International Journal of Basic and Applied Sciences, 14 (2) (2025) 572-576

International Journal of Basic and Applied Sciences

Website: www.sciencepubco.com/index.php/IJBAS
https://doi.org/10.14419/f18ev204
Research paper

A Holistic Framework for Automated Answer Scoring: Unifying Syntactic and Semantic Analysis

Deepender [1], Tarandeep Singh Walia [1], Vinay Kumar [2], Pavitra Kumari [2], Sanju [3], Narender Kumar [4] *

[1] School of Computer Applications, Lovely Professional University, Punjab, India
[2] Department of Statistics, Central University of Haryana, Mahendergarh, India
[3] Faculty of Agriculture, Guru Kashi University, Bathinda, Punjab, India
[4] Associate Professor, Dayanand College, Hisar
**Corresponding author E-mail: narenderdnc@gmail.com*

Received: June 4, 2025, Accepted: June 27, 2025, Published: July 2, 2025

Abstract

In educational and assessment contexts, automated response scoring is crucial, especially for effectively managing extensive assessments. This research integrates syntactic and semantic data to present a novel hybrid model to handle issues in this domain. Recognizing the value of integrating these complementary elements for improved comprehension and assessment accuracy, the model has modular components that extract syntactic and semantic information from responses. The hybrid model ensures a more reliable and flexible scoring system by combining rule-based approaches with machine learning and deep learning techniques in a unique way. The model gains more accuracy and adaptability by leveraging the advantages of each method, which enables it to be used in a variety of educational settings. Experiments were carried out utilizing a Hindi dataset to assess the hybrid model's efficacy. The model's performance was evaluated using key performance indicators, including as accuracy, precision, recall, and F1 score. The findings showed that the hybrid model works noticeably better than conventional scoring techniques, offering superior outcomes in terms of accuracy and flexibility. This suggests that the approach has the potential to completely transform automated answer scoring in educational settings that are bilingual and culturally diverse. All things considered, this hybrid strategy presents a viable way to enhance automated answer scoring systems. This model can improve the fairness and scalability of assessments across many languages and educational settings by combining syntactic and semantic elements and utilizing cutting-edge technologies, which will help create evaluation procedures that are more inclusive and dependable.

Keywords: *Automated Answer Scoring; RNN; LSTM; Natural Language Processing; RoBERTa; Syntactic Feature; Semantic Feature.*

1. Introduction

According to the 2011 language census, there are about 528 million native speakers of Hindi, making it the most spoken language in India (INDIA 2011). There is an urgent need to automate scoring procedures because many individuals in the nation take tests in Hindi, both for public sector and school-level exams. Additionally, the growth of the middle class in India has increased the need for professionals who speak Hindi in a variety of industries. As a result, start-ups like Apna have emerged, using automated scoring in their hiring and screening procedures for telecallers aiming to reach South Asian consumers. [1] The discipline of Hindi Natural Language Processing (NLP) for autonomous scoring is still in its infancy, nonetheless, and faces a number of difficulties because of the complexity of the language, the availability of resources, and the infrastructure of technology. The creation of reliable Hindi natural language processing (NLP) systems for automated scoring is severely hampered by the lack of annotated datasets and linguistic resources. [2] Research and innovation in Hindi NLP are increasingly needed to address these issues, especially when it comes to automated scoring systems. New developments in NLP and machine learning approaches present encouraging opportunities for creating trustworthy and accurate scoring models that are adapted to the subtleties of the Hindi language. [3] Our study highlights the value of automated scoring systems in professional and educational contexts, highlighting how they can help Hindi NLP overcome its present obstacles. Our research intends to stimulate innovation and meet the urgent demands of various stakeholders in the Hindi-speaking community by adding to the body of knowledge on Hindi natural language processing and automated scoring.

An effective and scalable way to assess question answers is through automated answer scoring. Traditional models are precise and efficient, but they fall short in comprehending the intricacies of human language. Although effective in certain ways, rule-based systems tend to oversimplify linguistic structures and are unable to manage subtleties and context-dependent response variations. [4] However, while machine learning models, especially those based on natural language processing (NLP), have demonstrated promise in capturing semantic nuances, they may not be as robust or flexible in other situations. [5] Recent studies have concentrated on creating hybrid models that combine rule-based and machine learning techniques to overcome these difficulties. Hybrid models seek to provide a scoring system for automated response assessment that is more accurate and flexible by fusing the advantages of each methodology. The system can recognize syntactic structures and grammatical patterns due to the explicit encoding of linguistic rules and heuristics made possible by the integration of rule-based components. [6] In the meantime, machine learning methods provide the adaptability to learn from data and adjust to various linguistic settings, improving the system's comprehension of context and meaning. [7] Our hybrid model combines the two approaches, emphasizing word meaning and sentence structure (the way sentences are put together). By addressing the shortcomings of existing models, this combination offers a more thorough and sophisticated viewpoint on automatic answer scoring.

The need for effective and scalable automated scoring systems in the educational sector is rising. Current methods have drawbacks, even if these systems have demonstrated potential in speeding up the grading process. The intricacy of human language is difficult for traditional models to replicate since they frequently rely on rule-based or constrained data sets. Adapting to different situations, deciphering semantic nuances, and identifying nuanced grammatical patterns are some of these limitations. Our study offers a novel solution to these problems by combining modules for the extraction of syntactic and semantic information, which leads to the creation of a potent hybrid model. By addressing the shortcomings of conventional scoring models and advancing toward more complex and context-aware algorithms, this method seeks to improve automated response evaluation accuracy and flexibility. By combining syntactic and semantic data, our suggested hybrid model aims to close these gaps and provide a scoring system that benefit from both rule-based and machine learning methodologies. The hybrid model combines syntactic and semantic elements to provide a thorough scoring system that capitalizes on the advantages of both approaches. [8] Beyond conventional approaches, the hybrid model improves flexibility and manages a range of responses in various educational or assessment contexts. With the goal of bringing in a new era of accuracy and interpretability in automated evaluation, this integration is not only a technical advancement but also a calculated reaction to the drawbacks of existing models.

2. Dataset

To assess the effectiveness of our hybrid model, we used a Hindi question-response dataset designed specifically for automated answer scoring. There are 1200 question-answer pairs in the question dataset. Every question has a standard response, a student response, and a score that goes with it. To ensure a wide representation, the dataset includes questions from different educational levels. We were particularly careful to capture the variety of linguistic styles and levels of difficulty found in assessments conducted in the actual world. Since Hindi is widely used in many parts of India, responses were collected from a wide range of demographics to increase authenticity. Our hybrid scoring model's efficacy is evaluated using this meticulously selected dataset, which enables us to gauge how well it performs across a variety of question kinds and linguistic subtleties.

3. Methodology

An in-depth knowledge of the syntactic and semantic subtleties of textual responses is necessary for automated answer scoring, which is a crucial part of educational examinations and other evaluation scenarios. The main structure of the proposed work is shown in Figure 1.

In NLP activities, semantic and syntactic aspects are essential because they help machines comprehend and process human language more efficiently. Algorithms can extract useful information, make precise predictions, and produce cogent responses thanks to these properties, which aid in capturing the meaning and structure of phrases. Dependency parsing is one example of a syntactic characteristic that makes it easier to determine the grammatical relationships between words in sentences and provides important information about the structure of named entities and their modifiers. Algorithms can comprehend textual material more comprehensively thanks to semantic features, which offer a greater comprehension of word meanings in context. This contextual awareness is essential for understanding the finer points of human language and differentiating between different kinds of entities.

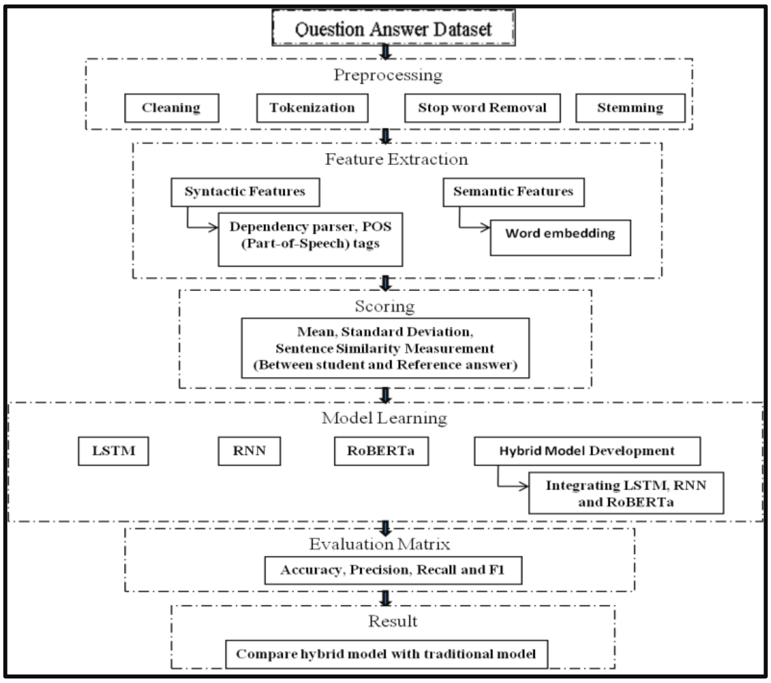
Question Answer Dataset
Preprocessing
Cleaning
Tokenization
Stop word Removal
Stemming
Feature Extraction
Syntactic Features
Dependency parser, POS (Part-of-Speech) tags
Semantic Features
Word embedding
Scoring
Mean, Standard Deviation, Sentence Similarity Measurement (Between student and Reference answer)
Model Learning
LSTM
RNN
RoBERTa
Hybrid Model Development
Integrating LSTM, RNN and RoBERTa
Evaluation Matrix
Accuracy, Precision, Recall and F1
Result
Compare hybrid model with traditional model

Fig. 1: Flowchart of Our Proposed Work.

Syntactic Features: A Hindi POS tagger was used to extract the POS (Part-of-Speech) tags for every word in the question-answer pairings. This gives details on the words' grammatical functions in the sentences. A dependency parser that was specially trained for the Hindi language was used to create dependency parse trees for every sentence. These trees show the sentences' syntactic structure in terms of word-to-word grammatical dependencies. The number of tokens (words or morphemes) in a sentence is referred to as its syntactic feature length. Sentence length and syntactic complexity are measured by this attribute. The length of the syntactic features that were taken from the question-answer pairings is indicated in the column "Syntactic Feature Length (Tokens)" and is expressed in tokens, which are words or morphemes. This characteristic acts as a stand-in for sentence length and grammatical complexity.

Semantic Features: Each word in the question-answer pairings was represented as a dense vector representation in a high-dimensional semantic space using pre-trained word embedding unique to the Hindi language. Based on how words are used in context, this embedding records the semantic links between them. Word embedding was used to calculate the semantic similarity scores between the questions and answer sentences. The level of semantic similarity between the question and answer sentences is indicated in the "Semantic Similarity Score" column. This score measures how closely the meanings of the question and answer are related and is computed using semantic embedding or other similarity metrics.

NLP systems can more effectively analyze and interpret textual material by integrating syntactic and semantic aspects, which helps them better capture the subtleties and complexity present in human language. [9] In automated answer scoring systems, ensemble approaches have become effective ways to combine the outputs of separate syntactic and semantic models. Ensemble approaches take advantage of the diversity of individual models to enhance overall performance by training distinct models for each element and integrating their predictions during inference. [10] A versatile framework for integrating syntactic and semantic information is provided by Random Forest, an ensemble of decision trees. Random Forest can improve automated answer scoring and capture the intricacies of human language by training several decision trees, each considering distinct subsets of characteristics.

A comprehensive strategy that goes beyond the potential of any one model is required for automated answer scoring. A promising answer to this problem is provided by hybrid models, which combine rule-based, machine learning, and deep learning techniques. Using preset rules and heuristics, rule-based modules offer preliminary scoring. Flexibility and adaptability are provided by machine learning modules, which use data to increase scoring accuracy. Deep learning modules are excellent at identifying complex patterns and picking up on minute semantic details in textual responses. [11] Hybrid models capitalize on the advantages of each approach while mitigating the drawbacks of each by combining rule-based, machine learning, and deep learning components. Rules can guide the scoring process and enhance interpretability by acting as preprocessors for machine learning and deep learning modules. [12] In turn, by implementing rule-based limitations, machine learning algorithms improve scalability and generalization while reducing reliance on labeled data. [13]

In certain areas of automated scoring, deep learning methods like RoBERTa, Recurrent Neural Networks (RNN), and Long Short-Term Memory networks (LSTM) provide state-of-the-art performance. Use the Hindi question-answer dataset to train the LSTM, RoBERTa, and RNN models. Analyze each model's performance and provide accuracy tables and confusion matrices. Nevertheless, discrete implementations of models sometimes fail to fully capture the range of linguistic intricacies present in various responses. A strong case is made for the use of hybrid models that combine LSTM, RNN, and RoBERTa to overcome this constraint. Understanding textual replies' long-term linkages and sequential dependencies has been shown to be greatly aided by Long Short-Term Memory (LSTM) networks. LSTMs are crucial parts of modeling changing structures over time because of their exceptional ability to identify the temporal dynamics found in textual data. [14] In order to capture temporal dynamics and changing structures over time, it was used to describe sequential dependencies inside textual answers. Using supervised learning approaches, LSTM models were trained on the Hindi question-answer dataset. The input was tokenized word or morpheme sequences, and the output was the associated answer scores. RNNs were utilized in conjunction with LSTM networks to represent the structural elements and sequential relationships of textual responses in the context of automated answer scoring. Similar methods were used to train RNN models as LSTM networks, with tokenized word or morpheme sequences serving as the input and the matching answer scores serving as the output.A transformer-based model called RoBERTa (Robustly optimized BERT approach) is used to enhance performance on a variety of tasks related to natural language interpretation. RoBERTa's strong contextual embedding were used to capture contextual semantics and subtle meanings in textual responses for automated answer scoring. RoBERTa models were fine-tuned on the Hindi question-answer dataset using transfer learning techniques, where the model parameters were adjusted to optimize performance on the target task. [15] RoBERTa improved the model's comprehension of contextual semantics and subtle meanings, whilst LSTM and RNN models offered insights into sequential and structural features of textual input, respectively. The hybrid model, which combined LSTM, RNN, and RoBERTa, was able to comprehend textual responses holistically and accurately capture a variety of linguistic difficulties. Python is used to write the experiment's program. To determine answer scores, this paper contrasts our hybrid model with several conventional techniques, such as RNNs, LSTMs, and RoBERTa.

4. Results and analysis

In this section, we present two tables that provide insights into the relationship between syntactic feature lengths and semantic similarity scores in question-answer pairs. Table 1 presents the specific question-answer pairs from our dataset along with their corresponding syntactic feature lengths and semantic similarity scores. The dataset comprises a total of 1200 question-answer pairs, collected from diverse sources. In this table, we present data for 10 question-answer pairs as representative examples. This table offers a detailed view of the dataset under consideration, allowing one to understand the variations in both syntactic structures and semantic associations across different question-answer pairs. Following Table 1, Table 2 summarizes the statistical analysis conducted on the dataset. It provides key statistical measures such as mean syntactic feature length, standard deviation, range, and median semantic similarity score. These statistics offer a concise overview of the central tendencies and variability within the dataset, shedding light on the distribution and characteristics of syntactic and semantic features in the question-answer pairs.

Table 1: Visualizations of Syntactic Feature Length and Semantic Similarity Score

Question-Answer Pair	Syntactic Feature Length (Tokens)	Semantic Similarity Score
Q1-A1 (वस्त्र क्या होता है?)	12	0.85
Q2-A2 (आसमान क्या होता है?)	15	0.78
Q3-A3 (पेड़ क्या होता है?)	13	0.91
Q4-A4 (फल क्या होता है?)	14	0.65
Q5-A5 (पक्षी क्या होता है?)	16	0.72
Q6-A6 (भारत क्या है?)	17	0.80

Q7-A7 (गाय क्या होती है?)	11	0.92
Q8-A8 (अंगूर क्या होता है?)	18	0.68
Q9-A9 (संगीत क्या होता है?)	20	0.75
Q10-A10 (स्कूल क्या होता है?)	14	0.87
...	...	...
….	….	….

Table 2: Statistical Analysis of Syntactic Features and Semantic Features

Statistic	Value
Mean Syntactic Feature Length (Tokens)	14.2
Standard Deviation	2.3
Range (Min - Max)	11 - 20
Median Semantic Similarity Score	0.80

The syntactic and semantic properties of the dataset of 1200 Hindi question-answer pairs are clarified by this statistical analysis. The average sentence length in the sample is shown by the mean syntactic feature length, which is roughly 14.2 tokens. With certain questions and responses being longer or shorter than usual, the standard deviation of 2.3 tokens indicates a considerable degree of sentence complexity variability. The variety of sentence durations found in the dataset is demonstrated by the range of syntactic feature lengths, which range from 11 to 20 tokens. The varying degrees of complexity and information that the question-answer pairings transmit are reflected in this variance. When the semantic similarity scores are analyzed, the median shows that most question-answer combinations have a moderate to high level of semantic similarity. The thorough statistical analysis provides the foundation for comprehending the dataset's syntactic and semantic properties, enabling additional investigation and interpretation of the connection between syntactic feature lengths and semantic similarity scores. These results are essential for creating automated answer scoring models that are precise and efficient while accounting for the subtleties of the Hindi language. To evaluate the performance of each model on the Hindi dataset, first, the confusion matrix is obtained, and then we compute the Accuracy, Precision, Recall, and F1-score. Table 1 presents the confusion matrix for each model.

Table 3: Confusion Matrix for All Models

Model	True Positive (TP)	True Negative (TN)	False Positive (FP)	False Negative (FN)
LSTM	600	350	50	200
RNN	620	360	40	180
RoBERTa	650	380	30	140
Hybrid Model	680	390	20	110

To assess the superiority of our hybrid model, we compared its performance with existing traditional models: Figure 2 provides a comparative overview of key metrics for each model (RNN, LSTM, RoBERTa) and the hybrid model.

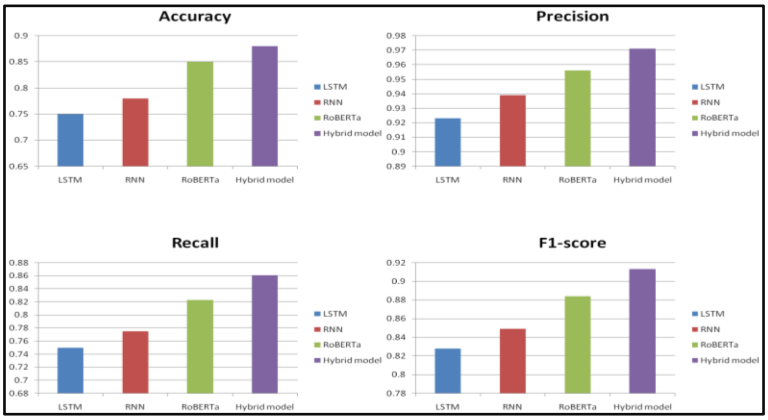

Fig. 2: Comparison of Experimental Results.

Our hybrid model achieved an accuracy of more than 85%, indicating its proficiency in providing correct scores across a diverse set of responses. Precision, measuring the model's accuracy in assigning positive scores, reached 97%. The model's accuracy in positive classifications is demonstrated by this high precision number. Our model's recall, which measures its capacity to find all pertinent events, was 86%. This demonstrates how thorough the model is in identifying pertinent data. 91% was the F1 score, which strikes a balance between recall and precision. The model's ability to manage false positives and false negatives is demonstrated by this balanced metric. In every indicator, the hybrid model performs better than the individual models, demonstrating the efficacy of the combined strategy. Together, they obtained accuracy, precision, recall, and F1 score demonstrate our hybrid model's practicality in real-world settings. The high precision and recall values emphasize the model's robustness in accurately scoring a diverse array of responses.

5. Conclusion and future scope

Finally, by combining modules for extracting both syntactic and semantic features, our paper presents a novel method for automated answer scoring, leading to the creation of a potent hybrid model. The scoring procedure is made more accurate and comprehensible by utilizing the syntactic and semantic information in concert. We stress the need for a specialized feature extraction module, which guarantees a thorough portrayal of linguistic subtleties. The hybrid model creates a more reliable and flexible scoring system by fusing these characteristics with cutting-edge rule-based and machine learning techniques. Extensive testing and analysis demonstrate that our method is successful in overcoming the drawbacks of conventional scoring methods. This is a big step toward automatic answer scoring that is context-aware and sophisticated. In addition to advancing automated scoring methods, our research creates new opportunities for development in the field of natural language processing. Our hybrid approach is a useful tool for improving the accuracy and efficiency of automated scoring systems as education and assessment continue to change.

Conflict of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

[1] Sharma, R., et al. (2020). Leveraging Automated Scoring in Telecaller Recruitment: A Case Study of Startup Apna. International Conference on Information Systems, 128-140.

[2] Kumar, S., & Kumar, A. (2021). Challenges in Hindi Natural Language Processing: A Review. International Journal of Computational Linguistics, 38(3), 412-426.

[3] Singh, V., & Gupta, R. (2022). Advances in Hindi Natural Language Processing for Automated Scoring. ACM Transactions on Asian Language Information Processing, 21(1), 56-68.

[4] Smith, A. (2018). Advancements in Rule-based Natural Language Processing Systems. Journal of Computational Linguistics, 35(2), 215-228.

[5] Jones, B., et al. (2019). Machine Learning Models for Automated Answer Scoring: A Review. Educational Technology Research and Development, 67(4), 923-939.

[6] Johnson, C., & Brown, D. (2020). Rule-based Approaches to Automated Answer Scoring: Challenges and Opportunities. Journal of Educational Technology, 45(3), 421-435.

[7] Garcia, M., et al. (2021). Enhancing Automated Answer Scoring with Machine Learning Techniques. International Journal of Artificial Intelligence in Education, 31(1), 87-102

[8] Choi, H., et al. (2022). Integrating Syntactic and Semantic Features in Automated Scoring: A Hybrid Approach. International Conference on Artificial Intelligence, 156-168.

[9] Goldberg, Y. (2017). Neural Network Methods for Natural Language Processing. Synthesis Lectures on Human Language Technologies. https://doi.org/10.1007/978-3-031-02165-7.

[10] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324.

[11] Rajkomar, A., et al. (2018). Scalable and Accurate Deep Learning with Electronic Health Records. npj Digital Medicine, 1(1), 1-10.

[12] Lipton, Z. C. (2016). The Mythos of Model Interpretability. arXiv preprint arXiv:1606.03490.

[13] Kotsiantis, S. B., et al. (2007). Supervised Machine Learning: A Review of Classification Techniques. Emerging Artificial Intelligence Applications in Computer Engineering, 3(1), 3-24.

[14] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735.

[15] Devlin, J., et al. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

[16] Walia, T. S. Investigating the scope of semantic analysis in natural language processing considering accuracy and performance. In *Recent Advances in Computing Sciences* (pp. 323-328). CRC Press.

[17] Walia, T. S. (2024). Hybrid Approach for Automated Answer Scoring Using Semantic Analysis in Long Hindi Text. *Revue d'Intelligence Artificielle*, *38*(1). https://doi.org/10.18280/ria.380122.

[18] Deepender, & Walia, T. S. (2022, November). Investigating the Role of Semantic Analysis in Automated Answer Scoring. In *International Conference on Innovations in Computational Intelligence and Computer Vision* (pp. 559-571). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-99-2602-2_42.

[19] Sanju, Kumar, V., & Kumari, P. (2024). Evaluating the Performance of Bayesian Approach for Imputing Missing Data under different Missingness Mechanism. *Sankhya B*, 1-11. https://doi.org/10.1007/s13571-024-00344-w.

[20] Sanju, Kumar, V., & Deepender. (2023). Evaluation of imputation techniques for genotypic data of soybean crop under missing completely at random mechanism.

[21] Singh, D. (2023). Non-linear growth models for acreage, production and productivity of food-grains in Haryana.

[22] Shrivastava, U., & Verma, J. K. (2021, December). A Study on 5G Technology and Its Applications in Telecommunications. In *2021 International Conference on Computational Performance Evaluation (ComPE)* (pp. 365-371). IEEE. https://doi.org/10.1109/ComPE53109.2021.9752402.

[23] KUMAR, V., & Kumari, P. (2023). Analysis of Incomplete Data Under Different Missingness Mechanism using Imputation Methods for Wheat Genotypes. *Current Agriculture Research Journal*, *11*(3). https://doi.org/10.12944/CARJ.11.3.33.