Advance Sustainable Science, Engineering and Technology (ASSET)
Vol. 7, No.4, October 2025, pp. 02504036-01 ~ 02504036-020
ISSN: 2715-4211 DOI: https://doi.org/10.26877/asset.v7i4.2586

Hybrid Expert System for Academic Stress Diagnosis Using Forward Chaining and Score Weighting

Indra Gunawan[1], Adhika Pramita Widyassari[1*], Ismail Yusuf Panessai[2], Jonathan Rante Carreon[3]

[1]Informatics Department, STT Ronggolawe, Jl. Kampus Ronggolawe No.1 Mentul, Indah, Cepu, Blora, Central Java 58315, Indonesia

[2]Faculty of Artificial Intelligence and Cyber Security, Universiti Teknikal Malaysia, Durian Tunggal, Melaka, Malaysia

[3]Faculty of Medical Technology, Huachiew Chalermprakiet University, Bang Chalong, Bang Phli District, Samut Prakan 10540, Thailand

*dikasari9@gmail.com

Abstract. Academic stress classification is a significant challenge in education, as previous approaches often rely on opaque models or require large training datasets. This study develops a hybrid expert system for academic stress classification using forward chaining and Certainty Factor (CF) score fallback. The system was tested on 100 student cases with the following label distributions: Mild (48), Moderate (37), and High (13), classified independently by three experts. Label validity was tested using pairwise Cohen's kappa, yielding a mean value of 0.8280. The system achieved 100% accuracy, a 32% improvement over the classical forward chaining baseline (68%). Statistical evaluation using Wilson score intervals demonstrated high consistency across all key metrics (accuracy, precision, recall, F1-score) with a 95% CI of [96.4%, 100%]. The system is designed with an explicit and auditable rule structure, enabling deterministic classification based on symptoms. Although validation results are high, the unbalanced label distribution opens up the potential for spectrum bias. Going forward, the system is planned to be tested across institutions, assessed for integration with counseling services, and compared with other hybrid approaches.

Keywords: academic stress, diagnosis, expert system, forward chaining, score weighting

(Received 2025-08-19, Revised 2025-09-16, Accepted 2025-10-30, Available Online by 2025-10-31)

1. Introduction

Student mental health is a critical issue in higher education, both globally and in Indonesia. Students are undergoing a complex life transition, where academic demands, social pressures, and uncertainty about

the future often lead to significant stress [1]. Unfortunately, many students are unaware of the severity of their stress and thus do not seek help until their condition worsens. Furthermore, the persistent stigma surrounding mental health counseling services discourages students from consulting professionals [2]. Early intervention is crucial to prevent more serious psychological impacts.

Unaddressed academic stress has been shown to directly impact student performance and academic success [3]. High levels of stress can reduce learning motivation, impair concentration, and make it difficult to complete academic assignments, including theses or final projects [4]. In many cases, increasing psychological stress without adequate support leads to burnout, delayed graduation, and even the decision to drop out of college [5]. National data highlights that psychological and academic factors are the dominant causes of high student dropout rates in Indonesia, with hundreds of thousands of cases annually [6]. This underscores the urgent need for an early detection system that can help students recognize their stress levels before they develop into more serious problems [7].

To address these challenges, expert systems have emerged as a promising technological solution [8], [9]. These systems enable an automated, knowledge-based diagnostic process [10], supporting the early detection and systematic reporting of psychological conditions [11]. In recent years, various studies have developed forward chaining-based expert systems to aid diagnostic and decision-making processes in various domains. For example, an expert system for detecting potassium deficiency in cocoa plants achieved 88% accuracy in leaf image classification based on RGB color channels [12], with outputs in the form of specific fertilizer dosage recommendations (Urea, SP-36, MOP) for plant recovery. In the health sector, forward chaining was used in the diagnosis of leptospirosis with 91.3% accuracy and an average inference time of 0.8 seconds [13], producing outputs in the form of severity classifications and medical follow-up recommendations. A heart disease diagnosis system combining forward chaining with fuzzy logic achieved 94.6% accuracy [14], with outputs in the form of risk classifications (low, high, risky) and therapy recommendations. In the FITTER Forum, expert systems play a role in providing guidance on injection site rotation, lipodystrophy detection, and appropriate injection device selection, with a structured, best-practice, education-based approach. [15]. Forward chaining is also considered superior in terms of privacy and auditability for medical expert systems, with a data breach risk of <1% compared to black-box models [16], making it a safe and transparent option for clinical applications. In education, the combination of forward and backward chaining has improved students' metacognitive scores by 18% through adaptive intelligent tutoring [17], with personalized learning strategies based on student responses. Meanwhile, in the financial domain, a forward chaining-based expert system can predict stock market trends with 89.2% accuracy and an RMSE of 0.034 [18], producing buy/sell signals and investment risk analysis.

While the forward chaining approach has proven effective in various contexts, these systems typically focus on domains with relatively stable and deterministic symptom structures. In contrast, expert systems for diagnosing student stress face more complex and dynamic challenges. Stress symptoms are often multidimensional and overlapping, influenced by psychosocial factors that are not always explicit [19]. Therefore, developing an expert system for this domain requires an inference approach that is not only rule-based but also accounts for symptom weighting, ambiguity, and the possibility of non-deterministic combinations. The system developed in this study integrates forward chaining with a score-based or certainty factor approach to handle the complexity of stress symptoms more flexibly and adaptively.

Several expert system studies have incorporated scoring, weighting, or certainty factors. For instance, [20] developed a rule-based system with CF to differentiate Bipolar Disorder and Major Depressive Disorder, using 17 clinical symptoms with an accuracy of 93.33%. [21] combined CF with spatial weighting to evaluate geological disaster vulnerability, resulting in a more accurate and interpretable evaluation. In another medical domain, [22] designed a web-based expert system to diagnose abdominal colic in infants using a combination of FC, CF, and interpolation, achieving 96% accuracy, higher than the Dempster-Shafer method. Meanwhile, [23] used Case-Based Reasoning (CBR) for lung cancer diagnosis, achieving 94.47%–100% accuracy on two different datasets. These studies demonstrate that incorporating weighting and CF can enhance the accuracy and flexibility of expert systems in handling

variations in psychological symptoms, strengthening their role as initial screening tools.

However, while these studies show the effectiveness of scoring and weighting, they still have limitations in addressing the complexity of non-deterministic and overlapping symptoms common in student academic stress. The systems developed by [20], [21], and [22] rely on direct matching, and [23] uses a similarity-based approach dependent on historical data. None of these studies explicitly combines a rule-based inference mechanism with a score-based fallback mechanism to dynamically handle input mismatches.

This is the novelty of our study: by integrating a rule-based forward chaining engine with a confidence-based score weighting mechanism, our expert system can provide a flexible diagnosis of stress levels even when user input does not explicitly meet the rules' premises. This hybrid approach not only improves the accuracy and transparency of the inference process but also allows the system to remain operational and reliable under ambiguous or partial input conditions.

Based on this background, the purpose of this study is to design and evaluate an expert system based on forward chaining and adaptive belief-based score weighting to diagnose student stress levels. This system was developed to fill the gap in previous research by offering a hybrid approach that is more responsive to the multidimensional, overlapping, and often non-explicit nature of stress symptoms.

2. Methods

This expert system is designed using a hybrid approach that combines a rule-based forward chaining inference engine with a certainty factor (CF)-based score-weighting fallback mechanism. The primary goal of this approach is to improve diagnostic accuracy and flexibility, particularly when dealing with non-explicit, overlapping, or partial symptom input.

2.1. Symptoms and Weight

a. Symptom Collection

A total of 60 symptoms of academic stress were collected through a literature review and interviews with psychologists (experts), field observations, and open-ended interviews with students. The list of symptoms covered four main domains: physiological (e.g., muscle pain, nausea, sleep disturbances), emotional (e.g., anxiety, frustration, loss of motivation), cognitive (e.g., difficulty thinking, forgetfulness, time disorientation), and behavioral (e.g., procrastination, avoidance, pacing) [24].

b. Number and Qualifications of Experts

The assessment was conducted by three experts:

- Experts 1 and 2: Clinical psychologists with >5 years of experience in student counseling.
- Expert 3: Student affairs expert with >8 years of experience in academic guidance and observing student behavior.

c. Determination of Trust Score (Certainty Factor)

Each expert assigned a certainty factor (CF) value to each symptom [25], which is the level of confidence that the symptom indicates mild (T1), moderate (T2), or severe (T3) stress. The CF value assessment in this study refers to an expert-based heuristic approach as described by [26]. To maintain uniformity of perception among experts, the CF value guideline shown in Table 1 was used. This guideline explains the level of confidence regarding the relevance of the symptom to academic stress and provides a clinical description for each value range.

Table 1. Certainty Factor (CF) Value Guide Table

Nilai CF	Expert Confidence Level	Clinical Description of Symptoms
0.9	Very confident	Symptoms are very distinctive, dominant, and almost always present in cases of severe stress
0.8	Confident	Symptoms are strong and frequent, highly relevant to academic stress

Nilai CF	Expert Confidence Level	Clinical Description of Symptoms
0.7	Moderately confident	Symptoms are common and relevant, but not exclusive to academic stress
0.6	Uncertain positive	Symptoms may be relevant, but may occur in other conditions or are inconsistent
0.4	Weak/atypical	Symptoms are nonspecific, occur infrequently, or do not sufficiently support a strong inference

Note: CF values <0.4 were not used as they were deemed insufficient to support expert system inference. This guideline was communicated to all experts prior to the assessment process to ensure consistency of perception and system validity.

CF values were determined through: independent assessment by each expert, aggregation of values (average), and consensus discussion for symptoms with significant differences. Initial CF values from each expert were compiled and analyzed using an average aggregation approach. For symptoms with significant differences, consensus discussion was conducted to harmonize perceptions. The final CF values were used as the basis for inference in the expert system. Table 2 below shows examples of ten of the 60 representative symptoms, along with the CF values from each expert, the aggregation results, and the final consensus scores. This table explicitly demonstrates the variation in assessments and the perception alignment process.

Table 2. Ten Representative Symptoms of the Results of Aggregation and Consensus of Certainty Factor (CF) Values

Code	Symptoms	CF Expert 1	CF Expert 2	CF Expert 3	Average CF	Consensus CF
J1	Sudden and intense feelings of anxiety	0.8	0.9	0.8	0.83	0.8
J5	Loss of motivation to study	0.9	0.9	0.8	0.87	0.9
J13	Frequent muscle aches/headaches	0.6	0.5	0.7	0.60	0.6
J18	Feeling lonely even when surrounded by people	0.9	0.8	0.9	0.87	0.9
J25	Reluctant to talk about your thesis	0.4	0.5	0.4	0.43	0.4
J34	Cold sweats while working on your thesis	0.9	0.9	0.9	0.90	0.9
J43	Decreased quality of work done	0.8	0.9	0.8	0.83	0.8
J52	Suicidal thoughts	0.4	0.5	0.4	0.43	0.4
J55	Feeling tense	0.9	0.8	0.9	0.87	0.9
J60	Constant procrastination	0.6	0.7	0.6	0.63	0.6

Note: Consensus CF values were determined through averaging and discussion for symptoms with significant differences.

d. Inter-Expert Agreement

To measure the consistency of classification between experts, a Cohen's kappa (κ) analysis was conducted on 60 academic stress symptoms. The results showed a κ value of 0.839, indicating a very high level of agreement and reliable classification validity. The calculation was based on the distribution of T1/T2/T3 classifications from each expert, with an actual agreement proportion of 90%.

The Cohen's kappa (κ) calculation in this study refers to the approach described by [27], which emphasizes that κ is a measure of inter-rater reliability for categorical data. The κ value is calculated based on the proportion of actual agreement and random expectations, as formulated by Cohen (1960) and developed in recent studies. Interpretation of the κ value follows international standards, where a value >0.75 indicates a very high level of agreement.

To provide a concrete illustration of the assessment process and classification results between experts, Table 3 displays the ten most representative symptoms of academic stress. These symptoms were selected based on their domain variation (physiological, emotional, cognitive, and behavioral), severity, and distribution of certainty factor (CF) values.

Table 3. Representative: CF Assessment and Inter-Expert Classification

Code	Symptoms	Expert 1	Expert 2	Expert 3	Consensus	Score CF
J1	Sudden and intense feelings of anxiety	T1	T1	T1	T1	0.8
J5	Loss of motivation to study	T3	T3	T3	T3	0.9
J13	Frequent muscle aches/headaches	T1	T3	T1	T1/T3	0.6
J18	Feeling lonely even when surrounded by people	T3	T3	T3	T3	0.9
J25	Reluctant to talk about your thesis	T2	T2	T2	T2	0.4
J34	Cold sweats while working on your thesis	T3	T3	T3	T3	0.9
J43	Decreased quality of work done	T2	T2	T2	T2	0.9
J52	Suicidal thoughts	T3	T3	T3	T3	0.4
J55	Feeling tense	T1	T1	T1	T1	0.9
J60	Constant procrastination	T1	T1	T1	T1	0.6

e. Distribution of Consensus Classifications Among Experts

To provide a quantitative overview of the expert system's scope, a summary of the final classifications of 60 academic stress symptoms was conducted. This classification reflects the consensus among experts following the assessment and discussion process. The following table 4 shows the number of symptoms categorized as mild (T1), moderate (T2), severe (T3), and mixed classifications due to differences in perception among experts. This classification distribution indicates that the expert system has balanced coverage across varying levels of academic stress intensity. The proportion of symptoms with mixed classifications also demonstrates the system's flexibility in handling ambiguity and differences in perception between experts.

Table 4. Distribution of Consensus Classification Among Experts

Consensus Classification	Number of Symptoms	Percentage (%)
T1	11	18.3%
T2	17	28.3%
T3	23	38.3%
T1/T2	2	3.3%
T1/T3	1	1.7%
T2/T3	3	5.0%
T1/T2/T3	3	5.0%
Total	**60**	**100%**

2.1. Knowledge Base

The knowledge base consists of 35 active rules, compiled based on the classification results and distribution of CF values from three experts (two psychologists and one student affairs expert), as well as references from academic psychology literature. Each rule consists of a validated combination of symptoms associated with stress levels (T1, T2, T3). The rules are structured in a forward chaining format, but the system does not rely entirely on explicit matching. The following table shows 10 representative rules out of the 35 active rules used in the expert system, including the resulting combinations of symptoms and stress levels. A complete list is available on (https://github.com/adhika-pramita/rule-stress-level/blob/main/rules-stress-level.xlsx).

Table 5. Ten representative rules in expert systems

No Rule	Rule (If And Then)	Stress Level
1	J1, J2, J3	Mild
5	J6, J9, J10, J15, J16, J21, J23	Moderate
10	J10, J24, J36, J40, J43	High
11	J16, J45, J4, J7, J9, J14, J18, J20, J22, J37, J40, J56	High
21	J9, J16, J26	Mild
22	J2, J36, J44	Mild
23	J4, J6, J7, J15, J18, J23, J41, J53	Moderate
31	J1, J2, J3, J6, J11, J13, J16	High
32	J4, J17, J22, J36	Mild
35	J13, J21, J1, J2, J16, J23, J29	Moderate

Conflict resolution occurs when more than one rule is met, with the system selecting a diagnosis based on the priority of intensity or number of matching symptoms. In a hybrid approach, conflicts are resolved through a CF scoring mechanism. Knowledge base maintenance is performed periodically through rule audits, structural revisions, and the addition of new rules based on actual case data and expert input.

2.2. Forward Chaining Inference Mechanism

This expert system uses a forward chaining approach as its primary inference pathway, a reasoning process that moves from user-provided facts (in the form of symptoms) to a conclusion (stress level). Rule premises are combinations of symptoms confirmed by educational psychology experts and serve as the knowledge base, while their consequences are classifications of stress levels: mild, moderate, or severe. These rules form the primary basis of the system's reasoning process.

When a user selects a set of symptoms, the system searches for rules whose premises match the combination of symptoms. If a matching rule is found, the system immediately assigns a stress level based on the consequences of that rule. If no explicit match is found, the system resorts to a fallback mechanism based on certainty factor (CF) score weighting, which is explained in the next subsection.

2.3. Score-Weighted Fallback Mechanism

If no identically matching rule is found in the forward chaining process, the system will resort to a score-weighted fallback mechanism using Certainty Factor (CF) values. This mechanism is designed to handle cases where the symptom input does not fully satisfy the rule's premises but remains relevant to the expert-defined stress classification. Each symptom in the knowledge base has a CF value between 0.4 and 0.9, determined by consensus of three experts. This value reflects the symptom's contribution to each stress level: mild, moderate, or severe. Scoring Steps:

1. Selected Symptom Identification

   The system collects all symptoms selected by the user:

$$[G = \{g_1, g_2, ..., g_n\}] \quad (1)$$

2. Summing CF Scores per Domain
   For each stress level (D \in {T1, T2, T3}), the system calculates a total CF score as follows:
   $$[CF_D = \sum_{g_i \in D} CF(g_i)] \quad (2)$$
   Where:
   (CF_D) is the total score for level (D)
   (g_i) is the i-th symptom belonging to level (D)

3. Diagnosis Determination
   The system assigns a stress level based on the level with the highest total CF score:
   $$[\text{Stress Level} = \arg\max_{D \in \{T1, T2, T3\}} CF_D] \quad (3)$$
4. Resolution: No Diagnosis
   if there are no relevant symptoms If no symptoms are detected or inputted or selected, the system outputs: [\text{Stress Level} = \text{No diagnosis}]
5. Ambiguity Resolution
   If two levels have the same CF score, the system selects the lower stress level first.
   This approach follows the principle of clinical caution, where initial intervention should begin with the mildest level. If the condition does not improve, intervention can be gradually increased.
   The priority order used is:
   $$[\text{Mild} > \text{Moderate} > \text{High}] \quad (4)$$

Case Study Resolution
A user selects the following six symptoms (Table 6):

Table 6. Case Study

Symptoms	Stress Level	Score CF
J2	Mild	0.7
J9	Moderate	0.8
J16	Moderate	0.6
J22	Mild	0.5
J28	High	0.9
J40	High	0.8

Stress Score Calculation:
• Mild (T1):
[CF_{T1} = 0.7 + 0.5 = 1.2]
• Moderate (T2):
[CF_{T2} = 0.8 + 0.6 = 1.4]
• High (T3):
[CF_{T3} = 0.9 + 0.8 = 1.7]
Diagnosis Results:
Because the highest score is at the **High** level (**1.7**), the system determines:
[\text{Stress Level} = \text{High}]

Ambiguity Case Example
If the user selects symptoms that result in a score of:
• Mild CF = 1.5
• Moderate CF = 1.5
• High CF = 1.2
Then the system will choose: [\text{Stress Level} = \text{Mild}]

Because the system follows the precautionary principle and chooses the lower stress level when the score is tied

The inference flow of the developed expert system is visually explained through the diagram in Figure 1. It begins with forward chaining rule matching and continues with a fallback scoring mechanism if no identical matching rule is found. This diagram illustrates the integration between the deterministic approach and score-based weighting in the system's reasoning process.

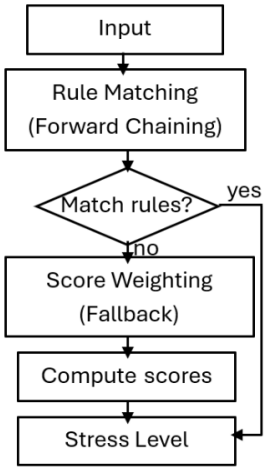

Figure 1. The Steps of the Hybrid Forward Chaining and Score

2.4. Data and Participants

a. Inclusion and Exclusion Criteria

This study involved participants recruited using a purposive sampling technique, with the following inclusion criteria:

- Active students of STT Ronggolawe from the Informatics, Mechanical Engineering, Civil Engineering, and Electrical Engineering study programs
- Alumni who have graduated within the past year
- Aged between 18 and 30 years
- Willing to complete the questionnaire in full

Exclusion criteria included:

- Respondents who did not complete the questionnaire
- Duplicate or invalid data

b. Participant Demographics

A total of 100 respondents participated in this study, with the following demographic distribution (Table 7):

Table 7. Demographic Distribution

Characteristics	Category	**Number (n)**	**Percentage (%)**
Gender	Male	68	68%
	Female	32	32%
Study Program	Informatics	31	31%
	Mechanical Engineering	27	27%
Age Range	Civil Engineering	29	29%
Characteristics	Electrical Engineering	13	13%
Gender	18–30 years old	100	100%

c. Ethics Procedures and Approval

This research has obtained official permission from STT Ronggolawe through an institutional research permit. Participants were recruited voluntarily and provided access to the online questionnaire link without coercion, with the assumption that completing the questionnaire indicated implicit consent to participate. Before completing the questionnaire, all respondents received an informal explanation of the research objectives and procedures through an introduction in a Google Form. The Student Stress Symptoms Questionnaire was accessed through the Google Form link (https://forms.gle/8ar4aEtyz4qURRsHA). The collected data was then exported and input into a web-based hybrid expert system, which combines forward chaining and weighted scoring mechanisms. Before being processed by the system, all responses were pre-assessed by an educational psychology expert. The collected data is kept confidential and will only be used for academic analysis purposes.

d. Handling Missing Data

All incoming data was checked for completeness and validity. Data that was incomplete, duplicate, or did not meet the inclusion criteria was excluded from the analysis. This study found no significant missing data, so all 100 respondents were deemed valid and used in the expert system inference process.

2.5. Validation

Expert system validation was conducted through a functional verification approach, focusing on the conformity of the system's decisions to expert classifications and auditing the inference logic. The system did not undergo a statistical training process like machine learning models (k-fold, training set, test set, or model fitting are irrelevant for expert systems); the knowledge base was pre-established through expert consensus, and inference was performed deterministically based on rules and certainty factor (CF) values.

A total of 100 cases from respondents were first assessed by three independent experts. Each expert assigned a stress level classification (mild, moderate, high) or (T1, T2, T3) based on the combination of symptoms selected by the respondents. After the expert assessments were completed, all cases were run through the hybrid expert system. The system's classification results were then compared with the classifications from the three experts. Validation was conducted based on the majority agreement principle, namely:

- If the system's output matches at least two of the three experts, the case is declared valid.
- If the system's output differs from the majority of experts, the case is declared invalid, and an audit was conducted of:
    - Rules active in inference
    - Distribution of certainty factor (CF) values

2.6. Evaluation

Expert system performance evaluation is conducted by measuring the conformity of the system's output to expert classifications, using standard classification metrics adapted for the context of functional verification. Although the system does not undergo statistical training, these metrics are still used to provide a quantitative overview of the system's accuracy and consistency.

a. Evaluation Metrics

Evaluation metrics are calculated based on a tabulation of valid and invalid cases, with reference to the classification results of three experts. Validation is performed based on the majority principle (at least two experts agree), and the metrics are calculated as follows:

- Accuracy
  The proportion of total cases where the system's output agrees with the majority of experts.
- Precision
  The proportion of the system's correct classifications among all classifications provided by the system for a given stress level.

• Recall
The proportion of cases that should have been classified as a given stress level that were successfully recognized by the system.
• F1 Score
The harmonic mean of precision and recall, used to assess the balance between classification accuracy and completeness.

b. Confidence Interval

To provide a statistical estimate of the stability of a metric, a Confidence Interval (CI) is calculated. The range of accuracy, precision, recall, and F1 score values with a specific confidence level (e.g., 95%), using the Wilson score interval approach.

3. Results and Discussion

3.1. Expert Assessment and Distribution of Stress Levels

Before testing the hybrid system, which combines a rule-based forward chaining mechanism with a certainty factor (CF) fallback scoring mechanism, 100 cases were manually assessed by three independent experts. Each expert assigned a stress level based on the symptoms present in each respondent. Decisions were made through majority classification (≥2 experts agreeing on the same level). The results of these decisions were used as the primary reference in the system validation process. Figure 2 below presents the distribution of expert classifications for the 100 cases based on stress level.

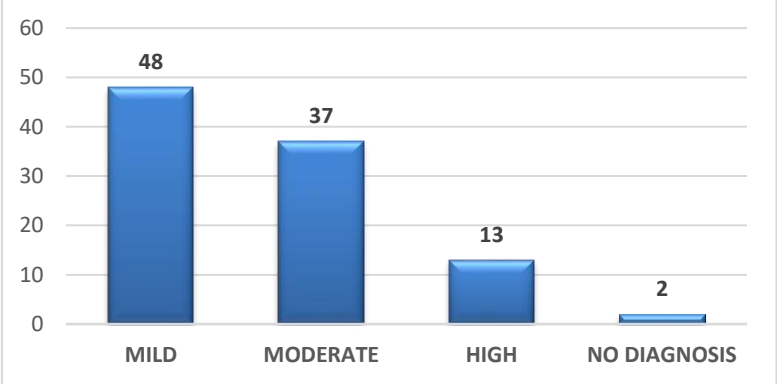

Figure 2. Distribution Graph of Initial Classification by Experts Based on Stress Level

The visualization shows that the majority of cases were classified as Mild (48%), followed by Moderate (37%), and High (13%). Two cases did not meet the diagnostic criteria because no symptoms were reported or selected, and were therefore classified as "No Diagnosis." Table 8 below presents five examples from 100 cases to demonstrate how majority agreement classification was determined based on individual expert assessments.

Table 8. Expert Assessment and Majority Agreement Results on Five Representative Cases

Id case	Stress symptoms	Expert 1	Expert 2	Expert 3	Majority Agreement
1	J10, J13, J15, J17, J21, J26, J36, J56	Moderate	Mild	Mild	Mild
9	J2, J10, J12, J32, J36, J51	High	High	Moderate	High
30	J56, J32, J54, J57	Moderate	Moderate	Moderate	Moderate
39	-	No diagnosis	No diagnosis	No diagnosis	No diagnosis
88	J36, J44	Mild	Mild	Mild	Mild

3.2. Expert Validation of System Diagnosis Results

To ensure the accuracy and reliability of the developed diagnostic system, a validation process was conducted by three independent experts, namely by comparing the system diagnosis and the expert assessment of 100 cases (respondents). Table 9 presents five of the 100 cases, which showed symptoms, the system diagnosis, the expert diagnosis (majority agreement results), and the final validation status.

Table 9. Expert Validation of System Diagnosis (5 Case Examples)

Id case	Stress symptoms	System Diagnosis	Expert Diagnosis	Conclusion
1	J10, J13, J15, J17, J21, J26, J36, J56	Mild	Mild	Valid
9	J2, J10, J12, J32, J36, J51	High	High	Valid
30	J56, J32, J54, J57	Moderate	Moderate	Valid
39	-	No diagnosis	No diagnosis	Valid
88	J36, J44	Mild	Mild	Valid

Based on Table 9, there was 100% agreement between the system's diagnosis and the expert assessment in all cases. This indicates that the system consistently interpreted the combination of symptoms according to clinical practice. In cases where no symptoms were reported (e.g., Case 39), the system did not provide a diagnosis, further confirming its accuracy and reliability.

3.3. Performance Comparison of Hybrid vs. Forward Chaining Methods

To assess the effectiveness of the hybrid approach, a comparative evaluation was conducted on 100 expert-validated test cases. Table 10 presents a detailed comparison of the two methods.

Table 10. Comparison Performance

Category	Hybrid (Forward Chaining + Scoring)			Forward Chaining		
	Number of Correct Diagnoses	Number of Wrong Diagnoses	Accuracy	Number of Correct Diagnoses	Number of Wrong Diagnoses	Accuracy
Mild	48	0	100%	34	14	70.8%
Moderate	37	0	100%	28	17	75.7%
High	13	0	100%	4	1	30.8%
No Diagnosis	2	0	100%	2	0	100%
Total Valid	100	0	100%	68	32	68.0%

Tables 11 and 12 compare the precision, recall, and F1 scores between the forward chaining method and the hybrid method (forward chaining + fallback scoring). It can be seen that the hybrid method achieves a perfect score (1.000) in all categories, while the forward chaining method shows lower and more variable performance, especially in the High category, which has an F1 score of only 0.444.

Table 11. Diagnostic Performance Comparison per Category

Category	Precision (FC)	Recall (FC)	F1-Score (FC)	Precision (Hybrid)	Recall (Hybrid)	F1-Score (Hybrid)
Mild	0.708	0.708	0.708	1.000	1.000	1.000
Moderate	0.757	0.622	0.683	1.000	1.000	1.000
High	0.308	0.800	0.444	1.000	1.000	1.000
No Diagnosis	1.000	1.000	1.000	1.000	1.000	1.000

Table 12. Performance Comparison Summary of Both Approaches

Metrix	Forward Chaining	Hybrid (FC + Score)	Difference (Δ)
Accuracy	0.680	1.000	+0.320
Precision	0.693	1.000	+0.307
Recall	0.783	1.000	+0.217
F1-Score	0.709	1.000	+0.291

The evaluation results showed a significant performance difference between the classic forward chaining method and the hybrid approach, which combines forward chaining with score weighting. With the forward chaining method, the system was only able to provide a diagnosis for 68 out of 100 test cases, with an overall accuracy of 68%. Quantitatively, the hybrid approach provided a 32% increase in accuracy. These results confirm that handling non-explicit or partial input is a key factor in improving expert system performance, especially in the psychological domain, where symptom data can be incomplete or ambiguously structured.

3.4. Ablation: Evaluation of Forward Chaining vs Hybrid

Ablation was performed to isolate the contribution of each expert system component. The main focus was to compare system performance using only deterministic forward chaining versus a hybrid system that combines a rule-based forward chaining mechanism with a fallback scoring certainty factor (CF). Table 13 displays the distribution of inferences based on the method from 100 cases with varying symptom distributions.

Table 13. Distribution of Inferences Based on Method

Inference Method	Number of Case Valid	Invalid	Local Accuracy
Forward Chaining			
Identical Rules	25	25	0	100.0%
Partial Rule Matching	73	41	32	56.2%
No Symptoms Selected	2	2	0	100.0%
Total Forward Chaining	**100**	**68**	**32**	**68.0%**
Hybrid (FC + Scoring)			
Identical Rules	25	25	0	100.0%
Failback Scoring Mechanism	73	73	0	100.0%
No Symptoms Selected	2	2	0	100.0%
Total Hybrid	100	100	0	100.0%

In the hybrid system, once the identical rule is not met, the system immediately activates the CF scoring fallback mechanism. There is no partial matching process as in classic forward chaining. Therefore, all 73 cases that do not match identically are immediately processed through CF scoring, resulting in valid and accurate classifications. With this approach, the hybrid system successfully classified all 100 cases, including 32 that previously failed forward chaining. Local accuracy increased from 68% to 100%, without any loss of validation.

3.5. Robustness Validation and Inference Audit

a. Robustness to Premise Incompleteness

The hybrid system was tested on 32 cases that failed to be classified by pure forward chaining. Incomplete premises in these cases caused the rule base to be inactive, requiring the system to rely on the CF fallback scoring mechanism. As a result, all cases were validly classified, demonstrating the

hybrid system's high resilience to partial or incomplete input. Table 14 shows the distribution of stress levels generated by the fallback scoring mechanism.

Table 14. Distribution of Inference Fallback Scoring

Selected Stress Level	Number of Cases	Percentage
Mild (T1)	11	34.4%
Moderate (T2)	10	31.3%
High (T3)	11	34.4%
Total	**32**	**100%**

The distribution of inference results shows that the CF fallback scoring produces a balanced classification. It is not biased towards either the mild or the severe levels. This strengthens the validation that the hybrid system is able to handle incomplete premises proportionally.

b. Inference Audit Against Expert References

Validation was conducted to ensure that the hybrid system not only successfully classifies cases that failed forward chaining but also produces inferences that are consistent with the majority expert classification and can be explicitly audited. To this end, an audit was conducted on all 32 cases processed through the CF fallback scoring mechanism. This audit compared the system's classification results with the predetermined expert classification (the majority decision of three independent experts), which is presented in table 15.

Table 15. Results of the Hybrid System Inference Audit Against Expert References

Selected Stress Level	Number of Cases	Percentage	Consistent with experts
Mild (T1)	11	34.4%	100%
Moderate (T2)	10	31.3%	100%
High (T3)	11	34.4%	100%
Total	**32**	**100%**	**100%**

This table shows that all inferences generated by the hybrid system through the CF fallback scoring mechanism are consistent with the majority expert classifications. No contradictory cases were found, thus system validation can be declared complete and scientifically justifiable.

c. Inference Trail Audit

Thirty-two cases were successfully classified by the hybrid system and demonstrated consistency with the previously established expert classifications. These cases are: 2, 3, 9, 11, 12, 13, 15, 24, 25, 32, 41, 43, 44, 45, 46, 48, 50, 53, 54, 56, 57, 58, 60, 64, 69, 70, 74, 75, 82, 88, 89, and 90. To strengthen the validation of the hybrid system, Table 16 below displays five of the 32 cases, along with the distribution of CF scores across three stress levels, system inferences, and matching status against expert classifications.

Table 16. Hybrid System Inference Audit Against Expert References (5 Representative Cases)

Case	Active Symptoms	Mild (T1)	Moderate (T2)	High (T3)	Highest Score	System Inference Status (hybrid)	Inference Status (Expert)	Validation
3	J2, J6, J29, J35, J16, J38, J46	J2(0.7)+ J35(0.9) =**1.6**	J2(0.7)+J6(0. 7)+J16(0.7)+ J29(0.6)=**2.7**	J2(0.7)+J6(0 .7)+J38(0.4) +J46(0.6)=**2 .4**	2.7	Moderat e	Moderat e	Valid
15	J1, J3, J5, J8, J6, J9, J10, J17, J18, J20, J22, J23, J24, J28, J29, J40, J50, J56	J1(0.8)+ J3(0.7)+ J9(0.6)+ J17(0.8) +J22(0. 8)+J56(0.9)=**4.6**	J3(0.7)+J6(0. 7)+J9(0.6)+J 10(0.7)+J22(0.8)+J23(0.6) +J28(0.4)+J2 9(0.6)+J50(0. 8)=**5.9**	J5(0.9)+J8(0 .9)+J9(0.6)+ J10(0.7)+J1 8(0.9)+J20(0.8)+J24(0.8)+J40(0.8)= **6.4**	6.4	High	High	Valid
45	J2, J26, J36	J26(0.8) +J36(0. 8)=**1.6**	J2(0.7)=**0.7**	J2(0.7)=**0.7**	1.6	Mild	Mild	Valid
60	J2, J1, J10, J5, J8, J14, J26	J1(0.8)+ J26(0.8) =**1.6**	J2(0.7)+J10(0.7)=**1.4**	J2(0.7)+J5(0 .9)+J8(0.9)+ J10(0.7)+J1 4(0.8)=**4.0**	4.0	High	High	Valid
89	J36, J1, J2, J13, J16, J28, J46	J1(0.8)+ J13(0.6) +J36(0. 8)=**2.2**	J2(0.7)+J16(0.7)+J28(0.4) =**1.8**	J2(0.7)+J46(0.6)=**1.3**	2.2	Mild	Mild	Valid

The hybrid system demonstrates high robustness in the face of incomplete premises. The CF scoring fallback mechanism not only saves cases but also maintains the accuracy and transparency of inferences. With an explicit audit trail, this system is suitable for use in scientific validation contexts and expert system-based educational applications.

d. Robustness to Low Scores

The hybrid system does not set a minimum threshold for classification. Inference is performed by selecting the stress level with the highest CF score, regardless of its value. Therefore, the system still issues a diagnosis even if the highest score is below a certain value, as long as there is a clear difference in scores between levels. To ensure that the system does not produce weak or inconclusive diagnoses, an analysis of cases with low scores was conducted. The results showed that even though some cases had a highest score below 3.0, the system still produced classifications that were consistent with expert judgment and stable against input disturbances (presented in Table 17).

Table 17. Robustness to Low CF Scores

Case	Highest Score	Selected Level	Consistent with Experts	Validation
3	2.7	Moderate	Yes	Valid
45	1.6	Mild	Yes	Valid
89	2.2	Mild	Yes	Valid

The three cases above had the highest scores <3.0, yet they still produced valid classifications that were consistent with expert references. This demonstrates that the hybrid system relies not on absolute thresholds, but rather on the distribution of scores across levels.

e. Noise Simulation: Adding Random Phenomena

The purpose of this simulation is to test whether the hybrid system still produces stable inferences when the input symptoms are contaminated" by random symptoms irrelevant to the rule base. Table 18 presents noise simulations for five example cases.

Table 18. Noise Simulations

Case	Active Symptoms	Additional Symptoms	New High Score	Initial Inference	Inference After Noise	Status
3	J2, J6, J29, J35, J16, J38, J46	J3	3.4	Moderate	Moderate	Stable
15	J1, J3, J5, J8, J6, J9, J10, J17, J18, J20, J22, J23, J24, J28, J29, J40, J50, J56	J16 and j30	7.0	High	High	Stable
45	J2, J26, J36	J4 and J5	2.4	Mild	Mild	Stable
60	J2, J1, J10, J5, J8, J14, J26	J9	4.6	High	High	Stable
89	J36, J1, J2, J13, J16, J28, J46	J31	2.2	Mild	Mild	Stable

Results: The addition of noise did not significantly affect the distribution of CF scores. The system continued to select the same stress level, demonstrating robustness to random input.

f. Symptom Drop Simulation: Robustness to Premise Incompleteness

This simulation tests whether the hybrid system still produces valid inferences when one or more highly weighted symptoms are removed from the input. Five representative cases were selected from the fallback scoring group, and the removal of dominant symptoms was performed to assess the impact on the CF score distribution and classification results (presented in Table 19).

Table 19. Symptom Drop Simulation

Case	Active Symptoms	Symptoms Removed	New High Score	Initial Inference	Inference After Noise	Status
3	J2, J6, J29, J35, J16, J38, J46	J2 (0.7)	2.0	Moderate	Moderate	Stable
15	J1, J3, J5, J8, J6, J9, J10, J17, J18, J20, J22, J23, J24, J28, J29, J40, J50, J56	J10 (0.7)	5.7	High	High	Stable
45	J2, J26, J36	J36 (0.8)	0.8	Mild	Mild	Stabil
60	J2, J1, J10, J5, J8, J14, J26	J5 (0.9) and J14 (0.8)	2.3	High	High	Stable
89	J36, J1, J2, J13, J16, J28, J46	J13 (0.6)	1.8	Mild	Moderate	No Stable

Four of the five cases still produced the same inferences even after the dominant symptom was removed. Case 89 showed a change in inference from Mild to Moderate after the removal of J13 (0.6), which shifted the score distribution between levels. This indicates that the hybrid system remains stable in producing inferences even when the input is incomplete. Classification changes only occur if the removed symptom has a significant impact on the score distribution between levels. Thus, the system demonstrates robustness to incomplete premises, while remaining sensitive to key, defining symptoms.

3.6. Reliability: Inter-Expert Agreement and System Evaluation Stability

The reliability of an expert system depends not only on the accuracy of its inferences, but also on the consistency of the reference labels used for validation and the stability of the evaluation metrics. Therefore, reliability analysis is conducted in two dimensions: (1) the level of agreement between experts as sources of reference labels, and (2) the stability of the hybrid system evaluation metrics through statistical analysis.

a. Inter-Expert Agreement

To ensure that the majority expert classification could be used as a valid reference, an agreement analysis was conducted between three independent experts using the pairwise Cohen's kappa approach. Because kappa is only applicable to two raters, the value was calculated for each pair of experts and averaged. The distribution of classifications from each expert across the 100 cases showed high consistency, as shown in Figure 3, and the results of the Kappa calculations are presented in Table 20.

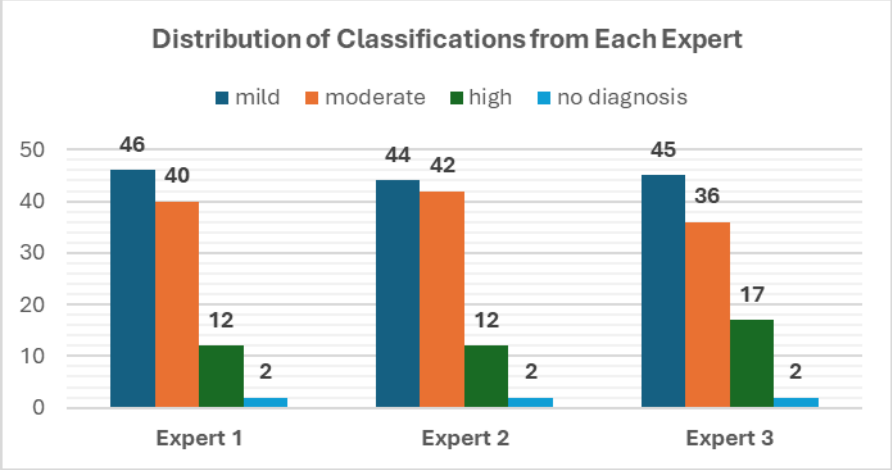

Figure 3. Distribution of Classifications from Each Expert

Table 20. Kappa Calculation Results

Expert Pairs	Same Agreement	Po	Pe	Kappa (κ)	Interpretation
Expert 1 vs. 2	88	0.88	0.3951	0.8017	Very Good Agreement
Expert 1 vs. 3	90	0.90	0.3974	0.8347	Very Good Agreement
Expert 2 vs. 3	91	0.91	0.3966	0.8476	Very Good Agreement
Average	**89.7**	**0.8967**	**0.3964**	**0.8280**	**Very Good**

Based on Altman's (1991) [27] interpretation, a value between 0.81–1.00 indicates very good agreement. With an average kappa value of 0.8280, the majority expert classification can be used as a highly reliable reference for evaluating hybrid systems. This level of consistency strengthens the system's validity in producing inferences that meet expert standards.

b. Stability of System Evaluation Metrics

In addition to high accuracy, a reliable expert system must demonstrate stable performance when tested against data variations or resampling. To achieve this, a confidence interval (CI) analysis was performed on the main evaluation metrics: accuracy, precision, recall, and F1-score. Because the hybrid system produced classifications that fully matched the expert's reference for 100 validation cases, all evaluation metrics were scored at 100%. However, to demonstrate that the system's performance was not only high but also statistically consistent, the Wilson score interval approach was used to calculate the CI for extreme proportions. The results are shown in Table 21.

Table 21. Results of the Confidence Interval (CI) Calculation

Metric	Value	95% CI
Akurasi	100%	[96.4%, 100%]
Precision	100%	[96.4%, 100%]
Recall	100%	[96.4%, 100%]
F1-Score	100%	[96.4%, 100%]

The narrow confidence interval indicates that the system's performance is not only high but also statistically stable. This indicates that the hybrid system is reliable in a variety of classification scenarios, including resampling of case data.

3.7. Risks and Ethics

The hybrid expert system developed in this study demonstrated high accuracy across 100 validation cases. However, system evaluation relies not only on technical performance but also on awareness of the methodological limitations and ethical implications of its use. Because this system did not undergo a training process like machine learning approaches, the term "overfitting" is not used technically. Instead, it is important to note that the system was tested on a limited number of cases and in a homogeneous institutional context, so its external validity cannot be ascertained.

Furthermore, the label distribution exhibited a class imbalance, with only 13 cases classified as "High" out of 100. This imbalance has the potential to impact the system's sensitivity to severe stress, especially when applied to populations with varying prevalence. Therefore, balancing the case distribution and cross-population validation are crucial steps in further system development.

From an ethical perspective, the use of expert systems in stress classification has significant consequences. While systems can aid early identification, automated classification should not be used as the sole basis for psychological interventions, educational decisions, or clinical assessments. Systems should be positioned as transparent and auditable tools, with interpretation still involving human professionals. The principles of precaution, informed consent, and protection against psychological impact must be an integral part of the implementation of this system.

3.8. Comparative Literature: Methodological Positioning in the Hybrid Diagnosis Landscape

The system developed in this study uses a hybrid approach that combines rule-based forward chaining as the primary inference pathway, with a fallback to Certainty Factor (CF) scores to handle uncertainty or cases not fully defined by the rules. This approach is designed to maintain logical transparency while providing flexibility in dealing with symptom variations. Although this study does not include an empirical comparison with other hybrid approaches such as pure CF systems or Case-Based Reasoning (CBR), the methodological positioning of this system can be described descriptively (Table 22).

Table 22. Comparison of General Characteristics of Hybrid Methods and Comparative Methods

Aspects	Certainty Factor (CF)	Case-Based Reasoning (CBR)	This System (Forward Chaining + CF Fallback)
Primary Inference Path	CF score per rule	Similarity to previous cases	Explicit rules (forward chaining)
Uncertainty Handling	Native CF scoring	Similarity score between cases	CF fallback if rules are insufficient
Logical Transparency	Medium (depending on rule structure)	Low (based on historical cases)	High (explicit rules + directed fallback)
Auditability	Limited	Low	High (rule documentation and fallback path)

Aspects	Certainty Factor (CF)	Case-Based Reasoning (CBR)	This System (Forward Chaining + CF Fallback)
Data Dependability	Low	High (requires a representative case base)	Low (not based on historical training)
Adaptability	High if rule is flexible	High if the case base is large	Modular and extensible

This approach allows the system to produce deterministic inferences when rules are sufficient, and still provide meaningful output when ambiguity arises, without sacrificing transparency or auditability. With an explicit rule structure and documented CF fallbacks, the system offers a balance between logical clarity and inference flexibility. Quantitative comparisons with other approaches are planned as a next step, to test the system's robustness in broader, more heterogeneous contexts.

4. Conclusion

This study successfully designed and evaluated a hybrid expert system for classifying stress levels based on psychological symptoms, with the primary inference pathway being rule-based forward chaining and a fallback mechanism using Certainty Factor (CF) scores. Key verified contributions include: an explicit and documented rule-based structure, fully consistent classification with expert references across 100 validation cases, and high system auditability through documentation of logic and symptom distribution. However, this study has limitations that should be explicitly noted. The evaluation was conducted at only one institution, with a relatively small sample size and an unbalanced label distribution, particularly in the severe stress category. This opens up the possibility of undetected spectrum bias, thus the system's external validity cannot be confirmed.

Future work plans include testing the system on cross-institutional data to assess the robustness of inferences across population variations. Furthermore, the system's integration with counseling services in educational settings is designed to support more responsive and data-driven interventions. Finally, an open benchmark against other hybrid approaches, such as pure CF and CBR, will be conducted to assess the relative merits of this system quantitatively and methodologically.

Acknowledgements

The author would like to express his gratitude to the Ministry of Higher Education, Science, and Technology (Kemdiktisaintek) of the Republic of Indonesia, through the novice lecturer research scheme, funding for Fiscal Year 2025, for fully supporting the implementation of this research.

We would like to thank the students, alumni, and institutional staff of STT Ronggolawe, the research subjects, for their participation and data support. We also appreciate the contributions of colleagues and other parties who assisted in the collection, analysis, and preparation of this paper.

References

[1] M. Mofatteh, "Risk factors associated with stress, anxiety, and depression among university undergraduate students," *AIMS Public Health*, vol. 8, no. 1, pp. 36–65, 2021, https://doi.org/10.3934/publichealth.2021004

[2] M. Huenergarde, "College Students' Well–Being: Use of Counseling Services," *Am. J. Undergrad. Res.*, vol. 15, no. 3, Dec. 2018, https://doi.org/10.33697/ajur.2018.023

[3] Z. Xu, "The Impact of Online Resources on Reducing Mental Health Stigma among International College Students: A Case Study," *Lect. Notes Educ. Psychol. Public Media*, vol. 58, no. 1, pp. 141–147, July 2024, https://doi.org/10.54254/2753-7048/58/20241770.

[4] A. Karttunen, A. Hakkarainen, and L. Holopainen, "Associations between School Burnout, Perceived Learning Difficulties, and Delayed Graduation from Upper Secondary Education: A Longitudinal Study," *Int. J. Educ. Psychol.*, vol. 12, no. 3, pp. 289–306, Oct. 2023, https://doi.org/10.17583/ijep.10637.

[5] R. Yusof, N. H. Mohamed Harith, A. Lokman, M. F. Abdul Batau, Z. Mohd Zain, and N. H. Rahmat, "A Study of Perception on Students' Motivation, Burnout and Reasons for Dropout," *Int. J. Acad. Res. Bus. Soc. Sci.*, vol. 13, no. 7, p. Pages 403-432, July 2023, http://dx.doi.org/10.6007/IJARBSS/v13-i7/17187.

[6] Ahsanu Bil Husna, Asyifa Salsabila Rahmi, and Nur Ilmya Nugraha Ningrum Irfandi Putri, "Exploring the Root Causes of Burnout Syndrome among College Students: A Systematic Literature Review of Contributing Factors," *J. Promkes*, vol. 13, no. SI1, pp. 246–260, Jan. 2025, https://doi.org/10.20473/jpk.V13.ISI1.2025.246-260.

[7] N.B. Mahesh Kumar, T. Chithrakumar, T. Thangarasan, J. Dhanasekar, and P. Logamurthy, "AI-Powered Early Detection and Prevention System for Student Dropout Risk," *Int. J. Comput. Exp. Sci. Eng.*, vol. 11, no. 1, Jan. 2025, https://doi.org/10.22399/ijcesen.839.

[8] D. B. Olawade, O. Z. Wada, A. Odetayo, A. C. David-Olawade, F. Asaolu, and J. Eberhardt, "Enhancing mental health with Artificial Intelligence: Current trends and future prospects," *J. Med. Surg. Public Health*, vol. 3, p. 100099, Aug. 2024, https://doi.org/10.1016/j.glmedi.2024.100099.

[9] O. Ali, W. Abdelbaki, A. Shrestha, E. Elbasi, M. A. A. Alryalat, and Y. K. Dwivedi, "A systematic literature review of artificial intelligence in the healthcare sector: Benefits, challenges, methodologies, and functionalities," *J. Innov. Knowl.*, vol. 8, no. 1, p. 100333, Jan. 2023, https://doi.org/10.1016/j.jik.2023.100333.

[10] H. Henderi, F. Al Khudhorie, G. Maulani, S. Millah, and V. T. Devana, "A Proposed Model Expert System for Disease Diagnosis in Children to Make Decisions in First Aid," *INTENSIF J. Ilm. Penelit. Dan Penerapan Teknol. Sist. Inf.*, vol. 6, no. 2, pp. 139–149, Aug. 2022, https://doi.org/10.29407/intensif.v6i2.16912.

[11] S. I. Oguoma, K. K. Uka, C. A. Chukwu, and E. C. Nwaoha, "An Expert System for Diagnosis and Treatment of Mental Ailment," *OALib*, vol. 07, no. 04, pp. 1–22, 2020, https://doi.org/10.4236/oalib.1106166.

[12] M. T. Hafizal *et al.*, "Implementation of expert systems in potassium deficiency in cocoa plants using forward chaining method," *Procedia Comput. Sci.*, vol. 216, pp. 136–143, 2023, https://doi.org/10.1016/j.procs.2022.12.120.

[13] M. D. Sinaga, F. Tambunan, C. J. M. Sianturi, A. Syahputra, F. Tahel, and S. Aliyah, "An Expert System for Diagnosing Leptospirosis Disease Using Forward Chaining and Bayes Theorem," in *2019 7th International Conference on Cyber and IT Service Management (CITSM)*, 2019, pp. 1–4. https://doi.org/10.1109/CITSM47753.2019.8965338.

[14] T. Mazhar *et al.*, "A Novel Expert System for the Diagnosis and Treatment of Heart Disease," *Electronics*, vol. 11, no. 23, p. 3989, Dec. 2022, https://doi.org/10.3390/electronics11233989.

[15] D. C. Klonoff *et al.*, "Advance Insulin Injection Technique and Education With FITTER Forward Expert Recommendations," *Mayo Clin. Proc.*, vol. 100, no. 4, pp. 682–699, Apr. 2025, https://doi.org/10.1016/j.mayocp.2025.01.004.

[16] C. Wang, J. Zhang, N. Lassi, and X. Zhang, "Privacy Protection in Using Artificial Intelligence for Healthcare: Chinese Regulation in Comparative Perspective," *Healthcare*, vol. 10, no. 10, p. 1878, Sept. 2022, https://doi.org/10.3390/healthcare10101878.

[17] M. Abdelshiheed, J. W. Hostetter, X. Yang, T. Barnes, and M. Chi, "Mixing Backward- with Forward-Chaining for Metacognitive Skill Acquisition and Transfer," in *Artificial Intelligence in Education*, M. M. Rodrigo, N. Matsuda, A. I. Cristea, and V. Dimitrova, Eds., Cham: Springer International Publishing, 2022, pp. 546–552, http://dx.doi.org/10.48550/arXiv.2303.12223.

[18] S. Kamley, S. Jaloree, and R. S. Thakur, "Forecasting of Major World Stock Exchanges Using Rule-Based Forward and Backward Chaining Expert Systems," in *Quality, IT and Business Operations: Modeling and Optimization*, P. K. Kapur, U. Kumar, and A. K. Verma, Eds., Singapore: Springer Singapore, 2018, pp. 297–306. https://doi.org/10.1007/978-981-10-5577-5_23.

[19] J. Rodríguez-Arce, L. Lara-Flores, O. Portillo-Rodríguez, and R. Martínez-Méndez, "Towards an anxiety and stress recognition system for academic environments based on physiological features," *Comput. Methods Programs Biomed.*, vol. 190, p. 105408, July 2020, https://doi.org/10.1016/j.cmpb.2020.105408.
[20] M. H. Zolfagharnasab *et al.*, "A novel rule-based expert system for early diagnosis of bipolar and Major Depressive Disorder," *Smart Health*, vol. 35, p. 100525, Mar. 2025, https://doi.org/10.1016/j.smhl.2024.100525.
[21] J. Wang *et al.*, "Refined micro-scale geological disaster susceptibility evaluation based on UAV tilt photography data and weighted certainty factor method in Mountainous Area," *Ecotoxicol. Environ. Saf.*, vol. 189, p. 110005, 2020, doi: https://doi.org/10.1016/j.ecoenv.2019.110005.
[22] H. Soetanto, Painem, and M. K. Suryadewiansyah, "Optimization of Expert System Based on Interpolation, Forward Chaining, and Certainty Factor for Diagnosing Abdominal Colic," *J. Comput. Sci.*, vol. 20, no. 2, pp. 191–197, Feb. 2024, https://doi.org/10.3844/jcssp.2024.191.197.
[23] A. B. Tofighi, A. Ahmadi, and H. Mosadegh, "A novel case-based reasoning system for explainable lung cancer diagnosis," *Comput. Biol. Med.*, vol. 185, p. 109547, 2025, doi: https://doi.org/10.1016/j.compbiomed.2024.109547.
[24] J. Rodríguez-Arce, L. Lara-Flores, O. Portillo-Rodríguez, and R. Martínez-Méndez, "Towards an anxiety and stress recognition system for academic environments based on physiological features," *Comput. Methods Programs Biomed.*, vol. 190, p. 105408, July 2020, https://doi.org/10.1016/j.cmpb.2020.105408.
[25] K. Febrianto, E. D. Udayanti, B. V. Indriyono, W. Mahmud, and I. Zahari, "Expert System for Detection of Diseases in Layers Using Forward Chaining and Certainty Factor Methods," *J. Masy. Inform.*, vol. 14, no. 2, pp. 80–95, Nov. 2023, https://doi.org/10.14710/jmasif.14.2.52266
[26] I. Azmi, G. Gunawan, and S. Anandianskha, "Application of expert system using certainty factor method to identify diseases in rice plants," vol. 12, no. 4, 2024, https://doi.org/10.35335/mandiri.v12i4.280.
[27] M. Li, Q. Gao, and T. Yu, "Kappa statistic considerations in evaluating inter-rater reliability between two raters: which, when and context matters.," *BMC Cancer*, vol. 23, no. 1, p. 799, Aug. 2023, https://doi.org/10.1186/s12885-023-11325-z.