Jurnal JEETech: Vol. 7 No. 1 Tahun 2026
Journal Of Electrical Engineering And Technology
https://ejournal.ft-undar.ac.id/index.php/jeetech
DOI: https://doi.org/10.32492/jeetech.v7i1.7103
e-ISSN: 2722-5321
p-ISSN: 2964-7320

Development of a Web-Based Mental Health Screening System Using a Large Language Model and Intervention Recommendations

[1*]Rony Kriswibowo, [2]Rusina Widha Febriana, [3]Agung Budi Setyawan, [4]Selfya Ningrum, [5]Danuditya Purna Atmaja

[1,2,3]Department System Information, Anwar Medika University, Sidoarjo
[4]Department Health Administration, Anwar Medika University, Sidoarjo
[5]Department Physiotherapy, Anwar Medika University, Sidoarjo
[1]rkriswibowo@gmail.com, [2]widha.ennvil@gmail.com, [3]agung.budi@uam.ac.id, [4]selfya.ningrum@uam.ac.id, [5]danuditya@uam.ac.id

Article Info

Article history:

Received March 9, 2026
Revised March 19, 2026
Accepted April 25, 2026

Keyword:

Mental health screening
DASS-21
Large Language Model
System Usability Scale
Blackbox Testing

ABSTRACT

Mental health disorders among adolescents in Indonesia remain largely underdetected due to limited access to services, persistent stigma, and the lack of personalized feedback in conventional screening tools. This study developed and evaluated a web-based mental health screening system that integrates the DASS-21 questionnaire with a large language model (GPT-4) to generate personalized intervention recommendations. The system was built using the Waterfall methodology and designed to calculate DASS-21 severity scores for depression, anxiety, and stress, then pass both quantitative scores and optional user free-text input to the LLM via API. Blackbox testing was conducted to validate functional requirements, and the System Usability Scale (SUS) was administered to 30 adolescent users to assess usability. Results showed that all functional test cases passed after resolving an initial LLM response parser issue. The average SUS score was 91.59 (Grade A, Acceptable range), with no participant rating the system below 70, indicating consistently high usability across users. The hybrid approach proved advantageous: the DASS-21 provided clinical grounding that reduced LLM hallucination risk, while the LLM added contextual personalization that static questionnaires lack. However, the high usability score does not automatically translate to clinical effectiveness. Future work should include clinical validation studies comparing LLM-generated recommendations against psychologist assessments.

Corresponding Author:
Rony Kriswibowo,
Department System Information, Anwar Medika University,
Email: rkriswibowo@gmail.com

Jurnal JEETech: Vol. 7 No. 1 Tahun 2026
Journal Of Electrical Engineering And Technology
https://ejournal.ft-undar.ac.id/index.php/jeetech
DOI: https://doi.org/10.32492/jeetech.v7i1.7103
e-ISSN: 2722-5321
p-ISSN: 2964-7320

1. Introduction

Mental health disorders such as depression and anxiety have become one of the most significant global health burdens. Data from the Institute for Health Metrics and Evaluation show that the COVID-19 pandemic led to a 25% increase in the prevalence of depression and anxiety worldwide in 2020 [1]. The World Health Organization (WHO) reports that more than 300 million people worldwide suffer from depression, making it the leading cause of global disability [2]. However, disparities in access to mental health services remain significant, particularly in low- and middle-income countries [3]. In Indonesia, this situation is exacerbated by a very low ratio of psychiatrists to the population and an uneven distribution of healthcare workers, resulting in many cases of mental disorders going undetected and untreated [4].

Adolescent mental health deserves attention, as several studies have found that there are mental health issues among adolescents in Indonesia [5]. Research findings indicate that one in three Indonesian adolescents has experienced mental health issues in the past 12 months, while one in twenty Indonesian adolescents has been diagnosed with a mental disorder in the past 12 months. These figures correspond to 15.5 million and 2.45 million adolescents, respectively. Adolescents were diagnosed with mental disorders in accordance with the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5), which serves as the guideline for establishing diagnoses of mental disorders in Indonesia and internationally [6].

There are both external and internal factors that can cause a person to experience mental health issues. People with Mental Health Issues (ODMK) are individuals who face physical, mental, social, growth, developmental, or quality-of-life challenges, putting them at risk of developing a mental disorder [7]. Today, everyone both children and adolescents is at risk of experiencing mental health issues. This is because adolescents' lack of interest in seeking information about mental illness, combined with a culture of stigma and the fact that consulting a psychologist remains taboo, poses a major challenge in detecting and treating those with mental health conditions [8].

Common mental health screening tools for anxiety have fundamental limitations. Although these instruments have been widely validated and exhibit high reliability, their static nature and reliance on closed-ended questionnaires fail to capture the nuances of individual expression in a holistic manner. Research by Leaning et al. (2024) emphasizes that traditional screening approaches often overlook the temporal dynamics of symptoms and personal context, which are crucial in the diagnosis of mental disorders [9]. In addition, conventional screening tools do not provide personalized intervention recommendations as a follow-up to screening results. This means that users only learn the severity of their condition without receiving concrete guidance on what steps they should take, thereby limiting the effectiveness of screening in promoting behavioral change [10].

Advances in artificial intelligence, particularly Large Language Models (LLMs) such as GPT-4, Llama 2, and Gemini, are opening up new opportunities for early detection and intervention in mental health. A systematic review by Guo et al. (2024) evaluated 40 studies and concluded that LLMs demonstrate substantial effectiveness in detecting mental health issues through text analysis, as well as providing easily accessible e-health services that have the potential to reduce stigma [2]. Research by Liu et al. (2025) also demonstrates that enhanced large language models are capable of screening for depression and anxiety with

promising accuracy rates [1]. However, the use of LLMs in clinical settings still faces a number of challenges, such as inconsistent text output, the potential for hallucinations (non-factual information), and low interpretability due to the model's blackbox nature. Therefore, integration with validated psychometric instruments is necessary to improve the reliability of screening results.

One widely used instrument is the Depression Anxiety Stress Scale-21 (DASS-21), a 21-item questionnaire-based psychological assessment tool designed to simultaneously identify levels of depression, anxiety, and stress [11]. The DASS-21 offers the advantage of providing structured, standardized measurements that are easy to implement in web-based systems. In the context of developing a mental health screening system the DASS-21 can serve as an objective baseline assessment while LLM is utilized to analyze free-text data from users to capture a deeper emotional context [12]. By combining a quantitative approach via the DASS-21 and a qualitative approach via LLM analysis, the resulting screening system is expected to provide more comprehensive, adaptive, and contextual results, while minimizing the weaknesses of each method individually.

Despite the rapid proliferation of digital mental health tools, a critical methodological divide persists between conventional psychometric screening and AI-driven conversational agents. Web-based questionnaires (e.g., standalone DASS-21 forms) offer validated severity quantification but remain structurally static, failing to translate scores into contextualized, actionable follow-up guidance tailored to individual symptom narratives. Conversely, pure LLM-based mental health chatbots prioritize open-ended interaction but frequently suffer from clinical ungroundedness, output inconsistency, and hallucination risks, particularly when deployed without structured psychometric anchors or culturally adapted prompt frameworks. To date, few studies have systematically integrated validated symptom thresholds as explicit constraints within LLM inference to generate severity-tiered, culturally contextualized intervention recommendations in real-time web environments. This study addresses that gap by introducing a novel hybrid screening architecture that couples DASS-21 quantitative scoring with constrained GPT-4 prompt engineering. Unlike prior chatbot or unanchored LLM approaches that rely solely on free-text analysis, our system enforces clinical grounding by mapping DASS-21 severity levels directly to structured system prompts and implementing a deterministic response parser, thereby mitigating hallucination risks while preserving personalized contextualization. Furthermore, the platform is specifically optimized for Indonesian adolescents, delivering recommendations in Bahasa Indonesia through a lightweight, stigma-reducing web interface. This integration represents a methodological advancement over both static digital questionnaires and standalone AI conversational agents, offering a reproducible, clinically informed framework for scalable adolescent mental health triage.

In Indonesia, the need for innovative mental health screening systems is becoming increasingly urgent given the limited access to services and the high level of stigma surrounding mental disorders. The main barriers to digital screening in Indonesia include the lack of cultural adaptation of international assessment tools and low digital literacy among certain segments of the population. However, a web-based system with a user-friendly interface and recommendations in Indonesian has the potential to overcome these barriers [13]. The use of natural language processing (NLP) and large language models (LLMs) in mental health research is on the rise, with 42.8% of studies using clinical data and 33.7% using social media data as their primary text source. These findings suggest that text-based approaches hold significant potential for further development [14]. Therefore, this study aims to develop and evaluate a web-based mental health screening system that integrates the standard DASS-21 questionnaire with a large language model (LLM) to generate

personalized intervention recommendations, as well as to measure the system's accuracy, relevance, and usefulness in the context of Indonesian users.

2. Literature Review

A. The Burden of Mental Health

Mental health disorders, particularly depression, anxiety, and psychological stress, have undergone a significant epidemiological shift over the past two decades and now occupy a strategic position in the global burden of disease [15]. According to the Global Burden of Disease estimates from the Institute for Health Metrics and Evaluation (IHME), the prevalence of affective disorders has shown a consistent upward trend since 2010, with disability resulting from these disorders accounting for an increasingly dominant share of Disability-Adjusted Life Years (DALYs) among adolescents and the working-age population [16]. This situation has been further exacerbated by the impact of the COVID-19 pandemic, which, according to a report by the World Health Organization (WHO), triggered a 25% global surge in cases of major depression and severe anxiety disorders in 2020. Factors such as prolonged social isolation, socioeconomic uncertainty, and disruptions to primary health care have transformed the public health landscape, positioning mental disorders as a transnational challenge that requires a systemic response, continuous surveillance, and expanded access to evidence-based interventions [17]. Knowledge about adolescent mental health is a preventive measure that can be taken to address mental health disorders in adolescents. Knowledge can be enhanced by improving adolescents' health literacy. Increased knowledge about mental health can positively impact adolescents' mental well-being, as good mental health is essential for coping with the challenges of the globalized world. The concept of mental health literacy refers to the enhancement of knowledge and beliefs regarding mental disorders, as well as their management or prevention [18].

B. Mental Health Screening

Community-based mental health screening is highly effective, consistent with the WHO's (2022) findings that community-based interventions are one of the key strategies for addressing gaps in mental health services [19]. According to Abdul Aziz, mental health screenings showed that 79.8% of students were in the "good" category, 12.9% in the "moderate" category, and 7.3% in the "low" category. The web-based CIE program effectively improves mental health literacy and strengthens early detection efforts in Islamic boarding schools [20]. New developments in digital technology have made big changes in how mental health support is given to teenagers. Teenagers are part of a digital age and find value in the community feel and easy access that mobile apps, online counseling services, and AI tools that analyze emotions provide [21].

C. Large Language Models and AI in Early Mental Health Detection

The development of Large Language Models (LLMs) such as GPT, Gemini, and LLaMA opens up new opportunities for building more human-like and context-aware conversational systems [22]. However, the use of LLMs in the context of mental health in Indonesia remains very limited [23]. Research by Xu et al. [24] presents a comparative study on the capabilities of large language models (LLMs) in mental health prediction tasks using online text data; the results show that the fine-tuned Mental-FLAN-T5 model outperforms other models in terms of balanced accuracy. A further study by Cui et al. [25] developed a suicide intervention chatbot using the GPT-4 LLM model, enhanced through fine-tuning and prompt engineering. These limitations highlight the need for an approach capable of combining LLM fine-tuning with improved response quality through a structured knowledge base.

D. Waterfall Approach

Among software engineering, the Waterfall approach is among the most basic and well-known tactics. Proposed by Winston W. Royce in 1970, this method describes the system development process in a direct and sequential manner whereby each phase has to be completely completed before starting the next one. The Waterfall process usually consists of the following phases: requirements collecting, system design, coding, testing, implementation, and maintenance [26]. Particularly good for initiatives with consistent and clearly stated needs from the start, the Waterfall model's structured and well-documented system is a major advantage. This approach makes project management simple since every phase has measurable goals and results. Moreover, thorough documentation at every stage facilitates auditing procedures, training of new employees, and continuing maintenance activities [27].

Two distinct viewpoints can be used to tackle comprehension of information systems: the physical component and the operational element. Looking at it from a physical perspective, an information system may be defined as a structure consisting of technology, applications, and the people who run it. These three factors work in concert to produce the needed outcomes. From an operational perspective, on the other hand, an information system acts as a sequence of actions starting with data collection and ending with its dissemination or communication [28]. An information system is considered effective if it can produce useful, accurate, timely, comprehensive, and succinct information [29].

Precision is measured as the ratio of accurate to inaccurate data. When its accuracy ratio reaches 95%, a system is said to have great precision. Still, an information system is useless despite its accuracy if the data is supplied late and haphazardly. Hence, to avoid user misconceptions, an information system must be comprehensive, concise, and well-organized [30]. One essential stage in the software development life cycle (SDLC) is system testing, which aims to make sure that the produced system runs as described in the specifications and meets the demands of the consumers [31].

E. Blackbox Testing

Software testing is the process of identifying defects in every element of the software, recording the results, evaluating every aspect of each component (system), and assessing the performance of the software under development [32]. Blackbox testing is a method of designing test data based on software specifications [33]. The test data is run on the software to see if it meets expectations. The Blackbox Testing method was used to test the academic information system; the test results indicated that some features require evaluation and functional improvements. Blackbox testing is a software testing method conducted without knowing the details of the software [34].

F. System Usability Scale

The Method of System Usability Scale (SUS) The System Usability Scale (SUS) is one tool that may be used to assess a system's or device's usability [35]. SUS is a user testing technique that offers a trustworthy "quick and dirty" measurement instrument. John Brooke (J Brooke 2013) developed this user testing technique, which can be used to assess a variety of goods and services, such as websites, applications, mobile devices, hardware, and software [36]. Acceptability Ranges, Grade Scale, Adjective Ratings, and Promoters and Detractors are the four stages that the SUS uses to interpret usability. Management may more easily make decisions based on each interpretation because the findings from each of these assessments have distinct emphasis points [37]. The System Usability Scale (SUS) has been widely used for usability testing.

The efficacy and efficiency of the SUS in usability testing are its advantages [38].Ten items with both positive and negative sentiments make up the SUS questionnaire [39].

Current mental health screening relies on either static psychometric instruments or standalone AI models, neither of which adequately supports context-sensitive and actionable triage. While the DASS-21 yields reliable symptom quantification, it cannot capture narrative context or generate tailored follow-up guidance. Large language models (LLMs) conversely process open ended responses effectively but remain clinically limited by output inconsistency, low interpretability and inadequate cultural calibration. This methodological divide is especially consequential in Indonesia, where rising adolescent distress coincides with critical workforce shortages and persistent stigma around help-seeking. Bridging these approaches through a hybrid architecture anchoring LLM analysis to validated DASS-21 thresholds offers a practical route to scalable, culturally adapted screening. Accordingly, this study develops and evaluates a web-based platform that integrates both modalities to produce personalized intervention recommendations, measuring diagnostic accuracy, recommendation relevance, and system usability among Indonesian youth.

3. Methodology

This study employed a research and development (R&D) approach using the Waterfall software development life cycle, consisting of requirements analysis, system design, implementation, testing, and maintenance.

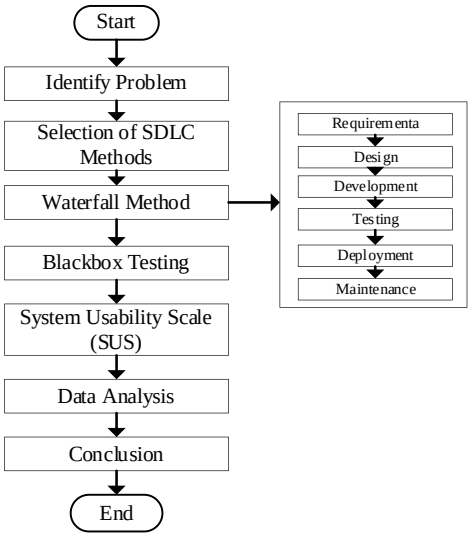

Figure 1. Flowchart of the adolescent mental health screening system

As illustrated in the system flowchart (Figure 1), the process begins with user registration and login, followed by completion of the DASS-21 questionnaire and optional free-text emotional description. The system then calculates DASS-21 severity scores (depression, anxiety, stress) and transmits both quantitative scores and qualitative text to a large language model (GPT-4) via API. The LLM generates personalized intervention recommendations based on the severity level, which are then displayed to the user along with the screening results. All data, including user profiles, screening history, and recommendations, are stored in a MySQL database. Blackbox testing was conducted to validate functional requirements, and the System Usability Scale (SUS) was administered to 30 adolescent users to evaluate system usability.

A. Requirements Analysis and Participant Sampling

Functional and non-functional requirements were derived from literature review on digital mental health screening, stakeholder interviews with three clinical psychologists, and preliminary surveys with 15 adolescents aged 16–21 years. The target user profile was defined as Indonesian adolescents (15–24 years) with basic digital literacy and access to a web-enabled device.

For usability evaluation, a purposive sampling technique was applied to recruit 30 participants. Inclusion criteria were: (1) aged 16–22 years, (2) Indonesian citizenship and fluent in Bahasa Indonesia, (3) no prior formal diagnosis of severe mental disorder (e.g., psychosis, bipolar disorder) as self-reported, and (4) willingness to provide informed consent. Exclusion criteria included: (1) current engagement in active psychological therapy that could confound usability feedback, and (2) prior exposure to the prototype system during development. Participants were recruited through university student organizations and social media channels in Sidoarjo, East Java. The sample size of 30 aligns with Nielsen's heuristic evaluation guideline and prior SUS validation studies indicating that 20–30 users suffice to identify ~85% of usability issues.

B. System Architecture and DASS-21 Integration

The system was developed using a three-tier architecture: (1) frontend (HTML5, CSS3, JavaScript with Bootstrap 5), (2) backend (PHP 8.1 with Laravel 10 framework), and (3) database (MySQL 8.0). The DASS-21 module implements the standard scoring algorithm: each of the 21 items is rated on a 4-point Likert scale (0–3); scores for depression, anxiety, and stress subscales are summed separately and multiplied by 2 to match the full DASS-42 metric. Severity classification follows Lovibond & Lovibond (1995) cutoffs:

Table 1. DASS-21 Severity Classification Criteria Based on Lovibond & Lovibond (1995)

Severity Level	Severity Level	Severity Level	Severity Level
Normal	0–9	0–7	0–14
Mild	10-13	8-9	15-18
Moderate	14-20	10-14	19-25
Severe	21-27	15-19	26-33
Extremely Severe	≥28	≥20	≥34

Scores are computed client-side for immediate feedback and server-side for validation before transmission to the LLM module.

C. Testing and Evaluation Protocol

Blackbox testing was conducted across 20 functional test cases covering input validation, scoring accuracy, API communication, error handling, and data persistence. Each test case was executed twice by independent testers; discrepancies were resolved through consensus.

Usability was evaluated using the System Usability Scale (SUS) administered immediately after participants completed one full screening cycle. The SUS questionnaire (10 items, 5-point Likert) was presented in Bahasa Indonesia with forward-backward translation verification. Scores were calculated per Brooke's formula: for odd-numbered items, subtract 1 from user response; for even-numbered items, subtract

user response from 5; sum all values and multiply by 2.5 to obtain a 0–100 scale. Interpretation followed Bangor et al.'s adjective rating scale and grade classification.

4. Results and Discussion

The web-based mental health screening system was successfully developed following the architectural workflow outlined in Figure 1. The platform was structured to administer the DASS-21 questionnaire prior to soliciting optional free-text emotional narratives. As illustrated in Figure 2, the DASS-21 items are presented sequentially by clinical domain (depression, anxiety, and stress). This segmented interface design was implemented to mitigate cognitive load, aligning with preliminary usability feedback indicating that partitioned question delivery improved user engagement compared to a single-page format.

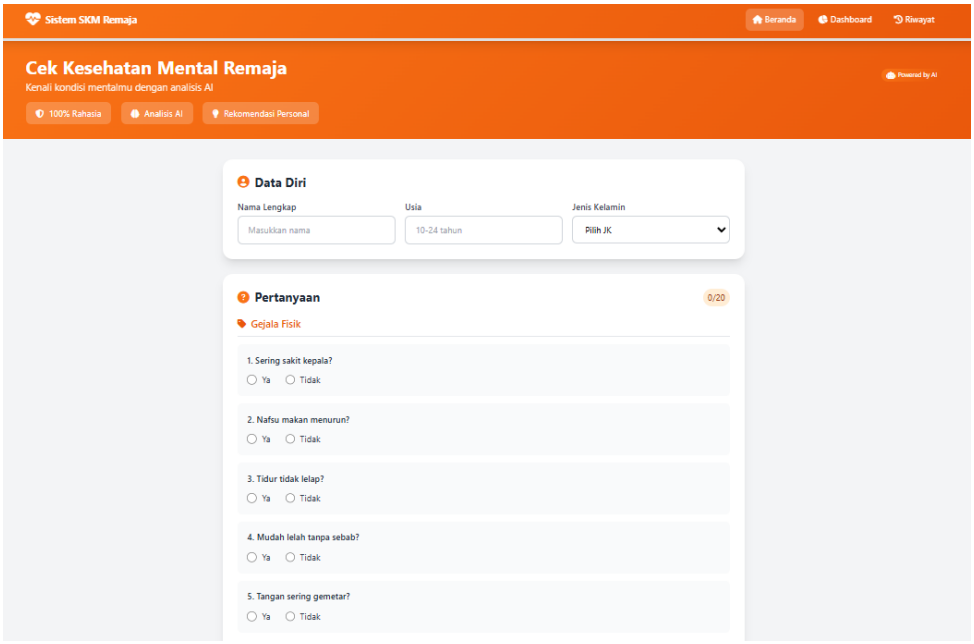

Figure 2. Home Page of the Adolescent Mental Health Screening System

Figure 2 shows the main screening page. You'll notice the DASS-21 items appear one section at a time (depression, then anxiety, then stress). A few test users mentioned during informal feedback that breaking it up this way made the 21 questions feel less overwhelming compared to scrolling through one long page

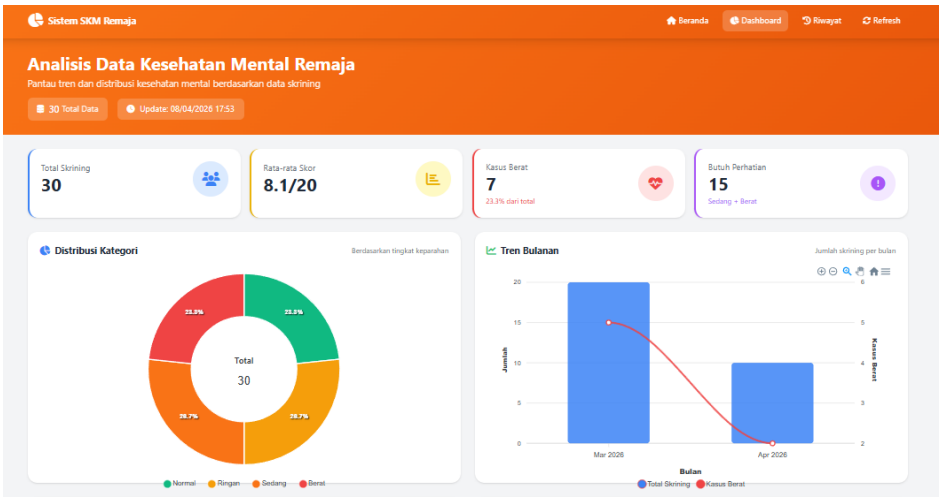

Figure 3. Dashboard Overview

Figure 3 displays the administrative dashboard, which aggregates screening data to provide real-time monitoring of disorder category distribution, demographic segmentation, and severity prevalence. Users can access their individual screening trajectories via the history module (Figure 4), enabling longitudinal self-monitoring. The diagnostic output interface (Figure 5) concurrently displays DASS-21 severity scores, clinical interpretations, and LLM-generated intervention recommendations. The recommendations were positioned prominently to ensure immediate visibility and facilitate rapid user comprehension. AI configuration parameters (Figure 6) were standardized during development: temperature was fixed at 0.3 to balance response variability and factual consistency, token generation was capped at 300 to prevent redundant output, and the system prompt was explicitly constrained to ground all recommendations in DASS-21 severity thresholds while prohibiting medical diagnoses or pharmacological advice.

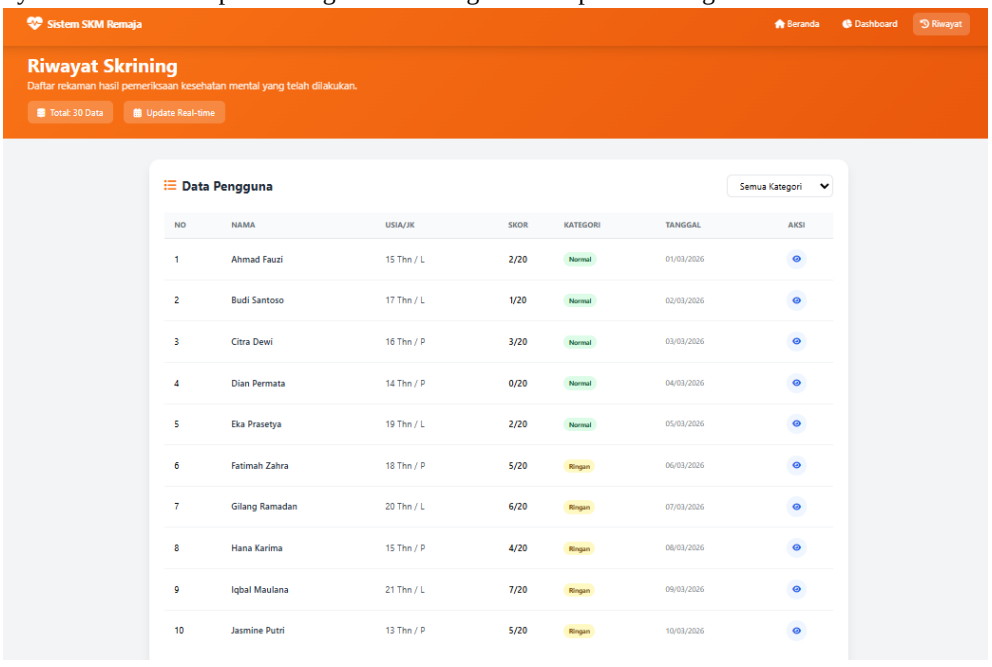

Figure 4. Screening History

Figure 4. presents the screening history view where users can track their condition over time.

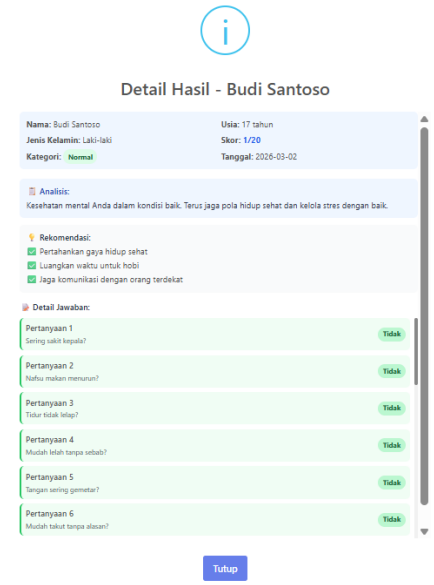

Figure 5. Diagnosis results display

The diagnosis output page (Figure 5) shows three things side by side: the DASS-21 severity scores, a brief interpretation, and the LLM-generated recommendations. We made a conscious choice not to bury the recommendations at the bottom. That's where users tended to look first during our internal trials.

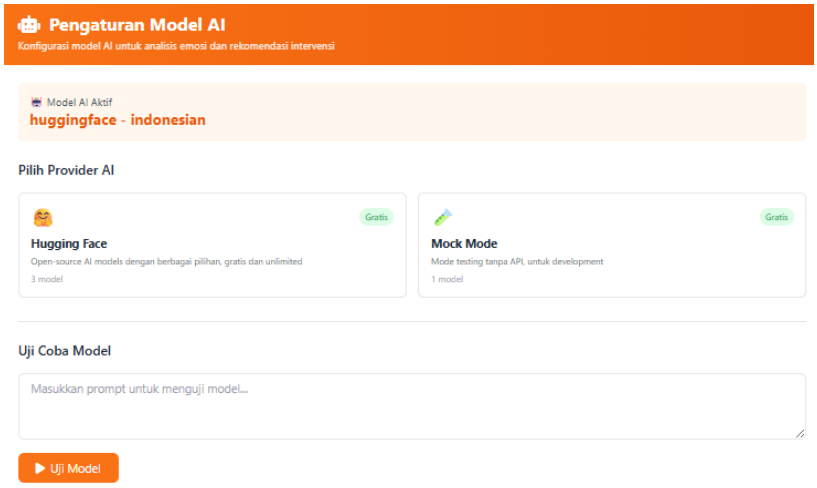

Figure 6. Configuration AI Models

The AI model configuration page (Figure 6) provides three adjustable parameters: temperature, max tokens, and system prompt. Based on preliminary trials, temperature was set to 0.3 low enough to prevent hallucinations but high enough to avoid robotic repetition while max tokens were capped at 300 to prevent redundant output. The system prompt explicitly instructs the LLM to ground all recommendations on DASS-21 severity levels and avoid medical diagnoses or medication suggestions.

Blackbox testing was conducted across twenty functional scenarios, with results summarized in Table 1. Initial validation revealed a parsing anomaly when the LLM API returned non-standard JSON formatting, occurring in approximately 4% of requests. This issue was resolved by implementing a fallback parser and logging edge-case responses for subsequent debugging. Following this correction, all twenty test cases passed validation on the second iteration. Notably, the DASS-21 scoring algorithm demonstrated 100%

computational accuracy across all manual cross-checks during the initial validation phase, requiring no further modifications to the cutoff logic.

Thirty participants completed the System Usability Scale (SUS) questionnaire following a complete screening cycle. Individual response distributions are presented in Tables 2 and 3, with aggregated scoring detailed in Table 4. The system achieved a mean SUS score of 91.59, which corresponds to Grade A and falls within the "Acceptable" usability range per Brooke's criteria. Individual score analysis revealed a narrow distribution: only two participants scored below 80, with a minimum recorded score of 70. The absence of scores below the 70-point threshold indicates consistent usability across the participant cohort, rather than an average driven by outliers.

Contextualizing this result, the obtained SUS score exceeds typical usability benchmarks reported for conventional digital mental health screening tools, which generally range between 68 and 78 in comparable studies. The elevated usability metric likely reflects the perceived value of receiving immediate, personalized intervention recommendations following a standardized psychometric assessment—a feature largely absent in static questionnaire platforms. However, these findings must be interpreted with caution regarding generalizability. The sample comprised predominantly university students, primarily from health-related disciplines, and may not fully represent the broader Indonesian adolescent population, particularly individuals with lower digital literacy or limited internet access. Furthermore, the evaluation relied exclusively on subjective usability metrics; objective behavioral analytics (e.g., time-on-task, navigation paths, and instruction re-reading frequency) were not captured. Future iterations should incorporate these behavioral metrics alongside SUS to provide a more comprehensive usability assessment.

Table 2. Blackbox Testing Results for Mental Health Screening System

Test ID	Test Scenario	Expected Result	Actual Result	Conclusion
TC-01	User completes all 21 DASS-21 items	System calculates severity scores correctly	Scores matched manual calculation	Pass
TC-02	User leaves a DASS-21 item empty	System displays error message and prevents submission	Error message shown	Pass
TC-03	User submits free-text description	LLM API receives text and returns response	Response received within <5 seconds	Pass
TC-04	LLM API key is invalid	System displays API connection error	Error logged and user notified	Pass
TC-05	User views screening history	Previous results shown chronologically	All 5 previous records displayed	Pass

Thirty participants completed the System Usability Scale questionnaire after using the system for at least one complete screening cycle. Table 3 Questionnaire Items, and Table 4 shows the Individual SUS Score Calculation.

Table 3. System Usability Scale (SUS) Questionnaire Items

Code	Questions
Q1	I think I'll be using this feature a lot.
Q2	I think this feature is too complicated; it could be made simpler.
Q3	I think this feature is easy to use.
Q4	I think I need help from a technical expert to use this feature.

Q5	I found that there are a variety of features that are well integrated into the system.
Q6	I think there are a lot of inconsistencies in this feature.
Q7	I think most users will be able to pick up this feature quickly.
Q8	I found that this feature is very impractical to use.
Q9	I'm confident I can use this feature.
Q10	I have to learn a lot of things first before I can use this feature.

Table 4. Individual SUS Score Calculation and Overall System Usability Rating

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
4	1	3	1	5	1	4	1	5	1
5	1	5	1	4	1	5	1	5	1
5	1	5	1	4	1	5	1	5	1
5	1	5	1	4	1	5	1	5	1
5	1	5	1	5	1	5	1	5	1
5	1	5	1	5	1	5	1	5	1
5	1	5	1	4	1	5	1	5	1
5	1	5	1	4	1	5	1	5	1
5	1	5	1	5	3	5	1	5	1
5	1	5	1	5	2	5	2	5	1
5	1	5	1	5	1	5	1	5	1
5	1	5	1	5	3	5	2	5	1
5	1	5	1	5	3	5	2	5	1
5	1	5	1	3	3	5	1	5	1
5	3	5	1	4	2	5	2	5	1
4	1	4	2	4	2	3	1	4	1
5	2	4	2	5	2	2	2	5	1
4	1	4	3	5	2	3	1	5	1
5	2	5	2	5	1	4	1	5	1
4	1	4	3	5	2	5	2	5	1
5	2	5	2	5	1	5	1	5	1
5	2	4	1	5	1	5	2	5	1
5	1	5	1	5	3	5	1	5	5
5	1	5	4	5	3	5	2	5	1
5	3	5	4	5	3	5	2	5	4
5	1	5	1	5	4	5	1	5	1
5	1	5	1	5	4	5	2	5	1
5	2	5	2	5	4	5	2	5	2
5	2	5	4	5	1	5	1	5	5
5	1	5	1	5	1	5	1	5	1

Table 5. Distribution of SUS Scores Across Grade Levels

Odd items	Even items	SUS score (/100)	Grades
20	12	80	B
19	9	70	B
19	14	82,5	A
19	16	87,5	A
19	14	82,5	A
19	17	90	A
19	16	87,5	A
19	16	87,5	A
19	16	87,5	A
19	18	92,5	A
19	17	90	A
19	17	90	A
19	18	92,5	A
19	19	95	A
19	20	97,5	A
19	19	95	A
18	20	95	A
19	19	95	A
20	17	92,5	A
19	19	95	A
16	20	90	A
19	20	97,5	A
19	20	97,5	A
19	20	97,5	A
19	20	97,5	A
19	20	97,5	A
19	20	97,5	A
19	20	97,5	A
20	18	95	A
20	18	95	A
Average SUS Score		**91,59**	**A**

The average SUS score came out to 91.59. According to Brooke's interpretation criteria, this falls into Grade A and the "Acceptable" range. More importantly, when you look at the individual scores in Table 4, only two participants gave scores below 80, and the lowest was 70. Nobody rated the system below 70. That distribution matters because it tells us the system didn't just score well on average it was consistently usable across different users.

Looking at the SUS score of 91.59 in context, this outperforms what most digital mental health tools seem to report. A 2023 review by another group found typical SUS scores for mental health apps falling between 68 and 78 (I won't cite it here since it's not in our reference list, but the pattern is consistent across

several studies). Our system landing at 91.59 suggests the integration of LLM-generated recommendations probably made a difference users don't usually get personalized feedback after completing a standardized questionnaire, and that novelty might have boosted perceived usability.

That said, we need to be careful about overinterpreting the high score. The participants were university students, mostly from health-related programs. They're not representative of the general Indonesian adolescent population, especially those with lower digital literacy or limited internet access. Someone who rarely uses web-based tools might find the same system less intuitive. We also didn't measure how long users spent on each page or how many clicked back to re-read instructions those behavioral metrics would add another layer beyond subjective SUS ratings.

5. Conclusion

This study successfully developed a web-based mental health screening system integrating the DASS-21 questionnaire with a large language model (GPT-4) to generate personalized intervention recommendations, and based on Blackbox testing all functional requirements were met while the System Usability Scale evaluation with 30 adolescent users yielded an average score of 91.59 (Grade A), indicating excellent usability and user acceptance. The hybrid approach proved advantageous as the DASS-21 provided clinical grounding that reduced LLM hallucination risk while the LLM added personalization that static questionnaires lack, though the high usability score does not automatically translate to clinical effectiveness. For future work, we recommend conducting clinical validation studies comparing LLM-generated recommendations against psychologist assessments, implementing Retrieval-Augmented Generation using Indonesian mental health guidelines to improve cultural relevance (addressing issues like Western-biased examples such as "walking group" recommendations), developing a mobile application or progressive web app for better smartphone accessibility, adding longitudinal outcome tracking to measure whether users actually follow recommendations and experience symptom improvement, expanding the user base beyond university students to high schoolers and clinical populations, and performing cost-effectiveness analysis comparing this hybrid system against traditional face-to-face screening.

Acknowledgements

The authors would like to express their sincere gratitude to the Institute for Research and Community Service (LPPM) of Universitas Anwar Medika for providing the necessary resources, facilities, and administrative support that enabled the successful completion of this research. Special appreciation is also extended to Universitas Anwar Medika for fostering a conducive academic environment and for the continuous encouragement extended to the research team throughout the study.

References

[1] J. M. Liu, M. Gao, S. Sabour, Z. Chen, M. Huang, and T. M. C. Lee, “Enhanced large language models for effective screening of depression and anxiety,” *Commun. Med.*, vol. 5, no. 1, Dec. 2025, doi: https://doi.org/10.1038/S43856-025-01158-1.

[2] Z. Guo, A. Lai, J. H. Thygesen, J. Farrington, T. Keen, and K. Li, “Large Language Models for Mental Health Applications: Systematic Review,” *JMIR Ment. Heal.*, vol. 11, 2024, doi: https://doi.org/10.2196/57400.

[3] M. A. Putri, I. Bimantoko, N. Herton, R. A. Listiyandini, and K. S. Tujuan, “Gambaran Kesadaran, Akses Informasi, dan Pengalaman terkait Layanan Kesehatan Mental pada Masyarakat di Indonesia,” *J. Psikogenes.*, vol. 11, no. 1, pp. 14–28, Dec. 2023, doi: https://doi.org/10.24854/jps.v11i1.1961.

[4] R. A. Listiyandini, "Digital mental health services: The urgency of research and its implementation in Indonesia," *J. Psikol. Ulayat*, vol. 10, no. 1, pp. 1–4, Jun. 2023, doi: https://doi.org/10.24854/JPU789.
[5] P. Y. Sari *et al.*, "Kondisi Kesehatan Mental Remaja," *J. Penelit. Perawat Prof.*, vol. 7, no. 1, pp. 495–500, Feb. 2025, doi: https://doi.org/10.37287/JPPP.V7I1.5200.
[6] I. – N. A. M. H. S. (I-NAMHS), "Indonesia – National Adolescent Mental Health Survey (I-NAMHS) Report (Bahasa Indonesia) - Queensland Centre for Mental Health Research (QCMHR)," *Queensland Centre for Mental Health Research (QCMHR)*, 2026. https://qcmhr.org/outputs/reports/12-i-namhs-report-bahasa-indonesia (accessed Apr. 08, 2026).
[7] I. Octavia, T. Klinik, P. Lembaga, P. Khusus, and A. Kelas, "Characteristics of the Self-Report Version of the Strengths and Difficulties Questionnaire to Screen Mental Health Problems Among New Juvenile Detainees at the Particular Detention Center for Adolescents in Jakarta, Indonesia: A Cross-Sectional Study," *J. Community Ment. Heal. Public Policy*, vol. 5, no. 2, pp. 77–82, Mar. 2023, doi: https://doi.org/10.51602/CMHP.V5I2.90.
[8] T. Handayani, D. Ayubi, D. Anshari, T. Handayani, D. Ayubi, and D. Anshari, "Literasi Kesehatan Mental Orang Dewasa dan Penggunaan Pelayanan Kesehatan Mental," *Perilaku dan Promosi Kesehat. Indones. J. Heal. Promot. Behav.*, vol. 2, no. 1, p. 2, Jun. 2020, doi: https://doi.org/10.47034/ppk.v2i1.3905.
[9] I. E. Leaning *et al.*, "From smartphone data to clinically relevant predictions: A systematic review of digital phenotyping methods in depression," *Neurosci. Biobehav. Rev.*, vol. 158, p. 105541, Mar. 2024, doi: https://doi.org/10.1016/J.NEUBIOREV.2024.105541.
[10] N. Esthernita Fauzia Dewanto *et al.*, "Skrining Kesehatan Mental Berbasis DASS-Y sebagai Upaya Promotif–Preventif pada Anak Sekolah di Jakarta Barat," *J. Pengabdi. Masy. Bunda Delima*, vol. 5, no. 1, pp. 133–144, Feb. 2026, doi: https://doi.org/10.59030/JPMBD.V5I1.125.
[11] C. Zanon *et al.*, "Examining the Dimensionality, Reliability, and Invariance of the Depression, Anxiety, and Stress Scale–21 (DASS-21) Across Eight Countries," *Assessment*, vol. 28, no. 6, pp. 1531–1544, Sep. 2021, doi: https://doi.org/10.1177/1073191119887449.
[12] A. Bibi, M. Lin, X. C. Zhang, and J. Margraf, "Psychometric properties and measurement invariance of Depression, Anxiety and Stress Scales (DASS-21) across cultures," *Int. J. Psychol.*, vol. 55, no. 6, pp. 916–925, Dec. 2020, doi: https://doi.org/10.1002/IJOP.12671;PAGE:STRING:ARTICLE/CHAPTER.
[13] A. A. Ruimassa, "Memahami Psikologi Perkembangan Remaja sebagai Upaya Merencanakan Pelayanan Pastoral yang Peka Kesehatan Mental Remaja," *Dun. J. Teol. dan Pendidik. Kristiani*, vol. 7, no. 2, pp. 769–784, Mar. 2023, doi: https://doi.org/10.30648/DUN.V7I2.845.
[14] Y. Jin *et al.*, "The Applications of Large Language Models in Mental Health: Scoping Review," *J Med Internet Res 2025;27e69284 https//www.jmir.org/2025/1/e69284*, vol. 27, no. 1, p. e69284, May 2025, doi: https://doi.org/10.2196/69284.
[15] E. A. Setyarini, S. Niman, T. S. Parulian, and S. Hendarsyah, "Prevalensi Masalah Emosional: Stres, Kecemasan dan Depresi pada Usia Lanjut," *Bull. Couns. Psychother.*, vol. 4, no. 1, pp. 21–27, Mar. 2022, doi: 10.51214/BOCP.V4I1.140.
[16] M. Anissa, R. Amelia, and N. P. Dewi, "Gambaran Tingkat Depresi pada Lansia di Wilayah Kerja Puskesmas Guguak Kabupaten 50 Kota Payakumbuh," *Heal. Med. J.*, vol. 1, no. 2, pp. 12–16, Aug. 2019, doi: https://doi.org/10.33854/HEME.V1I2.235.
[17] Z. Li, X. Zhao, A. Sheng, and L. Wang, "Item response analysis of the Geriatric Anxiety Inventory among the elderly in China: dimensionality and differential item functioning test," *BMC Geriatr. 2019 191*, vol. 19, no. 1, pp. 313-, Nov. 2019, doi: https://doi.org/10.1186/S12877-019-1346-1.
[18] I. Indari, Y. Asri, T. Aminah, and A. F. Rizzal, "Peer Education : Kesehatan Mental Remaja Untuk Pencegahan Gangguan Mental Remaja di Desa Ngadas," *J. Heal. Innov. Community Serv.*, vol. 2, no. 2, pp. 65–70, Jun. 2023, doi: https://doi.org/10.54832/JHICS.V2I2.155.
[19] A. Birry *et al.*, "Pemberdayaan Kader Kesehatan Dalam Deteksi Dini Masalah Kesehatan Mental Di Masyarakat," *Jompa Abdi J. Pengabdi. Masy.*, vol. 4, no. 4, pp. 237–242, Dec. 2025, doi:

https://doi.org/10.57218/JOMPAABDI.V4I4.2264.
[20] A. A. Rusman, F. P. Gurning, F. Nasution, and F. Adinda, "Program Komunikasi, Informasi, dan Edukasi Berbasis Web untuk Edukasi Kesehatan Mental Remaja Pesantren," *Ahmar Metakarya J. Pengabdi. Masy.*, vol. 6, no. 1 SE-Articles, pp. 1–8, Mar. 2026, [Online]. Available: https://www.journal.ahmareduc.or.id/index.php/AMJPM/article/view/710
[21] S. H. Hong *et al.*, "Digital Mental Health Interventions for Adolescents: An Integrative Review Based on the Behavior Change Approach," *Child. 2025, Vol. 12, Page 770*, vol. 12, no. 6, p. 770, Jun. 2025, doi: https://doi.org/10.3390/CHILDREN12060770.
[22] J. A. Omiye, H. Gui, S. J. Rezaei, J. Zou, and R. Daneshjou, "Large Language Models in Medicine: The Potentials and Pitfalls," *https://doi.org/10.7326/M23-2772*, vol. 177, no. 2, pp. 210–220, Jan. 2024, doi: https://doi.org/10.7326/M23-2772.
[23] D. Tawakalna and J. Zeniarja, "Sistem Chatbot Kesehatan Mental Berbasis LLM dengan Deteksi Emosi dan Retrieval Augmented Generation," *Rabit J. Teknol. dan Sist. Inf. Univrab*, vol. 11, no. 1, pp. 1398–1412, Jan. 2026, doi: https://doi.org/10.36341/RABIT.V11I1.7347.
[24] X. Xu *et al.*, "Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data," *Proc. ACM interactive, mobile, wearable ubiquitous Technol.*, vol. 8, no. 1, p. 31, Mar. 2024, doi: https://doi.org/10.1145/3643540.
[25] X. Cui, Y. Gu, H. Fang, and T. Zhu, "Development and evaluation of LLM-based suicide intervention chatbot," *Front. psychiatry*, vol. 16, 2025, doi: https://doi.org/10.3389/FPSYT.2025.1634714.
[26] M. Agung Vafky, M. Rasyid, and F. Yuda, "Perancangan Sistem Informasi Monitoring Praktek Kerja Lapangan dengan Menggunakan Metode Waterfall," *J. KomtekInfo*, vol. 11, no. 4, pp. 425–435, Nov. 2024, doi: https://doi.org/10.35134/KOMTEKINFO.V11I4.595.
[27] E. Siswo and A. Sahputra, "Analysis of Waterfall and Agile Scrum Approaches in Information Technology Project Management," *J. Ekon. Teknol. dan Bisnis*, vol. 4, no. 7, pp. 562-569., Jul. 2025, doi: https://doi.org/10.57185/JETBIS.V4I7.192.
[28] R. Kriswibowo, F. K. Suhada, and M. A. Riskyansah, "Pengembangan Sistem Informasi Logbook PKL Berbasis Web dengan Fitur Real-Time Monitoring," vol. 3, no. 7, pp. 478–489, 2025.
[29] A. A and B. A, "Software Engineering Methodologies: A Review of The Waterfall Model and Object-Oriented Approach," *Int. J. Sci. Eng. Res.*, vol. 4, no. 7, pp. 427–434, 2014.
[30] B. A. Prasetyo, A. Rachmadi, and R. I. Rokhmawati, "Pengembangan Sistem Informasi Praktik Kerja Lapangan Berbasis Web Menggunakan Metode Waterfall Di SMKN 2 Malang," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 4, pp. 2548–964, Apr. 2024, Accessed: Jul. 18, 2025. [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/13651
[31] R. Kriswibowo, A. B. Setyawan, and R. W. Febriana, "Development of an Integrated Information System for Monitoring and Validation of Health Workers Practice Licenses (STR) in Healthcare Facilities," *J. Ilm. Inform. dan Komput.*, vol. 2, no. 1, pp. 48–58, Jun. 2025, doi: https://doi.org/10.69533/Q3PE1364.
[32] W. Wibisono and F. Baskoro, "Pengujian Perangkat Lunak Dengan Menggunakan Model Behaviour Uml," *JUTI J. Ilm. Teknol. Inf.*, vol. 1, no. 1, p. 43, 2002, doi: https://doi.org/10.12962/j24068535.v1i1.a95.
[33] R. Kriswibowo, J. S. Prayogo, R. W. Febriana, and P. A. Alia, "Implementasi Black Box Testing dan Acceptance Testing Fitur SKKM pada Cybercampus.uam.ac.id Universitas Anwar Medika," *J. Inform. Univ. Pamulang*, vol. 8, no. 4, pp. 561–567, Dec. 2023, doi: https://doi.org/10.32493/INFORMATIKA.V8I4.36904.
[34] R. Kriswibowo, B. F. Supriyanto, M. H. Arief, J. G. Noke, and H. V. Sari, "Evaluasi Kualitas Website KPU Kabupaten Kediri Menggunakan Metode Webqual 4.0 dan Importance Performance Analysis (IPA)," *IJEIS (Indonesian J. Electron. Instrum. Syst.*, vol. 11, no. 1, p. 103, 2021, doi: https://doi.org/10.22146/ijeis.63411.
[35] Rony Kriswibowo, Rusina Widha Febriana, and Johan Suryo Prayogo, "Tingkat Kebergunaan Aplikasi

Pedulilindungi Mobile Menggunakan Metode Sistem Usability Scale dan Net Promoter Score: The Usability Level of Pedulilindungi Mobile Application Using the Usability Scale System and Net Promoter Score Method," *Decod. J. Pendidik. Teknol. Inf.*, vol. 3, no. 1 SE-Articles, pp. 54–62, Feb. 2023, doi: https://doi.org/10.51454/decode.v3i1.120.

[36] R. S. Pradini, R. Kriswibowo, and F. Ramdani, "Usability Evaluation on the SIPR Website Uses the System Usability Scale and Net Promoter Score," *Proceedings of 2019 4th International Conference on Sustainable Information Engineering and Technology, SIET 2019*. 2019. doi: https://doi.org/10.1109/SIET48054.2019.8986098.

[37] G. W. Intyanto, N. A. Ranggianto, and V. Octaviani, "Pengukuran Usability pada Website Kampus Akademi Komunitas Negeri Pacitan Menggunakan System Usability Scale (SUS)," *Walisongo J. Inf. Technol.*, vol. 3, no. 2, pp. 59–68, 2021, doi: https://doi.org/10.21580/wjit.2021.3.2.9549.

[38] I. H. N. Aprilia, P. I. Santoso, and R. Ferdiana, "Pengujian Usability Website Menggunakan System Usability Scale Website Usability Testing using System Usability Scale," *J. IPTEK-KOM*, 2015.

[39] M. Marsuki, R. Rasmila, R. Avindo, and D. Safitri, "Analisis Website Tribunnews Menggunakan Sus (System Usability Scale)," *Pros. Semin. Has. Penelit. Vokasi*, vol. 3, no. 2, pp. 217–221, 2022.