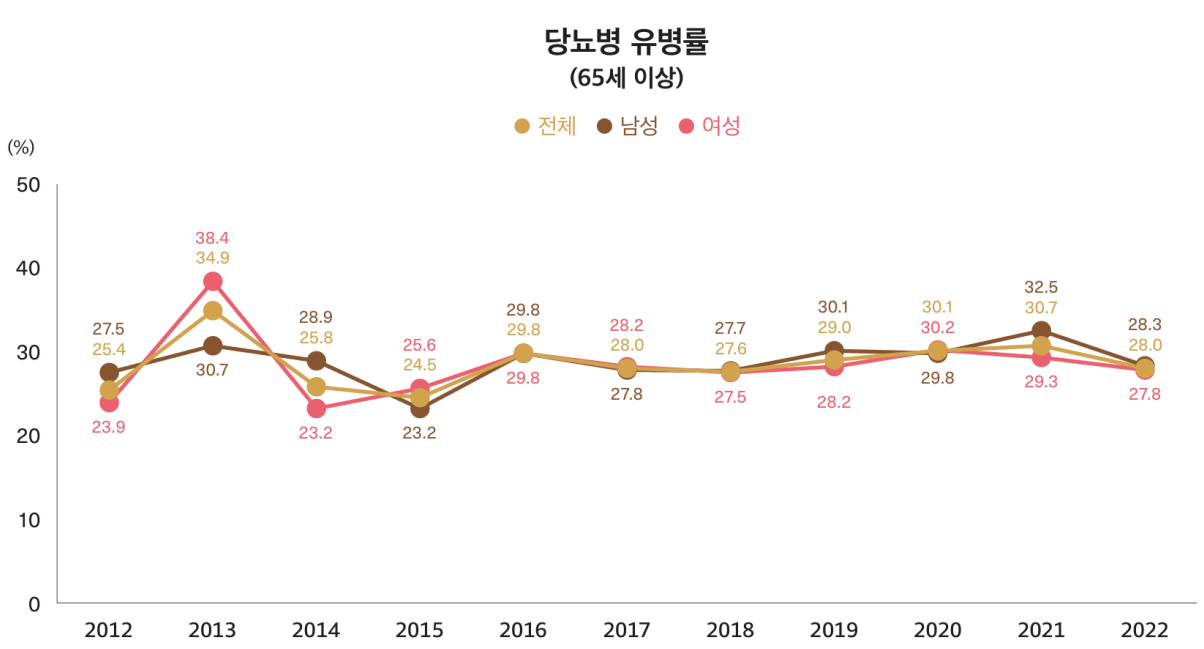
RNA 기반 당뇨 진단을 통한 사후 치료와 선제적 예방의 통합 서비스 개발

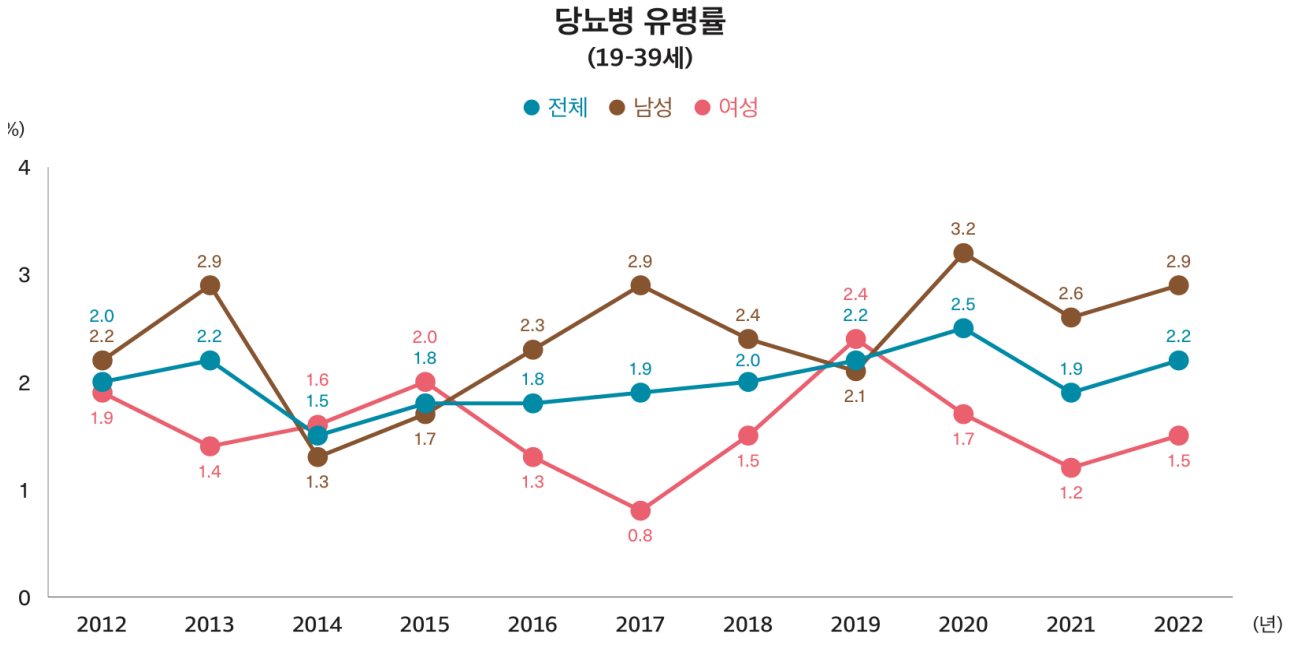
푸른콩 김대엽, 김호중, 이연우, 한성주

Nvidia AI 솔루션 경진대회

주제 선정 배경

당뇨병: 공복혈당≥126 mg/dL 또는 당뇨병 약물치료 (인슐린 포함) 또는 의사진단 또는 당화혈색소≥6.5%





당뇨 진단이 늦어지면 합병증 발생 위험이 높아져 의료비 증가와 삶의 질 악화를 초래함

RNA 진단을 통해 당뇨를 조기에 발견하여 생활 습관 개선을 통해 예방하는 것을 목표로 한다.

출처: DIABETES FACT SHEET IN KOREA 2024

Direct Medical Costs for Patients with Type 2 Diabetes and Related Complications: A Prospective Cohort Study Based on the Korean National Diabetes Program

목적 및 필요성

Persona A – 숨겨진 고위험군

기본 정보
34세, 사무직
규칙적인 건강검진 참여

현행 검사 결과: 모두 정상
공복혈당 95mg/dL
혈압, 지질, 비만 모두 정상

Insight
혈당, 혈압 등 임상데이터로 판단했을 때 정상 군에 속하는 경우
겉보기에는 정상이지만 RNA는 조용히 진행중인 발병 초기 신호를 감지

Persona A – 방치되었을 경우

기본 정보
혈당 조절 악화 - 당뇨 전단계 진입
동반질환 고혈압 발생

경제적 영향
매년 140만원의 의료비 지출
건강보험심사평가원 통계 중

Insight
혈당, 혈압 등 임상데이터로 판단했을 때 정상 군에 속했던 경우
RNA 기반 예측으로 미리 개입했다면 막을 수 있었던 발병

당뇨의 증상과 관리

제1형 당뇨
자가면역 질환. 내 몸의 면역 세포가 췌장의 베타세포를 적으로 오인해 파괴
인슐린 자체가 생성되지 않음

제2형 당뇨
인슐린은 분비되는데, 세포가 인슐린 신호를 무시하거나 적게 반응
초기에는 인슐린을 더 많이 뿜어내며 버티지만 결국 췌장이 기능이 떨어짐

솔루션 가치

Persona A와 같은 고위험군 1명을 조기 예방할 때 얼마만큼의 경제적 효과가 발생하는가?

$$\text{B/C ratio} = \frac{\text{발병확률(model precision 0.8)} \times \text{예방 성공률(DPP 0.58)} \times \text{손실비용(합병증 없는 당뇨 환자 기준 3년 의료비 4,188,300원+연속혈당측정기 3,000,000원)}}{\text{진단비 200,000원 + 예방 관리비(DPP 3년 집중관리비용 기준 2,048,975원)}}$$

B/C ratio = 1.48

Persona A – 숨겨진 고위험군

기본 정보
34세, 사무직
규칙적인 건강검진 참여

현행 검사 결과: 모두 정상
공복혈당 95mg/dL
혈압, 지질, 비만 모두 정상

Insight
혈당, 혈압 등 임상데이터로 판단했을 때 정상군에 속하는 경우
겉보기에는 정상이지만 RNA는 조용히 진행중인 발병 초기 신호를 감지

최소 1.48배의 미래 손실 방지 효과

Benefit(편익)
3년간 미래 손실 방지 금액

Cost(비용)
3년간 예방 투자비용 (진단비 + 예방 관리비)

항목	기존 혈당검사	RNA기반 분석
발견시점	혈당 상승 이후	혈당 상승 이전
발견 가능 단계	2, 3단계	1단계
놓치는 사람	정상처럼 보이는 고위험군	조기 포착 가능
개입	치료 중심	예방 중심

RNA기반 당뇨 조기진단의 효과성

아키텍쳐

RNA 기반 분류

GSE164416
GSE25724
GSE81608
GSE164416
GSE86468
GSE86469

데이터 셋

RNA 데이터 분석

ML VS Deep Learning

Bio Informatics

React

React Web/App → FASTAPI → SQLite DB

AI Models
YOLO - best.pt
PaddleOCR

- Food/Drink 자동 인식
- Kalman Filter 기반 다중 객체 추적
- OCR 기반 성분 분석 + 위험도 산출

Docker - container

Object Detection+Nutrition Analysis

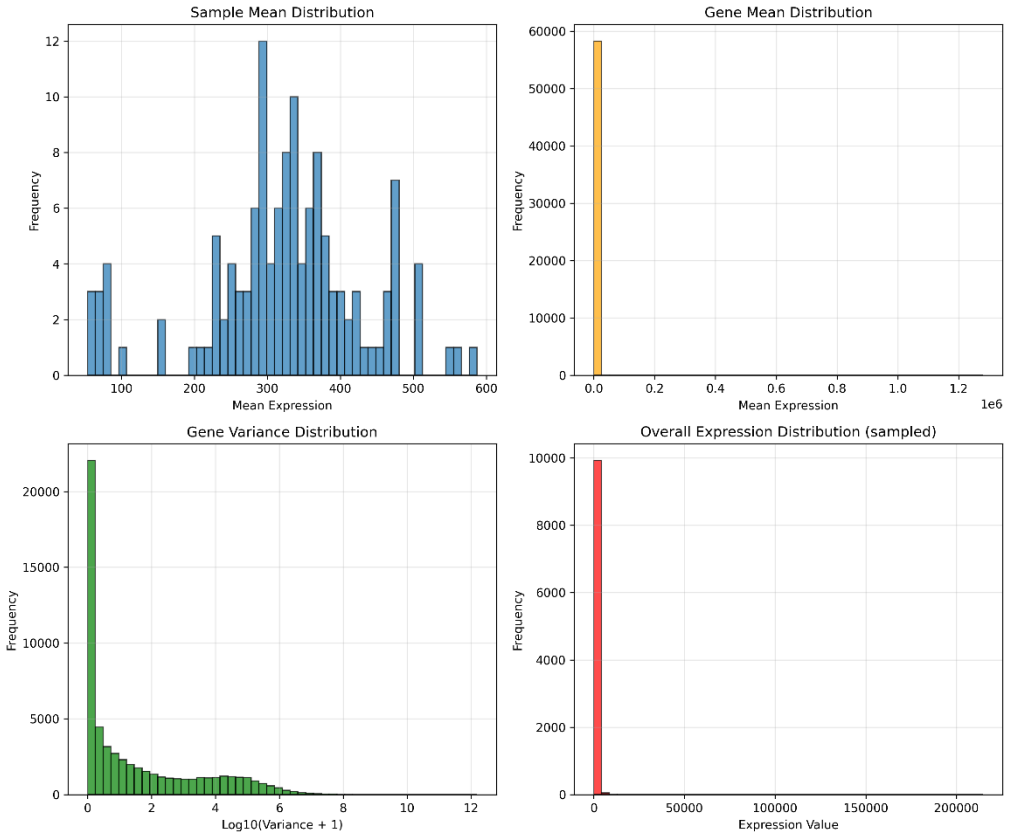
음식/음료/약 감지(YOLO)

- 트래킹 및 디바운싱(Kalman)
- 성분 정보 파싱
- 라벨 텍스트 인식(OCR)
- ArUco 기반 칼로리 계산

→ 당뇨 위험도 산출

GSE 데이터 - Gene Expression Omnibus

GSE 데이터셋은 NCBI가 운영하는 세계 최대의 유전체 데이터 아카이브인 GEO에 등록된 하나의 연구 단위를 식별하는 고유 ID



GPL

-> 어떤 기술로 측정했는가?

Affymetrix 마이크로어레이 칩인지, Illumina HiSeq과 같은 RNA-Seq 장비인지

GSM

-> 개별 샘플이 무엇인가

당뇨 환자 1번의 혈액 샘플', '정상인 2번의 췌장 조직 샘플'과 같이 실험의 최소 단위

Series matrix file

-> 개별 샘플이 무엇인가

해당 샘플에서 해당 유전자가 얼마나 발현되었는지를 나타내는 '숫자'

→ RNA 기반 당뇨 조기 진단

RNA는 생명체의 유전 정보를 단백질로 합성하는 데 관여하는 핵산의 일종

데이터 소개

사용한 데이터 셋

GSE164416

대상 조직: 전혈 (Whole Blood)

플랫폼: RNA-Seq (Illumina NovaSeq 6000)

구성: 제2형 당뇨병 환자와 정상 대조군의 혈액 전사체 데이터.

GSE76894

대상 조직: 췌장 랑게르한스섬 (Human Islets)

플랫폼: RNA-Seq (Illumina HiSeq 2000)

구성: 제2형 당뇨(T2D) 19명 vs 정상(ND) 84명 (총 103 샘플)

GSE25724

대상 조직: 췌장 랑게르한스섬 (Human Islets)

플랫폼: Microarray (Affymetrix)

구성: 제2형 당뇨(T2D) 6명 vs 정상(ND) 7명

GSE86468

대상 조직: 췌장 랑게르한스섬 (Human Islets)

플랫폼: RNA-Seq

구성: 당뇨(T2D) 및 정상(Control) 기증자의 췌장 조직.

GSE81608

대상 조직: 췌장 랑게르한스섬 (Human Islets)

플랫폼: RNA-Seq (Illumina HiSeq 2000)

구성: 제2형 당뇨(T2D) vs 정상(ND)

GSE86469

대상 조직: 인간 췌장 랑게르한스섬 (Human Islets)

플랫폼: RNA-Seq (Illumina HiSeq 2500)

구성: 제2형 당뇨병환자와 비당뇨 기증자의 췌장 조직 전사체

Bio Informatics

gene	control_mean	igt_mean	t2dm_mean	fc_control_igt	fc_igt_t2dm	fc_control_t2dm	p_control_igt	p_igt_t2dm	p_control_t2dm	trend
ENSG00000151834	60.2857143	35.97368421	4.256410256	-0.729052014	-2.814349021	-3.543401035	0.050892	7.07E-06	8.84E-09	decreasing
ENSG00000081181	727.952381	686.8157895	240.0512821	-0.083802338	-1.512682132	-1.59648447	0.722828	3.14E-10	3.40E-06	decreasing
ENSG00000172575	1091.7619	1038.763158	467.0769231	-0.071724153	-1.151437397	-1.22316155	0.742245	7.85E-08	6.87E-06	decreasing
ENSG00000165195	1212.47619	1057.815789	500.025641	-0.196694188	-1.079495271	-1.276189458	0.353829	1.00E-08	7.13E-06	decreasing
ENSG00000145888	166.238095	156.2105263	82.35897436	-0.089205698	-0.915288387	-1.004494085	0.754084	0.002581	1.38E-05	decreasing
ENSG00000050165	176.047619	451.1842105	530.1282051	1.352773174	0.232149528	1.584922702	0.001393	0.312508	1.44E-05	increasing
ENSG00000255829	22.7142857	21.39473684	10.76923077	-0.082596701	-0.928139684	-1.010736385	0.707345	3.12E-05	2.27E-05	decreasing
ENSG00000069424	54.6190476	90.97368421	154.2307692	0.725642091	0.755121514	1.480763605	0.009056	0.001011	3.18E-05	increasing
…	…	…	…	…	…	…	…	…	…	…

Progressive gene changes

Control → IGT → T2DM 순서대로 점진적으로 발현량이 변하는 유전자들

Control : 정상인(정상 대조군)

IGT : 당뇨병 전단계 (정상과 당뇨병 사이)

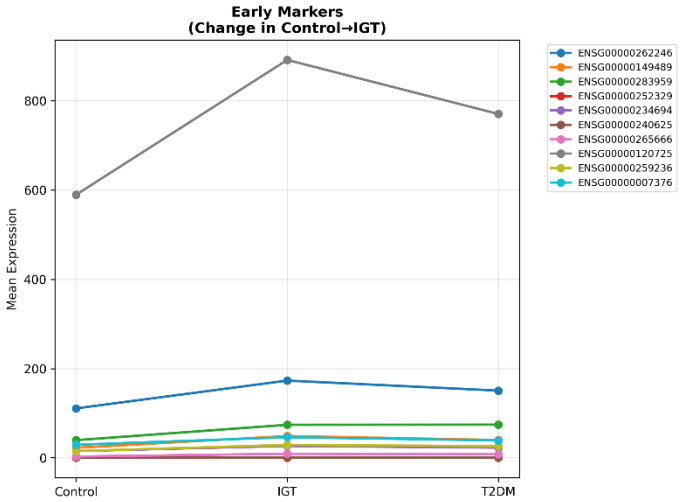
T2DM : 질병이 완전히 확립된 단계

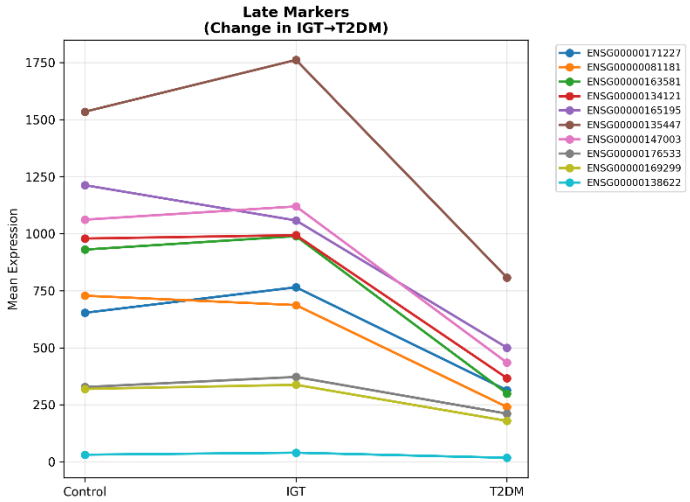
질병 진행의 분자적 증거

Increasing pattern: 염증, 스트레스 반응 유전자들이 많음. 췌장이 계속 스트레스 받으면서 염증 신호가 점점 커짐

Decreasing pattern: 인슐린 분비, 포도당 대사 관련 유전자들. 췌장 베타세포 기능이 점점 떨어지는 증거.

Bio Informatics - Early vs Late markers





Control ----[Early Change]----> IGT ----[Late Change]----> T2DM

Early Markers
용도: 조기 선별 검사
타이밍: 아직 IGT도 아닌 정상인을 검사

Late Markers
용도: 질병 중증도 평가, 치료 효과 모니터링
타이밍: 조기 당뇨 환자

Early marker

Control과 IGT를 비교한 p-value가 0.05보다 작음
Control에서 IGT로 갈 때 fold change 절댓값이 0.5보다 큼
IGT와 T2DM을 비교한 p-value가 0.1보다 큼

Late marker

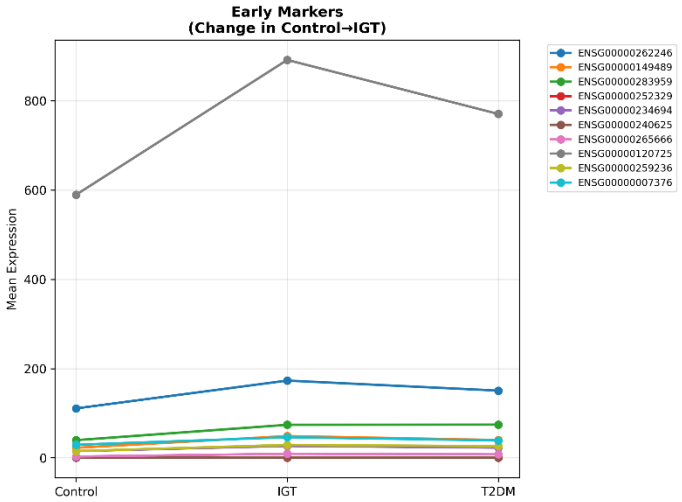
Control과 IGT를 비교한 p-value가 0.1보다 큼
IGT와 T2DM을 비교한 p-value가 0.05보다 작음
IGT에서 T2DM으로 갈 때의 fold change 절댓값이 0.5보다 큼

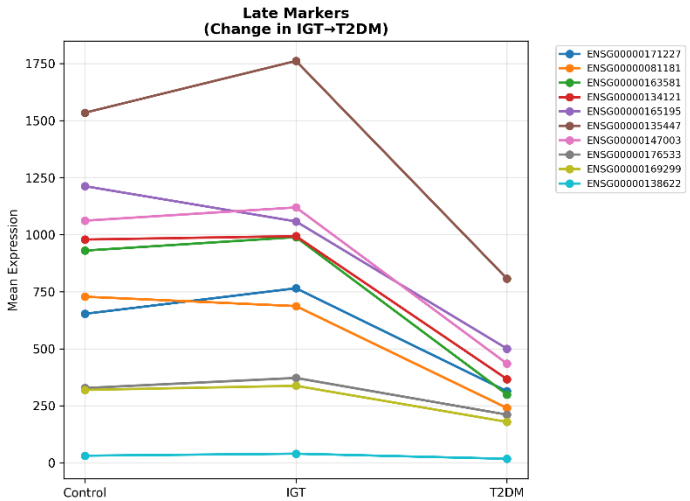
gene	fc_control_igt	fc_igt_t2dm	p_control_igt	p_igt_t2dm	control_mean	igt_mean	t2dm_mean
ENSG00000262246	0.638917	-0.1993	0.000152	0.13903	110.8095	173.1053	150.641
ENSG00000149489	1.08388	-0.28216	0.000155	0.127755	22.47619	48.76316	39.92308
ENSG00000283959	0.88524	0.006873	0.000322	0.970473	39.61905	74.02632	74.38462
ENSG00000252329	-0.85561	0.3083	0.000347	0.172639	1.285714	0.263158	0.564103
ENSG00000234694	0.782584	-0.21741	0.000349	0.180402	15.04762	26.60526	22.74359
ENSG00000240625	0.729352	0.021419	0.000516	0.923942	0.142857	0.894737	0.923077

- gene: 유전자 ID
- fc_control_igt: Control → IGT fold change (log2)
- fc_igt_t2dm: IGT → T2DM fold change (log2)
- p_control_igt: Control vs IGT p-value
- p_igt_t2dm: IGT vs T2DM p-value
- control_mean: Control 그룹 평균 발현량
- igt_mean: IGT 그룹 평균 발현량
- t2dm_mean: T2DM 그룹 평균 발현량

우리가 분석하는 샘플은 총 98개
Control 그룹이 21명 GT(내당능 장애) 그룹이 38명 그리고 T2DM(제2형 당뇨병) 그룹이 39명

Bio Informatics - Early vs Late markers

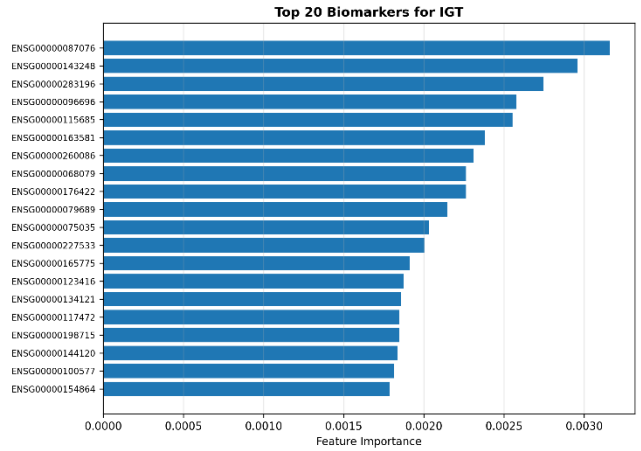


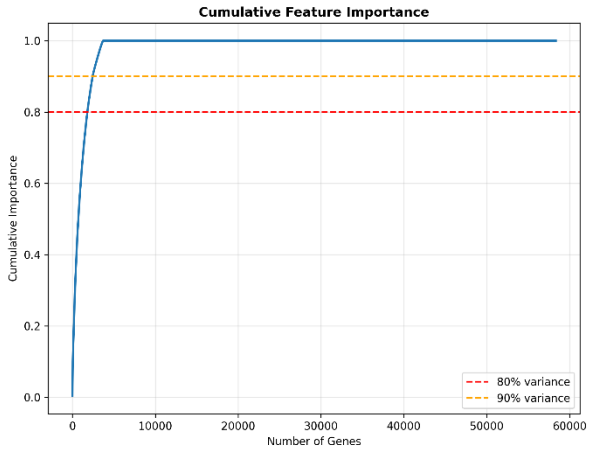


Control ----[Early Change]----> IGT ----[Late Change]----> T2DM

Early Markers
용도: 조기 선별 검사
타이밍: 아직 IGT도 아닌 정상인을 검사

Late Markers
용도: 질병 중증도 평가, 치료 효과 모니터링
타이밍: 조기 당뇨 환자





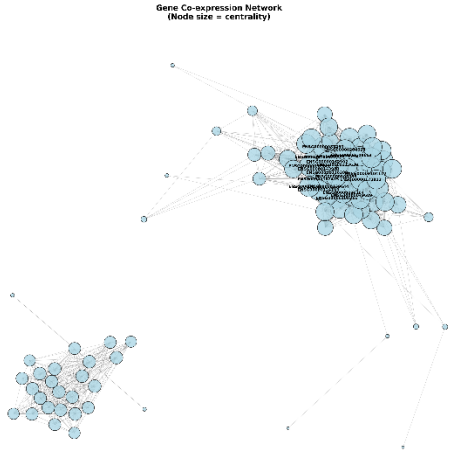
Control ----[Early Change]----> IGT ----[Late Change]----> T2DM

Random Forest Classifier가 5000개 유전자를 학습해서
IGT를 가장 잘 구분하는 유전자를 찾아낸 결과

IGT를 가장 잘 구분하는 유전자가 0~5000개 사이에 몰려서 존재

우리가 분석하는 샘플은 총 98개 정도
Control 그룹이 21명 GT(내당능 장애) 그룹이 38명 그리고 T2DM(제2형 당뇨병) 그룹이 39명

Bio Informatics



유전자 공동발현 네트워크

유전자들 간의 발현 패턴 유사도를 계산해서 네트워크로 연결

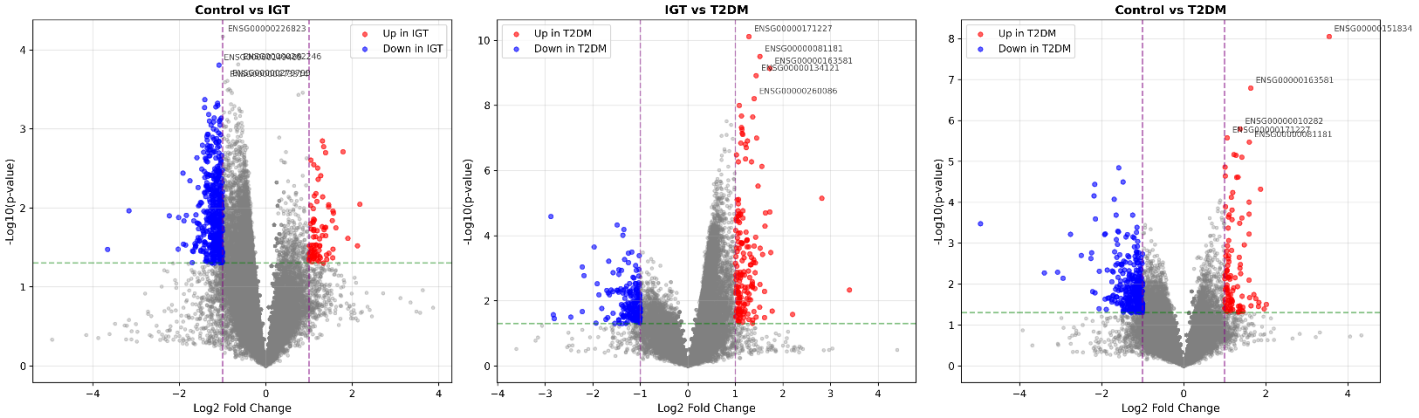
유전자 A와 B가 항상 같이 높아지거나 낮아지면 (|correlation| > 0.7), 둘을 선으로 연결

2만 개 유전자 중에서 가장 분산이 큰 상위 100개 유전자를 선택하는 것으로 시작

degree centrality, 즉 연결 중심성 각 유전자가 몇 개의 다른 유전자와 연결되어 있는지 수

betweenness centrality, 중개 중심성 네트워크에서 다른 노드들을 연결하는 경로에 얼마나 자주 등장하는지를 측정

당뇨의 심화에 따라 주요 RNA는 강한 상관관계를 가지고 변화함을 알 수 있다.



Control vs IGT, IGT vs T2DM, Control vs T2DM

x축은 fold change, y축은 p-value의 음의 로그값

X축: 변화의 양
Y축: 통계적 신뢰도

오른쪽 위 , 왼쪽 위 - 핵심 타겟

통계적으로도 확실하고, 발현량도 크게 변화한 유전자들입니다.

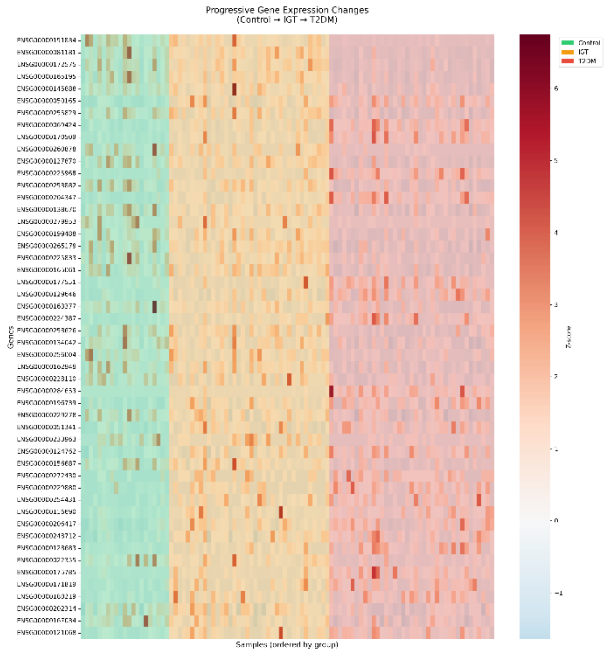
가운데 아래 노이즈

변화량도 적고, 통계적 의미도 없는 대다수의 유전자, "회색 지대"

위쪽 중앙 & 양쪽 아래

위쪽 중앙: 통계적으로는 유의미하지만 변화량이 너무 미미한 유전자
양쪽 아래: 변화량은 큰데 통계적으로 유의하지 않은 유전자

Bio Informatics



Progressive Heatmap

Z-score 정규화: 각 유전자의 평균을 0, 표준편차를 1로 맞춤

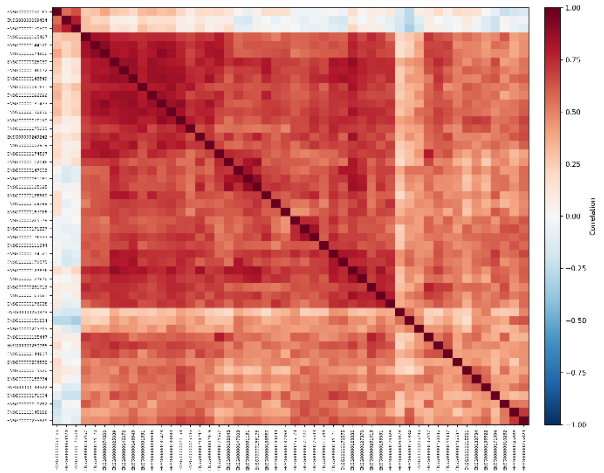
가장 극적으로 변하는 상위 50개 유전자가 세로로 나열

상위 progressive genes들의 발현 패턴을 색깔로 표현

왼쪽에서 오른쪽으로 진해지는 유전자: Control → T2DM으로 가면서 점점 발현이 증가.
왼쪽에서 오른쪽으로 진해지는 유전자: 발현 감소. 베타세포 기능 유전자

Control -> IGT -> T2DM

정상 -> 당뇨전단계 -> 당뇨



Correlation network heatmap

Pearson Correlation Coefficient: 두 변수 X, Y의 선형 관계 강도를 -1에서 1 사이로 표준화한 값

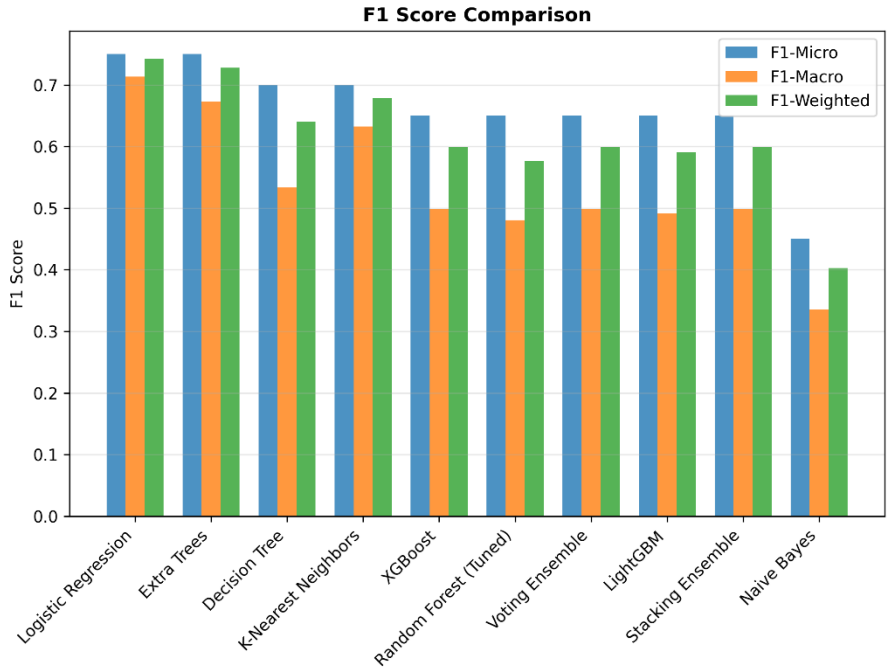
가장 극적으로 변하는 상위 50개 유전자

정상에서 당뇨전단계를 거쳐 당뇨병으로 향하는 RNA패턴이 매우 명확하고
상위 n개의 유전자가 매우 강한 영향력을 보임을 확인 할 수 있다.

RNA 기반 Model 학습

ML Model 성능 결과

Rank	모델	Test Acc	F1-Macro	ROC-AUC	Overfit
1	Logistic Regression	0.75	0.714	0.833	0.250
2	Extra Trees	0.75	0.673	0.834	0.250
3	Decision Tree	0.70	0.533	0.904	0.300
4	K-Nearest Neighbors	0.70	0.632	0.827	0.018
5	XGBoost	0.65	0.499	0.888	0.350
6	Random Forest	0.65	0.481	0.869	0.350
7	Voting Ensemble	0.65	0.499	0.860	0.350
8	LightGBM	0.65	0.492	0.917	0.350
9	Stacking Ensemble	0.65	0.499	0.872	0.350
10	Naive Bayes	0.45	0.336	0.623	0.332



Extra Trees Logistic Regression LightGBM K-Nearest Neighbors

↳ 우수한 성능과 안정성을 보이는 4가지 모델을 딥러닝과 비교해보자.

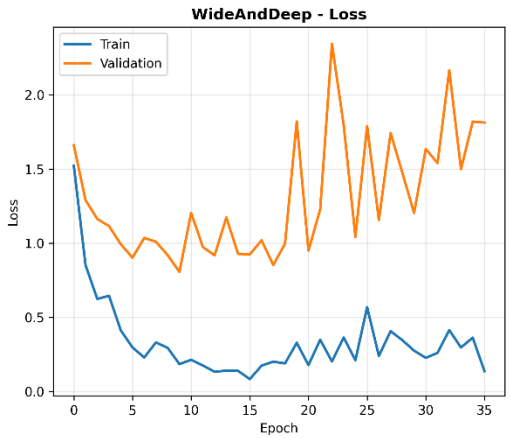
↳ 앞선 분석을 기반으로 데이터의 양과 다양성을 늘리면 더욱 더 좋은 성능을 낼 수 있을 것이라 기대할 수 있다.

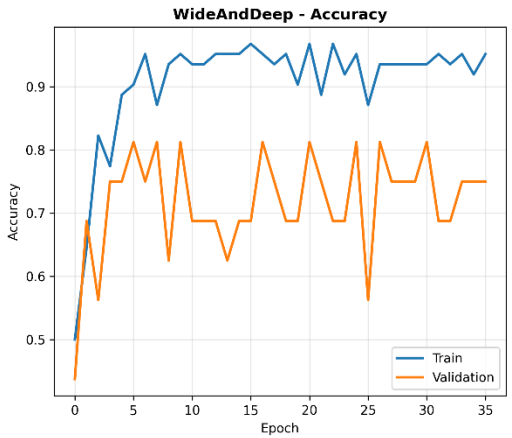
RNA 발현에 따른 분류 문제에서 어떤 모델이 가장 좋은 성능을 보일 것인가?

RNA 기반 Model 학습

Bio Informatics

Rank	Model	Accuracy	F1-Macro	F1-Weighted	ROC-AUC
1	WideAndDeep	0.65	0.580	0.646	0.834
2	Ensemble	0.55	0.000	0.552	0.000
3	AttentionMLP	0.50	0.416	0.499	0.788
4	DeepResNet	0.45	0.343	0.412	0.782

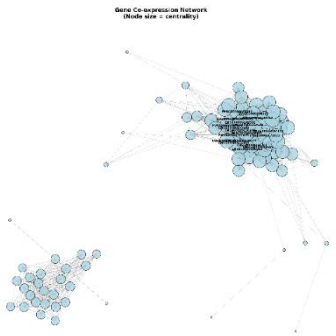




Extra Trees Logistic Regression LightGBM K-Nearest Neighbors WideAndDeep

↳ 우수한 성능과 안정성을 보이는 4가지 모델을 WideAndDeep 비교해 보았을 때 ML모델의 성능이 앞도적

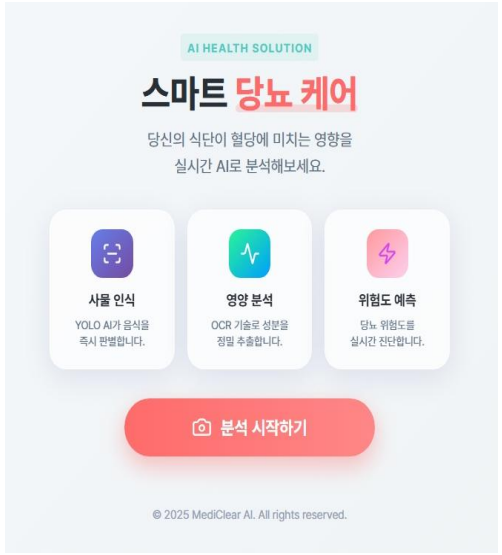
↳ 앞선 분석을 보았을 때 RNA의 발현은 그래프의 형태를 띈다. 다양한 데이터를 기반으로 GNN을 도입 시 좋은 성능이 기대된다.



RNA 발현에 따른 분류 문제에서는 ML모델이 가장 좋은 성능을 이끌어 냄

RNA Seq추출 이후 ML 모델을 통한 분류 -> CV 기반 통한 사후 치료와 선제적 예방의 통합 서비스

Vision 모듈 목표



React + FastAPI

Vision Algorithm

Food/Drink 자동 인식

과일 → 커스텀 YOLO
음료 → COCO 기반 모델로 Drink 클래스 구성

실시간 안정화

Kalman Filter 기반 다중 객체 추적
디바운싱 + 재검증으로 검출 흔들림·깜빡임 제거

크기·거리 기반 칼로리 추정

ArUco 마커 기반 거리 스케일 추정
객체 크기 기반 칼로리 근사 계산

OCR 기반 성분 분석 + 위험도 산출

성분표(당류/탄수화물/지방 등) 파싱
당뇨 위험도 점수 계산

$$K_k = P_{k|k-1}H^\top (HP_{k|k-1}H^\top + R)^{-1}$$

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k(z_k - H\hat{x}_{k|k-1})$$

$$P_{k|k} = (I - K_kH)P_{k|k-1}$$

Kalman Filter 란?

잡음이 섞인 불확실한 측정값들을 이용해, 시스템의 진짜 상태(State)를 가장 최적으로 추정(Estimate)해내는 알고리즘

무엇을 찾아야하는가?

제2형 당뇨 예방의 중심은 "식습관 관리"

정제 탄수화물: 흰쌀, 흰빵, 라면·국수, 감자 등

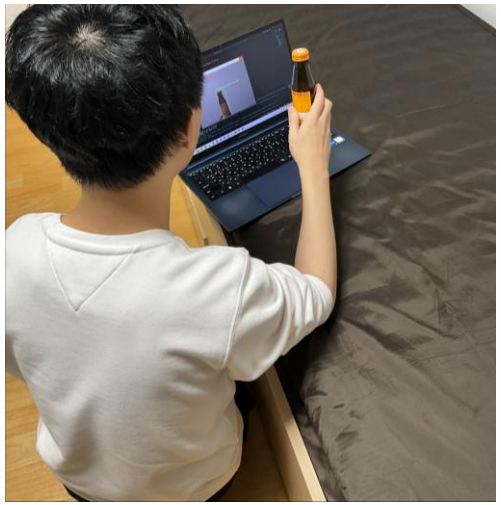
가당 음료: 탄산음료 달달한 과일주스 스낵, 인스턴트

인슐린 과다 + 규칙적인 식사 + 탄수화물 계산이 필수

Paddle OCR

PaddleOCR은 마트폰 앱 내에서 '영양 성분표' 같은 복잡한 실생활 이미지를 빠르고 정확하게 읽어낼 수 있는 OCR 솔루션

Vision 모듈 목표



세연 2학사 한성주 학생의 테스트
비타 500을 들고있다.

Vision Algorithm

Food/Drink 자동 인식

과일 → 커스텀 YOLO
음료 → COCO 기반 모델로 Drink 클래스 구성

실시간 안정화

Kalman Filter 기반 다중 객체 추적
디바운싱 + 재검증으로 검출 흔들림·깜빡임 제거

크기·거리 기반 칼로리 추정

ArUco 마커 기반 거리 스케일 추정
객체 크기 기반 칼로리 근사 계산

OCR 기반 성분 분석 + 위험도 산출

성분표(당류/탄수화물/지방 등) 파싱
당뇨 위험도 점수 계산

$$K_k = P_{k|k-1}H^\top (HP_{k|k-1}H^\top + R)^{-1}$$

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k(z_k - H\hat{x}_{k|k-1})$$

$$P_{k|k} = (I - K_kH)P_{k|k-1}$$

Kalman Filter 란?

잡음이 섞인 불확실한 측정값들을 이용해, 시스템의 진짜 상태(State)를 가장 최적으로 추정(Estimate)해내는 알고리즘

무엇을 찾아야하는가?

제2형 당뇨 예방의 중심은 "식습관 관리"

정제 탄수화물: 흰쌀, 흰빵, 라면·국수, 감자 등

가당 음료: 탄산음료 달달한 과일주스 스낵, 인스턴트

인슐린 과다 + 규칙적인 식사 + 탄수화물 계산이 필수

Paddle OCR

PaddleOCR은 마트폰 앱 내에서 '영양 성분표' 같은 복잡한 실생활 이미지를 빠르고 정확하게 읽어낼 수 있는 OCR 솔루션

비젼모델 버전

버전별 주요 변경 사항

포토그래퍼 22학번 데이터사이언스학부 이연우 학생
모델 22학번 소프트웨어학부 한성주 학생

V0.1 – Drug / Food 프로토타입

YOLO 기반의 정지 이미지 탐지를 구현하여 음식과 약품 구분의 기술적 타당성(PoC)을 검증

정지된 사진을 통해서만 분석이 가능하다는 근본적인 제약이 존재

V0.2 – FoodDetection Live

카메라 스트림 입력을 도입하여 정지 영상에서 실시간 탐지 기능으로 전환

프레임 단위로 탐지가 이루어져 바운딩 박스 떨림이 심하고 기록의 신뢰도가 낮음

V0.3 – YOLO + Kalman Filter 안정화

칼만 필터(Kalman Filter)를 도입하여 객체의 움직임을 예측하고 안정적으로 추적

여전히 성분표가 아닌 제품의 겉모습 기준으로만 위험도를 판단한다는 한계가 있음

V1.0 – FoodDetection_OCR 1차

음료 영역(ROI)에 PaddleOCR을 적용하여 텍스트 추출을 처음으로 시도했습니다.

단일 영역만 크롭하는 방식이라 둥근 병의 곡면이나 중앙에 위치한 텍스트가 누락됨

V1.1 – Center-Merge OCR (듀얼 ROI)

전체 영역과 중앙 영역을 각각 OCR 한 뒤 결과를 병합하는 알고리즘을 적용

이중 연산으로 속도가 느려지고, 저해상도나 작은 글씨 인식에는 물리적 한계 존재

V1.2 – MultiCrop OCR + 영양·칼로리

MultiCrop 및 CLAHE 보정을 통해 라벨 텍스트 회수율을 극대화

제공되는 칼로리는 추정치이며, 정밀한 제품 DB 연동은 차기 버전(V2) 과제로 남겨두었습니다.

Q & A

푸른콩 김대엽 김호중 이연우 한성주