Journal of The Korean Data Analysis Society (October 2025), 27(5), 1501-1515.
https://doi.org/10.37727/jkdas.2025.27.5.1501

의생물학 Tabular 데이터에서 딥러닝과 전통적 머신러닝의 성능 비교*

송채린[1], 이나현[2], 곽일엽[3]

요 약

딥러닝은 영상·자연어 처리 분야에서 혁신적 성과를 거두었으나, 전자의무기록(EMR)과 유전체 검사 결과처럼 표 형식(Tabular)으로 저장되는 의생물학 데이터 영역에서는 여전히 그래디언트 부스팅 계열의 전통적 머신러닝(ML)이 주류로 활용된다. 본 연구는 공개 의생물학 데이터 5종(표본 수 303~212,691건, 수치·범주형 변수 혼합)에 대해 동일 전처리와 Optuna 기반 하이퍼파라미터 탐색을 적용하고, XGBoost, LightGBM, CatBoost 3개의 ML 모델과 ResNet, FT-Transformer, TabNet, Tab-Transformer 4개의 딥러닝 모델의 분류 성능과 계산 효율성을 체계적으로 평가한다. 실험 결과, 1만 표본 미만의 소·중규모 데이터 세트에서는 ML 모델이 일관되게 높은 성능과 빠른 학습 속도를 보였으며, 최대 25.2%p 높은 정확도를 달성하였다. 딥러닝 모델은 대용량 표본(20만 이상)에서 ML 모델과 대등하거나 근소 우위를 보였으나, feature 수 증가에 따른 계산 복잡도 급증으로 효율성이 현저히 저하되었다. 특히 소규모 데이터에서 딥러닝 모델의 효율성은 0.002-1.15 범위로 변동성이 컸으나, 트리 기반 모델(특히 LightGBM, XGBoost)은 안정적으로 높은 효율성을 유지하였다. 결론적으로, 소·중규모 의생물학적 Tabular 문제에는 그래디언트 부스팅 기반 ML 모델이 여전히 안전한 선택이며, 수십만 이상 표본과 충분한 연산 자원이 확보될 경우 Transformer 계열 DL 모델이 제한적이나마 성능 이득을 제공할 수 있음을 확인하였다.

주요 용어 : 의생물학 데이터, Tabular 데이터, 머신러닝, 딥러닝, 벤치마킹.

1. 서론

1.1. 연구 배경 및 목적

의료·생명과학 분야에서 활용되는 인공지능 연구의 상당 부분은 영상·유전체와 같은 비정형 데이터를 대상으로 하지만, 실제 임상 의사결정 과정에서는 전자의무기록(EMR)·검사 결과·생활 습관 정보 등 정형적인 표 형식(tabular) 데이터가 여전히 핵심 근거 자료로 사용된다. 이러한 데이터는 수치형과 범주형 변수가 복합적으로 얽혀 있다는 구조적 특성 때문에, 그동안은 로지스틱 회귀나 그래디언트 부스팅 계열 알고리즘과 같은 전통적 머신러닝 모델이 일관된 우수성을 보여 왔다

*이 성과는 2023년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No.RS-2023-00208284). 송채린과 이나현은 이 연구에 동등하게 기여하였음.
[1]06974 서울시 동작구 흑석로 84, 중앙대학교 응용통계학과 학사과정. E-mail: cleo7987@cau.ac.kr
[2]06974 서울시 동작구 흑석로 84, 중앙대학교 스마트시티학과 석사과정. E-mail: naa012@cau.ac.kr
[3](교신저자) 06974 대한민국 서울특별시 동작구 흑석로 84, 중앙대학교 응용통계학과, 부교수.
E-mail: ikwak2@cau.ac.kr
[접수 2025년 7월 19일; 수정 2025년 8월 28일; 게재 확정 2025년 9월 2일]

(Shwartz-Ziv, Armon, 2022; Grinsztajn et al., 2022). 반면, 이미지·자연어 처리에서 혁신을 이끈 딥러닝이 tabular 영역으로 확장되면서 TabNet, FT-Transformer 등 새로운 모델들이 제안되고 있으나, 의생물학 tabular 데이터에서의 성능 안정성은 아직 많이 연구되지 못했다.

의생물학 dataset은 표본 수가 수백 건에서 수십만 건까지 폭넓게 분포하며, 클래스 불균형과 범주형 변수 비율 또한 크게 달라진다. 이러한 이질적 조건에서 “딥러닝이 전통적 머신러닝을 언제, 어느 정도 대체할 수 있는가?”에 대한 정량적 근거가 부족한 상황이다. 따라서 현업 연구자와 임상의가 데이터 규모·특성에 맞추어 최적의 모델을 선택할 수 있도록, 체계적이고 재현 가능한 비교 연구가 필요하다.

따라서 본 연구는 표본 수가 다른 다섯 개 의생물학 Tabular 데이터를 대상으로, (1) 전통적 머신러닝 알고리즘 중 좋은 성능을 보이는 그래디언트 부스팅 기반 ML 모델 세 개(XGBoost, LightGBM, CatBoost)와 (2) 최근 제안된 tabular 딥러닝 아키텍처 네 개(ResNet, FT-Transformer, TabNet, Tab-Transformer)의 분류 성능을 동일한 전처리와 하이퍼파라미터 탐색 조건에서 체계적으로 비교와 분석을 한다. 각 모델은 Optuna 기반 탐색을 거친 최적 설정에서 세 차례 반복 학습하여 평균 정확도를 산출하였고, 추가적인 복잡도 지표로 학습 시간과 파라미터 규모를 기록하였다. 이를 통해 데이터 규모가 작을 때 머신러닝이 제공하는 안정적 우수성과 데이터가 충분히 클 때 딥러닝이 확보할 수 있는 잠재적 이득을 실증적으로 규명하고, 궁극적으로 의생물학 tabular 과제에서 데이터 크기·특성을 고려한 모델 선택 지침을 제시하고자 한다.

1.2. 연구 범위 및 방법

본 연구는 공개 의생물학 Tabular dataset 다섯 가지를 분석 대상으로 삼았다. 데이터 전처리는 최소 개입 원칙에 따라 수행하였다. 원본 데이터에서 결측치가 포함된 레코드는 삭제하였으며(1% 미만), 범주형 변수는 별도의 정수 인코딩 없이 문자열 상태로 유지한 채, 모델 내부 인코더(CatBoost의 내장 처리기, FT-Transformer의 EmbeddingEncoder 등)를 통해 자동으로 처리되도록 하였다. 이로써 전처리에서 발생할 수 있는 정보 손실이나 알고리즘 편향을 최소화하고, 각 모델이 범주형 데이터를 최적 방식으로 활용할 수 있도록 하였다. 수치형 변수는 표준화(z-score) 없이 원값을 유지해, 알고리즘 간 비교 시 불필요한 편향을 최소화했다. 성능 공정성을 확보하기 위해 Optuna 기반 하이퍼파라미터 탐색(50 trials, TPE sampler)을 전 모델에 적용하고, 탐색에서 얻은 최적 설정으로 3회 반복 학습 후 평균 정확도(accuracy)를 기록하였다. 딥러닝 계열은 channels, num_layers, batch_size, learning_rate, numerical_encoder_type 등을, 머신러닝 계열은 max_depth, n_estimators, learning_rate(GBM류) 등을 탐색 범위에 포함하였다.

2. 관련 문헌 고찰

Tabular Data는 딥러닝보다는 XGBoost(Chen, Guestrin, 2016), CatBoost(Dorogush, Ershov, Gulin, 2018), LightGBM(Ke et al., 2017) 등의 전통적인 ML이 강점을 보이는 분야로 알려져 있다. 하지만, 최근 다양한 딥러닝 기반 방법들이 등장하며 기존의 판도를 바꾸려는 시도가 이어지고 있다.

2.1. Tabular Data를 위한 딥러닝 기반 방법들의 동향

전통적인 머신러닝 모델의 대안으로, Tabular 데이터의 고유한 특성을 학습하기 위해 이미지나 자연어 처리에서 성공을 거둔 딥러닝 아키텍처를 도입하려는 연구가 활발히 진행되고 있다.

초기에는 다층 퍼셉트론(MLP)에 잔차 연결(residual connection)을 적용한 ResNet 기반 모델이 제안되었다. 이후 결정 트리의 작동 방식에 착안하여 해석 가능성을 높인 TabNet(Arik, Pfister, 2021)이나, 자연어 처리의 판도를 바꾼 트랜스포머 아키텍처를 도입한 Tab-Transformer(Huang et al., 2020) 및 FT-Transformer(Gorishniy et al., 2021)와 같은 모델들이 등장했다.

이러한 모델들은 각기 다른 철학으로 Tabular 데이터의 복잡한 상호작용을 학습하고자 하였으나, 데이터의 규모나 특성에 따른 성능 편차가 보고되고 있어 체계적인 비교 연구는 여전히 요구되는 상황이다(Gorishniy et al., 2021; Shwartz-Ziv, Armon, 2022).

국내에서도 머신러닝 및 딥러닝 모델들을 Tabular 데이터와 접목하는 연구들이 활발하게 이루어지고 있다(Nam et al., 2017; Gong, Kim, 2023; Choi et al., 2023). 특히 단순한 적용뿐만 아니라, 데이터 규모별로 성능 차이를 확인하려는 연구도 수행되었다(Yoon, Choi, Kim, 2022).

2.2. 의생물학 분야에서 Tabular Deep Learning 적용

의생물학 분야는 전자의무기록(EMR), 유전체 데이터, 임상시험 결과 등 방대한 양의 Tabular 데이터를 포함하고 있어, 딥러닝 기술의 주요 시험대가 되고 있다. 딥러닝 모델은 수많은 변수 간의 비선형적이고 복잡한 상호작용을 학습하여 질병 예측, 환자 예후 분석, 맞춤형 치료법 제안 등에서 인간의 한계를 넘어설 잠재력을 가진다(Ching et al., 2018).

실제로 최신 Tabular 딥러닝 모델을 의생물학 데이터에 적용한 연구들이 활발히 진행되고 있다. 예를 들어, 일부 연구에서는 중환자실(ICU) 환자의 EMR 데이터를 TabNet에 적용하여 사망률이나 특정 질병 발생 위험을 예측하고, 모델의 해석 가능 기능을 통해 예측에 영향을 미친 주요 임상 지표를 제시하기도 했다. 또한, 트랜스포머 기반 모델들은 시간에 따라 기록되는 환자의 시계열 데이터(longitudinal data)를 분석하여 질병의 진행 과정을 모델링하거나, 특정 약물에 대한 환자의 반응을 예측하는 연구에 활용되고 있다(Rasmy et al., 2021).

하지만 이러한 성공 사례에도 불구하고, 의생물학 분야에서 딥러닝 모델의 적용은 여전히 여러 도전 과제에 직면해 있다. 데이터의 규모가 작거나 클래스 불균형이 심한 경우, 딥러닝 모델의 성능이 불안정해지거나 과적합(overfitting)될 위험이 크다(Choy et al., 2018). 또한, 모델의 예측 과정을 완전히 설명하기 어려운 '블랙박스' 특성은 생명과 직결되는 의료 현장에서의 수용성을 낮추는 요인이 된다(Holzinger et al., 2019).

따라서 본 연구와 같이, 다양한 규모와 특성을 가진 의생물학 dataset을 대상으로 최신 딥러닝 모델과 전통적 머신러닝 모델의 성능을 체계적으로 비교·검증하는 것은 매우 중요하다. 이는 현장의 연구자와 임상의에게 특정 문제에 가장 적합한 모델을 선택할 수 있는 실증적 근거를 제공함으로써, 인공지능 기술의 임상적 활용을 앞당기는 데 기여할 수 있을 것이다.

3. 모델링 방법론

3.1. Gradient Boosting 기반 머신러닝 모델

본 연구에서 주요 비교 대상으로 삼은 전통적 머신러닝 모델은 Gradient Boosting 계열의 XGBoost, LightGBM, CatBoost 세 가지이다. 이들은 모두 트리 기반의 앙상블 학습 구조를 공유하며, 각기 다른 최적화 및 정규화 전략을 통해 표 형식(Tabular) 데이터에서 과적합을 완화하고 변수 중요도를 직관적으로 제공하는 등, 딥러닝에 비해 높은 해석 가능성과 실용성을 갖는 것이 특징이다. XGBoost는 Gradient Boosting 프레임워크에 정규화 항, 조기 종료, 분할 후보의 사전 정렬 등 최적화를 접목함으로써 예측 성능과 학습 속도 모두를 개선하였다.

LightGBM은 히스토그램 기반 분할 및 Leaf-wise 트리 성장 전략을 채택하여 대규모 dataset에 대한 고속 학습과 정확도 향상을 동시에 달성하며, GPU 가속과 다중 클래스 분류도 지원한다. CatBoost는 순차적(target) 인코딩 및 ordered Boosting 기법을 통해 범주형 변수에 대한 별도 전처리 없이도 우수한 성능을 보이며, 클래스 불균형 및 고차원 범주 처리에도 강인한 성능을 나타낸다. 이처럼 서로 다른 모델 구조와 규제 메커니즘을 지닌 Gradient Boosting 기반 모델들을 동일한 파이프라인에서 평가함으로써, 데이터 규모 및 특성 변화에 따른 전통적 ML 계열의 성능 스펙트럼을 정량화하고, 이후 제시될 딥러닝 계열 모델과의 상대적 성능 비교를 위한 기준점을 확보하였다.

3.2. Tabular 데이터를 위한 딥러닝 기반 모델

딥러닝은 이미지와 자연어 처리 분야에서 혁신적인 성과를 이뤄냈지만, 변수 간 상호작용이 비교적 단순한 Tabular 데이터 환경에서는 기존 머신러닝 기법에 비해 뚜렷한 성능 우위를 보여주지 못했다. 그러나 최근에는 수치형·범주형 변수가 공존하고, 표본 수가 불균형하며, 변수 간 관계가 희소한 Tabular 데이터의 특성을 보다 정밀하게 반영하는 전용 딥러닝 아키텍처들이 등장하고 있다. 이들은 기존의 한계를 극복하며, 특정 조건에서 Gradient Boosting 계열 모델을 능가하는 가능성을 보여주고 있다.

본 연구는 대표적인 네 가지 모델인 ResNet(Arik, Pfister, 2021), FT-Transformer(Gorishniy, 2021), TabNet(Arik, Pfister, 2021), Tab-Transformer(Huang et al., 2020)을 실험 대상으로 삼아 성능과 계산 효율을 비교하였다.

3.2.1. ResNet

Arik et al.(2021)로부터 제안된 ResNet은 Residual Block을 도입한 신경망 구조로, 깊은 네트워크에서도 학습이 안정적으로 이루어지도록 한다. 각 블록은 기본적인 MLP 구조에 Skip Connection을 추가해 정보 손실이나 기울기 소실 문제를 완화한다. 블록 내부에서는 Batch Normalization과 선형 변환 후 ReLU를 거치고 Dropout의 과정을 거쳐 입력과 더해져 출력된다. ReLU의 이런 구조는 성능 저하 없이 네트워크를 깊게 쌓는 것을 가능하게 하며, 보편적이고 강력한 성능을 보인다.

3.2.2. FT-Transformer

FT-Transformer는 Gorishniy(2021)로부터 제안되었으며, feature 간 상호작용이 중요한 Tabular 환경에서 높은 표현력과 성능을 보이는 모델이다. 각 feature를 개별적으로 임베딩하는 Feature Tokenizer를 사용해 feature를 토큰처럼 처리하고, 그 토큰들이 Multi-Head Self-Attention을 거치며 서로 다른 관점에서 Attention을 수행함으로써 feature 간의 관계를 학습한다. 이를 통해 모델의 표현력도 향상되며 FT-Transformer는 Attention Maps를 활용해 효율적으로 feature importance도 평가가 가능해 해석 가능성 측면에서도 강점을 가진다.

3.2.3. TabNet

TabNet은 Arik, Pfister(2021)이 제안했다. 표본마다 중요한 feature가 달라지는 Tabular 데이터 환경에서 특히 효과적으로 작동하며, 모델 성능과 학습 효율성 모두에서 강점을 지닌다. TabNet은 각 결정 단계에서 Sequential Attention을 사용해 feature 간의 상호작용을 단계적으로 반영하고, sparse feature selection을 통해 중요한 feature만 선택하는 마스크를 학습한다. 이러한 작동 방식 덕분에 end-to-end 학습이 가능하며, 선택된 feature의 중요도를 기반으로 모델의 해석 가능성까지 확보가 가능하다. 또한, self-supervised learning을 활용해 입력 feature의 일부를 마스킹하고 이를 예측하는 방식으로 학습함으로써, 누락된 feature를 효과적으로 보완하고 label 데이터가 적을 때에도 성능을 높일 수 있다.

3.2.4. Tab-Transformer

Huang et al.(2020)로부터 제안된 Tab-Transformer는 범주형 데이터에 특화된 Transformer 기반 모델으로, 각 범주형 feature를 임베딩한 후, self-attention을 적용해 feature 간 관계를 학습한다. 기존 임베딩 방식과 달리, 각 feature의 고유성을 유지하면서도 feature 간 의미를 명확하게 구별할 수 있으며, 각 범주 값의 문맥적 의미도 함께 반영한다. 이를 통해 범주형 feature 간의 복잡한 상호작용을 효과적으로 포착한다. 또한, 학습된 Attention Weights를 통해 feature 중요도 분석 및 해석 가능성도 제공한다.

이처럼 네 모델은 Residual Connection, Self-Attention, sparse feature selection 등 서로 다른 설계 철학을 통해 Tabular 데이터의 구조적 제약을 극복하고자 하였다. 이러한 모델들은 각각 Tabular 데이터의 특성과 학습 환경에 따라 장단점이 다르며, 본 연구는 이를 실험적으로 비교하고 분석하고자 한다.

4. 연구 데이터

4.1. 실험에 사용된 의생물학 dataset

해당 모델들이 실제 의학 데이터에서도 좋은 효과를 보이는지 검정하기 위해, 데이터 크기에 차등을 두어 총 Kaggle에서 5가지의 Open Dataset을 이용했다.

4.1.1. Thyroid Cancer Risk dataset

Thyroid Cancer Risk dataset은 총 212,691개의 표본과 17개의 feature로 구성된 대규모 의생물학 dataset이다. 이 dataset은 갑상선암 발생 위험 요인을 평가하기 위해 수집되었으며, 주요 Feature는 환자의 인구통계학적 정보(Age, Gender, Country, Ethnicity), 임상 기록(Family History, Radiation Exposure, Iodine Deficiency, Diabetes 등), 생활 습관 요인(Smoking, Obesity) 및 갑상선 호르몬 수치(TSH_Level, T3_Level, T4_Level)로 이루어져 있다. Target(Y) 변수는 최종 진단 결과인 Diagnosis(Benign: 양성, Malignant: 악성)이며, 이는 갑상선암의 존재 여부를 나타낸다.

Table 1. Summary of biomedical tabular datasets used in this study

Name of the dataset	Number of samples	Number of features
Thyroid Cancer Risk	212,691	17
Alzheimer's Prediction	74,283	25
Breast Cancer	4,024	16
Heart Failure Prediction	918	12
Heart Disease(UCI)	303	14

4.1.2. Alzheimer's Prediction dataset

Alzheimer's Prediction dataset은 74,283개의 표본과 25개의 feature로 구성되어 있으며, 알츠하이머병 발생 위험 요인을 분석하기 위해 다양한 국가(20개국)로부터 수집된 데이터이다. 주요 Feature는 인구통계학적 변수(Country, Age, Gender, Education Level), 생활 습관 변수(BMI, Physical Activity, Smoking Status, Alcohol Consumption), 의료적 변수(Diabetes, Hypertension, Cholesterol Level), 유전적 요인(Genetic Risk Factor) 등을 포함한다. Target(Y) 변수는 Alzheimer's Diagnosis(알츠하이머병 진단 여부)로 정의되며, 0(진단되지 않음)과 1(진단됨)로 구분된다.

4.1.3. Breast Cancer dataset

Breast Cancer dataset은 미국 SEER(국립암연구소) 프로그램의 2017년 11월 업데이트 데이터를 기반으로 하며, 2006년부터 2010년 사이 Infiltrating duct carcinoma와 lobular carcinoma로 진단된 여성 환자 4,024명의 정보를 포함한다. 주요 feature로는 나이(Age), 인종(Race), 결혼 상태(Marital Status), 종양 병기(T Stage, N Stage, A Stage), 조직학적 등급(Grade, differentiate), 종양 크기(Tumor Size), 호르몬 수용체 상태(Estrogen Status, Progesterone Status) 등이 포함된다. Target(Y) 변수는 환자의 생존 상태(status: 생존/사망)로 설정하였으며, 환자의 생존 개월 수(Survival Months) 또한 부가적으로 제공된다.

4.1.4. Heart Failure dataset

Heart Failure dataset은 심혈관 질환(CVD)에 의한 심부전 발생을 예측하기 위한 dataset으로, 총 918개의 표본과 12개의 feature로 구성되어 있다. 주요 feature는 환자의 나이(Age), 성별(Sex), 흉통 유형(ChestPainType), 안정 시 혈압(RestingBP), 혈중 콜레스테롤 수치(Cholesterol), 공복혈당(FastingBS), 운동 중 협심증(ExerciseAngina), 최대 심박수(MaxHR) 등 심혈관 건강 상태를 반영하는 변수들로 구성된다. Target(Y) 변수는 HeartDisease로, 1(심장질환 있음), 0(정상)으로 이진 분류된다.

4.1.5. Heart Disease dataset

Heart Disease dataset은 UCI Machine Learning Repository에서 제공하는 고전적인 심장병 예측 dataset으로, 총 303개의 표본과 14개의 feature를 사용한다. 주요 feature는 환자의 나이(age), 성별(sex), 흉통 유형(cp), 안정 시 혈압(trestbps), 혈중 콜레스테롤 수치(chol), 운동 중 협심증(exang), 최대 심박수(thalach) 등으로 구성된다. Target(Y) 변수는 target으로, 심장질환의 존재 여부를 0(질병 없음)부터 1~4(질병 있음)까지 구분하나, 본 연구에서는 이진 분류(0: 질병 없음, 1: 질병 있음)로 재정의하여 사용하였다.

4.2. 탐색적 데이터 분석

4.2.1. 기초 통계분석

데이터 구조에 대한 이해도를 높이기 위해 각 데이터의 주요 변수에 관한 간단한 시각화를 진행했다.

Figure 1 왼쪽의 Thyroid Cancer 데이터에서 진단 결과 분포를 보면 양성(Benign) 환자가 전체의 약 75%를 차지하여 다수였으며, 양성(Malignant) 환자는 약 25%로 나타났다. 이는 데이터가 비교적 균형 잡혀 있으나, 양성 환자가 우세함을 시사한다. Figure 1 오른쪽의 TSH 수치의 분포를 살펴본 결과, 0~10 μIU/mL 구간에 걸쳐 고르게 분포하고 있으며, 뚜렷한 치우침이나 이상치는 관찰되지 않는다.

Figure 2 왼쪽의 Alzheimer's 진단 여부 분포를 보면, 진단되지 않은(No) 환자가 약 58%를 차지하고 진단된(Yes) 환자가 약 42%를 차지하여, 비교적 클래스 균형이 양호한 것으로 나타난다. 그림 2 오른쪽의 연령 분포를 살펴본 결과, 환자들은 50세부터 95세 사이에 고르게 분포하고 있으며, 고령층(특히 70세 이상) 환자가 다수를 차지하는 경향을 보인다.

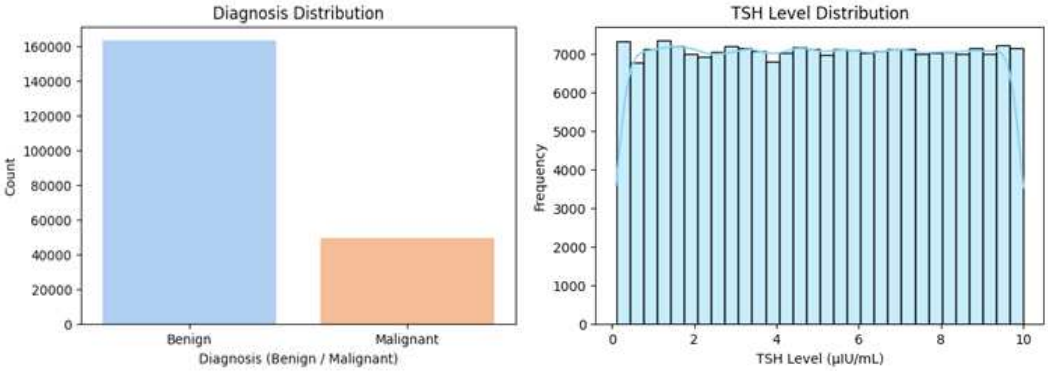

Figure 1. Thyroid Cancer Diagnosis (left) and TSH level (right) distribution

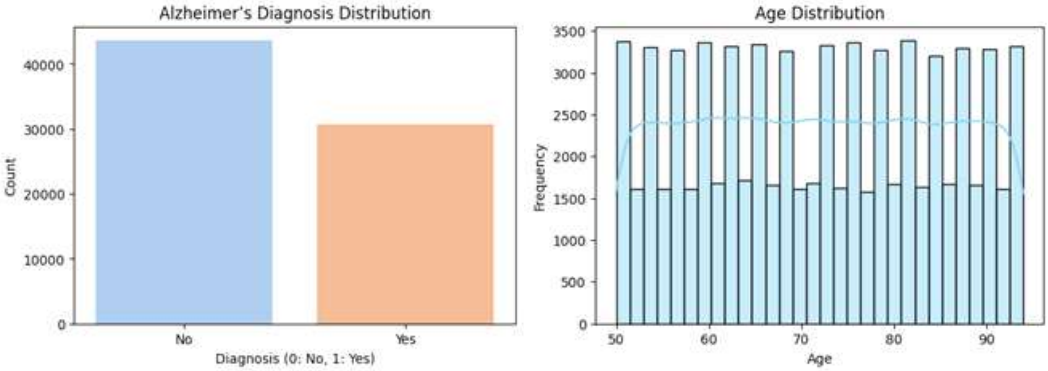

Figure 2. Alzheimer's Diagnosis (left) and Age (right) distribution

Figure 3 왼쪽의 생존 상태(Status) 분포를 보면, 생존 환자가 약 85%로 대다수를 차지하였고, 사망 환자는 약 15%로 나타나 전체적으로 클래스 불균형이 존재하는 것으로 확인된다. Figure 3 오른쪽의 종양 크기(Tumor Size)의 분포를 분석한 결과, 20mm 이하의 작은 종양을 가진 환자가 많았으며, 종양 크기가 클수록 환자 수는 급격히 감소하는 경향을 보인다.

Figure 4 왼쪽의 심장질환 여부(Heart Disease) 분포를 보면, 질병이 없는 환자(0)보다 질병이 있는 환자(1)가 약간 더 많은 비율을 차지하여, 비교적 균형 잡힌 클래스 분포를 나타난다. Figure 4 오른쪽의 나이(Age) 분포를 살펴본 결과, 환자 연령은 대체로 50세를 중심으로 정규분포에 가까운 형태를 보였으며, 심장질환 고위험군인 50~60대 환자가 가장 많은 것으로 나타난다.

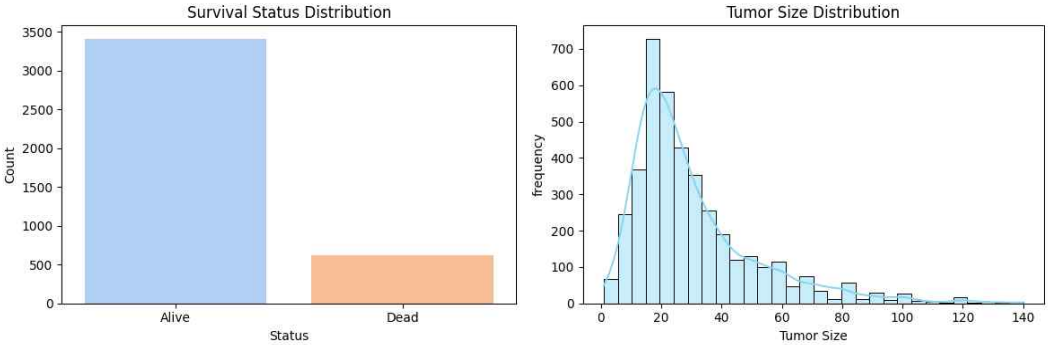

Figure 3. Breast Cancer Survival Status (left) andumor Size (right) distribution

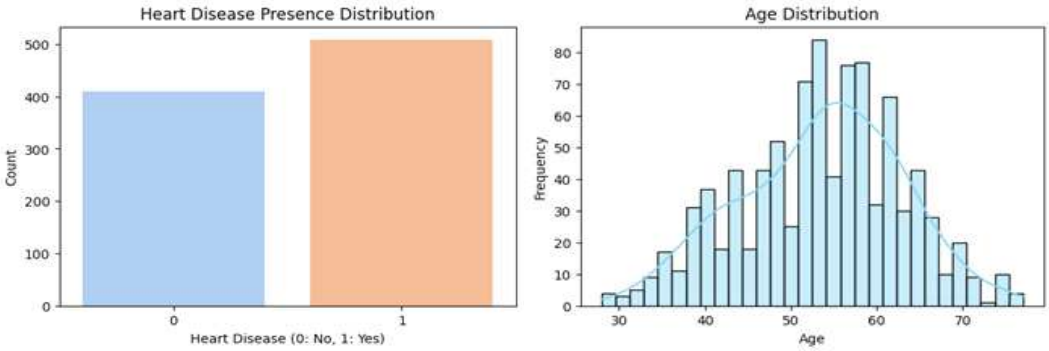

Figure 4. Heart Failure Presence (left) and Age (right) distribution

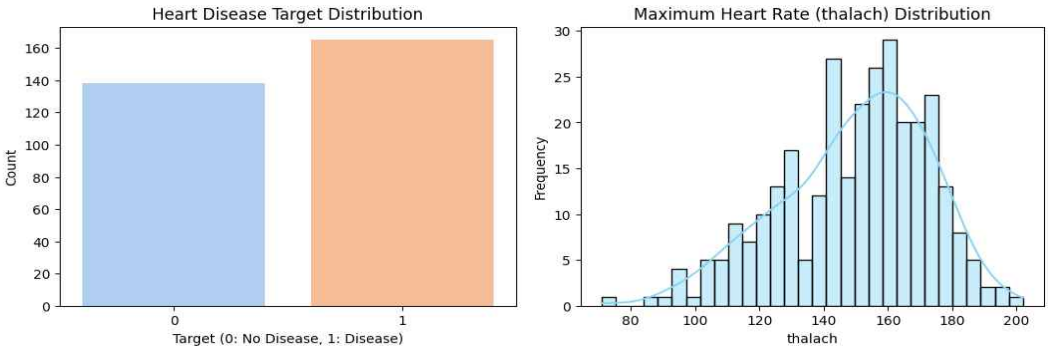

Figure 5. Heart Disease (left) and Maximum Heart Rate (right) distribution

Figure 5 왼쪽의 심장질환 여부(target) 분포를 보면, 질병이 없는 환자(0)보다 질병이 있는 환자(1)가 소폭 더 많은 비율을 차지한다. Figure 5 오른쪽의 최대 심박수(thalach) 분포를 분석한 결과, 140~170 bpm 구간에 환자들이 가장 많이 분포하였으며, 전체적으로는 약한 정규분포 형태를 따르는 것으로 나타났다.

5. 실험

5.1. 실험 설정

모든 모델은 NVIDIA Quadro RTX A6000 GPU(CUDA)와 CPU를 모두 지원하는 공통 코드베이스 위에서 실행되었으며, 학습 단계에서는 주로 GPU를 사용해 연산 효율과 처리 속도를 극대화하였다. 모든 실험은 PyTorch 환경에서 수행되었고, 머신러닝 기반 모델은 scikit-learn 패키지를 활용하여 구현하였으며, 딥러닝 기반 모델은 Tabular 데이터에 최적화된 구조들을 포함하여 torch_frame 라이브러리를 통해 실험 환경을 일관되게 구성하였다. 보다 자세한 실험 구성 정보는 Table 2에 기재되어 있다.

모든 딥러닝 모델의 학습에는 Optuna를 활용하여 하이퍼파라미터 최적화를 추가로 수행하였으며, 조정된 주요 하이퍼파라미터에는 channel, num_layers, batch size, learning rate, num_heads, encoder_pad_size, attention_dropout, ffn_dropout 등이 포함된다. 이를 통해 각 모델의 성능을 극대화하고 공정한 비교를 가능하게 하였다. 모델별로 하이퍼파라미터 튜닝을 4회 수행하고, 각 설정에 대해 모델을 각각 다른 분할 seed를 활용해 4회 반복 학습하여 얻은 정확도(accuracy) 점수의 평균을 분석에 활용하였다.

5.2. 주요 결과

6가지의 머신러닝 모델과 4가지의 딥러닝 모델을 각각 학습하여 실험한 결과는 Table 3에 제시되어 있다.

Table 3은 하이퍼파라미터 튜닝을 적용하지 않은 상태에서 각 모델의 성능을 비교한 결과이다. 모든 모델은 서로 다른 시드(seed)를 사용하여 4회 반복 실험을 수행하였으며, 평균 정확도(mean accuracy)와 표준편차(standard deviation)를 산출하여 모델의 성능과 안정성을 평가하였다.

Heart Disease(303×14), Heart Failure(918×12), Breast Cancer(4024×16)처럼 1만 표본 이하의 소규모 데이터 세트에서는 CatBoost, LightGBM, XGBoost와 같은 머신러닝 기반 모델들이 전반적으로 더 우수한 성능을 보였다. 예를 들어, Breast Cancer에서는 CatBoost(0.9012), LightGBM(0.8932), XGBoost(0.8974)가 ResNet(0.8822)이나 FT-Transformer(0.8611)보다 높은 정확도를 기록했다. 이는 데이터가 충분히 크지 않은 상황에서 딥러닝 모델이 안정적인 학습을 수행하기 어렵기 때문으로 해석된다. 또한 머신러닝 모델들이 상대적으로 낮은 표준편차(s.d) 값을 기록하며 성능의 안정성 측면에서도 우위를 보였다.

Alzheimer(74,283×25)와 같은 중간 규모 데이터 세트에서는 FT-Transformer가 ResNet이나 TabNet보다 우수한 성능(0.7212)을 보였으나, 최종적으로는 LightGBM(0.7274)과 CatBoost(0.7266)가 여전히 가장 높은 정확도를 기록했다. 즉, 데이터가 수만 개 수준으로 커질 경우 일부 딥러닝 모델도 경쟁력을 가지지만, 머신러닝 모델이 여전히 강력한 성능을 유지함을 확인할 수 있다.

Table 2. Experimental setups

Categories	Value / Metric
Split ratio	Train : Validation : Test = 6 : 1 : 3
Seed	11, 22, 33, 42 (four runs)
Epochs / Repetitions	15 epochs / 4 rep
Search trials	50(Optuna, TPE sampler)
Batch size	128, 256, 512 (search space)
Channels	128, 256, 512 (search space)
Num layers	2 – 6 (search space)
Learning rate	1e-5 ~ 1e-3 (log-uniform, search space)
Numerical encoder	linear, linearbucket, linearperiodic (search space)
Evaluation Metric	Accuracy (on validation set)
Device	NVIDIA Quadro RTX A6000 GPU

Table 3. Model performance before hyperparameter tuning

Dataset \ Models		ResNet	FT-T	TabNet	Tab-T	CatBoost	LightGBM	XGBoost
Thyroid Cancer	mean	0.8298	0.8255	0.8264	0.7672	0.8275	0.8270	0.8258
	s.d	0.002	0.001	0.001	0.000	0.004	0.004	0.001
Alzheimer	mean	0.7139	0.7212	0.6730	0.5464	0.7266	**0.7274**	0.7143
	s.d	0.002	0.001	0.001	0.093	0.002	**0.002**	0.001
Breast Cancer	mean	0.8822	0.8611	0.8870	0.8609	**0.9012**	0.8932	0.8974
	s.d	0.023	0.022	0.004	0.021	**0.009**	0.005	0.005
Heart Failure	mean	0.8370	0.8279	0.7799	0.8288	**0.8658**	0.8526	0.8514
	s.d	0.014	0.021	0.014	0.022	**0.026**	0.023	0.020
Heart Disease	mean	0.5906	0.5714	0.5522	**0.8050**	0.7940	0.7720	0.7775
	s.d	0.047	0.084	0.104	**0.017**	0.033	0.027	0.030

Table 4. Model performance after hyperparameter tuning

Dataset \ Model		ResNet	FT-T	TabNet	Tab-T	CatBoost	LightGBM	XGBoost
Thyroid cancer	mean	0.8271	0.8272	0.8271	0.8218	0.8257	0.8257	**0.8273**
	s.d	0.001	0.001	0.001	0.011	0.002	0.002	**0.002**
Alzheimer	mean	0.7150	0.7227	0.7126	0.6783	0.7291	0.7283	**0.7285**
	s.d	0.003	0.001	0.005	0.007	0.002	0.001	**0.001**
Breast cancer	mean	0.8942	0.8986	0.8914	0.8462	**0.9122**	0.9031	0.9080
	s.d	0.002	0.006	0.001	0.011	**0.006**	0.006	0.007
Heart Failure	mean	0.8460	0.8397	0.8397	0.8369	**0.8746**	0.8446	0.8707
	s.d	0.012	0.019	0.007	0.010	**0.023**	0.026	0.026
Heart Disease	mean	0.5769	0.5742	0.7885	0.7280	**0.8440**	0.8258	0.7895
	s.d	0.029	0.113	0.040	0.093	**0.018**	0.026	0.039

가장 대규모였던 Thyroid Cancer(212,691×17) 데이터 세트에서는 모델 간 성능 차이가 사실상 없었다. Tab-Transformer를 제외한 모든 모델이 0.8250~0.8290 구간에서 수렴하는 결과를 보였으며, ResNet(0.8298)이 가장 높은 성능을 기록했다. 이는 충분한 데이터가 제공될 경우, 딥러닝 모델이 머신러닝 모델과 동등하거나 오히려 약간 더 높은 성능을 발휘할 수 있음을 시사한다.

Table 3은 각 모델의 기본 구조에서의 성능을 비교한 것이며, Table 4는 모델별 하이퍼파라미터 튜닝 후 최적 성능을 보여준다. 대규모 데이터 세트인 Thyroid Cancer에서는 모든 모델이 유사한

성능을 보이거나 오히려 하이퍼파라미터 튜닝 후 성능 평균 정확도가 미미하게 떨어지는 것을 관찰할 수 있었다. 이는 데이터가 충분히 큰 경우, 기본 설정만으로도 안정적인 학습이 가능함을 시사한다.

유사하게 Alzheimer dataset에서도 튜닝 전후의 성능 변화가 미미한 것으로 나타났다. 한편, Breast Cancer dataset에서는 대부분의 모델에서 튜닝 후 성능 향상이 확인되었다. Heart Failure 및 Heart Disease와 같은 소규모 dataset에서는 일부 모델에서 성능이 소폭 개선된 반면, 일부 모델은 눈에 띄게 성능이 저하되기도 하였다. 이는 dataset의 표본 수가 적을수록 모델이 오버피팅에 더 민감해져, 오히려 튜닝이 성능 저하로 이어질 가능성이 있음을 보여준다.

Table 5는 각 모델의 효율성 지표($efficiency = accuracy \div \mathrm{sec}onds$, hyperparameter 튜닝하기 전 기준)를 비교한 결과를 제시한다. 값이 높을수록 예측 성능 대비 계산 효율이 우수함을 의미한다. 전반적으로 LightGBM은 4개의 데이터 세트에서 압도적인 효율성을 보여, 제한된 계산 자원 환경에서 특히 유리한 것으로 나타났다. XGBoost 또한 상대적으로 높은 효율성을 확보하였다. 딥러닝 기반 모델들(ResNet, FT-T, TabNet, Tab-Transformer)의 효율성은 데이터 세트 규모와 특성에 따라 상당한 변동성을 보였다. 소규모 데이터 세트에서는 상대적으로 높은 효율성을 달성하였으나, 중대규모 데이터 세트에서는 현저한 효율성 저하를 나타냈다.

구체적으로, Heart Failure Prediction 데이터 세트에서 FT-T(0.552), TabNet(0.578), Tab-Transformer(0.503)는 XGBoost 모델에 이어 상당한 효율성을 보였으며, Heart Disease 데이터 세트에서도 FT-T(1.039), Tab-Transformer(1.150)가 트리 기반 모델에 준하는 효율성을 달성하였다. 이는 소규모 데이터에서 딥러닝 모델의 초기화 및 학습 오버헤드가 상대적으로 적고, 복잡한 비선형 패턴 학습 능력이 효과적으로 발휘되기 때문으로 해석된다.

반면, 대규모 데이터 세트인 Thyroid Cancer와 Alzheimer에서는 모든 딥러닝 모델이 0.002-0.040 범위의 낮은 효율성을 기록하였다. 특히 Tab-Transformer는 대규모 데이터 세트에서 260-331초의 학습 시간을 요구하여 효율성이 급격히 저하되었다. 이는 Attention 메커니즘 기반 모델의 계산 복잡도가 데이터 크기에 따라 제곱 수준으로 증가하는 특성에 기인한 것으로 분석된다.

특히 주목할 점은 CatBoost의 효율성-정확도 trade-off이다. CatBoost는 모든 데이터 세트에서 가장 낮은 효율성(0.022-0.043)을 보였으나, 이는 모델의 고유한 기술적 특성에 기인한다. CatBoost는 ordered boosting과 정교한 범주형 변수 처리를 위해 각 표본마다 서로 다른 모델을 학습하며, 이 과정에서 계산 복잡도가 기하급수적으로 증가한다. 또한 기본 설정에서 보수적인 학습률과 엄격한 조기 종료 조건을 사용하여 학습 시간이 18-41초로 다른 트리 기반 모델(LightGBM: 0.06-0.82초, XGBoost: 0.34-0.93초) 대비 상당히 길다. 그러나 이러한 계산 오버헤드는 대부분의 데이터 세트에서 최고 수준의 예측 정확도를 달성하는 것으로 상쇄된다. 이는 정확도를 최우선으로 하는 응용 분야에서는 CatBoost의 낮은 효율성이 충분히 정당화될 수 있음을 시사한다.

이러한 결과는 딥러닝 기반 모델들이 소규모 복잡 패턴 데이터에서는 계산 비용 대비 합리적인 성능을 제공하나, 데이터 규모 증가 시 트리 기반 모델 대비 효율성이 급격히 저하되는 특성을 명확히 보여준다. 따라서 제한된 계산 자원 환경에서는 데이터 세트의 크기와 복잡성뿐만 아니라 정확도와 효율성 간의 우선순위를 종합적으로 고려한 모델 선택이 필요하다.

Table 5. Model efficiency (accuracy divided by execution time)

Dataset \ Model		ResNet	FT-T	TabNet	Tab-T	CatBoost	LightGBM	XGBoost
Thyroid Cancer	efficiency	0.0099	0.0046	0.0057	0.0023	0.0307	**2.5900**	0.8880
	time (sec)	84.08s	180.83s	143.52s	331.79s	26.95s	0.32s	0.93s
Alzheimer	efficiency	0.0396	0.0134	0.0204	0.0021	0.0363	0.8880	**0.9650**
	time (sec)	17.85s	54.06s	32.94s	260.46s	20.02s	0.82s	0.74s
Breast Cancer	efficiency	0.1800	0.1960	0.1970	0.1390	0.0219	**3.5700**	1.2990
	time (sec)	4.93s	4.40s	4.50s	6.20s	41.42s	0.25s	0.69s
Heart Failure	efficiency	0.2530	0.5520	0.5780	0.5030	0.0362	**14.5000**	2.5000
	time (sec)	3.33s	1.47s	1.35s	1.65s	23.92s	0.06s	0.34s
Heart Disease	efficiency	0.1900	1.0390	0.7360	1.1500	0.0429	**12.9000**	1.4700
	time (sec)	3.13s	0.55s	0.75s	0.70s	18.52s	0.06s	0.53s

6. 결론 및 논의

본 연구에서는 다양한 규모의 의생물학 Tabular dataset을 이용하여 전통적인 머신러닝 모델과 딥러닝 기반 모델의 성능을 비교하였다.

6.1. 주요 연구 결과

첫째, 데이터 규모가 모델 성능에 미치는 영향이 명확히 확인되었다. 1만 표본 미만의 소·중규모 dataset에서는 CatBoost, LightGBM, XGBoost와 같은 그래디언트 부스팅 기반 모델이 일관되게 우수한 성능을 보였으며, 최대 25.2%p의 정확도 우위를 달성하였다. 반면, 20만 이상의 대규모 dataset에서는 딥러닝 모델이 전통적 머신러닝과 대등하거나 근소한 우위를 보여, 충분한 데이터가 확보될 경우 딥러닝 모델의 잠재력이 발휘됨을 확인하였다.

둘째, 효율성 분석 결과 feature 수와 sample 수의 복합적 영향이 중요한 요인으로 나타났다. 딥러닝 모델들은 고차원 대용량 데이터(Thyroid Cancer: 17 features, Alzheimer's: 25 features)에서 계산 복잡도가 급격히 증가하여 효율성이 0.002-0.040 수준까지 저하된 반면, 저차원 소규모 데이터에서는 상대적으로 높은 효율성(0.14-1.15)을 보였다. 이는 딥러닝 모델의 attention 메커니즘과 파라미터 규모가 데이터 차원에 민감하게 반응하기 때문으로 해석된다.

셋째, 모델별 장단점이 확인되었다. Tab-Transformer는 소규모 데이터에서 경쟁력을 보였으나 확장성 측면에서 한계를 드러냈으며, TabNet과 FT-Transformer는 중간 규모에서 상대적으로 안정적인 성능을 보였다. 트리 기반 모델 중에서는 LightGBM이 모든 데이터 세트에서 일관되게 높은 효율성을 유지하여 제한된 계산 자원 환경에서 가장 실용적인 선택으로 평가되었다.

6.2. 실용적 함의

의생물학 연구 환경에서는 대부분의 dataset이 수천에서 수만 건 규모에 머물러 있으며, 계산 자원과 시간적 제약이 존재한다. 본 연구 결과는 이러한 현실적 조건에서 그래디언트 부스팅 기반 모델이 여전히 최적의 선택임을 시사한다. 특히 임상 의사결정 지원 시스템이나 실시간 진단 도구 개발에서는 높은 정확도와 빠른 처리 속도가 동시에 요구되므로, LightGBM이나 XGBoost가 적합한 솔루션이 될 수 있다.

그러나 대규모 population study나 다기관 연구를 통해 수십만 이상의 표본이 확보되고 충분한 GPU 자원이 있는 환경에서는 FT-Transformer나 TabNet과 같은 딥러닝 모델이 제한적이나마 성능 이득을 제공할 수 있다. 이는 특히 복잡한 생체 신호나 다차원 유전체 데이터 분석에서 유용할 것으로 기대된다.

6.3. 연구의 한계 및 향후 방향

본 연구의 주요 한계는 분석 대상이 5개의 공개 데이터 세트에 국한되었다는 점이다. 의생물학 데이터의 다양성을 고려할 때, 시계열 데이터, 고차원 유전체 데이터, 불균형이 심한 희귀질환 데이터 등 다양한 특성을 가진 dataset에 대한 추가 검증이 필요하다.

또한 본 연구에서는 정확도와 효율성에 초점을 맞추었으나, 의료 분야에서 중요한 해석가능성과 불확실성 정량화(uncertainty quantification) 측면에서의 비교 분석이 부족하다. 향후 연구에서는 SHAP values나 LIME과 같은 설명 가능한 AI 기법을 통한 모델 해석성 평가와 베이지안 접근법을 통한 예측 불확실성 분석이 포함되어야 할 것이다.

References

Arik, S. Ö., Pfister, T. (2021). TabNet: Attentive interpretable tabular learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8), 6679-6687. DOI: https://doi.org/10.1609/aaai.v35i8.16826

Chen, T., Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P. M., Zietz, M., Hoffman, M. M., Xie, W., Rosen, G. L., Lengerich, B. J., Israeli, J., Lanchantin, J., Woloszynek, S., Carpenter, A. E., Shrikumar, A., Xu, J., Cofer, E. M., … Greene, C. S. (2018). Opportunities and obstacles for deep learning in biology and medicine. *Journal of the Royal Society*, Interface, 15(141), 20170387. DOI: https://doi.org/10.1098/rsif.2017.0387

Cho, H. B., Baek, B., Kim H. J., Kim, Y. S., Park, B. (2023). Development of Defect Classification Model for Smart Factory Data Using Machine Learning. *Journal of the Korean Data Analysis Society*, 25(5), 1637-1651. (In Korean) DOI: https://doi.org/10.37727/jkdas.2022.24.3.1149

Choy, G., Khalilzadeh, O., Michalski, M., Do, S., Samir, A. E., Pianykh, O. S., Geis, J. R., Pandharipande, P. V., Brink, J. A., Dreyer, K. J. (2018). Current Applications and Future Impact of Machine Learning in Radiology. *Radiology*, 288(2), 318–328. https://doi.org/10.1148/radiol.2018171820

Gorishniy, Y., Rubachev, I., Khrulkov, V., Babenko, A. (2021). Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34, 18932-18943.

Grinsztajn, L., Oyallon, E., Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on tabular data?. *Advances in Neural Information Processing Systems*, 35, 507-520.

He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770-778.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4), e1312.

Huang, X., Khetan, A., Cvitkovic, M., Karnin, Z. (2020). Tabtransformer: Tabular data modeling using contextual embeddings. *ArXiv Preprint ArXiv*:2012.06678.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ..., Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30.

Kong, J. S., Kim, Y. (2023). A Study on the Utilization Strategies of Subjective Questions in Public Agency Surveys Using Deep Learning Techniques. *Journal of the Korean Data Analysis Society*, 25(6), 2187-2200. (In Korean) DOI: https://doi.org/10.37727/jkdas.2023.25.6.2187

Nam, S. H., Oh, M., Kim, S. K., Kang, C., Kim, K. K., Choi, S. B. (2017). Comparison of Machine Learning Models for Classification into User-oriented Groups. *Journal of the Korean Data Analysis Society*, 19(5), 2501-2507. (In Korean) DOI: https://doi.org/10.37727/jkdas.2017.19.5.2501

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 31.

Rasmy, L., Xiang, Y., Xie, Z., Tao, C., Zhi, D. (2021). Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digital Medicine*, 4(1), 86.

Shwartz-Ziv, R., Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, 81, 84-90. DOI: https://doi.org/10.1016/j.inffus.2021.11.011

Yoon, H. G., Choi, S. B., Kim, T. (2022). Discussion on Education Big Data Analytics using Deep Learning Models. *Journal of the Korean Data Analysis Society*, 24(3), 1149-1157. (In Korean) DOI: https://doi.org/10.37727/jkdas.2022.24.3.1149

Benchmarking Deep Learning vs. Traditional Machine Learning on Biomedical Tabular Data*

Chaerin Song[1], *Nahyun Lee*[2], *Il-Youp Kwak*[3]

Abstract

While deep learning has achieved revolutionary success in image and natural language processing, traditional gradient boosting-based machine learning (ML) models still dominate in the biomedical domain for tabular data. This study systematically evaluates the performance and efficiency of three ML models (XGBoost, LightGBM, CatBoost) and four deep learning (DL) models on five public biomedical datasets, applying identical preprocessing and hyperparameter tuning. Experimental results show that for small to medium-sized datasets (under 10,000 samples), ML models consistently demonstrated superior performance and speed. On large-scale datasets (over 200,000 samples), DL models showed comparable performance but with significantly decreased efficiency as the number of features increased. In conclusion, gradient boosting-based ML models remain a robust choice for most biomedical tabular problems, while Transformer-based DL models may offer limited benefits only when applied to very large datasets with sufficient computational resources.

Keywords : Biomedical Data, Tabular Data, Deep Learning, Benchmarking.

*This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (No.RS-2023-00208284). Chaerin Song and Nahyun Lee contributed equally to this work.

[1]Undergraduate Student, Department of Applied Statistics, Chung-Ang University, Seoul 06974, Korea. E-mail: cleo7987@cau.ac.kr

[2]Master Student, Department of Smart Cities, Chung-Ang University, Seoul 06974, Korea. E-mail: naa012@cau.ac.kr

[3](Corresponding Author) Associate Professor, Department of Applied Statistics, Chung-Ang University, Seoul 06974, Korea. E-mail: ikwak2@cau.ac.kr

[Received 19 July 2025; Revised 28 August 2025; Accepted 2 September 2025]