JJBS

Volume 19, Number 2,June 2026
ISSN 1995-6673
Pages 273 – 281
https://doi.org/10.54319/jjbs/190210

Jordan Journal of Biological Sciences

Machine Learning Approach to Identify and Evaluate the High protein-yielding Varieties of *Glycine max*

Tina Negi[1], Manu Pant[1,*], Kumud Pant [1] Arvind Singh Negi[2], Pankaj Nainwal[3], Abhishek Kumar[4]

[1]*Department of Biosciences, Graphic Era (Deemed to be University), Clement Town, Dehradun, Uttarakhand, India;* [2] *School of Agriculture, Graphic Era Hill University, Clement Town, Dehradun, Uttarakhand, India;* [3] *School of Pharmacy, Graphic Era Hill University, Clement Town, Dehradun, Uttarakhand, India*

Received: July 11, 2025; Revised: November 17, 2025; Accepted: November 26, 2025

Abstract

Soybeans are valued worldwide as legumes with exceptional health benefits, high nutritional value, and protein content. Though well-studied in many regions, there is limited information on the nutritional status of indigenous soybeans cultivated in the Uttarakhand region of the western Himalayan range in India, despite their prominence in the traditional food system. The present study aimed to assess protein content in indigenous soybean germplasms and use a machine learning algorithm to classify the nutritionally superior soybean samples. 50 samples were collected from different regions of the hilly areas of the state and evaluated for protein content using the Kjeldahl method, followed by an HPLC analysis to ascertain the amino acid content in germplasms with high protein content. Machine learning method-the Decision Tree algorithm (J48) in WEKA (Waikato Environment for Knowledge Analysis) software was used to analyse the attributes like location, seed traits, temperature, and nutrient content. The ten-fold cross-validation achieved 94 % accuracy in classifying germplasms according to the protein content, as shown in the confusion matrix. The study showed that maximum protein content in the analyzed samples ranges from 36.96% to 43.56%, mostly in samples collected at an elevation of 345 m – 1021 m, and 1699 m to 2044 m, latitudinal range of 28° to 30.9°, and longitudes 77.5° to 79.52°. HPLC analysis confirmed histidine as the most prominent amino acid, with the highest amount being 0.688 g/100g.

Keywords: Bhat, *Glycine max*, Food security, amino acids profiling, Machine learning.

1. Introduction

Soybean (*Glycine max*) is a leguminous crop (40-50% content) with significantly high protein, fat content, and is considered a potential solution to global food security (Keatinge et al., 2011, Tanwar and Goyal, 2021; Florentino et al., 2022). The protein content is attributed to its nitrogen- fixing capacity (Carter et al., 2016; Cafaro et al., 2020) that peaks during flowering and pod filling (Ciampitti et al., 2021; Kubar et al., 2021). Considering the high protein content, soy offers scope for diverse soy-based dietary and other consumable products to meet the nutritional needs and highlighting soybeans as an excellent source to address the issues of food security (Thrane et al., 2017; Qin et al., 2022). Soybeans also provide all nine essential amino acids, including histidine, isoleucine, lysine, methionine, phenylalanine, threonine, and tryptophan, supporting human health and development (Singer & Zhang, 2020; Hu et al., 2023). This further enhances the need for inclusion of soybean in the diet since the essential amino acids need to be taken from the diet; while non-essential amino acids can be produced internally (Thanapornpoonpong et al., 2008; Holeček 2020; Lopez & Mohinuddin, 2024).

Uttarakhand is a hilly state in Northern India where the native soybean is called ‘bhat’ and is a very popular cuisine. However, there is a scarcity of information on the nutritional profiling of this crop. This has led to lack of general awareness on the need for planned cultivation and utilization of soybean in the region. Plant-based proteins are considered a cheaper and more sustainable source of protein with high nutritional and nutraceutical potential. Asians typically prefer plant-based protein in several traditional cuisines due to its long shelf life and high nutritional value (Qin et al.,2022). In recent times, there has been a growing trend towards vegan and vegetarian lifestyles, and increased consumption of soy has been observed due to its high protein concentration (Tamang, 2012). A variety of products, such as sausages, bacon, burger patties, etc., containing soy protein are available in the market and readily consumed by consumers due to their low cost, and health benefits like less risk of cardiovascular ailments (Akintola et al., 2022; Chetachukwu et al., 2022). In this scenario, a focused study on indigenous soybean germplasms and analysis of protein content become important to identify superior germplasms and utilise them as a cost-effective protein source for a healthy diet, addressing issues of food security.

* Corresponding author. e-mail: manupant@geu.ac.in.

The present study was conducted to establish the protein profile of soybean germplasm from the Garhwal region of Uttarakhand state in India and further assess the amino acid content in the samples with high protein content.

2. Materials and Methods

2.1. Protein estimation using Kjeldahl method

In this study, a total of 50 soybean cultivars (approximately 100 gm of each sample) were collected from different Uttarakhand regions and sun-dried for 3-4 days. The sample was further processed to prepare analytical sample (fine dry powder) of particle size 0.8-1mm using electric grinders. Estimation of Crude protein and nitrogen content was performed using the Kjeldahl method. The basic principle involves the digestion of the material in H_2SO_4 and conversion of protein 'N' to ammonium sulfate $(NH_4)_2SO_4$ at an elevated boiling point in the presence of potassium sulphate K_2SO_4 and copper (Cu) catalyst to enhance the reaction rate. Ammonia is released and subsequently steam distilled into boric acid under alkaline conditions, and then quantified by titration with standardized Hydrochloric acid (HCl). The methodology involves two steps: digestion and distillation. Block digester was switched on and heated up to 420°C.2 g 'bhat' sample (10 gm) with sulphuric acid H_2SO_4 and potassium sulphate and copper sulphate as catalysts (7.0 g K_2SO_4 + 0.8 g $CuSO_4$) were added into the digestion tubes. This process of digestion was carried out for 50-60minutes. The samples were cooled down for another 15-20 min, and they were proceeded further for the distillation process. Digested Kjeldahl tubes were connected to the distillation unit along with a 40% NaOH solution in the alkali tank. The unit contains an Erlenmeyer titration flask (capacity: 500 ml) filled with 30ml boric acid (H_3BO_3) solution along with an indicator. As soon as the distillation process begins the procedure of the automatic titration system immediately starts the titration inside the Kjeldahl tube. It is steam distilled until ≥150 ml of distillate is collected in the titration flask. H_3BO_3 acts as an adsorbent receives the distillate from the distillation unit and facilitates titration to estimate the quantity of nitrogen and protein content in the sample using 0.1000M HCl standard solution. The whole procedure was carried out automatically by using a steam distiller with automatic titration. The total Nitrogen content present in the crude sample is thus calculated using equation 1 and further crude protein % is estimated using equation 2 mentioned below.

Equation 1

$$Kjeldahl\ nitrogen\ \% = \frac{\left(Vs - V_B\right) \times M \times 14.01}{W \times 6.38}$$

Equation 2

$$Crude\ \ protein\ \% = \%\ Kjeldahl\ N \times F$$

where V_S = Volume (ml) of standardized acid used to titrate a test
V_B = Volume (ml) of standardized acid used to titrate reagent blank; M = HCl molarity; 14.01 = Molecular weight of N
W = Weight (g) of test sample; 6.38 = Factor to convert mg/g to percent; and F = 6.38 (Factor to convert N to protein)
Goulding et al.,2020; Licon, 2022)

2.2. Amino acid analysis

The top five performing samples based on their protein content were evaluated for their amino acid profiles using the HPLC technique with a UV-Vis fluorescence detector. All the amino acids were prepared with 0.25mM/L concentration (Conc.) standards, followed by crushing of 'bhat' samples into fine powder. The sample was then digested using HCl acid overnight. 100 mg sample was weighed and added with 300µl of 6N HCl in digestion bottles and kept in the oven at 110-120° for 24 hours. After 24 hours, the tubes were taken out of the oven and added with 2ml of 0.1N HCl solution and filtered with a 0.22µm syringe filter. A final volume of 1.0µl was injected into the HPLC, and the program for the amino acid analysis was set and run under the conditions mentioned in Table 1 (Lamp et al., 2018).

Table 1 Different chromatographic parameters involved in separation of amino acids using HPLC

Column	Phenomenex Gemini-NX C18 3µm C183µm,150mmL x3. 0mmI.D/Shim pack scepter C183µM,150MMLX3.0mmI		
Mobile Phase A	25mM DI potassium dihydrogen phosphate pH 7		
Mobile Phase B	Water-acetonitrile methanol (15:45:40v/v/v)		
Flow rate	0.7Ml/min		
Auto sampler temp	15°C		
Column temp	40°C		
Injection volume	1.0µl		
Rinse RO	Water: methanol (20:80v/v)		
Rinse R3	Water: acetonitrile (80:20) v/v)		
Detection conditions	Detector: Shimadzu fluorescence detector RF-20AXS		
Run time	50MIN		
Time program	TIME	A Conc%	B Conc%
	0.00	90.0	10.0
	30.0	55.0	45.0
	35.0	0	100.0
	42.0	0	100.0
	43.0	90.0	10.0
	50.0	90.0	10.0

2.3. Statistical analysis

The different values of the amino acids obtained for the respective samples were further analyzed to ascertain the significant difference among the samples. One-way ANOVA statistical analysis tool was used for the purpose. LSD was performed at a significance level of 0.05. P-values less than 0.05 were considered significant, while

those with p-values greater than 0.05 were considered non-significant. The post-hoc analysis was further conducted using python to determine statistically significant differences in amino acid concentrations among the surveyed villages. Tukey's HSD test was used to compare group means while controlling for family-wise error, making it appropriate for biochemical and nutraceutical datasets involving multiple group comparisons.

2.4. Machine learning analysis of protein content in seed through a decision tree algorithm implemented in WEKA

The data generated by Kjeldahl method was used to train the machine learning models. The dataset was loaded into WEKA's Explorer interface, and it was ensured that the data was formatted properly (Frank et al., 2016). The data was pre-processed by normalizing the data and handling any missing values. Feature selection was done to configure the J48 classifier by selecting it from the "Classify" tab and setting the appropriate parameters, such as the confidence factor and the minimum number of instances per sample. The data was divided into training and test sets after cross-validation. The J48 decision tree algorithm was trained by analysing metrics such as the confusion matrix, accuracy, precision, recall, and F-measure. The J48 decision tree algorithm implemented in WEKA was employed for the classification of seed samples obtained from different sites in Uttarakhand classified into high, medium, and low. The samples with the best protein and nitrogen content as well as attributes like the number of seeds in 10 grams were found to be decisive for the classifying samples into high, medium and low yielding. Other parameters that were included in the classifier were geographical, environmental, and physical factors like latitude, longitude, altitude, seed coat colour, average temperature from, average temperature to, protein percentage, shape, size and class. The Decision Tree algorithm in WEKA represents a very strong method of classification and regression in particular, useful for data analysis like that in our research for protein content estimation in seeds. This is a class of non-parametric supervised learning methods that construct models to predict the value of a target variable using simple decision rules inferred from the features of the data. Algorithms in WEKA include J48, which is a realization of C4.5, RandomTree, and REPTree. The process starts by importing data in ARFF or CSV format into WEKA, followed by preprocessing stages including missing values handling, normalization, and feature selection. The actual decision tree classifier is then selected from the Classify menu, configured for a confidence factor and pruning options, and trained on the dataset. It is evaluated using cross-validation and analyzed with the performance metrics and confusion matrices given by WEKA. Such a generated decision tree can be visualized, thus allowing interpretation of decision rules and feature importance. For seed classification, this model can classify the seed samples under classes such as "good" or "medium" according to various attributes such as percentage of nitrogen, protein percentage, and geographical features. Patterns and insights were taken from the findings indicating considerable attributes that distinguish seed quality.

3. Results and Discussion

3.1. Protein content

The present study aims to evaluate the protein content and amino acid profile in soybean by using the machine learning approach after obtaining data from the Kjeldal method of protein estimation and amino acid quantification using the HPLC technique. Overall Protein content among all the villages were observed to ranges from 36.13% (New Tehri) to 43.56 % (Gwad Khirsu) with nitrogen content ranging from 5.78% to 6.97% for the same villages. (Fig.1). Studies suggests that high nitrogen content increases the protein content in cereals (Kaufman et al.,2013) and similar results were observed in the current study. The highest concentration of protein was found in S15 Gwad Khirsu with 43.56% followed by Maral S27 with 43.18% indicating that soybean from Pauri Garhwal could be the highest potential source of protein. After S5 & S27 the other highest protein content carrying villages with protein percentage more than the checks were S43 Joshimath (42.64%), S5 Pratapnagar (42.63%) and S28 Sripur (42.31%) (Table 2). The top five samples with the highest protein content, along with the nitrogen content, were observed to surpass the checks with high margins and emerge as the promising sources of protein. The villages with the highest percentage of protein were found to have protein content ranging from 42.31% to 43.56 % w/w, which is useful for meeting the food security issues prevailing in hilly areas.

Table 2 The top five villages with the highest protein and nitrogen content along with their 11 validated attributes

S.no	1	2	3	4	5
Village	Gwad Khirsu	Maral	Joshimath	Pratapnagar	Sripur
Latitude	30°1	29°48	30°53	30°3	30°75
Longitude	78°52	78°36	79°51	78°29	78°6
Altitude(m)	1699.56	1650	1874.52	2620	1765.9
Seed coat colour	Yellowish white	Yellowish white	Greenish yellow	Yellowish white	Yellowish white
Average temp from (° c)	15° C	17° C	14° C	15° C	15° C
Average temp to (° c)	25° C	28° C	21° C	18° C	23° C
No of seeds in 10gm	97	60	97	86	80
Nitrogen %	6.97	6.91	6.82	6.82	6.77
Protein %	43.5625	43.1875	42.64	42.625	42.312
Seed Shape size	Round medium	Round large	Round large	Oval large	Oval medium
Class	Good	Good	Good	Good	Good

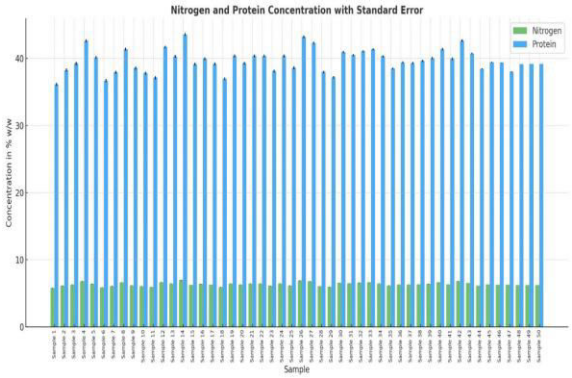

Figure 1. Nitrogen and Protein content in Soybean samples from different villages of Uttarakhand

According to studies, protein percentage in meat is approximately 18-25% (Ahmad et al., 2018; Widodo et al., 2022). On the other hand, pulses and legumes contain 18% to 40% protein, respectively (Erbersdobler et al.,2017). Studies performed by Conte et al. (2014) Gálová et al. (2019) Parveen et al. (2022) Varalaxmi et al. (2024) have reported the protein content ranging from 20-26% in pulses and 18-26 % in other cereals such as maize oats, barley, millets, rice, etc., which is considerably lower than the values observed in the present study. The present study shows a protein content of up to 43.56% in 'bhat', which is substantially higher than the reported content in previous studies. This study backs up Hoffman and Falvo (2004) research, where they found soybeans have a higher protein percentage of~ 40-50% and less carbohydrate than other legumes. Similarly, Dhull et al. (2023 reported a lower percentage of protein (24%) and a higher content of carbohydrates (63%), indicating a higher protein percentage in the present study. The reported amount of protein in the current study falls under the recommended dietary allowance of 0.8-1.2gm/kg weight for elderly people for proper body functioning and avoiding muscle mass loss, as suggested by Traylor et al. (2018)

3.2. Amino acid content

All the top five performing samples with the highest content of protein were further evaluated for the presence of amino acids. The samples were found to have all the essential and non-essential amino acids such as threonine(T) tryptophan (W) isoleucine(I), leucine(L), lysine(K) and histidine (H), arginine (R), aspartic acid(D), glutamine(Q), glycine(G) ornithine, gamma aminobutyric acid (GABA) and cysteine (C) except methionine (Table 3). Reported studies show that non-essential amino acids are synthesized inside the body but their synthesis is limited at the time of illness or stress, therefore, consumed through diet (Shine & Rostom,2021). The study reported the absence of methionine but detected cysteine, another sulfur-containing amino acid. Cysteine plays a significant role in various biological reactions. It helps in the formation of disulfide bonds, playing a key role in protein binding, systematic protein folding, and formation of tertiary and quaternary protein structures (Brosnan & Brosnan, 2006). Cysteine is an essential substrate for the formation of glutathione and taurine, which are biological antioxidants that protect the cells from cell damage and cell toxicity (Sereno et al., 2014). The current study revealed that Pratapnagar was seen to have nine amino acids with histidine as the most present amino acid, with a concentration of 0.516g/100g at retention time (Rt) of 6.216 (Fig 2). Gawad Khirsu was found to have a total of twelve amino acids (Fig 3). Maral had shown the presence of a total of eleven amino acids with histidine being the one with highest concentration at Rt 6.16 (Fig 4). Sripur similar to Gawad khirsu had twelve amino acids (Fig 5), while Joshimath showed the presence of ten amino acids (Fig 6). On comparing all five samples, S15 Gwad Khirsu and S28 Sripur were found to have the maximum count (12) of amino acids. In terms of total amino acid content, S15 Gawad Khirsu (at an elevation of 1699m) scored the highest concentration of 3.183±0.6 g/100g followed by Pratapnagar (at an elevation of 2620m) with 1.057±0.7g/100g content (Table 3). The study showed that samples at the highest altitude/low temperature and lowest altitude/high temperature were observed to have the highest total amino acid content. Almost all the samples were found to be enriched with the amino acid Histidine (H) with the highest content of 0.688g/100g from S15 Gawad Khirsu with a retention time peak of 6.104. Histidine is an essential amino acid that serves many benefits, including antioxidant and anti-inflammatory properties. It is known to reduce the effect of stress and anxiety and improve sleep quality (Thalacker-Mercer, 2020). Studies have suggested its potential role in treating rheumatoid arthritis and chronic renal failure (Holeček, 2020; Thalacker-Mercer, 2020).

There have been some studies (García et al., 1998; Lopez et al., 2024) where the content of different amino acids was evaluated, and it was found that the content of leucine, lysine, valine isoleucine phenylamine was found to be 8 g, 6.5 g,5g,5g,4g/100 g of protein which were relatively higher than the present study. Similarly, another recent study performed by Custodio-Mendoza et al., (2024) showed the amino acid content to be higher than in the present study. A study done by Kudełka et al., (2021) reported amino acids with values lower than those of the present study. Nizkii et al., (2020) performed a similar study and obtained results similar to those obtained in the current study.

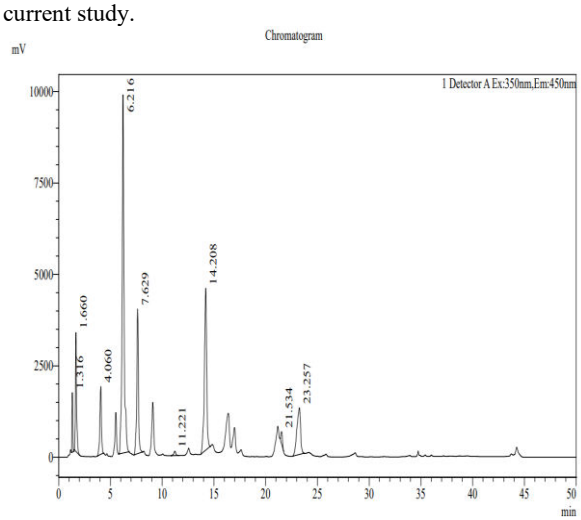

Figure 2 HPLC Chromatogram of S1pratapnagar showing the separation peaks of Aspartic acid (D), Glutamic acid (E), Histidine (H), Threonine (T), Tyrosine (Y), Tryptophan (W), Isoleucine (I), Arginine (R), Aminobutyric acid (GABA **)**

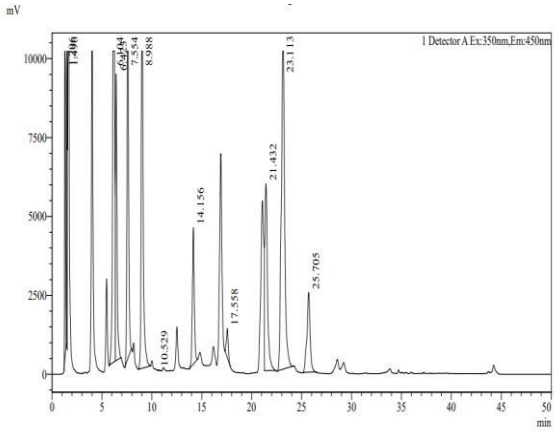

Figure 3 HPLC Chromatogram of S2 Gawad Khirsu showing the separation peaks of Aspartic acid (D), Glutamic acid (E), Histidine (H), Glycine(G), Threonine (T), Arginine (R), Alanine (A), Tyrosine (Y), Cysteine (C), Tryptophan (W), Isoleucine (I), Ornithine (ORN)

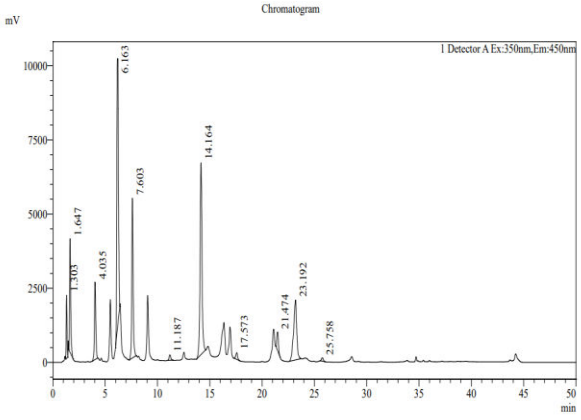

Figure 4 HPLC Chromatogram of S3 Maral showing the separation peaks of Aspartic acid (D), Glutamic acid (E) Histidine (H), Glycine (G), Tyrosine (Y), Cysteine (C), Tryptophan (W), Isoleucine (I), Ornithine (ORN), Arginine (R), Aminobutyric acid (GABA)

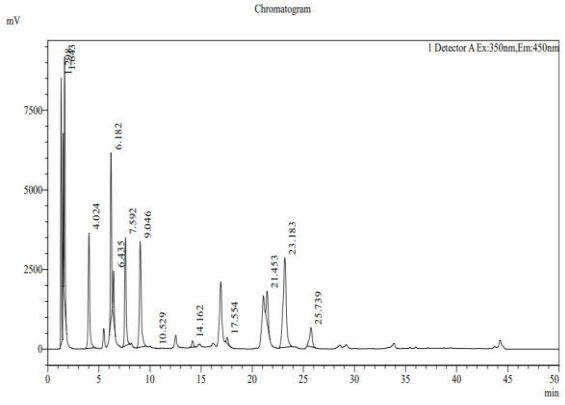

Figure 5 HPLC Chromatogram of S4 Sripur showing the separation peaks of Aspartic acid D, Glutamic acid E, Histidine (H), Glycine(G), Threonine (T), Arginine (R), Alanine (A), Tyrosine (Y) Cysteine (C), Tryptophan (W), Isoleucine (I), ornithine (ORN)

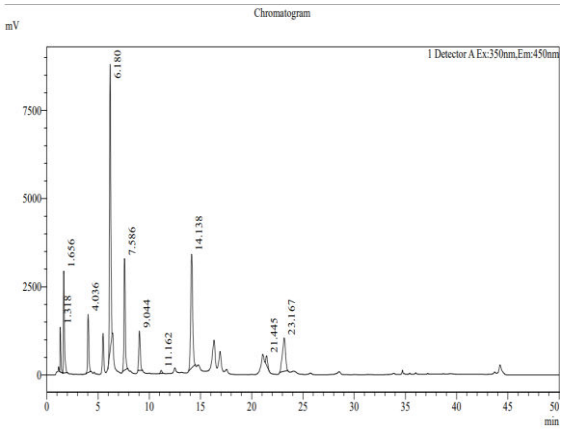

Figure 6 HPLC Chromatogram of S5 Joshimath showing the separation peaks of Aspartic acid (D), Glutamic acid (E), Histidine (H) Threonine (T), Alanine (A), Tyrosine (Y), Tryptophan (W), Isoleucine (I), Arginine (R), Aminobutyric acid (GABA)

Table 3 Amino acid profile of all the top 5 villages with highest concentration of protein

.		VILLAGES				
S.no	Amino acid	Pratapnagar g/100g Mean ± SD	Gawad Khirsu g/100g Mean ± SD	Maral g/100g Mean ± SD	Sripur g/100g Mean ± SD	Joshimath g/100g Mean ± SD
1	Aspartic acid	0.030±0.01	0.241±0.2	0.039±0.1	0.149±0.5	0.024±0.4
2	Glutamic acid	0.119±0.2	0.225±0.3	0.129±0.1	0.217±0.3	0.092±0.2
3	Histidine	0.516±0.1	0.688±0.3	0.398±0.2	0.167±0.1	0.284±0.1
4	Glycine	0.0000	0.071±0.3	0.038±0.2	0.0080.4	0.000
5	Threonine	0.182±0.12	0.516±0.2	0.000	0.138±0.1	0.117±0.02
6	Arginine	0.000	0.248±0.1	0.000	0.068±0.1	0.018±0.03
7	Alanine	0.000	0.000	0.000	0.0000.2	0.000
8	Tyrosine	0.164±0.2	0.139±0.1	0.201±0.2	0.006±0.4	0.103±0.1
9	Cysteine	0.000	0.016±0.3	0.004±0.2	0.003±0.1	0.0000.4
10	Tryptophan	0.011±0.1	0.353±0.2	0.021±0.3	0.037±0.2	0.010±0.2
11	Isoleucine	0.000	0.372±0.19	0.064±0.1	0.094±0.2	0.030±0.2
12	Ornithine	0.000	0.314±0.2	0.008±0.4	0.061±0.7	0.000
13	Asparagine	0.018±0.02	0.000	0.030±0.2	0.000	0.040±0.2
14	GABA	0.018±0.019	0.000	0.020±0.09	0.000	0.029±0.3
	Total	1.057±0.7	3.183±0.6	0.952±0.29	0.949±0.33	0.747±0.26
	Average	0.0755±0.05	0.2274±0.04	0.0680±0.02	0.0678±0.02	0.0533±0.01
	S.D.	YS	YS	YS	YS	YS
	LSD	0.05	0.05	0.05	0.05	0.05

Results of statistical analysis using ANOVA revealed a significant difference in amino acid concentration with P <0.05 value in 'bhat' samples collected from different locations. Post hoc test revealed arginine as the most significant amino acid. Multiple pairwise comparisons showed highly significant differences, particularly between Gawad Khirsu and Joshimath, Maral, Pratapnagar, and Sripur, with Gawad Khirsu consistently having lower Arginine levels ($p < 0.001$). The factors that contribute to the variation in amino acid content include genotype, environmental factors and soil conditions (Goldflus et al., 2006). With the change in temperature and elevation, the change in total amino acid concentration was observed, indicating the direct relation of abiotic factors like temperature and altitude in determining the concentration of amino acids in soybean 'bhat'. Mourtzinis et al. (2017) also reported such variations in protein and amino acid content with temperature. The present study confirms that temperature and altitude could be one of the reasons for variation in amino acid profile in soybean 'bhat', indicating them as essential parameters to determine the amino acid content. Another factor that could contribute to the variations in amino acid profile is its relationship with nitrogen. Reports suggest that high nitrogen levels increase the amino acid content. Applying nitrogen in the form of fertilizer can impact the amino acid composition and protein content in plants (Thanapornpoonpong, 2008); therefore, high content of nitrogen in Gawad Khirsu could be the most acceptable reasoning for high concentration of amino acids (Atanasova, 2008). Apart from these factors, there is a slight possibility of amino acid hydrolysis and modification due to acid (HCL) digestion during the extract preparation, which would have led to loss of amino acids.

3.3. WEKA analysis: Seed quality assessment through machine learning

The decision tree algorithm implemented in Weka software has been used in this study to analyze and understand the influence of various parameters or combinations of various parameters that can be decisive in determining the seed quality or classifying seeds based on different attributes into medium or good quality seed. With the J48 decision tree algorithm in Weka, an open-source data mining and machine learning tool, a classification accuracy of 94% was obtained for the seed data set on performing ten-fold cross-validation (Table 4). The classifier was trained with seed data obtained from 50 different geographical locations in Uttarakhand. The seed samples with protein content <= 39.95 were placed under the class medium (total instances or samples 29.0); the seed samples with protein content > 39.95 were considered as belonging to class good (total instances or number of samples 21.0).

Table 4. Stratified Cross-Validation Metrics for Model Performance Evaluation

Correctly Classified Instances	47 (94%)
Incorrectly Classified Instances	3 (6%)
Kappa statistic	0.8792
Mean absolute error	0.06
Root mean squared error	0.2449
Relative absolute error	12.2822%
Root relative squared error	49.546%
Total Number of Instances	50

The confusion matrix revealed that in 50 samples, all 21 samples designated as good variety seeds were correctly classified or cross-validated as good variety seeds, but 26 samples out of 29 medium variety seeds could only be classified correctly, and 3 were classified incorrectly, resulting into 94% validation accuracy estimated by the matrix (Fig 7). However, 94% validation is considered highly significant and highlights its reliability in class prediction of different quality groups of seeds. This high degree of accuracy has demonstrated the major influence of geographical, environmental and physiological characteristics in determining the quality of seeds. Factors including longitude, altitude, seed colour, seed shape, number of seeds in 10 gm, average temperature have emerged as pivotal factors in determining the seed quality. The results suggest that the use of these attributes can be further considered in similar studies for various other seed-related assessments (Fig 8).

The analysis reveals that 11 attributes (latitude, longitude, altitude, seed coat colour, average temperature, seed shape and size, number of seeds in 10 gm, nitrogen percentage, protein percentage, class of seed) selected for the analysis were effective in differentiating between the good and medium types of 'bhat' seeds, making them suitable for further similar analysis. The analysis also demonstrated the high degree of overlapping among the different samples, indicating the similar values of attributes shared between the different soybean cultivars.

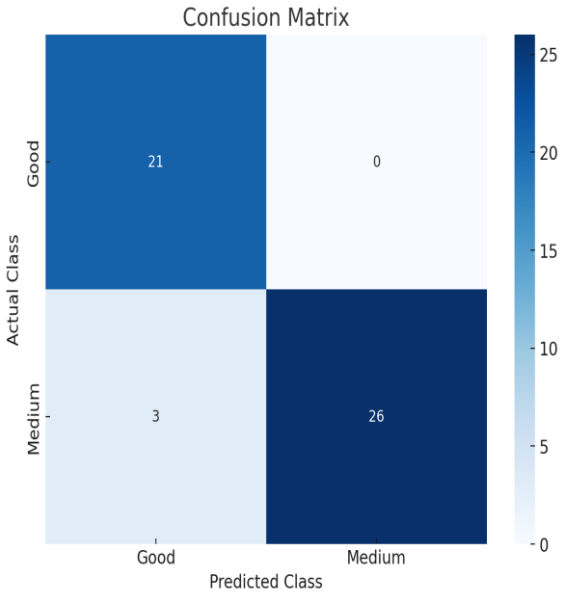

Figure 7 Confusion Matrix Showing Actual versus Predicted Class Distribution of the Model

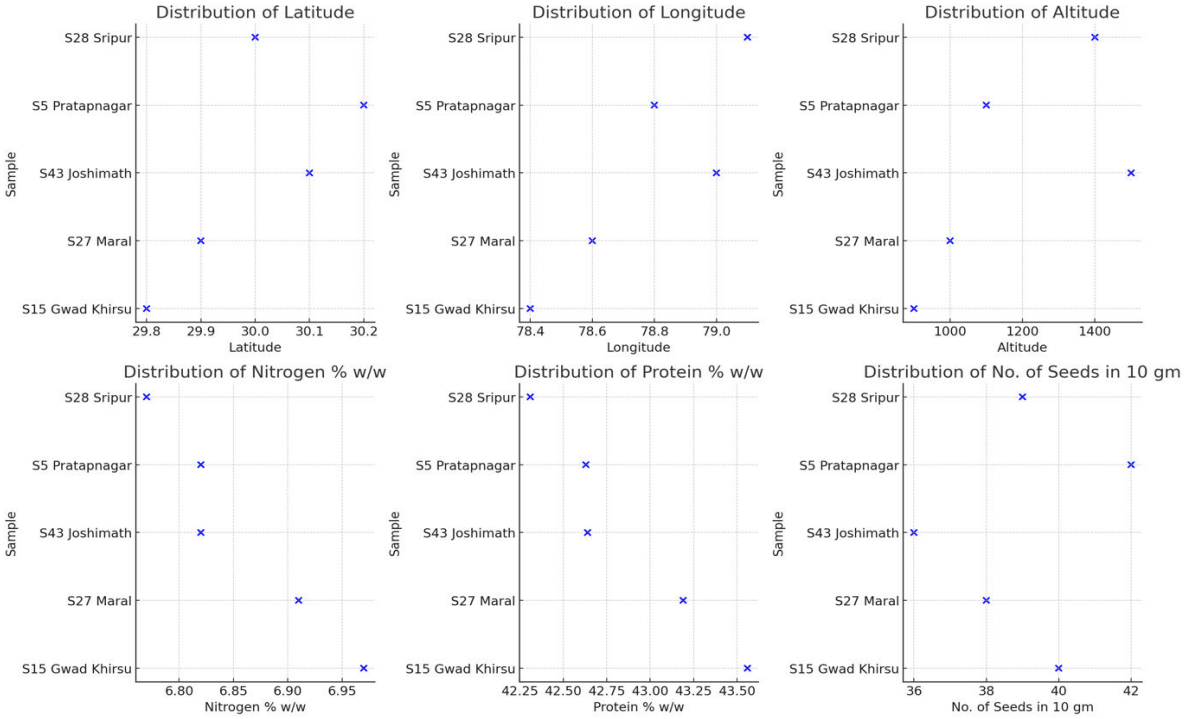

Figure 8. Distribution of bhat samples according to various attributes: latitude, longitude, altitude, nitrogen % w/w, protein % w/w and no of seeds in 10 gm

Results revealed that the maximum number of samples fell within the latitude range of 29.45-30.9° and the longitude range of 78.2°-79.52° in an altitudinal gradient ranging from 345-1021 m and 1699-2044 m. The results show that the protein content ranges from 39% to 40 %. The top five seed samples with the highest protein and nitrogen content were identified as S15 Gwad Khirsu, S27 Maral, S43 Joshimath, S5 Pratapnagar, and S28 Sripur, which ranged between 42.31% - 43.56% w/w (Table 2).

The machine learning approach used here enabled quick and efficient data analysis, pattern identification, and classification of seed samples based on various attributes, such as nitrogen and protein percentages. The decision tree model helps visualize rules and how features can be ranked in importance, which leads to seeds being classified as 'good' or 'medium'. Besides saving time, the auto-analysis provides a robust framework for decision-making by using data to identify the best seeds to be cultivated. Therefore, machine learning is an invaluable tool in agricultural research, enabling the development of sustainable, high-yield farming practices.

4. Conclusion

Machine learning algorithms were used to identify protein content in soybean seeds collected from Garhwal region of Uttarakhand, India. The study confirmed Sample 15 from Gwad Khirsu to have the highest protein content of 43.56% followed by the sample from Maral village, having 43.18% protein. All the essential amino acids were present in the analysed top five protein-containing samples, except for methionine. Histidine was further identified as the most prominent amino acid. The statistical analysis confirmed that the variations in nutrient content in the samples relate to environmental factors, such as temperature and altitude. The WEKA analysis found the best protein content (39-40%) in the samples collected from villages at altitudinal ranges of 345-1021m and 1699-2286 m. This approach can be used to analyse the effects of environmental factors on crop varieties, revealing patterns that are difficult to analyse beyond manual methods. Using advanced machine learning for seed quality and nutrient content offers new insights, promoting reproducibility, productivity, and quality in agriculture, hence ensuring global food security.

Acknowledgement

The authors would like to sincerely thank Graphic Era (Deemed to be University) for providing support and access to the Central Instrumentation Facility (CIF) lab at Graphic Era for conducting the necessary laboratory work. The authors sincerely acknowledge the financial grant received from the Uttarakhand State Council of Science and Technology (UCOST), Government of Uttarakhand.

Funding

The authors acknowledge the funding received by the Uttarakhand State Council of Science and Technology (UCOST) (Project No UCS&T/R&D-25/20-21/19327/1) for this study.

Conflicts of Interest

There are no competing interests that need to be disclosed by the authors.

Author's Contribution

TN: Methodology, data curation, validation, writing-original draft, MP: conceptualisation, experimentation, resource management, writing-review and editing, KP: investigation and formal analysis, ASN: resource management, project administration, investigation, PN: project administration, data curation, AK: Writing-review and editing, visualization. All authors have read and approved the final manuscript.

Ethical issues

There are no ethical issues that need to be disclosed by the authors.

References

Ahmad RS, Imran A, Hussain MB. 2018. Nutritional composition of meat. In ***Meat Sci. Nutr.***, edited by Arshad, M.S. **61(10.5772):** pp. 61-75.

Akintola CP, Finnegan D, Hunt N, Lalor R O,Neill S, Loscher C.2022. Nutrition Nutraceuticals: A Proactive Approach for Healthcare. In **Advances in Nutraceuticals and Functional Foods**, edited by Gopi, S., Balakrishna, P., Apple Academic Press, pp. 123-172

Atanasova E. 2008. Effect of nitrogen sources on the nitrogenous forms and accumulation of amino acid in head cabbage. *Plant Soil and Environ.*, **54(2):** 66.

Brosnan JT, Brosnan, M E.2006.The sulfur-containing amino acids: an overview. *J.Nutr.*, 136(6), 1636S-1640S.

Cafaro La Menza N, Monzon J P, Lindquist J L, Arkebauer, T J, Knops J M, Unkovich M,Grassini, P .2020. Insufficient nitrogen supply from symbiotic fixation reduces seasonal crop growth and nitrogen mobilization to seed in highly productive soybean crops. *Plant, Cell Environ*, **43(8):** 1958-1972.

Carter A M, Tegeder M.2016.Increasing nitrogen fixation and seed development in soybean requires complex adjustments of nodule nitrogen metabolism and partitioning processes. *Curr. Biol.*, 26(15), 2044-2051.

Chetachukwu S A, Tahergorabi R., Hosseini SV.2022. Proteins, peptides, and amino acids. **In Nutraceutical and Functional Food Components**, 2nd edition, edited by Galanakis, C.M., Academic Press, pp. 19-48. https://doi.org/10.1016/B978-0-323-85052-0.00014-3

Ciampitti I A, de Borja Reis, A F, Córdova S C, Castellano M J, Archontoulis SV, Correndo A A, & Moro Rosso L H.2021. Revisiting biological nitrogen fixation dynamics in soybeans.*Front. Plant Sci.*, **12:**727021.

Conte P, Paciulli M, Mefleh M, Boukid F.2024.Corn and barley protein concentrates: effects of fractionation and micronization on the chemical, functional, and thermal properties. *Eur. Food Res. Technol.*,1-11. https://doi.org/10.1007/s00217-024-04544-6

Custodio-Mendoza J A, Pokorski P, Aktaş H, Kurek M A.2024. Rapid and efficient high-performance liquid chromatography-ultraviolet determination of total amino acids in protein isolates by ultrasound-assisted acid hydrolysis. *Ultrason. Sonochem.*, **111:** 107082.

Dhull, S. B., Kinabo, J., & Uebersax, M. A. (2023). Nutrient profile and effect of processing methods on the composition and functional properties of lentils (Lens culinaris Medik): A review. *Legume Sci.*, ***5*(1):** e156.

Erbersdobler H F, Barth C A, Jahreis G.2017. Legumes in human nutrition. Nutrient content and protein quality of pulses. *Ernahrungs Umsch.*, **64(9):** 134-139. DOI: 10.4455/eu.2017.034

Florentino L H, Lima R N, Molinari M D.2022. Soybean functional proteins and the synthetic biology. **In Soybean-Recent Advances in Research and Applications,** edited by Ohyama, T., Takahashi, Y., Ohtake, N., Sato, T., & Tanabata, S., IntechOpen, https://doi.org/10.5772/intechopen.98162

Frank E, Hall M A, Witten I H.2016. The WEKA workbench. Online appendix for "Data mining: Practical machine learning tools and techniques".https://hdl.handle.net/10289/17096

Gálová Z, Palenčárová E, Špaleková A, Rajnincova D, Drábeková J.2019.Protein profiles of buckwheat, rye and oat during in vitro gastro-duodenal digestion. *J. Food Nutr. Res.*, **58(3):** 266-274

García M C, Marina M, Laborda F, Torre M.1998.Chemical characterization of commercial soybean products. *Food Chem.*, **62:** 325–331.

Goldflus F, Ceccantini M, Santos W.2006. Amino acid content of soybean samples collected in different Brazilian states: Harvest 2003/2004. *Braz. J. Poult. Sci.*, **8:** 105-111.

Goulding D A, Fox P F O, Mahony J A.2020. Milk proteins: An overview.In **Milk proteins**, 3rd edition, edited by Boland, M., Singh, H. , Academic Press, pp. 21-98.

Hoffman J R, Falvo M J.2004. Protein–which is best? *J. Sports Sci. Med.*, **3(3):** 118.

Holeček M.2020.Why are branched-chain amino acids increased in starvation and diabetes?. *Nutrients*, **12(10):** 3087.

Hu S, Liu C, & Liu X.2023.The beneficial effects of soybean proteins and peptides on chronic diseases. *Nutrients*, **15(8):** 1811.

Kaufman, R. C., Wilson, J. D., Bean, S. R., Presley, D. R., Blanco-Canqui, H., & Mikha, M. (2013). Effect of nitrogen fertilization and cover cropping systems on sorghum grain characteristics.*J.Agric. Food Chem.*, **61(24):** 5715-5719.

Keatinge J D H, Easdown W J, Yang R Y, Chadha M L, Shanmugasundaram S .2011. Overcoming chronic malnutrition in a future warming world: the key importance of mungbean and vegetable soybean. *Euphytica*, **180:** 129-141. https://doi.org/10.1007/s10681-011-0401-6

Kubar M S, Shar A H, Kubar K A, Rind N A, Ullah H, Kalhoro S A, Ansari M J .2021. Optimizing nitrogen supply promotes biomass, physiological characteristics and yield components of soybean (Glycine max L. Merr.). *Saudi J. Biol. Sci*, **28(11):** 6209-6217.

Kudełka W, Kowalska M, & Popis M.2021. Quality of soybean products in terms of essential amino acids composition. *Molecules*, **26(16):** 5071.

Lamp A, Kaltschmitt M,Lüdtke O .2018. Improved HPLC-method for estimation and correction of amino acid losses during hydrolysis of unknown samples.*Anal. Biochem.*, **543**:140-145.

Licon C.2022. Proximate and other chemical analyses.In **Encyclopedia of dairy sciences**, 3rd edition, edited by Paul, L.H. McSweeney, McNamara, J.P., Academic Press, pp. 521-529. https://doi.org/10.1016/B978-0-12-818766-1.00344-5.

Lopez M J, Mohiuddin S S.2024**. Biochemistry, essential amino acids.** In StatPearls Publishing, Treasure Island (FL),

Mourtzinis, S, Borg B S, Naeve S.L, Osthus J, Conley S P.2018. Characterizing soybean meal value variation across the United States: a swine case study.*Agron. J.*, **110(6):** 2343-2349.

Nizkii S, Kodirova G, Kubankova G.2020. Determining the amino acid composition of soybean proteins using IR scanners. *Int. J. Pharm. Res. Allied Sci.*, **9(2-2020):** 45-49.

Parveen S, Jamil A, Pasha I, Ahmad F.2022**.** Pulses: a potential source of valuable protein for human diet. In **Legume Research,** vol 2, edited by Jimenez-Lopez, C., Clemente, A.J., IntechOpen. 99980, https://doi.org/10.5772/intechopen.104821

Qin P, Wang T, Luo, Y.2022. A review on plant-based proteins from soybean: Health benefits and soy product development. *J. Agric. Food Res.*, **7,** 100265.

Sereno M, Gutiérrez-Gutiérrez G, Gómez-Raposo C, López-Gómez M, Merino-Salvador M, Tébar F. Z, Casado E .2014. Oxaliplatin induced-neuropathy in digestive tumors. *Crit. Rev. Oncol. Hematol.,* **89(1):**166-178.

Shine B, Rostom H.2021. Basic metabolism: proteins. *Surgery* (Oxford), **39(1):** 1-6.

Tamang J P.2012. Plant-based fermented foods and beverages of Asia. **Handbook of plant-based fermented food and beverage technology**, CRC Press, pp. 49-90.

Tanwar B, Goyal A.2021.**Oilseeds: health attributes and food applications,** Springer, Singapore,p 20220004233

Thalacker-Mercer A E, Gheller M E.2020. Benefits and adverse effects of histidine supplementation.*J. Nutr.*, **150:** 2588S-2592S.

Thanapornpoonpong S N, Vearasilp S, Pawelzik E, Gorinstein S .2008. Influence of various nitrogen applications on protein and amino acid profiles of amaranth and quinoa. *J. Agric. Food Chem.*, **56(23):** 11464-11470.

Thrane M, Paulsen, P V, Orcutt M W, Krieger T M.2017. Soy protein: Impacts, production, and applications. **In Sustainable protein sources**, edited by Nadathur, S.R., Wanasundara, J.P.D., Scanlin,L. Academic Press, pp. 23-45. https://doi.org/10.1016/B978-0-12-802778-3.00002-0

Traylor D A, Gorissen S H, & Phillips S M. 2018. Perspective: protein requirements and optimal intakes in aging: are we ready to recommend more than the recommended daily allowance? *Adv. Nutr*, **9(3):** 171-182.

Varalakshmi S, Singh N K, Pareek N, Senthilkumar V.2024. Assessing potential of teosinte in diversification of maize germplasm for kernel protein.*Genet. Resour. Crop Evol*, 1-14. https://doi.org/10.1007/s10722-024-02025-z

Widodo, W., Winaya, A., Zalizar, L., Anggraini, A. D., Zahoor, M., & Mel, M. (2022). Protein level efficacy in improving meat nutritional contents in cross-bred local chickens aged 0 month to 2 month. *J. Biol. Sci.*, **15(5):** 893-896.