Data Scientist/ Data Analyst Portfolio

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

Kwangmin Kim

2025-07-11

[자연어 처리를 활용한 Data Governance 시스템 단계적 구축]

프로젝트 개요

- 기간: 2024.11 ~ 진행중 (Phase 1 완료, Phase 2-3 진행중)
- 역할: Technical Lead & 데이터 표준화 담당자

주요 도전과제 및 해결방안

- 16개 부서 53개 DB의 데이터 불일치로 인한 데이터 활용도 저하 문제 해결
- PCR 분자진단 실험 데이터의 표준화 및 품질 관리 체계 구축
- 수작업 기반 데이터 검증을 자동화하여 효율성 향상 및 일관성 확보
- 도전과제 1: 복잡한 명명 규칙의 프로그래밍 구현
  - 문제: 14단계 영문 약어 생성 규칙을 포함한 200여개 세부 규칙의 자동화
  - 해결방안
    - 계층적 Rule Engine 설계
    - 성능 최적화: 벡터화된 문자 분류로 개별 처리 대비 3배 성능 향상
    - 확장 가능한 아키텍처: Abstract Base Class 기반 규칙 검증기 설계
- 도전과제 2: 대규모 데이터의 품질 평가 자동화
  - 문제: 수천 개 테이블/컬럼에 대한 실시간 품질 지표 산출
  - 해결방안
    - 명명법 일관성 비율(NCR), 메타데이터 완결성(MCR) 등 10+ 지표
    - 16개 부서별 준수율 비교 분석
    - 배치 처리 아키텍처: pandas 기반 대용량 데이터 처리 최적화
    - 다차원 품질 지표 체계: 일관성/완전성/정확성/유효성 4개 차원
    - 실시간 모니터링: Streamlit 대시보드를 통한 시각화
- 도전과제 3: 도메인 지식과 기술의 융합
  - 문제: PCR 분자진단 도메인 특화 용어와 일반 IT 표준의 조화
    - 관용어 vs 표준화 원칙
    - SI 단위계 vs 업계 관행
  - 해결방안
    - 12개 도메인 그룹 분류 체계 설계
    - 예외 처리 프레임워크: 관용어 허용 목록 및 우선순위 규칙
    - 업무 전문가와의 협업을 통한 도메인 사전 구축

솔루션 설계 및 전략

Phase 1: 규칙 기반 시스템 □ 완료 DataLoader → TokenProcessor → AbbreviationManager → RuleAnalyzer → ReportGenerator ↓ ↓ ↓ ↓ ↓ Excel 처리 토큰화/문자분류 약어생성/중복해결 규칙검증 품질보고서 Phase 2: 딥러닝 확장 □ 진행중

BiLSTM + Attention: 신규 용어 도메인 그룹 자동 분류 S-BERT: 유사 필드명 탐지 및 클러스터링 훈련 데이터: 기존 표준화 규칙 기반 자동 생성

Phase 3: 자동화 워크플로우 □ 계획중

Apache Airflow: ETL 파이프라인 자동화 실시간 품질 모니터링 및 알림 시스템

□ 핵심 설계 원칙

모듈화: 각 기능별 독립적 클래스 설계로 유지보수성 확보 확장성: Abstract Base Class 기반으로 새로운 규칙 추가 용이 성능: 대용량 데이터 처리를 위한 벡터화 및 배치 처리 재사용성: 다른 도메인 적용 가능한 범용 프레임워크 설계

기술 스택 및 요구 역량

- 데이터 처리: Python (pandas, numpy)
- 자연어 처리: Pytorch (BiLSTM + Attention, S-BERT, NLTK, konlpy)
- 시각화 & 모니터링: Streamlit, matplotlib/seaborn, plotly
- 워크플로우 자동화: Apache Airflow
- 기술적 역량
  - 복잡한 비즈니스 규칙의 알고리즘 설계 능력
  - 대용량 데이터 처리 최적화 경험
  - 딥러닝 모델 설계 및 훈련 데이터 생성 능력
- 업무적 역량
  - 데이터 표준화 체계 수립 (표준화 원칙 및 표준 사전 구축)
  - 도메인 전문가와의 협업 및 요구사항 분석
  - 표준화 정책 수립 및 이행 관리
  - 경영진 및 업무 담당자와의 커뮤니케이션 능력

결과 및 성과

- 정량적 성과: 데이터 품질 향상
  - 표준화 체계 수립: 0% → 100% (최초 구축)
  - 메타데이터 완전성: 29.6% → 100% (80.4% 개선)
  - 부서 간 데이터 일관성: 8.4% → 100% (91.6% 개선)
- 효율성 개선
  - 물리명 규칙 검증 시간: 수동 4시간 → 자동 2.3초 (95% 단축)
  - 약어 생성 정확도: 63% → 100% (37% 향상)
- 시스템 구축 성과
  - 표준화 프레임워크 구축
  - 품질 평가 프로그램 핵심 모듈 8개 클래스 구축
  - 표준화 세부 규칙 200여개 생성
  - 품질 지표 15개 개발 및 자동 산출 체계 구축

- 정성적 성과
  - 조직 차원: 16개 부서 통합 데이터 표준 확립
  - 데이터 기반 의사결정 지원 체계 마련
  - 표준화 정책 수립 및 거버넌스 체계 구축
- 기술적 성취
  - 복잡한 언어학적 규칙의 프로그래밍 구현
  - 한국어/영문 혼재 환경에서의 NLP 적용
  - 확장 가능한 품질 관리 프레임워크 설계

기대효과 및 장기 계획

- 단기 기대효과 (6개월)
  - 운영 효율성
  - 데이터 통합 작업 시간 50% 단축
  - 신규 시스템 구축 시 표준 적용 자동화
  - 데이터 품질 이슈 사전 예방 체계 확립
- 비즈니스 가치
  - 실험 데이터 신뢰성 향상으로 연구 효율성 증대
  - 부서 간 데이터 공유 활성화
  - 규제 대응 및 감사 준비 시간 단축

[알고리즘 안전성 인허가 통계 분석 보고서 작성 반자동화]

프로젝트 개요

- 북미 진단 시장 진출을 위한 알고리즘 안전성 검증용 통계 분석 문서 작성 반자동화
- 의료 장비 및 시약 제품의 글로벌 진출 시 각국 정부의 규제 사항 존재
  - 시약의 안정성 검증
  - 장비의 안정성 검증
  - **진단 알고리즘의 안정성 검증**
- EU: IVDR (In Vitro Diagnostics Regulation) 준수 필요
- 북미: FDA(미국) 및 Health Canada의 세계 최고 수준 엄격한 기준 충족 필요
- 기존 Software Engineering 방식보다 더 엄격한 **Advanced Testing** 요구
- 소프트웨어 및 알고리즘 규제 강화 추세

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

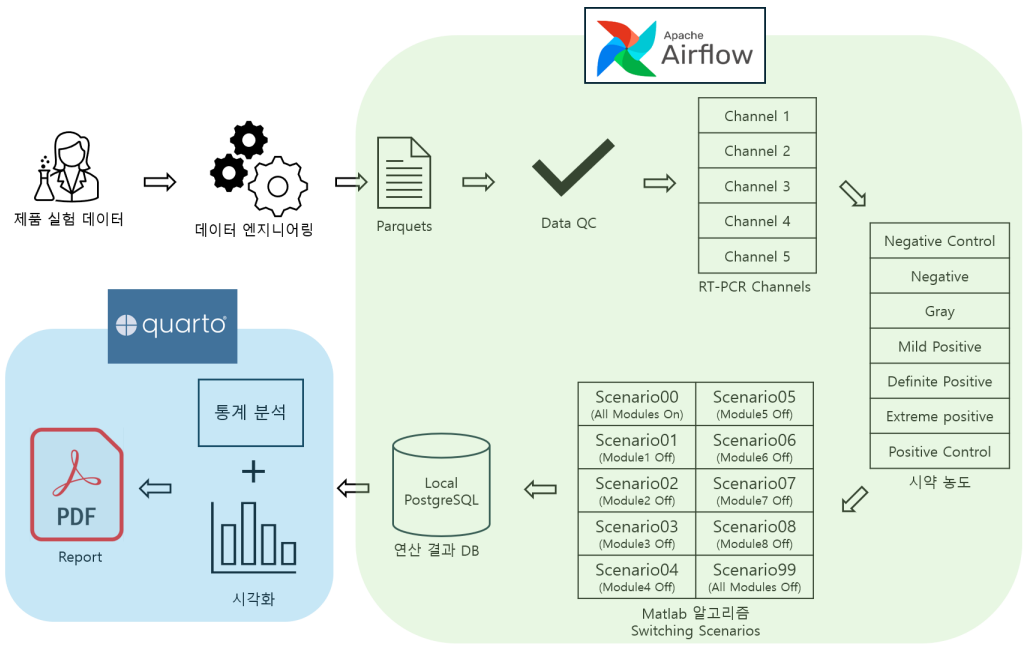


Figure 1: Data Pipeline

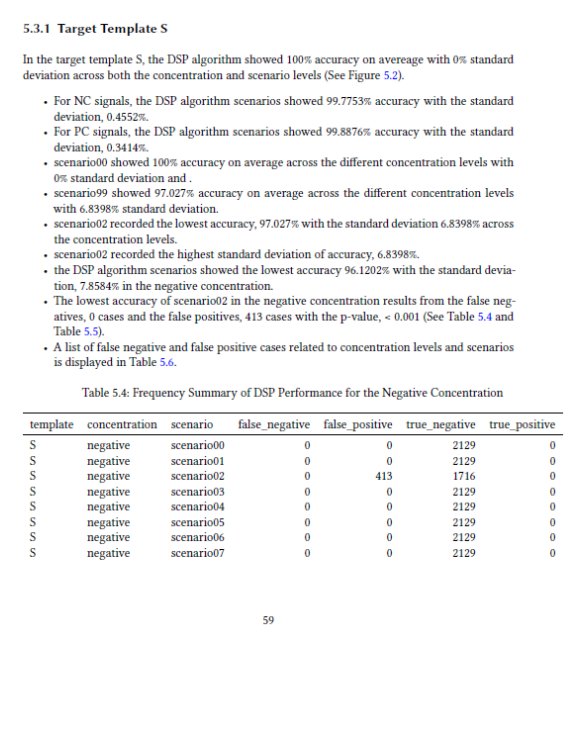
주요 도전과제 및 해결방안

- 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
  1. 오타 교정
  2. 결측치 처리
  3. 이상 데이터 처리
  4. 알고리즘 데이터 정합성 1차 검정
  5. 알고리즘 데이터 정합성 2차 검정

기술 스택 및 요구 역량

- 규제 지식: FDA Software Validation
- 통계 분석: Statistics (2-Way Repeated Measures ANOVA), Clinical Study Design
- 프로그래밍: R (Statistical Testing), Python (Engineering), Matlab (진단 알고리즘)
- 워크플로우: Apache Airflow
- 문서 자동화: Quarto
- 도메인 지식: Biology

결과 및 성과



5.3.1 Target Template S

In the target template S, the DSP algorithm showed 100% accuracy on avereage with 0% standard deviation across both the concentration and scenario levels (See Figure 5.2).

- For NC signals, the DSP algorithm scenarios showed 99.7753% accuracy with the standard deviation, 0.4552%.
- For PC signals, the DSP algorithm scenarios showed 99.8876% accuracy with the standard deviation, 0.3414%.
- scenario00 showed 100% accuracy on average across the different concentration levels with 0% standard deviation and .
- scenario99 showed 97.027% accuracy on average across the different concentration levels with 6.8398% standard deviation.
- scenario02 recorded the lowest accuracy, 97.027% with the standard deviation 6.8398% across the concentration levels.
- scenario02 recorded the highest standard deviation of accuracy, 6.8398%.
- the DSP algorithm scenarios showed the lowest accuracy 96.1202% with the standard deviation, 7.8584% in the negative concentration.
- The lowest accuracy of scenario02 in the negative concentration results from the false negatives, 0 cases and the false positives, 413 cases with the p-value, < 0.001 (See Table 5.4 and Table 5.5).
- A list of false negative and false positive cases related to concentration levels and scenarios is displayed in Table 5.6.

Table 5.4: Frequency Summary of DSP Performance for the Negative Concentration

template	concentration	scenario	false_negative	false_positive	true_negative	true_positive
S	negative	scenario00	0	0	2129	0
S	negative	scenario01	0	0	2129	0
S	negative	scenario02	0	413	1716	0
S	negative	scenario03	0	0	2129	0
S	negative	scenario04	0	0	2129	0
S	negative	scenario05	0	0	2129	0
S	negative	scenario06	0	0	2129	0
S	negative	scenario07	0	0	2129	0

59

(continued)

template	concentration	scenario	false_negative	false_positive	true_negative	true_positive
S	negative	scenario08	0	0	2129	0
S	negative	scenario99	0	413	1716	0

Figure 2: Report

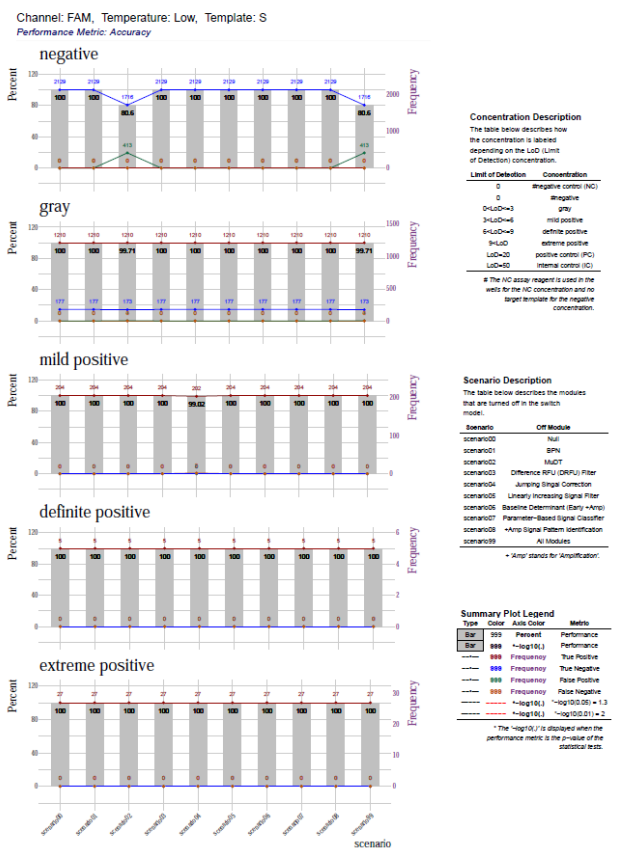


Figure 5.2: DSP Accuracy Performance in the Target Template S

Figure 3: Plot

- DSP Algorithm 결과물
  - FDA 제출용 verification & validation report 초안 완성
  - 문서화 시스템: 업무 소통 및 RDB 시스템 구축을 위한 자동화 시스템
  - 데이터 관리 시스템: Data Quality Control System 구축
  - FDA Validation Model: DSP 알고리즘 전용 Validation Model 확립
  - 특허 발명: FDA Validation Model 관련 특허 출원
  - 성능 평가 체계: 사내 최초 알고리즘 및 시약 제품 종합 성능 평가
  - 리스크 관리 통계 분석: 시약/장비 고유 효과 및 교란 요인 위험 관리 분석

기대효과

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

[진단 장비 QC 프로세스 자동화 및 알고리즘 고도화 프로젝트]

프로젝트 개요

- PCR 진단 시약을 타사 장비 공급업체의 장비에 넣어 검출 결과를 얻는다.
- 진단 서비스 결과의 정확도를 위해 **2 Phase 장비 QC 프로세스를 통해 장비의 성능을 평가**한다.
- 프로젝트의 목적: 1. 부정확한 **QC 알고리즘 개선** 2. 투입 리소스가 많은 **QC프로세스 과정을 간소화**시켜 현업의 부담을 경감
- 프로젝트 기간: 9개월
- **Two Step QC Process**
  - QC Step 1: 자사 시약에 맞게 장비간 **신호 Scale Calibration**
  - QC Step 2: 장비의 성능을 평가하여 **합격/불합격 분류** - 병목 현상 발생
  - 문제점
    - 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**

Figure 4: 기존 QC 프로세스

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 및 life cycle 관리
- **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

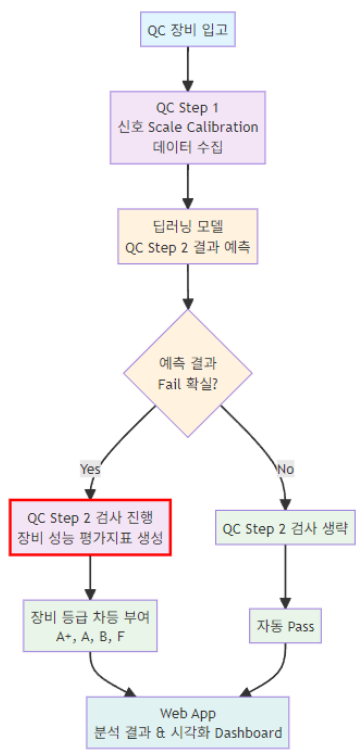


Figure 5: 개선된 QC 프로세스

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 (BiLSTM), 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

결과

- ETL 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 기반 합격/불합격 장비 분류, 장비 성능 4등급 부여
  - Pass (A+,A,B), Fail (F)

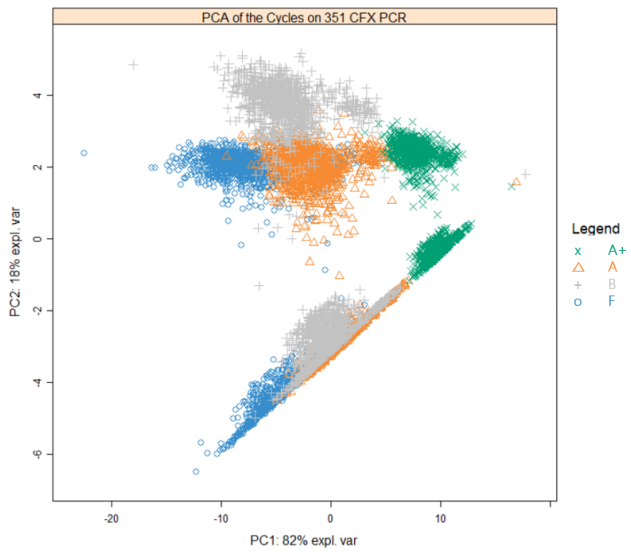


Figure 6: 장비 성능별 클러스터링

- BiLSTM을 활용한 Step 1 데이터를 통한 Step 2 결과 예측 모델 개방
  - 합격/불합격 분류 정확도: 99.3%
  - 장비 성능 등급 분류 정확도: 91.7%

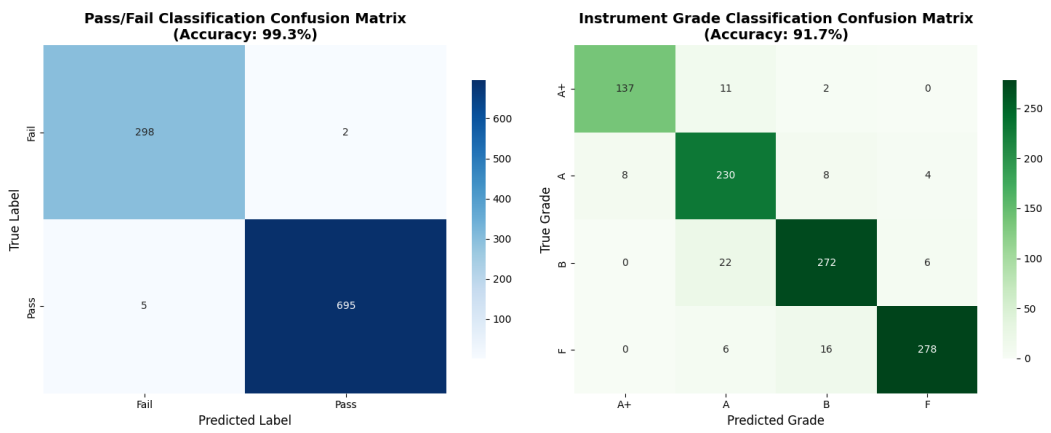


Figure 7: BiLSTM Confusion Matrix

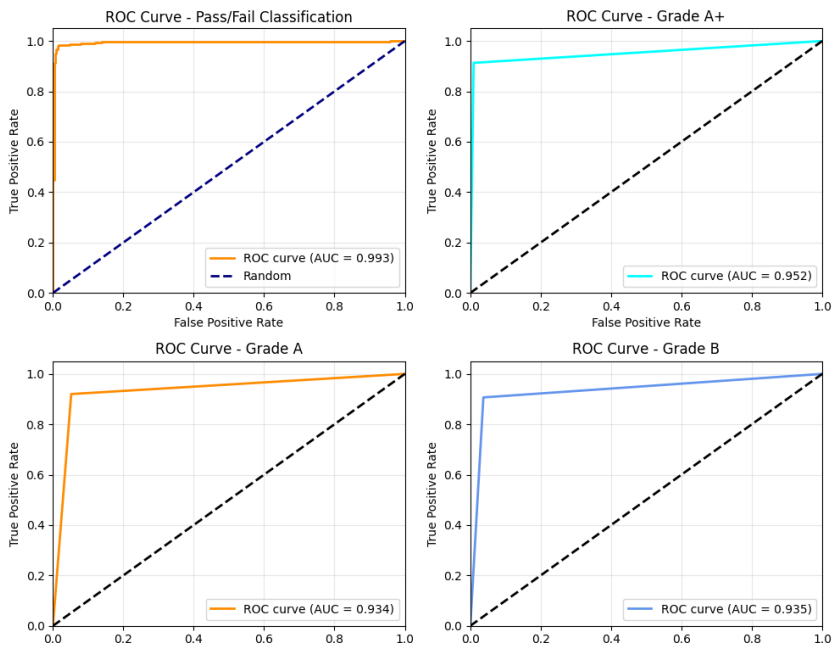


Figure 8: BiLSTM ROC Curve

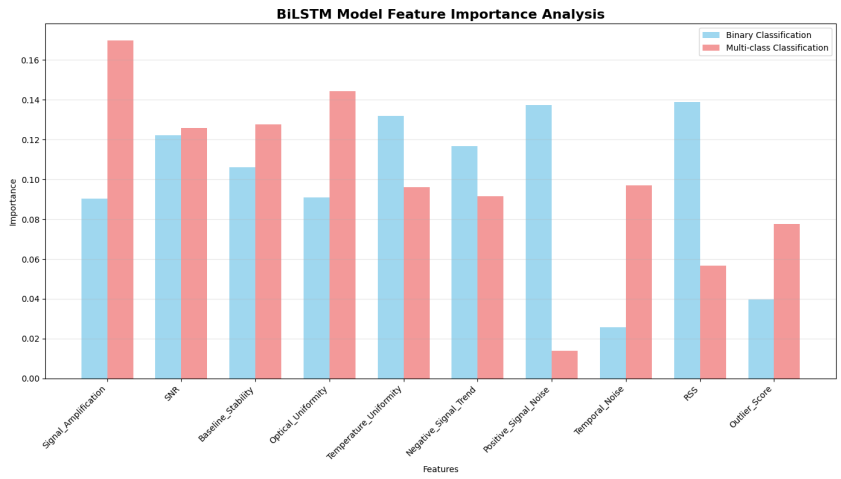


Figure 9: BiLSTM Importance

- Web App Dashboard Prototype 개발
  - 실무자가 데이터 업로드 하면 자동으로 분석 결과 제공
  - 시각화 및 데이터 관리 기능 제공

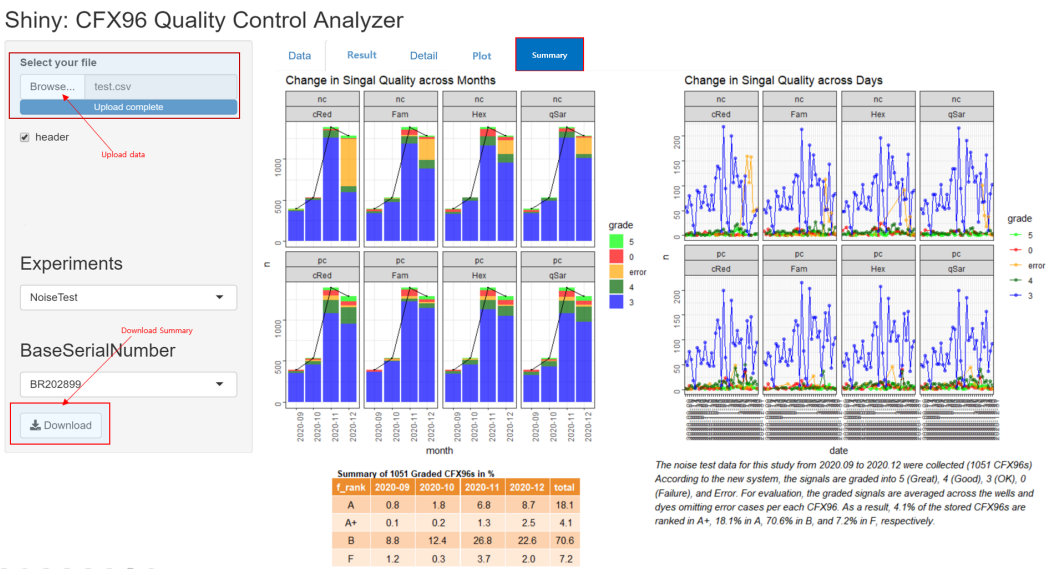


Figure 10: WebApp

- 총괄장 R&D 부문 우수상 수상
- 2개의 특허 발명을 출원

기대효과

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