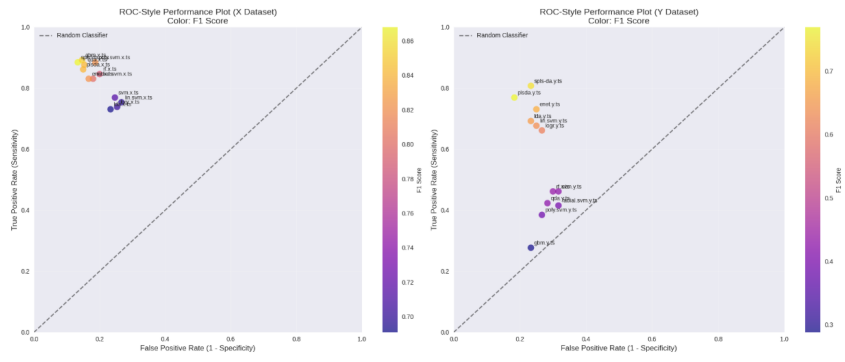

Figure 18: ML Classifier ROC(분류 성능 시각화)

6.6 성과 및 기대효과

- Columbia University Mailman School of Public Health 연례 연구 발표회 포스터 발표
- 연례 연구 경진대회에서 약 100명의 대학원생 중 상위 3명 선정, 상금 $1,000 및 학과장상 수상
- Columbia University Irving Medical Center 신경외과 정규직 Job Offer 획득
- 고차원 대사체 데이터에 대한 체계적 분석 파이프라인 구축
- 다학제 협업 연구 모델 및 분석 방법론 제안