Research article
Food Eng. Prog. 2026;30:260019
pISSN: 1226-4768, eISSN: 2288-1247
https://doi.org/10.13050/foodengprog.260019

산업식품공학
Food Engineering Progress

강력 밀가루의 품질 특성 기반 다변량 예측 모델을 이용한 용도 분류

임유정 · 김태훈 · 이형수*
대한제분㈜ 식품과학연구소

End-use classification of strong wheat flour using a multivariate predictive model based on quality characteristics

Yu-Jeung Lim, Tae-Hoon Kim and Hyeong Su Lee*
Food Research Institute, Daehan Flour Mills Co., Ltd., Incheon 22300, Korea

Abstract

Here, we evaluated the feasibility of classifying the end uses of strong commercial wheat flours using routinely measured physicochemical properties. Twenty flour samples were analyzed for moisture, ash, protein, water absorption, peak viscosity, particle size, and damaged starch; these samples were then assigned to three predefined end-use groups: lean fermented dough (n=4), non-fermented dough (n=9), and high-formula fermented dough (n=7). Principal component analysis and k-means clustering were performed after Z-score standardization. PC1 and PC2 explained 41.2% and 23.5% of the total variance, respectively, and the three clusters showed a 95.0% agreement with the predefined end-use labels. A softmax-based multinomial logistic regression model achieved an accuracy of 0.900, balanced accuracy of 0.833, macro-F1 score of 0.849, and macro-one-vs.-rest AUC of 0.997 under leave-one-out cross-validation. Application to four additional flour samples A identified lean fermented dough as the primary candidate for samples A–C and non-fermented dough for Sample D. These results suggest that routine quality measurements may provide an exploratory framework for preliminary end-use screening and the prioritization of subsequent product tests. However, further validation of our findings using larger independent datasets, multiple production lots, and standardized product tests remains necessary.

Keywords: Strong wheat flour, Principal component analysis, K-means clustering, Multinomial logistic regression

서 론

밀가루는 제빵 공정에서 구조 형성과 품질 발현에 핵심적인 역할을 하는 원료이며, 특히 강력분은 글루텐 네트워크를 형성하여 반죽의 가스 보유력, 성형성 및 최종 제품의 체적과 조직에 영향을 미친다(Shewry et al., 2002; Wieser, 2007). 전통적으로 밀가루의 품질 평가는 단백질 함량을 중심으로 이루어져 왔으며, 단백질 함량은 글루텐 형성 잠재력과 밀접하게 관련되어 반죽 물성과 제빵 성능을 예측하는 주요 지표로 활용되어 왔다(Dobraszczyk & Morgenstern, 2003). 그러나 동일하거나 유사한 단백질 함량을 갖는 밀가루 사이에도 반죽 특성과 최종 제품 품질의 차이가 발생할 수 있으며, 이는 단백질 함량 단일 지표만으로 가공 적합성을 충분히 설명하기 어렵다는 점을 시사한다(Wieser, 2007; Delcour & Hoseney, 2010).

이러한 차이는 글루텐 단백질의 조성 및 중합 상태와 같은 질적 특성뿐 아니라 전분의 물리화학적 특성, 입도 분포, 손상전분 함량 및 수분 흡수 특성 등 다양한 요인의 영향을 받는다(Lindeboom et al., 2004; Delcour & Hoseney, 2010). 제분 중 생성되는 손상전분은 정상 전분보다 높은 수화성과 효소 반응성을 나타내어 밀가루 흡수율과 반죽 특성에 영향을 미친다. 다만 손상전분의 영향은 제품군과 함량 수준에 따라 달라질 수 있으며, 과도한 손상전분은 반죽의

Received: Sep 2, 2026 / Revised: Oct 6, 2026 / Accepted: Oct 8, 2026 / Published: Nov 9, 2026
Corresponding author: Hyeong Su Lee, Food Research Institute, Daehan Flour Mills Co., Ltd., Incheon 22300, Korea
E-mail: hsses@naver.com

Copyright © 2026 Korean Society for Food Engineering.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

끈적임이나 최종 제품의 조직 저하를 유발할 수 있다(Barrera et al., 2007; Liu et al., 2014). Amylograph 최고점도는 밀가루 현탁액의 가열 중 점도 발현을 나타내며, 전분의 호화 특성뿐 아니라 α-amylase 활성, 현탁액 농도 및 분석조건의 복합적인 영향을 받는다(Posner & Hibbs, 2005).

이와 같이 강력분의 품질은 다수의 물리·화학적 변수로 구성되는 다변량 구조를 가지며, 각 변수는 독립적으로 작용하기보다 상호 연관되어 최종 품질에 기여한다. 따라서 단일 변수 중심의 경험적 평가만으로는 산업 현장에서 발생하는 복합적인 품질 변동을 충분히 설명하기 어렵다. 주성분분석(principal component analysis, PCA)은 상관된 다수의 변수를 서로 직교하는 소수의 축으로 축약하여 시료의 품질 구조와 변수의 기여 방향을 해석하는 데 유용하다(Jolliffe, 2002). K-means 군집분석은 사전에 정답 범주를 부여하지 않고 시료 간 유사성을 바탕으로 탐색적 집단을 형성하는 방법이며(Hartigan & Wong, 1979), 식품 품질 데이터의 시각화와 구조적 해석에도 활용되고 있다(Marti et al., 2015; Granato et al., 2018).

한편 산업 현장에서는 단순한 이화학적 품질 분류를 넘어 특정 제품군과 공정조건에 대한 밀가루의 상대적 적합성을 판단할 필요가 있다. 본 연구에서는 강력분 용도를 저배합 발효 반죽(lean fermented dough), 비발효 반죽(non-fermented dough) 및 고배합 발효 반죽(high-formula fermented dough)의 세 범주로 정의하였다. 저배합 발효 반죽은 효모를 이용하면서 당 및 유지의 배합 수준이 상대적으로 낮은 제품군으로 식빵, 바게트 및 일부 피자 도우 등을 포함하였다. 비발효 반죽은 효모 발효를 전제로 하지 않고 성형성, 점도 또는 퍼짐성이 중요한 면류, 쿠키 및 일부 비효모 크래커류를 포함하였다. 고배합 발효 반죽은 효모를 이용하면서 상대적으로 높은 수준의 당 또는 유지가 포함되는 제품군으로 단과자빵과 브리오슈류를 포함하였다(Cauvain & Young, 2007). 본 연구에서는 분류의 일관성을 확보하기 위한 조작적 기준으로 당 또는 유지 중 하나 이상의 배합량이 밀가루 중량 대비 10% 이상인 경우를 고배합으로 정의하였다. 이 기준은 보편적인 생화학적 분류 기준이 아니라 본 연구의 용도 분류 모델을 위해 사전에 설정한 기준이다.

다항 로지스틱 회귀(multinomial logistic regression)는 입력 변수에 따라 각 클래스에 속할 상대확률을 추정할 수 있는 지도학습 방법이다(Hosmer et al., 2013). 그러나 지도학습 모형의 타당성은 종속변수인 용도 라벨의 정의와 검증 절차에 크게 좌우된다. 본 연구에서는 PCA 및 K-means 분석 결과와 모델 예측결과를 확인하기 전에 각 제품 제조사가 제시한 주 용도와 대한제분 내부 제품 포지셔닝 기록을 검토하여 시료별 용도 라벨을 부여하였다.

따라서 본 연구에서는 시판 강력분 20종의 7개 이화학적 품질 특성을 이용하여 PCA와 K-means로 다변량 품질구조를 해석하고, 사전 부여한 용도 라벨을 기반으로 Softmax 다항 로지스틱 회귀모형을 개발한 후 내부 교차검증과 추가 시료 적용을 통해 용도 분류 가능성을 평가하고자 한다.

재료 및 방법

실험 재료

본 연구에는 국내·외에서 시판 중인 강력분 20종을 실험 재료로 사용하였으며, 제품의 단백질 및 회분 함량, 제분 특성, 이화학적 특성 및 제품 포지셔닝의 다양성을 고려하여 선정하였다. 제품명과 제조사는 분석 결과로부터 특정 상업 제품이 직접 식별되는 것을 최소화하기 위해 S1~S20으로 익명화하였다. 각 시료는 밀폐 용기에 보관하고 20±2℃에서 24시간 온도 평형한 후 분석하였다. 모형의 적용 가능성을 탐색하기 위해 학습자료에 포함되지 않은 시판 강력분 4종을 추가로 분석하고 Sample A~D로 구분하였다. 이들 추가 시료는 기준 용도 라벨과 독립적인 제품 시험 결과가 확보되지 않은 미지 시료로 사용하였다.

일반 성분 및 물성 분석

수분, 회분, 단백질, 흡수율, Amylograph 최고점도, 입도 및 손상전분을 동일 시료에서 3회 반복 측정하였고, 결과는 평균±표준편차로 나타내었다. 다변량 분석에는 각 시료의 반복 측정 평균값을 사용하였다(n=20).

수분 함량은 할로겐 방식 수분측정기(Moisture Analyzer ML-50, A&D Co., Japan)를 이용하여 시료 2 g을 105℃의 표준 가열모드에서 측정하였으며, 기기의 자동 종료조건에 따라 측정을 종료하였다.

회분 함량은 AACC Method 08-01.01을 참고하되 본 연구의 분석조건에 따라 시료 3 g을 620~630℃에서 3시간 회화하고 데시케이터에서 1시간 방랭한 후 측정하였다(AACC International, 2010).

단백질 함량은 듀마스 연소법(Dumas combustion method)에 기반한 질소분석기(Rapid N Exceed, Elementar, Germany)를 이용하여 AACC Method 46-30.01을 참고하여 측정하였다(AACC International, 2010). 측정된 총질소 함량에 질소-단백질 환산계수 5.7을 적용하였다. 단백질 함량과 Farinograph 흡수율은 14% 수분 기준으로 환산하였다.

흡수율은 Farinograph (Brabender GmbH, Germany)의 300 g bowl을 이용하여 AACC Approved Method 54-21.02의 constant flour weight 절차를 참고하여 30℃에서 측정하였다. 반죽 중심선이 500 Brabender Unit (BU)에 도달하는 데 필요한 수분량을 흡수율(%)로 정의하였다.

Amylograph 최고점도는 밀가루 현탁액의 가열 중 점도 발현 특

성과 α-amylase 활성의 복합적인 영향을 반영하는 지표로 측정하였다. Amylograph (Brabender GmbH, Germany)를 이용하여 AACC Approved Method 22-10.01을 참고하였으며, 14% 수분 기준 밀가루 65 g에 증류수 450 mL를 첨가하였다. 현탁액을 30℃에서 95℃까지 1.5℃/min의 속도로 가열하면서 75 rpm으로 교반하고, 가열 중 나타난 최고점도(peak viscosity)를 BU로 기록하였다.

입도 분포는 레이저 회절 입도분석기(Mastersizer 3000, Malvern Instruments Ltd., UK)의 건식분산 장치(Aero S)를 이용하여 분산압력 2.0 bar, 입자 굴절률 1.52, 흡광계수 0.1 및 obscuration 0.5~6%로 설정하였으며, 체적 기준 중앙입경인 Dv(50)을 대표 입도값으로 사용하였다.

손상전분은 SDmatic (Starch Damage Analyzer, CHOPIN Technologies, France)를 이용하여 AACC Approved Method 76-33.01에 따라 측정하였으며, 기기에서 산출한 AACC 환산 손상전분 함량(%)으로 표시하였다.

용도군의 사전 정의 및 라벨 부여

본 연구에서는 저배합 발효 반죽, 비발효 반죽 및 고배합 발효 반죽의 세 용도군(end-use groups)을 정의하였다. 각 시료의 사전 정의 용도 라벨(predefined end-use label)은 PCA, K-means 및 Softmax 모형의 분석 결과를 확인하기 전에 제품 제조사가 제시한 주 용도와 대한제분 내부 제품 포지셔닝 기록을 근거로 부여하였다.

복수 용도로 판매되는 제품은 제조사가 제시한 대표적인 1차 용도를 우선 적용하였으며, 제조사의 제시 용도와 내부 제품 포지셔닝 기록이 일치하지 않거나 분류가 불명확한 경우 연구진의 합의를 통해 최종 용도 라벨을 부여하였다. 저배합 발효군은 S3, S8, S12 및 S13(n=4), 비발효군은 S1, S2, S4~S7 및 S9~S11(n=9), 고배합 발효군은 S14~S20(n=7)으로 구성하였다.

데이터 전처리

변수별 단위와 분산 차이에 따른 영향을 제거하기 위해 7개 입력변수에 Z-score 표준화를 적용하였다(Jolliffe, 2002).

$$Z_{ij} = \frac{X_{ij} - \mu_j}{\sigma_j}$$

여기서 X_{ij}는 i번째 시료의 j번째 변수 값, μ_j와 σ_j는 각각 j번째 변수의 학습자료 평균과 표준편차이다. PCA와 최종 Softmax 모형에는 전체 학습자료 20종에서 산출한 평균과 표준편차를 적용하였다. LOOCV (leave-one-out cross-validation, 한 개 제외 교차검증)는 전체 시료에서 한 개를 검증용으로 제외하고 나머지 시료로 모델을 학습하는 과정을 모든 시료에 대해 반복하여 예측성능을 평가하는 방법이다. LOOCV에서는 정보 누출을 방지하기 위해 각 fold의 학습자료에서 평균과 표준편차를 다시 산출하였으며, 제외된 검증 시료에는 해당 fold에서 산출한 변환계수를 적용하였다. 추가 시료인 Sample A~D에는 전체 학습자료 20종에서 산출한 평균과 표준편차를 적용하였다.

주성분분석(PCA)

PCA는 Z-score 표준화된 7개 변수의 상관구조를 기반으로 수행하였다(Jolliffe, 2002). 각 주성분의 고유값과 설명분산비율을 산출하였으며, 시각적 해석에는 PC1과 PC2를 이용하였다. 두 주성분이 전체 변동을 모두 설명하지 않으므로 PC1-PC2 score plot은 자료의 주요 변동 경향을 탐색하는 목적으로 사용하였다. 변수의 기여 방향은 주성분 계수와 변수-주성분 간 loading correlation을 함께 확인하였다.

군집분석(K-means clustering)

PCA score plot과 군집구조를 동일한 2차원 공간에서 비교하기 위해 PC1-PC2 score를 이용하여 K-means 군집분석을 수행하였다(Hartigan & Wong, 1979). 따라서 본 군집분석은 전체 7차원 품질 공간의 유일한 최적 분할을 찾기 위한 분석이 아니라 PC1-PC2가 나타내는 주요 이화학적 변동 구조를 기술하기 위한 분석으로 정의하였다.

K=2-6에 대해 within-cluster sum of squares (WCSS)와 silhouette coefficient를 비교하였다(Rousseeuw, 1987). 군집 수는 silhouette coefficient만으로 결정하지 않고 WCSS의 변화, 군집당 시료 수, 분류의 단순성 및 산업적 해석 가능성을 종합적으로 고려하였다. 본 자료에서 K=3의 평균 silhouette coefficient는 0.437이었으며, K가 증가함에 따라 더 높은 값이 나타날 수 있었다. 따라서 K=3은 통계적으로 유일한 최적 군집 수가 아니라 PC1-PC2 공간에서 주요 이화학적 품질구조를 간결하게 기술하기 위한 탐색적 군집 수로 선택하였다.

Softmax 기반 다중 클래스 예측 모델

강력분의 용도 분류 가능성을 평가하기 위해 다항 로지스틱 회귀를 적용하였다(Hosmer et al., 2013). 종속변수는 사전 부여한 저배합 발효, 비발효 및 고배합 발효의 세 클래스였으며, 독립변수는 수분, 회분, 단백질, 흡수율, 최고점도, 입도 및 손상전분의 7개 Z-score를 사용하였다. 각 클래스 (k)에 대한 선형 예측식은 다음과 같다.

$$z_k = b_k + \sum_{j=1}^{7} \omega_{kj} x_j$$

여기서 z_k는 클래스 k의 선형 예측값, b_k는 절편, w_{kj}는 j번째 변수의 회귀계수, x_j는 표준화 입력값이다. 각 클래스에 속할 상대적 소속확률은 Softmax 함수로 계산하였다.

$$P(y = k \mid x) = \frac{e^{z_k}}{\sum_c e^{z_c}}$$

모델 학습에는 L2 정규화를 적용한 lbfgs solver를 사용하였으며, 정규화 강도는 C=1.0으로 고정하였다. Softmax 출력의 합이 1이 되는 것은 함수의 정의에 따른 것이며, 산출된 값은 본 연구의 학습 시료, 변수 범위 및 사전 정의된 세 용도 클래스 내에서의 상대적 클래스 소속확률로 해석하였다.

통계분석 및 모델 평가

Pearson 상관분석은 양측 검정으로 수행하였으며, 유의수준은 $p<0.05$로 설정하였다. K-means 군집과 사전 정의한 용도 라벨의 대응관계는 각 군집에서 가장 빈도가 높은 용도를 기준으로 산출한 일치율, adjusted Rand index (ARI) 및 normalized mutual information (NMI)을 이용하여 평가하였다. NMI는 두 분류의 엔트로피에 대한 산술평균 정규화 방식으로 산출하였다(Hubert & Arabie, 1985; Vinh et al., 2010).

Softmax 모형은 훈련자료의 재분류 성능과 내부 교차검증 성능을 구분하여 평가하였다. 표본 수가 20종으로 제한적이므로 LOOCV를 적용하였다. 각 fold에서는 검증 시료 1종을 제외한 학습자료만을 이용하여 변수별 평균과 표준편차를 산출하고 Z-score 표준화를 수행한 후 모형을 적합하였다.

모형의 성능은 accuracy, balanced accuracy, class-specific recall, macro-F1, confusion matrix 및 one-vs-rest 방식의 multiclass AUC (area under the receiver operating characteristic curve, ROC) 곡선 아래 면적로 평가하였다. AUC는 분류 모델이 양성 시료에 음성 시료보다 높은 예측점수를 부여하는 정도를 종합적으로 나타내는 판별력 지표로, 1에 가까울수록 판별력이 높고 0.5는 무작위 판별 수준을 의미한다. 본 연구에서는 각 용도군을 나머지 용도군과 구분하는 one-vs-rest 방식으로 클래스별 AUC를 산출한 뒤 산술평균한 macro-AUC를 사용하였다. 확률 예측 성능은 log loss와 multiclass Brier score를 이용하여 평가하였다. Multiclass Brier score는 각 시료에서 세 클래스의 예측확률과 one-hot 실제 라벨 간 제곱오차의 합을 구한 후 전체 시료에 대해 평균하여 산출하였다. AUC, log loss 및 Brier score는 각 LOOCV fold에서 얻은 out-of-fold 예측확률을 전체 시료에 대해 통합한 후 계산하였다. 모든 통계분석과 모델링은 Python 3.10.11의 pandas, NumPy 및 scikit-learn을 이용하여 수행하였다(Pedregosa et al., 2011).

결과 및 고찰

일반 성분 및 물성 분포 특성

강력분 20종의 일반 성분 및 물성 분석 결과는 Table 1에 제시하였다. 수분 함량은 12.9~14.7% 범위였으며 평균은 13.69%였다. 회분은 0.340~0.896%, 단백질은 10.60~13.50% 범위를 나타냈다. 흡수율은 57.8~71.5%, 최고점도는 167-1084 BU, Dv(50)은 30.7~65.0 μm, 손상전분은 3.86~7.03%의 분포를 보였다.

동일한 강력분 범주 안에서도 각 변수는 비교적 넓은 범위로 분포하였으며, 단일 품질지표만으로는 시료 간 특성 차이를 충분히 표현하기 어려운 다변량적 특성이 확인되었다. 이러한 차이는 원맥 구성, 제분조건 및 제품 설계 등의 복합적인 영향을 반영했을 가능성이 있으나, 본 연구에서는 개별 시료의 원맥 배합 및 상세 제분 조건을 직접 분석하지 않았으므로 이에 대한 인과관계는 판단하지 않았다.

특히 S5의 단백질 함량은 10.60%로 본 연구에 포함된 시료 중 가장 낮게 나타났다. 일반적으로 강력분은 비교적 높은 단백질 함량을 주요 품질 특성으로 가지지만, 상업용 밀가루의 제품 분류와 적용 용도는 단백질 함량뿐 아니라 반죽 특성, 제조사의 제품 설계 및 실제 적용 용도 등을 종합하여 설정될 수 있다. 본 연구에서는 단백질 함량을 기준으로 시료를 재분류하지 않고, 시판 제품의 제조사 제시 용도와 제품 포지셔닝을 기준으로 연구대상을 선정하였다. 따라서 S5의 낮은 단백질 함량은 본 연구에서 조사한 시판 강력분 간 품질 특성의 변이를 보여주는 사례이며, 단일 품질지표보다는 다변량 품질 정보를 함께 고려할 필요성을 보여준다.

Pearson 상관분석 결과(Fig. 1), 흡수율은 손상전분과 강한 양의 상관을 나타냈으며($r=0.742$, $p<0.001$), Amylograph 최고점도와도 양의 상관을 보였다($r=0.570$, $p<0.01$). 최고점도와 손상전분 사이에도 양의 상관이 확인되었고($r=0.499$, $p<0.05$), 회분과 손상전분은 음의 상관을 나타냈다($r=-0.457$, $p<0.05$). 반면 단백질과 흡수율 간 상관관계는 통계적으로 유의하지 않았다. 따라서 본 시료군에서 흡수율은 단백질 함량보다 손상전분 및 Amylograph 최고점도와 뚜렷한 동반 변화를 나타냈으며, 수분 흡수 특성을 단백질 함량만으로 설명하기 어렵다는 점을 보여준다.

한편 개별 시료에서는 단백질 함량과 Amylograph 최고점도가 반드시 동일한 방향으로 변화하지 않았다. 특히 S17과 S18의 단백질 함량은 각각 12.30%와 13.30%였으나 최고점도는 각각 167 및 195 BU로 비교적 낮게 나타났다. 본 연구에서도 단백질과 최고점도 사이에는 유의한 상관관계가 확인되지 않아($r=-0.393$, $p>0.05$), 최고점도의 변화를 단백질 함량만으로 설명하기는 어려웠다. Amylograph 최고점도는 가열 중 전분의 팽윤 및 호화 특성뿐 아니라 α-amylase 활성 등 다양한 요인의 영향을 받는 것으로 알려져

Table 1. General composition and physicochemical properties of 20 strong wheat flour samples

	Moisture (%)	Ash (%)	Protein (%)	Water absorption (%)	Peak viscosity (BU)	Particle size, Dv(50) (μm)	Damaged starch (%)
S1	13.5±0.1	0.340±0.005	12.00±0.12	67.6±0.4	673±12	41.9±0.5	4.47±0.08
S2	13.9±0.2	0.363±0.004	12.10±0.09	66.3±0.5	891±15	50.7±0.7	5.72±0.11
S3	13.2±0.1	0.507±0.008	13.30±0.15	64.3±0.3	609±11	50.8±0.6	5.33±0.09
S4	13.8±0.1	0.373±0.006	11.70±0.11	63.4±0.4	632±14	43.1±0.4	6.61±0.13
S5	14.0±0.2	0.420±0.007	10.60±0.08	59.9±0.6	584±10	45.2±0.5	4.70±0.07
S6	14.7±0.2	0.527±0.009	13.00±0.14	61.8±0.5	473±9	65.0±0.8	5.00±0.10
S7	13.7±0.1	0.457±0.005	11.10±0.10	63.5±0.4	1,084±22	61.0±0.9	4.99±0.09
S8	13.3±0.1	0.350±0.004	12.10±0.11	71.5±0.7	894±16	42.0±0.5	7.03±0.12
S9	14.7±0.2	0.354±0.005	12.00±0.13	63.0±0.5	934±19	44.0±0.6	5.52±0.11
S10	14.0±0.1	0.393±0.006	12.30±0.12	65.0±0.4	685±13	54.6±0.7	5.45±0.08
S11	13.8±0.2	0.384±0.005	11.90±0.09	64.4±0.5	761±14	50.3±0.6	5.57±0.10
S12	13.0±0.1	0.546±0.008	12.40±0.11	70.7±0.6	977±18	42.0±0.4	5.76±0.12
S13	12.9±0.1	0.380±0.004	12.70±0.14	66.5±0.4	931±17	39.9±0.5	5.51±0.09
S14	14.3±0.2	0.466±0.007	12.40±0.12	57.8±0.3	593±11	44.2±0.6	4.33±0.07
S15	13.3±0.1	0.490±0.008	12.30±0.10	59.8±0.4	767±15	46.0±0.5	4.02±0.08
S16	13.8±0.1	0.576±0.009	13.50±0.16	60.9±0.5	554±10	30.9±0.4	4.27±0.06
S17	13.0±0.1	0.519±0.007	12.30±0.11	60.0±0.4	167±5	35.0±0.4	4.27±0.08
S18	13.5±0.2	0.516±0.008	13.30±0.14	60.8±0.5	195±6	30.7±0.3	4.36±0.07
S19	13.9±0.1	0.483±0.006	12.30±0.12	59.1±0.4	658±12	36.5±0.5	3.86±0.06
S20	13.6±0.1	0.896±0.012	12.60±0.13	62.2±0.4	450±8	37.3±0.4	4.43±0.07

Values are presented as mean±SD of triplicate analytical determinations performed on the same commercial flour sample (n=3 analytical replicates). The triplicate determinations do not represent independent manufacturing lots. Samples were anonymized and coded as S1–S20.

있다(Posner & Hibbs, 2005). 따라서 S17과 S18에서 관찰된 낮은 최고점도는 단백질 함량과 직접적으로 연계하기보다는 전분 특성과 효소 활성 등을 포함한 품질 차이의 가능성을 함께 고려하여 해석하는 것이 적절하다. 다만 본 연구에서는 α-amylase 활성이나 전분의 세부 특성을 직접 분석하지 않았으므로, 이러한 요인이 해당 시료의 낮은 최고점도를 유발하였다고 단정할 수는 없다.

PCA 기반 강력분 품질 구조 해석

Z-score 표준화된 데이터를 기반으로 PCA를 수행한 결과, PC1과 PC2의 설명분산비율은 각각 41.2%와 23.5%였으며 누적 설명분산은 64.7%였다(Fig. 2). Loading correlation 분석 결과(Fig. 3), PC1은 최고점도(0.839), 손상전분(0.830), 흡수율(0.751) 및 입도(0.544)와 양의 관계를 보였고, 회분(−0.707)과 단백질(−0.529)은 반대 방향에 위치하였다. PC2는 수분(0.875)과 입도(0.565)가 양의 방향, 흡수율(−0.598)과 단백질(−0.437)이 음의 방향에 위치하였다. 이러한 loading 구조를 바탕으로 PC1은 흡수율, 최고점도 및 손상전분과 회분·단백질 수준의 대비가 나타나는 주요 변동 경향으로, PC2는 수분·입도와 흡수율·단백질의 대비가 나타나는 또 다른 주요 변동 경향으로 해석할 수 있었다.

다만 PC1과 PC2의 누적 설명분산은 64.7%로 전체 변동을 모두 설명하지 않으므로, 이러한 해석은 본 연구에서 관찰된 주요 이화학적 변동 구조를 기술하는 범위로 한정하는 것이 적절하다. 따라서 본 결과는 강력분의 품질 특성이 단백질 함량의 단일 지표보다는 여러 이화학적 특성이 결합된 다변량 구조를 가질 가능성을 보여주지만, PC1과 PC2만으로 전체 품질 특성을 설명한다고 해석해서는 안 된다. 또한 PCA 축의 방향과 loading 구조는 본 연구에 포함된 20개 시료와 7개 변수의 구성에 의존하므로, 다른 원맥, 제조 로트 또는 분석변수가 포함되는 경우 달라질 수 있다.

	Moisture (%)	Ash (%)	Protein (%)	Water absorption (%)	Peak Viscosity (BU)	Particle size (Dv(50))	Damaged starch (%)
Moisture (%)	1	−0.121	−0.162	−0.390	−0.008	0.404	−0.060
Ash (%)	−0.121	1	0.391	−0.313	−0.436	−0.253	−0.457*
Protein (%)	−0.162	0.391	1	−0.049	−0.393	−0.310	−0.200
Water absorption (%)	−0.390	−0.313	−0.049	1	0.570**	0.142	0.742***
Peak viscosity (BU)	−0.008	−0.436	−0.393	0.570**	1	0.400	0.499*
Particle size (Dv(50))	0.404	−0.253	−0.310	0.142	0.400	1	0.283
Damaged starch (%)	−0.060	−0.457*	−0.200	0.742***	0.499*	0.283	1

Fig. 1. Pearson correlation matrix of the seven physicochemical variables. The values represent Pearson's correlation coefficients. Colored cells indicate the strength and direction of the relationships (Red: positive, Grean: negative). BU, Brabender Units (for Peak Viscosity); Dv(50), median particle size. Statistical significance levels are indicated as follows: $^{*}p$(0.05, $^{**}p$(0.01, and $^{***}p$(0.001 (n=20, two-tailed test).

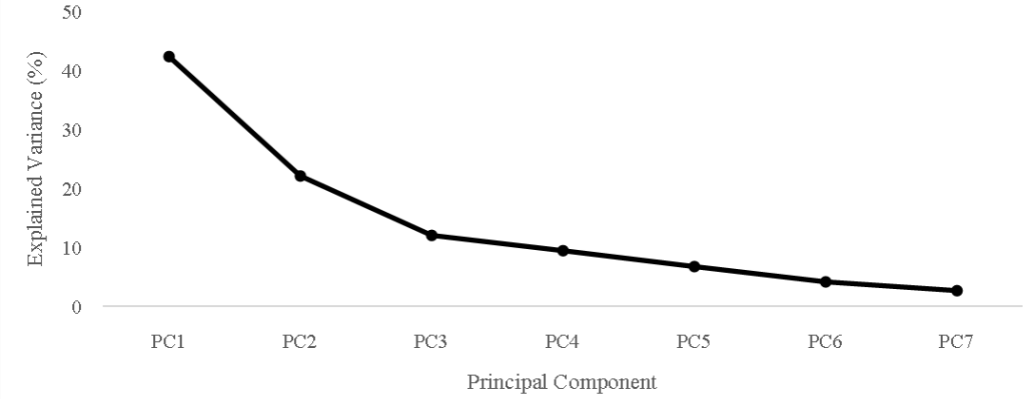


Fig. 2. Scree plot showing explained variance ratio of principal components. PC1 and PC2 explained 41.2% and 23.5% of the total variance, respectively, corresponding to a cumulative explained variance of 64.7% (n=20).

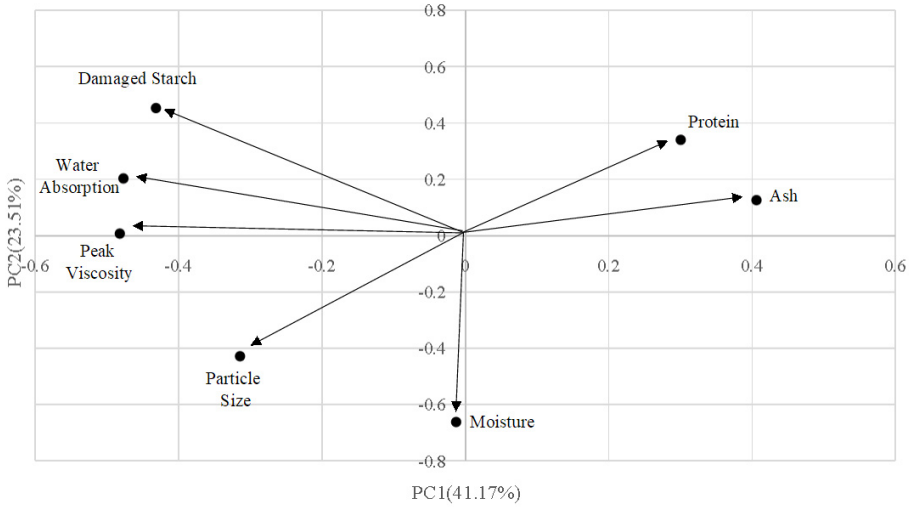


Fig. 3. PCA loading correlation plot of seven physicochemical variables. The vectors represent the correlations of each standardized physicochemical variable with PC1 and PC2 (n=20).

군집 분석 결과 및 품질군 특성

PC1-PC2 score 공간에서 K-means 군집분석을 수행한 결과, 세 개의 탐색적 품질군이 형성되었다(Fig. 4). K-means는 시료 간 거리를 기준으로 군집 내 제곱합을 최소화하는 비지도 군집분석 방법이며(Hartigan & Wong, 1979), 밀가루의 여러 품질변수로 구성된 다변량 구조를 탐색하는 데 활용될 수 있다(Marti et al., 2015).

Cluster 1은 S1, S3, S8, S12 및 S13으로 구성되었으며, 흡수율, 손상전분 및 최고점도가 상대적으로 높은 특성을 나타냈다. 따라서 이 군집은 본 연구의 시료 범위에서 수분 흡수와 전분의 점도 발현 특성이 두드러지는 품질군으로 해석되었다. 손상전분은 정상 전분보다 높은 수분 흡수능을 나타내므로 밀가루의 흡수율과 반죽 특성에 영향을 줄 수 있지만, 그 영향은 손상 정도와 제품 종류에 따라 달라질 수 있다(Barrera et al., 2007). 사전 정의한 용도 라벨과 비교한 결과, Cluster 1의 5종 중 S3, S8, S12 및 S13의 4종이 저배합 발효군에 해당하였고, S1은 비발효군에 해당하여, Cluster 1은 본 연구의 시료 범위에서 저배합 발효 용도군과 주로 대응되는 품질군으로 나타났다.

Cluster 2는 S14~S20의 7종으로 구성되었으며, 회분과 단백질이 상대적으로 높은 반면 흡수율, 최고점도, 입도 및 손상전분은 상대적으로 낮은 경향을 나타냈다. Cluster 2의 모든 시료는 사전 정의한 고배합 발효군에 해당하여, 해당 이화학적 특성의 조합이 본 연구의 고배합 발효 용도군과 높은 대응관계를 나타냈다.

Cluster 3에 포함된 모든 시료는 사전 정의한 비발효군에 해당하여, 해당 군집의 이화학적 특성이 본 연구의 비발효 용도군과 높은 대응관계를 나타냈다.

군집별 다수 용도를 기준으로 Cluster 1을 저배합 발효군, Cluster 2를 고배합 발효군, Cluster 3을 비발효군에 대응시켰을 때, 전체 20종 중 19종의 군집과 사전 정의 용도 라벨이 일치하여 95.0%의 일치율을 나타냈다. 또한 군집과 용도 라벨 간 일치도를 평가한 결과, adjusted Rand index는 0.855, normalized mutual information은 0.868로 나타났다. ARI는 우연에 의한 일치를 보정하여 두 분할 결과의 일치도를 평가하는 지표이며(Hubert & Arabie, 1985), NMI는 두 분류체계가 공유하는 정보를 각 분류의 엔트로피를 기준으로 정규화한 지표이다(Vinh et al., 2010).

이러한 결과는 제조사의 주 용도와 내부 제품 포지셔닝 자료를 바탕으로 사전 정의한 용도군이 이화학적 품질 특성만으로 형성된 비지도 군집구조와 높은 대응관계를 나타냈음을 보여준다. 즉, 수분, 회분, 단백질, 흡수율, 최고점도, 입도 및 손상전분의 조합이 본 연구에서 사전 정의한 용도군과 관련된 정보를 포함하며, 해당 품질 정보를 이용한 용도 예측모형 구축의 경험적 근거를 제공하였다. 다만 K-means 군집과 용도 라벨의 비교는 동일한 20개 시료에서 수행된 내부적 분석이므로 독립적인 타당성 검증으로 간주할 수 없다. 또한 K-means 군집은 PC1과 PC2가 설명하는 64.7%의 변동을 기반으로 형성된 이화학적 품질군이므로, 실제 제품 용도와 완전히 동일한 범주라기보다 사전 정의한 용도군을 설명하기 위한 기초적인 품질구조로 해석하는 것이 적절하다.

특히 비발효군으로 사전 분류된 S1이 저배합 발효군이 우세한 Cluster 1에 포함된 것은 사전 용도 라벨과 이화학적 군집구조가 모

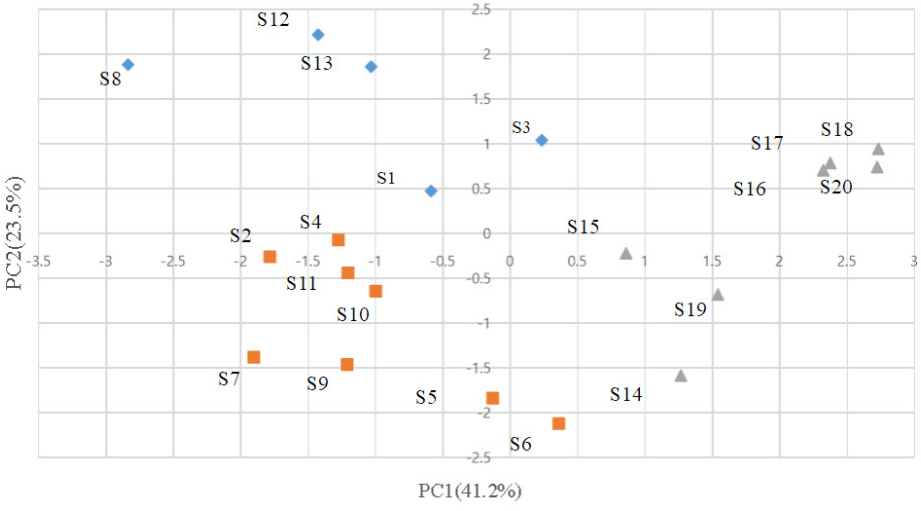


Fig. 4. PCA score plot with K-means clustering of 20 commercial strong wheat flour samples. The samples were assigned to three exploratory physicochemical clusters by K-means clustering in the PC1-PC2 score space. PC1 and PC2 explained 41.2% and 23.5% of the total variance, respectively, corresponding to a cumulative explained variance of 64.7%. Cluster numbers represent unsupervised physicochemical groups and were not assigned using the predefined end-use labels (n=20).

든 시료에서 완전히 일치하지 않음을 보여준다. 이는 일부 밀가루가 제조사의 주 용도와 다른 용도군에 가까운 이화학적 특성을 나타낼 가능성을 시사한다. 다만 군집의 불일치만으로 S1을 용도군 간 경계 시료로 단정할 수는 없다.

Softmax 기반 다중 클래스 예측 결과

사전 부여된 용도 라벨을 종속변수로 하고 수분, 회분, 단백질, 흡수율, 최고점도, 입도 및 손상전분의 표준화 값을 독립변수로 하여 Softmax 기반 다항 로지스틱 회귀모형을 구축하였다(Hosmer et al., 2013). 표준화된 입력변수에 대한 클래스별 회귀계수는 Table 2에 제시하였으며, 각 클래스별 선형 예측식은 다음과 같이 산출하였다.

$$z_{lean\ fermented\ dough} = -0.7259x_1 + 0.1035x_2 + 0.5550x_3 + 0.5879x_4 + 0.2864x_5 - 0.0176x_6 + 0.6029x_7 - 0.6728$$

$$z_{non-fermented\ dough} = +0.6636x_1 - 0.4839x_2 - 0.7305x_3 + 0.2351x_4 - 0.0899x_5 + 0.5475x_6 + 0.1081x_7 + 0.6577$$

$$z_{high-formula\ fermented} = 0.0623x_1 + 0.3803x_2 + 0.1755x_3 - 0.8231x_4 - 0.1965x_5 - 0.5299x_6 - 0.7110x_7 + 0.0151$$

여기서 z_1, z_2 및 z_3는 각각 저배합 발효, 비발효 및 고배합 발효 클래스의 선형 예측값이며, x_1=수분, x_2=회분, x_3=단백질, x_4=흡수율, x_5=최고점도, x_6=입도, x_7=손상전분의 Z-score이다. 각 클래스의 상대적 소속확률은 다음 Softmax 함수로 산출하였다.

$$P(y = k \mid x) = \frac{e^{z_k}}{\Sigma_c e^{z_c}}$$

훈련자료에 대한 재분류 결과 accuracy, balanced accuracy, macro-F1 및 macro one-vs-rest AUC는 모두 1.000으로 나타나, 적합된 모형이 학습자료의 용도 라벨을 모두 재분류하였다. 내부 예측성능을 평가하기 위해 각 시료를 한 번씩 제외하여 수행한 LOOCV에서는 accuracy 0.900, balanced accuracy 0.833, macro-F1 0.849 및 macro-AUC 0.997을 나타냈다. LOOCV 예측확률을 이용하여 산출한 log loss와 multiclass Brier score는 각각 0.260과 0.131 이었으며, 두 지표는 예측확률의 정확성을 평가하는 보완적 지표로 사용하였다(Gneiting & Raftery, 2007).

LOOCV에서 전체 20개 시료 중 18개가 사전 부여된 용도와 동일하게 분류되었다. 클래스별 recall은 저배합 발효군 0.500, 비발효군 1.00 및 고배합 발효군 1.00으로, 비발효군 9종과 고배합 발효군 7종은 모두 올바르게 분류된 반면 저배합 발효군은 4종 중 2종만 정확하게 분류되었다. 이러한 결과는 비발효군과 고배합 발효군이 본 연구의 이화학적 변수 공간에서 비교적 명확하게 구분된 반면, 저배합 발효군의 일부 시료는 인접 용도군과 중첩되는 특성을 나타냈을 가능성을 시사한다.

동시에 저배합 발효군은 다른 두 용도군에 비해 표본 수가 적어(n=4), 해당 용도군의 이화학적 특성 분포가 충분히 대표되지 않았을 가능성도 고려할 필요가 있다. 클래스 간 표본 수의 불균형은 모델 학습 과정에서 상대적으로 표본이 많은 용도군의 특성이 의사결정 경계 형성에 더 크게 반영되거나, 소수 용도군의 예측 결과가

Table 2. Standardized coefficients of the multinomial logistic regression model for end-use classification

	$z_{lean\ fermented\ dough}$	$z_{non-fermented\ dough}$	$z_{high-formula\ fermented\ dough}$
Moisture (x_1)	−0.7259	0.6636	0.0623
Ash (x_2)	0.1035	−0.4839	0.3803
Protein (x_3)	0.5550	−0.7305	0.1755
Water absorption (x_4)	0.5879	0.2351	−0.8231
Peak viscosity (x_5)	0.2864	−0.0899	−0.1965
Particle size (x_6)	−0.0176	0.5475	−0.5299
Damaged starch (x_7)	0.6029	0.1081	−0.7110
Intercept (b)	−0.6728	0.6577	0.0151

The coefficients indicate the relative contribution of each physicochemical variable to the class-specific linear predictor for the respective end-use class. The coefficients should not be interpreted as independent causal effects or absolute measures of variable importance. The class order and coefficients correspond to the linear predictors presented in the text.

개별 시료의 특성에 상대적으로 민감해지는 원인이 될 수 있다. 따라서 저배합 발효군의 낮은 recall은 단순히 용도군 간 이화학적 특성의 중첩만으로 해석하기보다는, 제한된 표본 수와 클래스 불균형에 따른 모델의 학습 불확실성이 함께 영향을 미쳤을 가능성을 고려하여 해석하는 것이 적절하다. 한편 높은 macro-AUC에도 저배합 발효군의 recall이 낮게 나타난 것은 AUC가 예측점수의 전반적인 순위 판별력을 평가하는 반면, recall은 최대 예측확률에 따른 최종 분류 결과를 평가하기 때문이다(Fawcett, 2006).

표준화 회귀계수의 방향은 군집분석에서 관찰된 주요 품질 특성과 대체로 일치하였으나, 입력변수 간 상관관계와 L2 정규화의 영향을 고려할 때 회귀계수의 부호와 크기를 개별 변수의 중요도 또는 인과적 효과로 해석해서는 안 된다(Bishop, 2006). 또한 본 결과는 제한된 수의 시료를 이용한 내부 교차검증에 기반하므로, 제시된 성능지표를 다양한 상업용 강력분에 대한 일반화된 예측성능으로 해석하는 데에는 주의가 필요하다. 본 모형은 실제 용도를 확정적으로 판정하기 위한 모델이라기보다, 이화학적 특성을 기반으로 잠재적인 용도군을 탐색하고 후속 제품시험의 우선순위를 설정하기 위한 탐색적 모델(exploratory model)로 해석하는 것이 적절하다. 향후 각 용도군의 시료 수를 확대하고 다양한 제조 로트 및 표준 제품시험 자료를 포함한 독립 시료를 이용하여 모델의 재현성, 일반화 가능성 및 확률 보정 수준을 검증할 필요가 있다.

추가 시료에 대한 모델 적용

용도를 사전에 지정하지 않은 시판 강력분 4종(Sample A~D)은 모델 학습에 사용하지 않은 미지 시료로 설정하고, 최종 도출된 예측식을 적용하여 모델의 외부 적용 가능성을 탐색하였다(Table 3). 각 시료는 하나의 용도로만 제시하지 않고 저배합 발효, 비발효 및 고배합 발효 클래스에 대한 확률벡터로 나타냈다. 이 확률은 각 시료의 표준화된 이화학적 특성으로부터 계산된 클래스별 선형 예측값을 Softmax 함수로 변환한 것으로, 학습자료의 변수 범위와 클래스 정의 내에서 각 용도군에 대한 상대적 소속 정도를 나타낸다(Bishop, 2006).

Table 3. Predicted end-use class-membership probabilities for four additional commercial strong wheat flours

	Probability for lean fermented dough (P1)	Probability for non-fermented dough (P2)	Probability for high-formula fermented dough (P3)
Sample A	0.5966	0.3789	0.0244
Sample B	0.5423	0.3971	0.0606
Sample C	0.6413	0.2845	0.0742
Sample D	0.3508	0.5776	0.0716

The values represent the predicted probabilities calculated by the multinomial logistic regression model. For each sample, the sum of probabilities (P1+P2+P3) equals 1.0. The class with the highest probability was presented as the model-predicted primary end-use candidate. These probabilities represent relative class membership within the scope of the training data and should not be interpreted as absolute probabilities of product success (n=4).
P1: Probability for lean fermented dough; P2: Probability for non-fermented dough; P3: Probability for high-formula fermented dough.

Sample A는 저배합 발효군의 예측확률이 0.5966으로 가장 높았으며, 비발효군과 고배합 발효군은 각각 0.3789와 0.0244로 나타났다. 따라서 저배합 발효군이 우선 용도로 제시되었으나, 비발효군에도 일정한 확률이 부여되어 두 용도에 관련된 이화학적 특성이 일부 중첩될 가능성을 보였다.

Sample B 역시 저배합 발효군의 확률이 0.5423으로 가장 높았지만, 비발효군의 확률이 0.3971로 나타나 1순위와 2순위 클래스의 차이가 0.1452로 가장 작았다. 이는 네 시료 중 Sample B의 용도 구분에 대한 상대적 불확실성이 가장 크며, 저배합 발효와 비발효 제품시험을 함께 실시할 필요성이 높은 시료임을 보여준다.

Sample C는 저배합 발효군의 예측확률이 0.6413으로 가장 높았고, 비발효군과 고배합 발효군의 확률은 각각 0.2845와 0.0742였다. 1순위와 2순위 클래스의 확률 차이는 0.3568로 네 시료 중 가장 크게 나타나, 다른 시료보다 저배합 발효군에 대한 상대적 소속이 뚜렷하였다.

Sample D는 비발효군의 예측확률이 0.5776으로 가장 높았으며, 저배합 발효군과 고배합 발효군은 각각 0.3508과 0.0716이었다. 이에 따라 Sample D는 비발효 용도를 우선적으로 검토하되, 저배합 발효 용도에 대한 적용 가능성도 함께 시험할 수 있는 시료로 해석하였다.

이러한 결과에서 최대 예측확률은 모형상 우선 검토할 용도군을 나타내고, 1순위와 2순위 확률의 차이는 용도 구분의 상대적 명확성을 판단하는 보조정보로 활용할 수 있다. 즉 Sample C와 같이 확률 차이가 큰 시료는 우세한 용도의 제품시험을 우선 실시하고, Sample B처럼 확률 차이가 작은 시료는 상위 두 용도에 대한 제품시험을 병행하는 방식으로 시험 설계를 차별화할 수 있다. 다만 예측 확률은 PCA score 공간에서 군집 중심까지의 거리를 직접 나타내는 값이 아니므로, 확률 차이가 작다는 사실만으로 해당 시료가 실제 군집 경계에 위치한다고 단정하기보다는 용도 분류의 불확실성이 상대적으로 높은 시료로 해석하는 것이 적절하다.

한편 네 시료 모두 고배합 발효군의 예측확률이 0.0244~0.0742로 낮게 나타났다. 이는 현재 모형에서 추가 시료의 고배합 발효군에 대한 상대적 소속확률이 낮게 산출되었음을 의미한다. 그러나 이 값은 추가 시료의 이화학적 특성뿐 아니라 학습자료의 클래스별 표본 수와 변수 분포에도 영향을 받으므로, 해당 시료의 실제 고배합 발효 제품 적합성이 낮다는 직접적인 근거는 되지 않는다.

확률벡터를 이용한 접근은 단순히 최대확률 클래스를 부여하는

것에서 나아가, 신규 밀가루에서 우선적으로 검토할 용도와 대안 용도를 함께 제시할 수 있다는 장점이 있다. 다만 본 모델은 신규 밀가루의 최종 용도 적합성을 확정하거나 실제 제품시험을 대체하기 위한 예측모형이라기보다, 제한된 이화학적 품질 정보를 바탕으로 잠재적인 용도군을 선별하고 후속 제품시험의 우선순위를 설정하기 위한 탐색적 도구(exploratory tool)로 해석하는 것이 적절하다. 특히 본 연구는 20종의 시판 강력분을 기반으로 구축된 소규모 데이터셋에 기초하므로, 모델이 제시하는 용도 클래스와 예측확률은 절대적인 용도 적합성이나 실제 제품 성능을 의미하지 않는다.

또한 Sample A~D에 대해서는 독립적인 제품시험 결과나 기준 용도 라벨을 확보하지 못하였으므로, 본 분석은 외부검증이 아니라 모델 적용 사례에 해당한다. 향후 표준 식빵, 면류 및 고배합빵 시험이나 모델 개발에 참여하지 않은 전문가의 독립적인 판정 결과와 비교하여 실제 용도 예측성능을 검증할 필요가 있다.

요 약

본 연구에서는 강력분의 용도 분류 가능성을 다변량 품질 정보로 평가하기 위해 시판 강력분 20종의 수분, 회분, 단백질, 흡수율, 최고점도, 입도 및 손상전분을 분석하고 PCA, K-means 군집분석 및 Softmax 기반 다항 로지스틱 회귀를 적용하였다. PCA 결과 PC1과 PC2는 각각 41.2%와 23.5%의 분산을 설명하여 누적 설명분산은 64.7%였다. PC1은 흡수율, 최고점도 및 손상전분, PC2는 수분과 입도의 기여가 크게 나타나 강력분 품질이 전분 기능성, 수분결합 및 입자 특성이 결합된 다차원 구조임을 보여주었다.

K-means 분석에서는 세 개의 이화학적 품질군이 형성되었으며, 사전 정의한 저배합 발효, 비발효 및 고배합 발효 용도군과 95.0%의 일치율을 나타냈다. Softmax 모형의 LOOCV 결과 accuracy는 0.900, macro-F1은 0.849, macro-AUC는 0.997이었다. 학습에 사용하지 않은 4종에 적용한 결과 Sample A-C는 저배합 발효군, Sample D는 비발효군의 확률이 가장 높았으며, 확률벡터를 통해 모형상 우선 용도와 용도 분류의 상대적 불확실성을 함께 확인할 수 있었다.

본 연구는 강력분의 품질군 분류와 확률 기반 용도 예측을 통합한 탐색적 분석체계를 제시하였다. 다만 본 모델은 제한된 수의 시판 강력분을 기반으로 구축되었으므로 실제 용도 적합성을 확정하기 위한 도구가 아니라, 신규 밀가루의 잠재적 용도군을 사전 탐색하고 후속 제품시험의 우선순위를 설정하기 위한 보조적 탐색 도구로 해석하는 것이 적절하다. 향후 다수 제조 로트와 표준 제품시험을 이용한 외부검증을 통해 모델의 일반화 가능성이 확보될 경우, 제품시험 설계 및 제품 포지셔닝을 지원하는 정량적 보조수단으로의 활용 가능성을 평가할 수 있을 것으로 판단된다.

ORCID

Yu-Jeung Lim https://orcid.org/0009-0008-7713-4417
Tae-Hoon Kim https://orcid.org/0009-0001-8791-2154
Hyeong Su Lee https://orcid.org/0000-0003-2670-7584

Conflict of interests

No potential conflict of interest relevant to this article was reported.

Acknowledgements

We would like to express our sincere gratitude to all members of Daehan Flour Mills Co., Ltd. who supported this research, and to Mr. Uemura, a Japanese flour milling R&D expert and advisor, for his valuable academic insights and practical guidance.

Data availability

Upon reasonable request, the datasets of this study can be available from the corresponding author.

Authorship contribution statement

Conceptualization: Lim YJ, Lee HS.
Data curation: Lim YJ.
Formal analysis: Lim YJ.
Methodology: Lim YJ.
Software: Lim YJ.
Validation: Lim YJ.
Investigation: Lim YJ.
Writing - original draft: Lim YJ, Kim TH, Lee HS.
Writing - review & editing: Lim YJ, Kim TH, Lee HS.

Ethics approval

Not applicable.

References

AACC International. 2010. Approved Methods of Analysis. 11th ed. Methods 08-01.01, 22-10.01, 46-30.01, 54-21.02, and 76-33.01. AACC International, St. Paul, MN, USA.

Barrera GN, Pérez GT, Ribotta PD, León AE. 2007. Influence of damaged starch on cookie and bread-making quality. Eur. Food Res. Technol. 225: 1-7.

Bishop CM. 2006. Pattern Recognition and Machine Learning. Springer, New York, NY, USA.

Cauvain SP, Young LS. 2007. Technology of Breadmaking. 2nd ed. Springer, New York, NY, USA.

Delcour JA, Hoseney RC. 2010. Principles of Cereal Science and Technology. 3rd ed. AACC International, St. Paul, MN, USA.

Dobraszczyk BJ, Morgenstern MP. 2003. Rheology and the breadmaking process. J. Cereal Sci. 38: 229-245.

Fawcett T. 2006. An introduction to ROC analysis. Pattern Recognit. Lett. 27: 861-874.

Gneiting T, Raftery AE. 2007. Strictly proper scoring rules, prediction, and estimation. J. Am. Stat. Assoc. 102: 359-378.

Granato D, Putnik P, Bursać Kovačević D, Santos JS, Calado V, Rocha RS, da Cruz AG, Jarvis B, Rodionova OY, Pomerantsev A. 2018. Trends in chemometrics: food authentication, microbiology, and effects of processing. Compr. Rev. Food Sci. Food Saf. 17: 663-677.

Hartigan JA, Wong MA. 1979. A K-means clustering algorithm. Appl. Stat. 28: 100-108.

Hosmer DW, Lemeshow S, Sturdivant RX. 2013. Applied Logistic Regression. 3rd ed. Wiley, Hoboken, NJ, USA.

Hubert L, Arabie P. 1985. Comparing partitions. J. Classif. 2: 193-218.

Jolliffe IT. 2002. Principal Component Analysis. 2nd ed. Springer, New York, NY, USA.

Lindeboom N, Chang PR, Tyler RT. 2004. Analytical, biochemical and physicochemical aspects of starch granule size. Starch. 56: 89-99.

Liu C, Li L, Hong J, Zheng X, Bian K, Sun Y, Zhang J. 2014. Effect of mechanically damaged starch on wheat flour, noodle and steamed bread making quality. Int. J. Food Sci. Technol. 49: 253-260.

Marti A, Ulrici A, Foca G, Quaglia L, Pagani MA. 2015. Characterization of common wheat flours (*Triticum aestivum* L.) through multivariate analysis of conventional rheological parameters and gluten peak test indices. LWT-Food Sci. Technol. 64: 95-103.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12: 2825-2830.

Posner ES, Hibbs AN. 2005. Wheat Flour Milling. AACC International, St. Paul, MN, USA.

Rousseeuw PJ. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20: 53-65.

Shewry PR, Halford NG, Belton PS, Tatham AS. 2002. The structure and properties of gluten: an elastic protein from wheat grain. Philos. Trans. R. Soc. B. 357: 133-142.

Vinh NX, Epps J, Bailey J. 2010. Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. J. Mach. Learn. Res. 11: 2837-2854.

Wieser H. 2007. Chemistry of gluten proteins. Food Microbiol. 24: 115-119.