논문 2026-2-12 http://dx.doi.org/10.29056/jsf.2026.06.12

손글씨 기반 알츠하이머병 탐지를 위한 확률 보정 및 임상적 유용성 평가

김영인*†

Calibration and Clinical Utility for Trustworthy Handwriting-Based Alzheimer's Disease Detection

Young-In Kim*†

요 약

손글씨 분석은 알츠하이머병의 비침습적 조기 발견 방법이다. 기존 DARWIN 데이터셋 연구들은 정확도나 AUC 같은 단순 성능 지표만 제시했을 뿐, 예측 확률의 신뢰성과 불확실한 사례 처리는 다루지 않았다. 본 연구는 6개 머신러닝 모델을 반복 5×5 교차검증으로 재평가하고, 4가지 확률 보정 기법과 불확실성을 정량화하는 몬드리안 유도 컨포멀 예측을 적용하였다. 실험 결과 랜덤 포레스트는 예측 성능이 가장 높았으나(AUC 0.957) 확률 보정 오차는 가장 컸다(ECE 0.192). 등위 회귀로 오차가 53% 감소하여(ECE 0.090) 통계적으로 유의미했으며($p<0.001$), 확신하기 어려운 사례에 판단을 유보하는 거부 옵션을 제공하였다. 결정곡선분석에서도 보정된 확률이 임상 의사결정에 더 큰 이득을 줄 수 있음을 확인하였다. 따라서 손글씨 기반 알츠하이머 진단 모델의 임상 활용에는 단순 성능 지표를 넘어 확률 보정과 불확실성 정량화가 함께 필요함을 확인하였다.

Abstract

Handwriting analysis is a non-invasive method for early detection of Alzheimer's disease. Prior DARWIN dataset studies reported only simple metrics such as accuracy and AUC, without addressing the reliability of predicted probabilities or the handling of uncertain cases. This study re-evaluates six machine learning models using repeated 5×5 cross-validation, applying four probability calibration techniques and Mondrian Inductive Conformal Prediction to quantify uncertainty. Random Forest showed the highest performance (AUC 0.957) but the largest calibration error (ECE 0.192). Isotonic regression reduced this by 53% (ECE 0.090), a significant improvement ($p<0.001$), and a rejection option withheld predictions for uncertain cases. Decision curve analysis showed that calibrated probabilities can yield greater net benefit for clinical decision-making. Thus, clinical use of such models requires probability calibration and uncertainty quantification, beyond simple metrics.

한글키워드 : 알츠하이머병, 손글씨 분석, DARWIN, 확률 보정, 컨포멀 예측, 의사 결정 곡선

keywords : Alzheimer's disease, Handwriting analysis, DARWIN, Probability calibration, Conformal prediction, Decision curve analysis

* 부산대학교 IT응용공학과
† 교신저자: 김영인(email: kimyi@pusan.ac.kr)

접수일자: 2026.06.01. 심사완료: 2026.06.06.
게재확정: 2026.06.20.

1. 서 론

알츠하이머병(Alzheimer's Disease, AD)은 중추신경계의 대표적인 진행성 신경퇴행성 질환으로, 전 세계 치매의 약 60~70%를 차지한다[1]. 임상 진단은 주로 신경학적 검사와 영상검사에 의존하기 때문에 비용이 높고 접근성이 낮으며, 특히 조기 단계에서 진단하기에 어려움이 있다. 이에 따라 비침습적이면서 저비용으로 쓸 수 있는 행동 기반 바이오마커(behavioral biomarker)에 대한 관심이 높아지고 있다. 그중에서도 손글씨(handwriting)는 미세운동과 시공간 인지 기능을 동시에 반영하므로, AD를 조기에 검출하기에 적합한 신호로 평가되고 있다.

이러한 측면에서 DARWIN(Diagnosis AlzheimeR WIth haNdwriting) 데이터세트[2]는 공개 이후 손글씨 기반 AD 분석 연구에서 널리 활용되는 데이터세트로 자리 잡았다. DARWIN은 피험자 174명(AD 환자 89명, 정상 대조군 85명)이 수행한 25개 손글씨·그림 태스크에서 18개의 운동학적·동역학적 특징을 추출해 총 450개 변수로 구성된다. 2022년 공개 이후 여러 연구가 DARWIN에 다양한 모델을 적용해 왔으나, 대부분은 정확도, F1, AUROC(Area Under the Receiver Operating Characteristic Curve) 같은 점 추정(Point Estimation) 지표만 제시했을 뿐, 모델이 출력하는 예측 확률의 신뢰성(probability calibration)과 불확실성 정량화(uncertainty quantification)는 다루지 않았다.

그러나 임상 의사결정 지원(Clinical Decision Support System, CDSS) 관점에서는 모델이 '얼마나 확신하는가'가 분류 결과 자체만큼 중요하다. 최근 STRATOS Initiative Topic Group 6의 지침[3]에는 이진 결과 예측 모형의 핵심 보고 항목으로 AUC, 보정 곡선(calibration plot), 결정곡선 분석(Decision Curve Analysis, DCA) 기반 순편익(net benefit), 결과 범주별 확률 분포 그래프를 권고하면서, 특히 보정과 임상적 유용성(clinical utility)을 평가의 중심 축으로 제시하였다. FUTURE-AI 합의 가이드라인[4]도 6대 원칙(Fairness, Universality, Traceability, Usability, Robustness, Explainability) 아래 30개 모범 사례에서 확률 보정과 불확실성 정량화를 신뢰할 수 있는 의료 AI의 핵심 요소로 제시하였다. 따라서 DARWIN 기반 손글씨 AD 진단 모델도 정확도뿐 아니라 확률 보정과 불확실성 정량화 측면에서 평가할 필요가 있다.

본 연구는 이를 위하여 다음과 같은 통합 프레임워크를 사용한다. 첫째, DARWIN에서 대표적인 베이스라인 모델 6종(Logistic Regression, SVM-RBF, Random Forest, XGBoost, LightGBM, MLP)을 동일한 반복 계층화 5×5-fold 교차검증 기법으로 재평가하였다. 둘째, 4종의 후처리 확률 보정(Platt[5], Isotonic[6], Beta[7], Temperature scaling[8])과 몬드리안(Mondrian) 유도 컨포멀 예측(Inductive Conformal Prediction, ICP)을 적용해 분류 성능과 보정 품질 사이의 절충 관계(trade-off)를 정량화하였다. 셋째, 결정곡선분석[9], 선택적 분류(selective classification)[10], SHAP(SHapley Additive exPlanations)[11] 분석을 수행하여 임상 유용성 측면에서 보정된 모델의 가치를 평가하였다.

본 연구의 주요 기여는 다음과 같다. (1) DARWIN 데이터세트에 확률 보정과 불확실성 정량화를 종합적으로 적용한 연구를 제시했으며, 6 모델 × 5 보정 방법 × 3 유의수준 × 다중 임상 평가지표를 단일 평가 프로토콜로 통합하였다. (2) Random Forest가 분류 성능(AUC=0.957)에서는 가장 우수하지만 보정 오차(Expected Calibration Error, ECE) 측면에서는 0.192로 가장 오보정되어 있음을 확인하고, 등위 보정으로

ECE를 53% 줄일 수 있음을 제시하였다. (3) 몬드리안 ICP를 통해 약 95%의 처리 비율(coverage)과 71~75%의 단일 클래스 예측(singleton) 비율을 기록함으로써, 임상 의사결정 과정에서 활용 가능한 거부 옵션(reject option)의 유효성을 입증하였다. (4) DCA와 SHAP 분석을 결합해 임상적 활용 시나리오를 제시했으며, STRATOS 및 FUTURE-AI 가이드라인의 권고사항을 충족하도록 하였다.

본 논문의 구성은 다음과 같다. 2장에서 관련 연구를 검토하고, 3장에서 제안하는 보정 및 컨포멀 예측 방법을 설명한다. 4장에서 실험 결과를 제시하고, 5장에서는 결론 및 향후 연구 방향에 대하여 기술한다

2. 관련 연구

Cilia 등[2]이 2022년 공개한 DARWIN 데이터세트는 피험자 174명에게 표준화된 손글씨 태스크 25개를 수행하게 하고, 태스크마다 시간·동작·압력 특징 18개를 추출한 자료로, 손글씨 기반 AD 진단에 널리 쓰인다. 이 연구는 Random Forest, KNN, LDA, SVM 등 여러 분류기를 태스크별로 학습한 뒤 앙상블하는 방식을 제안했으며, Random Forest 기준으로 정확도 약 91%, 민감도 83%를 보고하였다.

이후 다양한 모델이 적용되었다. Erdogmus와 Kabakus[12]는 1차원 특징을 2차원 이미지로 재배치한 CNN으로 정확도 90.4%를 보고했고, Sweidan 등[13]은 DeepSHAP, TSR, CoMTE 기반 설명가능성과 소프트 보팅(soft voting) 집계를 결합한 1D-CNN으로 피험자 단위 정확도 94.4%를 달성하였다. Mitra와 Rehman[14]은 ANOVA와 RFE 기반 특징 선택 후 스태킹 앙상블로 정확도 97.14%를 보고했고, Gong 등[15]은 Hybrid Transformer 기반 접근법을 제안하였다. Demircioglu Diren[16]은 ML 알고리즘 9개와 특징 선택 방법 7개를 비교해 SHAP 기반 SVM에서 정확도 96.2%를 달성했으며, Ho 등[17]은 Random Forest, Bagging, XGBoost, LightGBM 등을 SHAP 해석가능성과 결합해 보고하였다. 그러나 이들 선행 연구는 모두 점추정 지표(정확도, F1, AUC)와 SHAP·Grad-CAM 같은 정성적 해석가능성에 머물렀을 뿐, 예측 확률의 보정 품질이나 불확실성 정량화는 다루지 않았다.

확률 보정 분야에서는 Platt[5]이 SVM의 거리 출력에 로지스틱 회귀를 학습하는 시그모이드(sigmoid) 방식을 제안한 이래, Zadrozny와 Elkan[6]이 PAV 알고리즘 기반 등위 회귀(isotonic regression)를 이진 분류기 보정에 적용하고 이를 다중 클래스로 확장하는 방법을 제안하였다. Kull 등[7]의 Beta calibration은 beta 분포 가정에 기반한 모수적 모델로, 소표본 환경에서 강건하다. Guo 등[8]은 최근의 심층 신경망이 정확도가 높아져도 과확신(over-confidence) 경향을 보인다는 점을 실증하고, Temperature scaling이 매우 효과적임을 보고하였다.

컨포멀 예측(Conformal prediction)은 데이터의 특정 분포를 가정하지 않고도 통계적 정확도를 보장하는 예측 방법이다[18]. 의료 AI 분야에서는 Pereira 등[19]이 치매 환자군 분석에 이를 적용한 사례가 있으나, DARWIN과 같은 손글씨의 움직임 특징을 활용하여 신경계 질환을 진단하는 과정에 도입한 사례는 아직 없는 실정이다. 한편, Kladny 등[20]은 적은 데이터 환경에서 이 예측 방법이 제공하는 통계적 보장이 실제 임상 현장의 유용성과 차이가 있을 수 있다고 지적했다. 본 연구는 이러한 이론적 문제가 실제 데이터에서도 발생하는지 살펴보고자 한다.

3. 제안 방법

본 절에서는 확률 보정 및 컨포멀 예측 방법을 단계별로 설명한다. 전체 과정은 그림 1과 같으며, ① DARWIN 데이터세트 전처리, ② 6개 베이스라인 모델 학습, ③ 후처리 보정 및 몬드리안 ICP 적용, ④ 분류·보정·임상 유용성 평가 단계로 구성된다.

3.1 확률 보정 방법

확률 보정이란 모델의 예측 확률이 실제 발생 빈도와 일치하도록 만드는 절차로, 잘 보정된 분류기는 '80% 확신'이라고 출력한 사례 중 실제로 약 80%가 양성 클래스여야 한다. 본 연구에서는 다음과 같이 네 가지 후처리 보정 방법을 비교하고자 한다.

Platt scaling[5]은 분류기의 원시 출력값 p에 로지스틱 함수 $\sigma(a \cdot p + b)$의 스케일 파라미터 a와 편향 파라미터 b를 최적화하여 확률을 보정하는 방법이다. 구조가 단순하고 적용이 용이하다는 장점이 있으나, 이 방법은 원시 확률이 시그모이드 형태의 패턴을 따른다는 가정에 기반한다. 따라서 결정 트리 계열 모델처럼 원시 확률이 0 또는 1 근방으로 집중되는 양극화(polarization) 특성을 보이는 경우에는 보정 효과가 충분하지 않을 수 있다.

등위 회귀[6]는 데이터 분포에 대한 특정 가정 없이 데이터 자체의 패턴을 따르는 방법으로, 예측값의 상대적 순서를 유지하면서 데이터에 맞게 유연하게 보정 곡선을 조정한다.

Beta calibration[7]은 베타 분포를 따른다고 가정하고 파라미터를 추정하는 보정 방법으로, 두 개의 파라미터를 사용하며 확률값의 양 끝(0과 1 근방)에서 보정 효과가 두드러진다. 특히 소규모 데이터세트 환경에서는 등위 회귀에 비해 더 안정적인 추정 성능을 보인다.

Temperature scaling[8]은 모델이 내부적으로 계산하는 원시 점수(raw score) 영역에 단일 스칼라 값 T를 도입하여, 원시 점수를 T로 나눈 뒤 확률로 변환하는 방식으로 보정을 수행한다. 이 방법은 가장 높은 점수를 받은 클래스, 즉 최종 예측 결과를 그대로 유지하기 때문에 보정 전후의 분류 정확도가 변하지 않는다는 장점이 있다. 또한 신경망 모델에만 국한되지 않고, 트리 기반 모델에서도 모델이 출력한 확률 p를 역산하여 내부 점수로 되돌린 값에 동일하게 적용할 수 있다.

모델의 보정 성능을 정량적으로 평가하기 위해 ECE, Brier Score의 세 가지 지표를 사용한다. ECE는 예측 확률과 실제 정답 비율 간의 불일치를 측정하는 지표로, 값이 0에 가까울수록

모델이 잘 보정되어 있음을 의미한다. Brier Score는 예측 확률과 실제 이진 결과 간의 평균 제곱 오차로 0에서 1 사이의 값을 가지며, 값이

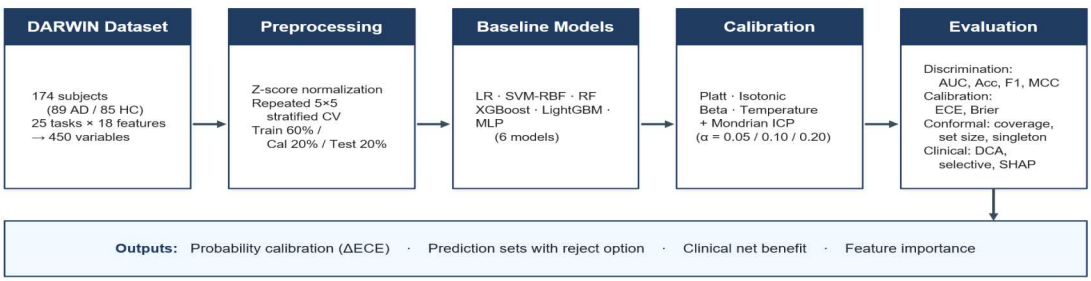

Fig. 1. Overall pipeline of the proposed probability-calibration and conformal-prediction framework.

그림 1. 제안하는 확률 보정 및 컨포멀 예측 프레임워크의 전체 파이프라인
Fig. 1. Overall pipeline of the proposed probability-calibration and conformal-prediction framework

낮을수록 예측의 정확도와 보정 품질이 모두 우수함을 나타낸다.

3.2 몬드리안 ICP

ICP는 모델 학습에 사용하지 않고 신뢰도 기준을 설정하기 위해 따로 떼어 놓은 보정 세트(calibration set)를 활용하여, 모델이 출력하는 예측 집합이 실제 정답을 포함할 확률을 통계적으로 보장하는 방법론이다. 본 연구에서는 클래스 불균형 문제에 강건한 몬드리안 ICP를 적용한다. 몬드리안 ICP는 모든 데이터를 한꺼번에 처리하지 않고 각 클래스를 따로따로 기준을 맞춰 조정함으로써, 특정 클래스의 샘플 수가 현저히 적은 상황에서도 어떤 클래스든 예측 결과가 일정 수준 이상 믿을 수 있도록 보장한다. 따라서 본 논문에서는 신뢰성이 있는 단일 클래스 예측 비율이 높을수록 모델이 이 샘플은 이 클래스다라고 확실하게 답하는 경우가 많음을, 공집합(empty, 모델이 판단을 거부하는 경우) 비율이 높을수록 모델이 어떤 클래스에도 해당하지 않는다고 판단해 예측 자체를 포기하는 경우가 많음을 의미한다.

3.3 임상적 유용성 평가

DCA는 모델이 실제 임상 현장에서 얼마나 유용한지를 평가하는 방법이다. 의사가 특정 확률 이상일 때 치료를 결정한다고 가정할 때, 모델을 사용하는 것이 "모든 환자를 치료하는 전략" 또는 "아무도 치료하지 않는 전략"보다 얼마나 더 이득인지를 순편익 지표로 정량화한다.

선택적 분류(Selective classification)는 모델의 예측 확률이 0.5보다 클수록 모델이 높은 확신을 가진 사례로 간주하여 자동으로 처리하고, 불확실한 사례는 전문의에게 넘기는 임상 워크플로를 같은 방식으로 따라 처리한다. 이를 위하여 확신이 큰 경우를 먼저 나타나도록 순서를 조정한 뒤, 높은 사례를 기준으로 평가한다.

마지막으로, SHAP 분석을 통해 각 임상 특징과 예측 태스크가 모델의 분류 결정에 기여하는 정도를 수치화함으로써, 모델의 판단 근거를 임상적으로 해석할 수 있도록 한다.

4. 실험 결과

4.1 데이터세트 및 실험 설정

본 연구는 UCI Machine Learning Repository에 공개된 DARWIN 데이터세트[2]를 사용한다. 데이터세트는 표 1과 같이 AD 환자 89명과 정상 대조군(Healthy Controls, HC) 85명, 총 174명으로 구성된다. 각 피험자는 표준화된 손글씨·그림 태스크 25개를 수행했고, 각 태스크에서 18개의 운동학적·시간적 특징을 추출해 총 450개의 입력 변수를 구성한다. 결측치는 없으며 클래스 비율은 51.1% : 48.9%로 거의 균형을 이루고 있다.

실험 방법은 다음과 같다. n=174의 소표본 한계를 고려해 반복 계층화 5-fold 교차검증

표 1. DARWIN 데이터세트 구성
Table 1. DARWIN dataset distribution

Class	Count	Ratio (%)	Description
AD (Patient, P)	89	51.1	Alzheimer's
HC (Normal, H)	85	48.9	Healthy
Total	174	100.0	25 tasks × 18 features = 450 vars

(Repeated Stratified K-fold CV, 5-fold × 5-repetition, 총 25회 평가)을 채택하였다. 각 fold에서 학습 분할(train portion) 내에서 다시 25%를 보정 세트로 분리하였다. 따라서 fold당 약 105명이 모델 컨포멀에, 약 35명이 보정 및 ICP에, 약 34명이 테스트에 쓰였다. 전처리 단계에서의 데이터 누수를 방지하고자, Z-score 정규화를 위한 통계적 파라미터(평균, 표준편차)의 추정은 훈련 분할에 한정하여 수행하였다. 이후 동일한 파라미터를 검증 및 테스트 분할에 적용함으로써 각 분할 간 정보 유출을 차단하였다. 학습 환경은 Windows 11의 WLS Ubuntu 22.04.5 LTS에서 Python 3.13.9, numPy 2.3.5, pandas 3.0.0, scikit-learn 1.8.0, XGBoost 3.2.0, LightGBM 4.6.0, SHAP 0.50.0을 사용하였다. 사용한 6개 베이스라인 모델의 주요 하이퍼파라미터는 표 2와 같다.

4.2 분류 및 보정 결과

표 3은 6개 베이스라인 모델의 보정 전(raw) 분류 성능과 보정 후 핵심 지표를 요약한다. 결과는 25개 fold의 평균 ± 표준편차로 제시한다. 보정 전 결과를 보면 Random Forest가 정확도 0.871, AUC 0.957로 가장 우수했으며, LightGBM(Acc 0.869, AUC 0.947)과 XGBoost(Acc 0.864, AUC 0.947)가 뒤를 이었다. MLP는 정확도 0.788로 가장 낮았다.

그러나 보정 측면에서는 정확도 순위와 정반대 패턴이 나타났다. Random Forest의 보정 전 ECE는 0.192로 모든 모델 중 가장 높았다. 이는 '정확도가 높은 모델일수록 보정이 잘 되어 있다'는 통념을 반증한다. 한편 LightGBM(ECE 0.118)과 XGBoost(ECE 0.121)는 트리 모델 중에서는 상대적으로 보정이 잘 되어 있었다. 그림 2의 신뢰성 도표(reliability diagram)가 이 패턴을

표 2. 베이스라인 모델의 하이퍼파라미터
Table 2. Hyperparameters of baseline models

Model	Hyperparameters
Logistic Regression	C = 1.0, L2 penalty, max_iter = 2000 (with StandardScaler)
SVM-RBF	C = 1.0, gamma = scale, probability = True (with StandardScaler)
Random Forest	n_estimators = 300, min_samples_leaf = 2, default depth
XGBoost	n_estimators = 300, max_depth = 4, learning_rate = 0.05, subsample = 0.8, colsample_bytree = 0.8
LightGBM	n_estimators = 300, num_leaves = 31, learning_rate = 0.05, subsample = 0.8, colsample_bytree = 0.8
MLP	hidden_layer_sizes = (64, 32), activation = ReLU, alpha = 1e-3, max_iter = 500, early_stopping (with StandardScaler)
Common	random_state = 42, n_jobs = -1, eval_metric = logloss; ICP $\alpha \in \{0.05, 0.10, 0.20\}$

시각적으로 보여 준다.

네 가지 후처리 보정 방법의 효과는 모델마다 상이하게 나타났다. 등위 회귀는 Random Forest의 ECE를 0.192에서 0.090으로 53% 감소시켜 가장 우수한 보정 성능을 보인 반면, Platt scaling은 ECE를 0.260으로 증가시켜 보정 품질을 오히려 저하시켰다(표 3). 이는 Random Forest가 트리 투표 비율에 기반한 예측 확률을 산출하는 특성상 양극화된 확률 분포를 형성하므로, 시그모이드 함수를 통한 추가적인 평탄화가 컨포멀하지 않기 때문이다. Temperature scaling은 정확도 및 AUC를 유지하면서 ECE를 0.115로 낮추어 Random Forest에 대한 안정적인 보정 대안으로 확인되었다. LightGBM과 XGBoost에서는 등위 회귀가 ECE를 각각 0.090과 0.096으로 감소시켜 유효한 보정 효과를 나타냈다. 한편, Logistic Regression과 SVM-RBF는 보정 전 ECE가 이미 0.13~0.15 수준으로 낮아 보정 적용에 따른 개선 폭이 상대적으로 제한적이었다.

4.3 보정 효과의 통계적 유의성

각 모델-보정 방법 조합이 실제로 보정 성능을 유의미하게 개선하는지 검증하기 위해, 25개 fold 데이터를 활용한 대응표본 Wilcoxon 부호순위 검정을 실시하고, 다중비교 문제를 제어하기 위해 모델별로 Holm-Bonferroni 보정을 적용하였다. 주요 결과는 표 4에 정리하였다. (각 모델 내 Holm-Bonferroni 보정 기준: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$)

표 3. 모델·보정 방법별 분류·보정 지표 (25개 fold 평균 ± 표준편차)
Table 3. Discrimination and calibration metrics by model and calibration method

Model	Calibration	Accuracy	AUC	ECE	Brier
LogReg	Raw	0.824±0.059	0.899±0.049	0.149±0.043	0.137±0.045
LogReg	Isotonic	0.813±0.070	0.883±0.049	0.126±0.043	0.140±0.039
SVM-RBF	Raw	0.834±0.072	0.904±0.062	0.132±0.045	0.124±0.047
SVM-RBF	Isotonic	0.809±0.089	0.881±0.068	0.111±0.052	0.134±0.048
Random Forest	Raw	0.871±0.059	0.957±0.026	0.192±0.037	0.111±0.017
Random Forest	Platt	0.867±0.056	0.957±0.026	0.260±0.039	0.157±0.011
Random Forest	Isotonic	0.863±0.047	0.934±0.040	0.090±0.042	0.098±0.036
Random Forest	Beta	0.867±0.053	0.957±0.026	0.153±0.034	0.094±0.021
Random Forest	Temperature	0.871±0.059	0.957±0.026	0.115±0.027	0.087±0.028
XGBoost	Raw	0.864±0.066	0.947±0.034	0.121±0.028	0.091±0.036
XGBoost	Isotonic	0.855±0.050	0.926±0.048	0.096±0.042	0.104±0.041
LightGBM	Raw	0.869±0.061	0.949±0.035	0.118±0.044	0.099±0.043
LightGBM	Isotonic	0.870±0.054	0.928±0.051	0.090±0.043	0.099±0.043
MLP	Raw	0.788±0.080	0.873±0.062	0.156±0.038	0.150±0.041
MLP	Isotonic	0.780±0.074	0.850±0.068	0.127±0.051	0.162±0.044

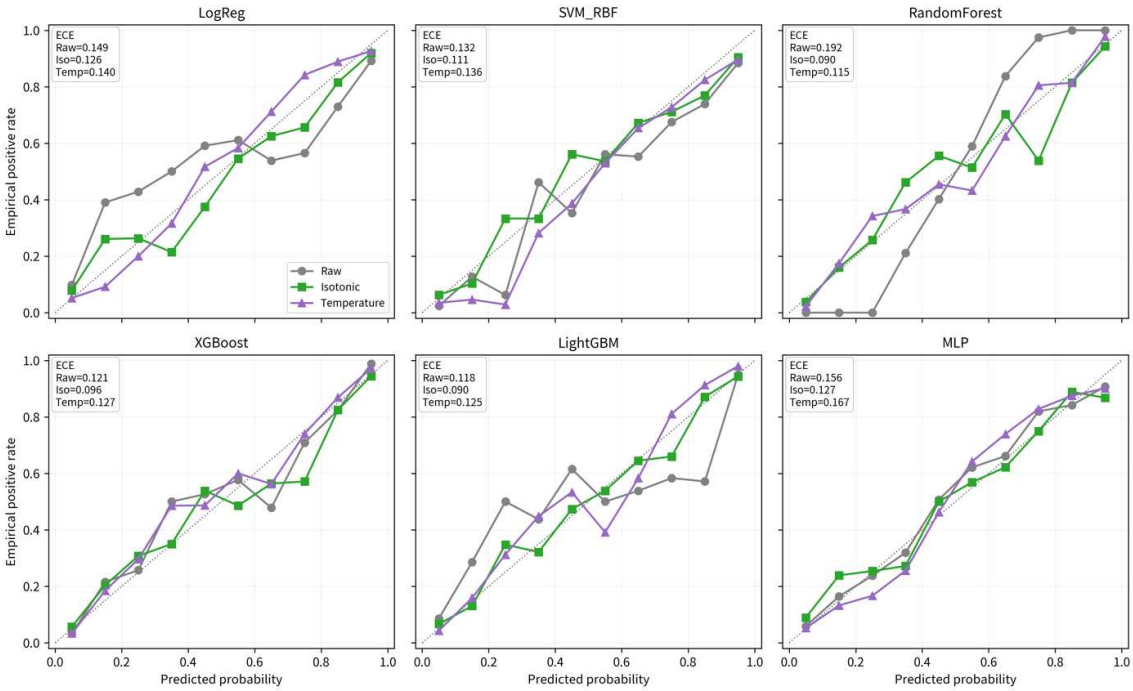

그림 2. DARWIN에 대한 6개 베이스라인 모델의 신뢰성 도표

Fig. 2. Reliability diagrams of six baseline models on DARWIN

표 4와 같이 Random Forest 모델에서는 Isotonic, Beta, Temperature scaling 세 가지 방법이 모두 ECE를 통계적으로 유의하게 감소시켰다(모두 $p < 0.001$). 특히 등위 회귀의 개선 효과가 가장 두드러져, ECE가 0.102포인트(약 53%) 감소하였다.

반면 Platt scaling은 Random Forest, XGBoost, LightGBM, MLP 네 모델 모두에서 ECE를 오히려 유의하게 증가시키는 역효과를 나타냈다. 이는 트리 기반 모델이 출력하는 확률 분포의 특성이 Platt scaling과 같은 시그모이드 함수의 S자 곡선 하나로 확률을 보정하는 방식과 맞지 않아 보정 후 확률이 더 부정확해져 컨포멀하지 않음을 실험적으로 보인 결과로, 의료 AI 개발 및 운용 실무에서 보정 방법 선택 시 중요한 고려 사항을 제시하였다.

4.4 컨포멀 예측 결과

몬드리안 ICP를 유의수준 $\alpha \in \{0.05, 0.10, 0.20\}$으로 설정하여 실험한 결과를 표 5에 제시하였다.

$\alpha = 0.10$일 때, Random Forest, XGBoost, LightGBM 등 트리 기반 모델들은 목표 커버리지($1-\alpha = 90\%$)를 상회하는 전체 테스트 사례의 약 95%에서 실제 정답 클래스가 예측 집합 내에 포함되었다. 예측 집합의 평균 크기는 1.25~1.29이었으며, 전체 사례의 71.2~75.4%에서 단일 클래스만을 포함하는 단일 클래스 예측이 도출되었다. 이는 전체의 약 25% 미만의 사례에서만 두 클래스를 모두 포함하는 불확실 예측이 발생하였음을 의미한다. 반면, 다층 퍼셉트론(MLP)은 동일한 유의수준에서 단일 클래스 예측 비율이 40.2%에 불과하여, 예측 집합이 얼마나 구체적이고 유용한 정보를 제공하는지의 관점에서 트리

표 4. 보정 효과에 대한 Wilcoxon 검정 (ECE, Holm-Bonferroni 보정)
Table 4. Wilcoxon paired test of calibration effects on ECE (Holm-Bonferroni adjusted)

Model	Method	Δ ECE	p (Holm)	Effect
Random Forest	Platt	+0.068	<0.001	Worsened ***
Random Forest	Isotonic	−0.102	<0.001	Improved ***
Random Forest	Beta	−0.038	<0.001	Improved ***
Random Forest	Temperature	−0.076	<0.001	Improved ***
XGBoost	Platt	+0.089	<0.001	Worsened ***
XGBoost	Isotonic	−0.026	<0.001	Improved ***
LightGBM	Platt	+0.074	0.001	Worsened **
LightGBM	Isotonic	−0.028	0.001	Improved **
MLP	Platt	+0.052	0.018	Worsened *
MLP	Isotonic	−0.029	0.018	Improved *

기반 모델에 비해 현저히 낮은 성능을 보였다.

α = 0.20일 때, 일부 사례에서 어떤 클래스도 포함하지 않는 공집합(empty set) 예측이 발생하였으며, 트리 기반 모델에서는 약 4에서 5%의 사례가 이에 해당하였다. 이러한 공집합 예측은 모델이 해당 사례에 대한 분류를 명시적으로 거부하는 것으로 해석할 수 있으며, 임상적 의사결정 상황에서 불확실성이 높은 사례를 전문가에게 재검토하도록 회부하는 거부 옵션(reject option)으로 활용할 수 있다.

4.5 임상적 유용성

그림 3의 DCA는 의사가 치료한다고 판단하는 기준 확률인 임상 의사결정 임계 확률(pt)이 변화함에 따라, 보정된 확률과 보정 전 확률(raw probability)의 순편익이 어떻게 달라지는지를 비교하였다.

Random Forest 모델의 경우, 보정되지 않은 원래 확률은 모델이 지나치게 높은 확신을 나타내기 때문에, p_t = 0.30~0.45 구간에서 순편익이 모든 환자를 일률적으로 양성으로 처리하는 'treat-all' 전략과 유사한 수준까지 하락하였다. 반면, 등위 회귀 또는 Temperature Scaling 방법으로 확률을 보정한 경우에는 동일한 구간에서도

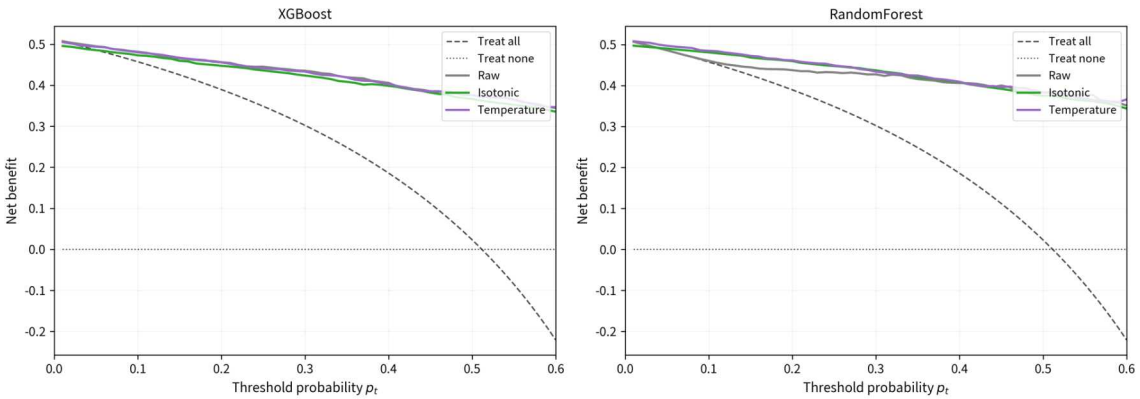

그림 3. XGBoost(좌)와 Random Forest(우)의 보정 전·보정 후 예측에 대한 DCA
Fig. 3. DCA comparing raw and calibrated predictions for XGBoost (left) and Random Forest(right)

표 5. 귀납적 몬드리안 컨포멀 예측 (평균 ± 표준편차)
Table 5. Inductive Mondrian Conformal Prediction (mean ± std)

Model	α	Coverage	Avg. set size	Singleton rate	Empty rate
LogReg	0.05	1.000±0.000	2.000±0.000	0.000±0.000	0.000±0.000
LogReg	0.1	0.961±0.034	1.512±0.148	0.488±0.148	0.000±0.000
LogReg	0.2	0.851±0.080	1.076±0.100	0.899±0.073	0.013±0.030
SVM_RBF	0.05	1.000±0.000	2.000±0.000	0.000±0.000	0.000±0.000
SVM_RBF	0.1	0.953±0.052	1.400±0.191	0.600±0.191	0.000±0.000
SVM_RBF	0.2	0.850±0.091	1.078±0.168	0.869±0.130	0.027±0.046
Random Forest	0.05	1.000±0.000	2.000±0.000	0.000±0.000	0.000±0.000
Random Forest	0.1	0.956±0.040	1.246±0.145	0.754±0.145	0.000±0.000
Random Forest	0.2	0.848±0.074	0.967±0.090	0.934±0.069	0.050±0.069
XGBoost	0.05	1.000±0.000	2.000±0.000	0.000±0.000	0.000±0.000
XGBoost	0.1	0.953±0.053	1.288±0.154	0.712±0.154	0.000±0.000
XGBoost	0.2	0.853±0.079	0.985±0.097	0.925±0.062	0.045±0.059
LightGBM	0.05	1.000±0.000	2.000±0.000	0.000±0.000	0.000±0.000
LightGBM	0.1	0.954±0.057	1.287±0.164	0.713±0.164	0.000±0.000
LightGBM	0.2	0.856±0.060	0.981±0.079	0.939±0.052	0.040±0.048
MLP	0.05	1.000±0.000	2.000±0.000	0.000±0.000	0.000±0.000
MLP	0.1	0.966±0.047	1.598±0.200	0.402±0.200	0.000±0.000
MLP	0.2	0.869±0.100	1.188±0.199	0.773±0.151	0.019±0.049

순편익이 일관되게 더 높게 유지되었다. 이는 확률 보정이 단순한 통계적 개선에 그치지 않고 실제 임상 의사결정의 유용성에 직접적인 영향을 미친다는 STRATOS 2025[3]의 권고사항을 실제 데이터 분석을 통해 확인하였다.

아울러, 선택적 분류 분석에서는 각 예측값의 확신도, 즉 |p − 0.5| 값을 기준으로 사례를 내림차순 정렬한 뒤, 확신도가 높은 상위 일정 비율의 사례만을 자동으로 처리할 때 실제로 얻을 수 있는 정확도를 평가하였다. 분석 결과, 등위 회귀로 보정된 Random Forest 및 XGBoost 모델은 확신도 기준 상위 30% 예측에서 96~97%의 정확

표 6. 평균 |SHAP| 합산 기준 상위 8개 태스크 (전체 데이터 재학습 XGBoost)
Table 6. Top 8 tasks ranked by aggregated mean |SHAP| (XGBoost refit on full data)

Rank	Task	Σ|SHAP|	Task content (description)
1	Task 23	1.342	Number sequence (1, 2, 3, ...)
2	Task 19	1.239	Cursive lowercase letter chain
3	Task 17	0.882	Word copying (in cursive)
4	Task 5	0.594	Two-pair sequence drawing
5	Task 8	0.568	Pentagons (visuospatial copy)
6	Task 3	0.438	Spiral drawing
7	Task 9	0.383	Clock drawing (memory)
8	Task 22	0.329	Sentence dictation

도를 기록하였다. 이는 전체 사례의 70%를 자동으로 처리하고 나머지 30%는 전문의가 직접 검토하는 AI가 명확한 사례를 자동 처리하고 불확실한 사례만 전문의가 검토하는 방식이 실제 임상에서 적용 가능한 수준임을 보여주었다.

4.6 SHAP 기반 해석가능성

전체 데이터로 재학습한 XGBoost 모델에 TreeSHAP[11]을 적용하여 각 특징(feature)이 예측에 기여하는 정도를 전역적(global)으로 분석하였으며, 그 결과를 표 6에 제시하였다.

TreeSHAP은 트리 기반 앙상블 모델의 구조를 직접 활용하여 SHAP 값을 정확하고 효율적으로 계산하는 알고리즘으로, XGBoost와 같은 트리 기반 모델에 적용 가능하다. 본 연구에서는 선택적 분류 분석에서 높은 예측 성능을 보인 XGBoost 모델에 수행하였다.

태스크별 진단 기여도를 분석한 결과, 23번, 19번, 17번, 5번, 8번 태스크 순으로 중요도가 높게 나타났다. 이는 태스크 8·15·17·19 및 태스크 23·19의 중요성을 보고한 Demircioglu Diren[16]의 선행 연구와 부분적으로 일치하는 결과이다. 특징 유형별로는 총 소요 시간(total_time), 공중 체류 시간(air_time), 평균 필압(pressure_mean)이 진단에 가장 크게 기여하는 것으로 확인되었다. 이러한 결과는 알츠하이머병(AD) 환자에게서 나타나는 운동 둔화(bradykinesia) 및 미세 떨림(micro-tremor) 현상을 반영하는 것으로 해석할 수 있다.

4.7 구성 요소 기여도 분석

프레임워크의 안정성을 검증하기 위해 네 가지 추가 실험을 수행하였다.

첫째, 태스크 수를 줄인 실험에서는 전체 25개 태스크를 14개로 축소하더라도 Random Forest의 AUC가 0.954에서 0.958로 사실상 동일하게 유지되었다. 나아가 선행 연구[16]에서 제안한 핵심 4개 태스크(8, 15, 17, 19)만 사용하는 경우에도 AUC 0.929를 기록하였다. 이 결과는 임상 현장에서 태스크 수를 대폭 줄인 간소화된 검사 절차만으로도 충분한 분류 성능을 확보할 수 있음을 시사한다.

둘째, 입력 특징 수를 줄인 실험에서는 SHAP 중요도 기준 상위 50개 특징만을 선택하였을 때 Random Forest AUC가 오히려 0.956에서 0.973으로 향상되었다. 이는 표본 크기가 174명에 불과한 소규모 데이터세트 환경에서 불필요한 특징이 모델 성능을 저하시키는 차원의 저주(curse of dimensionality)가 실제로 발생하고 있음을 보여 준다.

셋째, 보정 세트 크기를 변경한 실험에서는 비율을 15%, 25%, 35%로 달리 설정하더라도 ICP의 커버리지는 모두 0.93~0.97 범위 내에서 안정적으로 유지되었다. 그러나 예측 집합이 단일 레이블로 결정되는 단일 클래스 예측 비율은 단조롭게 변하지 않고 불규칙한 양상을 보였다.

넷째, 시간 계열 특징만을 사용한 실험에서는 필기 중 공중 이동 시간(air_time), 종이 접촉 시간(paper_time), 총 소요 시간(total_time)에서 파생된 75개 특징만으로도 Random Forest는 AUC 0.954를 유지하였다. 이는 시간 관련 특징이 진단에 가장 유효한 정보를 담고 있음을 나타내며, 앞서 수행한 SHAP 분석 결과와도 일치함을 확인하였다.

5. 결론

본 연구는 손글씨 기반 알츠하이머병 진단 데이터세트에 확률 보정과 컨포멀 예측 기법을 적용하여 6가지 머신러닝 모델을 동일한 기준으로 비교·평가하였다. 첫째, Random Forest가 분류 성능(AUC=0.957)은 가장 높았으나, 예측 확률의

신뢰도 오차(ECE=0.192)도 가장 커, 정확도와 확률 신뢰도가 반드시 비례하지 않음을 확인하였다. 둘째, 등위 보정은 Random Forest의 ECE를 53% 감소시키면서도 분류 정확도를 유지한 반면, Platt scaling은 트리 계열 모델의 신뢰도를 오히려 저하시켰다. 셋째, 몬드리안 ICP는 약 95% 커버리지와 71~75%의 단일 예측 비율을 달성해 임상 현장에서 활용 가능한 불확실성 정보를 제공하였다. 넷째, 결정곡선분석과 SHAP 분석 결과, 보정된 모델이 임상적 유용성과 해석 가능성 모두에서 우수하였다.

본 연구의 한계는 다음과 같다. 표본 수(n=174)가 작아 보정 집합 크기에 따른 불안정한 결과가 나타났으며, 이탈리아 단일 코호트 자료이므로 한국 노인 한글 손글씨 등 외부 집단으로의 일반화에는 추가 검증이 필요하다. 또한 이진 분류(AD vs. 정상)에 한정되어 경도인지장애(Mild Cognitive Impairment, MCI) 단계는 다루지 않았다. 향후에는 적응적 구간 분할 등 보완 지표를 함께 연구할 필요가 있다.

이 논문은 부산대학교 기본연구
지원사업(2년)에 의하여 연구되었음

참 고 문 헌

[1] World Health Organization, "Dementia", WHO Fact Sheet, Mar. 2025, https://www.who.int/news-room/fact-sheets/detail/dementia

[2] Nicole D. Cilia, Giuseppe De Gregorio, Claudio De Stefano, Francesco Fontanella, Angelo Marcelli, Antonio Parziale, "Diagnosing Alzheimer's disease from on-line handwriting: A novel dataset and performance benchmarking", Engineering Applications of Artificial Intelligence, vol.111, p.104822, May 2022, DOI : /10.1016/j.engappai.2022.104822

[3] Ben Van Calster, Gary S. Collins, Andrew J. Vickers, Laure Wynants, Kathleen F. Kerr, et al., "Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: overview and guidance", The Lancet Digital Health, vol.7, no.12, p.100916, Dec. 2025, DOI : /10.1016/j.landig.2025.100916

[4] Karim Lekadir, Alejandro F. Frangi, Antonio R. Porras, Ben Glocker, et al. (FUTURE-AI Consortium), "FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare", BMJ, vol.388, e081554, pp.1-22, Feb. 2025, DOI : https://doi.org/10.1136/bmj-2024-081554

[5] John C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods", in Advances in Large Margin Classifiers (A. J. Smola, P. L. Bartlett, B. Schölkopf, D. Schuurmans, eds.), MIT Press, 1999, pp.61-74, ISBN 978-0-262-19448-8

[6] Bianca Zadrozny, Charles Elkan, "Transforming classifier scores into accurate multiclass probability estimates", in Proc. 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2002, pp.694-699, DOI : https://doi.org/10.1145/775047.775151

[7] Meelis Kull, Telmo M. Silva Filho, Peter Flach, "Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration", Electronic Journal of Statistics, vol.11, no.2, pp.5052-5080, 2017, DOI : https://doi.org/10.1214/17-EJS1338SI

[8] Chuan Guo, Geoff Pleiss, Yu Sun, Kilian Q. Weinberger, "On calibration of modern neural networks", in Proc. 34th Int. Conf. on Machine Learning (ICML), PMLR

vol.70, 2017, pp.1321-1330, https://proceedings.mlr.press/v70/guo17a.html
[9] Andrew J. Vickers, Elena B. Elkin, "Decision curve analysis: A novel method for evaluating prediction models", Medical Decision Making, vol.26, no.6, pp.565-574, Nov. 2006, DOI:/10.1177/0272989X06295361
[10] Yonatan Geifman, Ran El-Yaniv, "Selective classification for deep neural networks", in Advances in Neural Information Processing Systems 30 (NeurIPS), 2017.
[11] Scott M. Lundberg, Su-In Lee, "A unified approach to interpreting model predictions", in Advances in Neural Information Processing Systems 30 (NeurIPS), 2017, pp.4765-4774, https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
[12] Pakize Erdogmus, Abdullah Talha Kabakus, "The promise of convolutional neural networks for the early diagnosis of Alzheimer's disease", Engineering Applications of Artificial Intelligence, vol.123, p.106254, Aug. 2023, DOI : /10.1016/j.engappai.2023.106254
[13] Jana Sweidan, Mounim A. El-Yacoubi, Anne-Sophie Rigaud, "Explainability of CNN-based Alzheimer's disease detection from online handwriting", Scientific Reports, vol.14, no.1, p.22108, 2024, DOI : https://doi.org/10.1038/s41598-024-72650-2
[14] U. Mitra, S. U. Rehman, "ML-powered handwriting analysis for early detection of Alzheimer's disease", IEEE Access, vol.12, pp.69031-69050, 2024, DOI: 10.1109/ACCESS.2024.3401104
[15] Changqing Gong, Huafeng Qin, Mounîm A. El-Yacoubi, "Hybrid Transformer for early Alzheimer's detection: Integration of handwriting-based 2D images and 1D signal features", arXiv:2410.10547, Oct. 2024, https://arxiv.org/abs/2410.10547
[16] Deniz Demircioglu Diren, "Design and validation of a hybrid machine learning model for Alzheimer's detection using handwriting data", American Journal of Alzheimer's Disease & Other Dementias, vol.40, p.15333175251374913, Sep. 2025, DOI : https://doi.org/10.1177/15333175251374913
[17] Ngoc T. N. Ho, Paulina Gonzalez, Gideon K. Gogovi, "Writing the signs: An explainable machine learning approach for Alzheimer's disease classification from handwriting", Healthcare Technology Letters, vol.12, e70006, Feb. 2025, DOI : https://doi.org/10.1049/htl2.70006
[18] Glenn Shafer, Vladimir Vovk, "A tutorial on conformal prediction", Journal of Machine Learning Research, vol.9, pp.371-421, 2008, https://www.jmlr.org/papers/v9/shafer08a.html
[19] Telma Pereira, Sandra Cardoso, Manuela Guerreiro, Alexandre de Mendonça, Sara C. Madeira, "Targeting the uncertainty of predictions at patient-level using an ensemble of classifiers coupled with calibration methods, Venn-ABERS, and conformal predictors: A case study in AD", Journal of Biomedical Informatics, vol.101, p.103350, Jan. 2020, DOI : https://doi.org/10.1016/j.jbi.2019.103350
[20] Klaus-Rudolf Kladny, Bernhard Schölkopf, Lisa. Koch, Christian F. Baumgartner, Michael Muehlebach, "A critical perspective on finite sample conformal prediction theory in medical applications", arXiv:2512.14727, Dec. 2025, https://arxiv.org/abs/2512.14727

저 자 소 개

김영인(Young-In Kim)

1996.2 명지대학교 컴퓨터공학과 박사
1996.4-2006.2 밀양대학교 컴퓨터공학부 교수
2007.8-2008.7 Univ. of Missouri 방문교수
2006.3-현재 : 부산대학교 IT응용공학과 교수
<주관심분야> 데이터베이스 시스템,
데이터마이닝, 기계학습