a. Research Project Title, Principal and Co-Investigators
Title: A Novel Soybean Selection Method for Tofu Production Using Machine Learning

Principle Investigator Minwei Xu, Assistant Professor, North Dakota State University
Co-Investigator Changhui Yan, Professor, North Dakota State University
Co-Investigator: David Bohem, Technical Manager, Northern Crops Institute

b. Research Overview and Objectives
Tofu is favored by people in East Asia and has been widely accepted by people from different areas of the world. Yield, texture, and protein content are major parameters used to evaluate the tofu quality. However, determining the parameters relies on tofu processing, which is time-consuming and labor-intensive. In addition, breeders/agronomists cannot get useful information from tofu parameters. Food scientists have related the tofu quality parameters to protein subunits of soybean seeds, such as 11S/7S ratio, 11SA3, and 11SA4 subunits. However, the profile of protein subunits has not been fully considered. North Dakota State University (NDSU) and Northern Crops Institute (NCI) plan to develop an alternative way to predict tofu quality based on the overall profile of soybean protein subunits. A high-resolution sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) will be employed to separate the protein subunit of 80 soybean seed varieties from East Asia and North America. Machine learning (ML) will be employed to construct a predictive model for the quality of tofu based on the SDS-PAGE images.

Objectives:

a). Classify soybean seeds based on their tofu quality, such as yield, texture, and protein content.
b). Test the protein subunit profile of soybean seeds
c). Build an ML model for predicting tofu quality based on soybean protein subunits
d). Evaluate the quality of soybean seeds from ND with the new ML model

c. Completed Work: Deliverables and/or Milestones.
Objective a). Classify soybean seeds based on their tofu quality, such as yield, texture, and protein content.

c.1 Sources of soybean seeds

Table 1 Origins of Soybean seeds for Comparative Evaluation of seeds from the United States and China

Country	Province/State	No. of soybean varieties
United States	North Dakota	98
United States	Minnesota	19
United States	California	2
China	Qinghai	1
China	Heilongjiang	12
China	Hubei	16
China	Shanxi	1
China	Hebei	1
China	Shandong	10

China	Hainan	1
China	Guizhou	1
China	Anhui	1
China	Beijing	4
China	Guangxi	1
China	Jiangxi	1
China	Zhejiang	5
China	Guangdong	2
China	Sichuan	1
China	Shanghai	1

In the United States, diverse states contribute to the soybean varieties, with significant numbers originating from North Dakota, Minnesota, and California. Specifically, in North Dakota, the soybean seeds were procured from the plant sciences department at NDSU, Fargo, USA. Similarly, in Minnesota, the soybean seeds were acquired from the University of Minnesota, Minnesota, USA. Soybean seeds were collected from 16 different provinces in China, covering latitudes from 20° to 48° and longitudes from 96° to 128°

c.2 Characteristics of Tofu collected from different sources in the United States and China

The study involved an examination of tofu prepared from 178 soybean varieties to assess the tofu quality parameters influenced by the sources. Utilizing hierarchical cluster analysis (HCA), as depicted in **Figure 1**, the tofu quality parameters were categorized into six distinct clusters, the details of which are presented in Table 2.

The results derived from HCA highlighted the division of tofu into six clusters, primarily based on the similarity of their water uptake, tofu yield, moisture, soymilk yield, brix, protein, firmness, gumminess, chewiness, springiness, cohesiveness, and resilience.

Cluster 1: This cluster has the widest ranges in water uptake and tofu yield, and high Brix levels, mainly comprising soybeans from China and some from North Dakota, USA. It shows high variability in firmness and resilience but lower in protein and moisture content.

Cluster 2: Characterized by narrower ranges in water uptake, tofu yield, and soymilk yield, and lower variability compared to Cluster 1. The tofu yield behavior notably differs from Cluster 1, and it mostly includes Chinese soybeans.

Cluster 3: Similar to Cluster 2 in many respects but with the lowest variability in water uptake and high variability in springiness and cohesiveness. It comprises soybeans from various Chinese provinces and some U.S. states.

Cluster 4: Exhibits the widest range in firmness and chewiness, and the highest gumminess, mostly comprising soybeans from the U.S. It differs significantly from other clusters in terms of tofu yield, resilience, and protein content.

Cluster 5: Similar to Cluster 1 in tofu yield and chewiness, but with the lowest variability in water uptake and the highest in gumminess. It also shows significant differences in moisture content.

Cluster 6: Shows the highest variability in water uptake and the lowest in springiness and cohesiveness. It has the lowest moisture content and is solely composed of U.S. soybeans.

Overall, the findings reveal significant differences and similarities among the clusters in various tofu and soymilk properties. This highlights the diverse potential of soybean varieties for tofu production, influenced by their geographical origin.

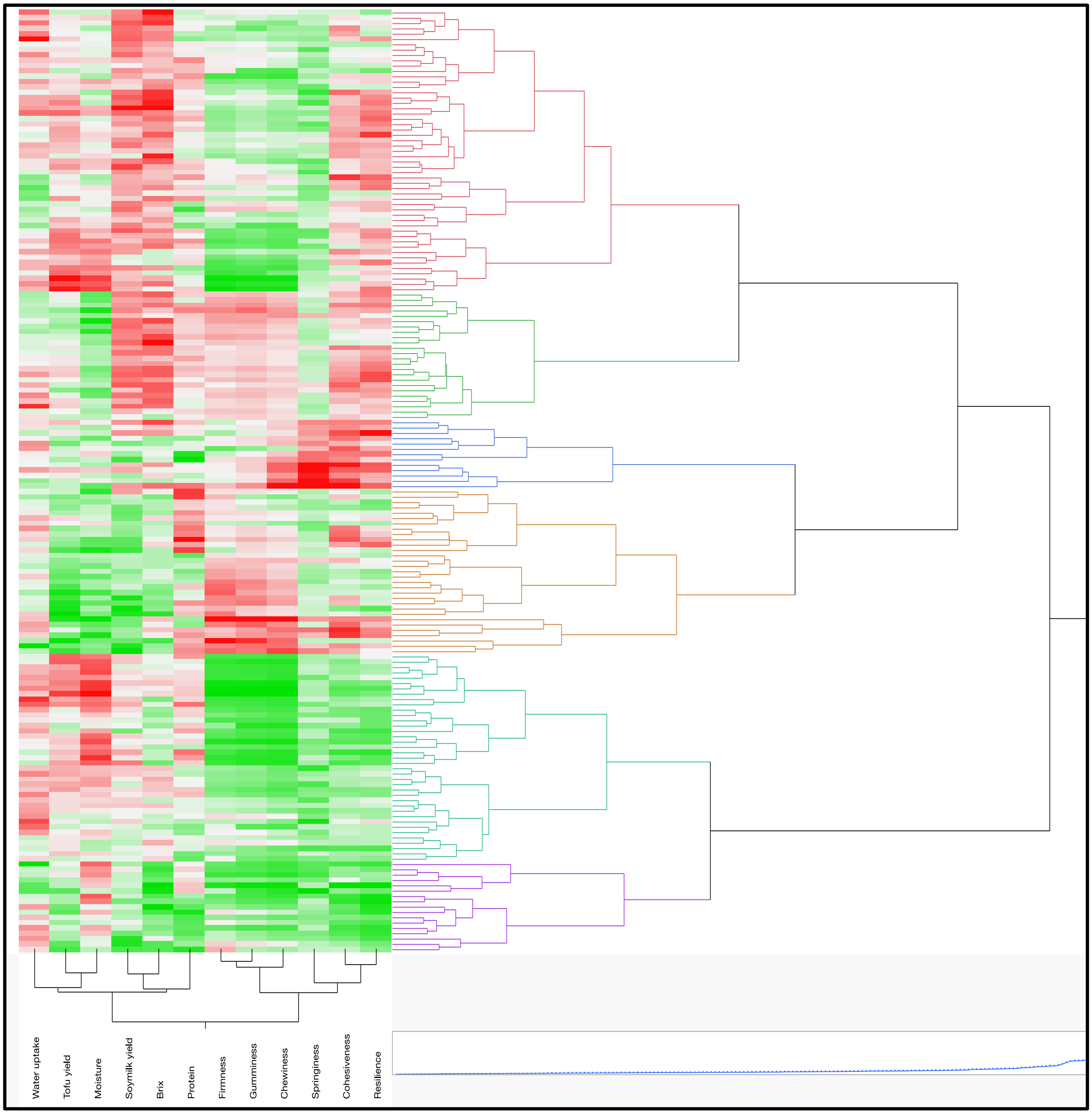

Figure 1 Clustering of tofu quality based on their sources using Hierarchical clustering analysis (HCA) The color of Cluster 1, 2, 3, 4, 5 and 6 are indicated with red, green, blue, orange, dark green and purple respectively

Table 2 Evaluation of Tofu Quality across Various Soybean Classes

	Cluster	Min	Max	Mean	C.V
Water uptake (kg/kg soybean)	1	1275.20	1501.40	1396.75±45.42ab	3.25

	2	1329.00	1486.20	1385.32±38.41cb	2.77
	3	1317.00	1439.60	1391.35±36.63bc	2.63
	4	1193.80	1437.60	1381.45±45.04bc	3.26
	5	1351.00	1483.60	1414.66±31.39a	2.22
	6	1137.00	1438.00	1364.44±78.46c	5.75
Tofu yield (kg/kg soybean)	1	557.00	762.60	631.13±44.59a	7.06
	2	527.40	623.00	580.44±27.77b	4.79
	3	515.20	631.60	576.48±32.38b	5.62
	4	450.20	609.80	519.18±40.39d	7.78
	5	553.40	726.40	624.74±40.82a	6.53
	6	492.00	599.40	545.32±32.33c	5.93
Soymilk yield	1	2970.40	3664.80	3290.45±133.66a	4.06
	2	3043.00	3460.00	3326.09±98.78a	2.97
	3	2405.00	3317.20	2993.52±291.25b	9.73
	4	2165.60	3279.80	2638.54±248.61d	9.42
	5	2593.00	3370.40	3074.66±133.64b	4.35
	6	2280.00	3093.20	2769.01±243.68c	8.80
Brix	1	7.10	8.50	7.79±0.34a	4.38
	2	6.80	8.50	7.89±0.35a	4.39
	3	6.70	8.20	7.40±0.47b	6.31
	4	5.90	7.60	6.95±0.42c	6.00
	5	6.40	7.70	7.23±0.35b	4.79
	6	5.60	7.30	6.28±0.43d	6.91
Firmness (g force)	1	3946.09	21559.55	11970.40±3694.24cd	30.86
	2	15142.30	28594.87	20914.85±3373.28b	16.13
	3	8115.42	21754.72	14611.57±3225.11c	22.07
	4	11849.00	44620.59	23860.06±8277.85a	34.69
	5	3725.04	17735.30	9393.72±3624.82d	38.59
	6	5021.51	23089.98	11680.89±4774.01cd	40.87
Springiness	1	0.86	1.09	0.95±0.04c	4.42
	2	0.95	1.27	1.00±0.07cb	7.16
	3	1.26	2.35	1.68±0.38a	22.66
	4	0.91	1.53	1.04±0.16b	15.27
	5	0.83	1.00	0.94±0.04c	4.23
	6	0.81	0.98	0.93±0.05c	5.04
Cohesiveness	1	0.60	0.81	0.68±0.04b	6.25

	2	0.63	0.78	0.70±0.04b	5.39
	3	0.71	0.85	0.77±0.04a	5.32
	4	0.57	0.82	0.68±0.07b	10.34
	5	0.51	0.65	0.58±0.04c	6.10
	6	0.46	0.66	0.55±0.05d	8.84
Resilience	1	0.16	0.32	0.24±0.03b	13.18
	2	0.21	0.31	0.26±0.03ab	10.04
	3	0.23	0.34	0.27±0.03a	11.03
	4	0.15	0.29	0.21±0.04c	16.82
	5	0.13	0.23	0.18±0.02d	13.28
	6	0.11	0.19	0.15±0.02e	15.72
Chewiness	1	2418.44	13779.49	7366.47±2350.40c	31.91
	2	11619.74	21492.02	14634.54±2644.83b	18.07
	3	8062.45	37698.82	19831.71±8027.94a	40.48
	4	6393.78	45702.32	16600.13±8208.06b	49.45
	5	1937.28	10189.28	4921.54±2049.43c	41.64
	6	2256.64	11154.95	5543.59±2278.63c	41.1
Gumminess	1	2500.78	14136.09	7683.23±2502.80c	32.57
	2	11725.25	21814.25	14574.93±2401.82a	16.48
	3	6358.28	18407.85	11380.47±2825.29b	24.83
	4	7180.66	32983.29	15554.06±5734.04a	36.87
	5	2013.39	10901.09	5264.49±2239.62d	42.54
	6	2329.11	11454.01	5935.39±2296.18cd	38.69
Protein	1	39.18	58.98	52.31±3.95ab	7.55
	2	50.12	59.41	54.18±2.14a	3.94
	3	34.94	60.21	49.53±6.92b	13.97
	4	41.25	67.10	54.69±6.51a	11.90
	5	42.29	60.61	52.18±4.23ab	8.11
	6	36.69	55.24	46.11±6.39c	13.87
Moisture	1	72.61	78.92	74.94±1.52b	2.03
	2	69.64	73.48	72.12±1.08d	1.50
	3	71.35	75.34	73.55±1.25c	1.70
	4	69.36	74.94	72.12±1.28d	1.78
	5	72.88	80.18	76.29±1.94a	2.55
	6	73.16	78.17	75.43±1.58ab	2.09

d. Progress of Work and Results to Date

Objective b). Test the protein subunit profile of soybean seeds

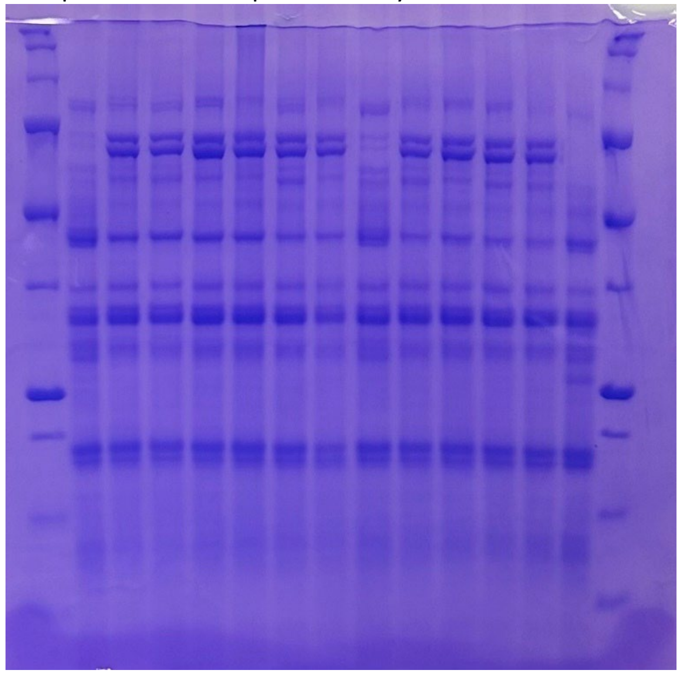

Figure 2 The SDS-PAGE profile of different soybean seeds

The SDS-PAGE is in progress. We have finished 80 varieties of protein subunit of soybean seeds. A big difference of protein subunits between different soybean cultivars have been identified.

Objective c). Build an ML model for predicting tofu quality based on soybean protein subunits

A matlab based algorithm has been developed to read those SDS-PAGE images automatically. It is a initial step to allow the ML model to access those image data.

e. Work to be Completed.

Objective d). Evaluate the quality of soybean seeds from ND with the new ML model

f. Other relevant information: potential barriers to achieving objectives, risk mitigation strategies, or breakthroughs.

There are no commercially available pre-cast gels, so we have to make the gels ourselves. The students are not very skilled, which results in the failure of some gels. This is a key reason why the SDS-PAGE experiment is progressing slower than the tofu process.

With continued practice, the students should become more skilled and eventually complete the task. However, this could result in a delay of a couple of months for this project

g. Summary

In summary, this study has provided a comprehensive understanding of how source influences soybean seed characteristics and, subsequently, tofu quality parameters. These insights have practical implications for the soybean and tofu industries, offering opportunities for product optimization and market differentiation based on the sourcing. Additionally, the study underscores the importance of considering both soybean seed characteristics and tofu quality attributes when aiming to produce tofu that aligns with consumer expectations and preferences