From User Reviews to Digital Innovation: Harnessing Feedback for Advancing Health Applications

Hissah S. Al-Atwi 1 , Salmah M. Al-Qarni 2 , Wejdan B. Al-Marwani 3 , Areej M. Al-Wabisi 4 , G a d a a S. Al-Atwi 5 , Omar . Asiri 6 , Fathi A. M u b a r a k i 7[a,1]

Abstract. This study investigates the integration of Artificial Intelligence (AI) techniques in analyzing user feedback on mobile health (mHealth) applications. By conducting a systematic literature review of 23 selected studies, it explores the effectiveness of sentiment analysis, machine learning models, and UX evaluation in improving mHealth usability and reliability. The findings indicate that larger datasets and balanced feature sets significantly enhance model performance. Moreover, the study highlights the need for culturally aware AI design, standard evaluation metrics, and hybrid models to improve digital health outcomes. Visual and comparative analyses support the proposed research directions and frameworks.

Keywords. Digital health applications, user experience, machine learning, thematic analysis, healthcare technology

1. Introduction

Integration of technology into healthcare has changed how medical services are delivered and accessed. The use of mHealth apps is perhaps the most striking example, since it allows users to manage chronic diseases, track medication intake, monitor fitness, arrange consultations, and even obtain telehealth services. Considering the global demand for healthcare, their role in improving access and relieving traditional health systems remains particularly noteworthy. As much as the technical design of mHealth apps is important, their real-life effectiveness relies on an intricate balance between the design and the users' changing needs across contexts. Furthermore, app success has now been defined in terms of usability, reliability, security, scalability, and meeting user expectations (Aljohani, 2025; Boudreaux et al., 2014). In addition, the concept of user experience (UX) extends beyond satisfaction into trust, participation, and the importance of the context. Such parameters deeply influence the ways people engage with andassess health tools provided digitally (Gerges & Elgalb, 2024). Evaluative data gleaned from user reviews and ratings systems is one of the richest, yet underexploited, resources for evaluative data. Reviews and ratings are reflective of satisfaction, technical issues, expectations, or even user frustrations. At the same time, reviews are problematic in terms of scale and structure since feedback is emotionally charged and often unbound, contradictory, or inconsistent (Saoane Thach, 2019). To provide solutions to these

[1] Corresponding Author: Author Name, Contact details.

complexities, researchers are increasingly adopting Artificial Intelligence (AI) techniques including Machine Learning (ML) and Natural Language Processing (NLP). These methods provide unparalleled access to sentiment analysis, issue classification related to the application, and analysis regarding specific features (Aljohani, 2025). Their algorithms are capable of classifying reviews into categories such as bugs, performance, usability, and security, thereby enabling continuous improvement in application development (Al Kilani et al., 2019). The primary issue that remains is the integration of sociocultural and linguistic parameters into the AI models across users' demographics. Differences in language, speech patterns, and sentiment expressions can result in over-misclassification or non-detection in pattern recognition and classification, indicating the need for ethnocentric AI design (Aljohani, 2025; Khan & Alotaibi, 2020).Furthermore, the absence of common criteria for assessing health applications leads to discrepancies in the evaluation of their usability and overall quality. Instruments like the Mobile App Rating Scale (MARS) and ISO/IEC 25010 provide frameworks for evaluating enrichment techniques, but their use is not widespread (Wang et al., 2019). From a policy perspective, the combination of user- generated review analytics into formal evaluation systems could enhance evaluation processes and assist healthcare institutions in identifying and recommending reliable applications. Equally, analysis of user feedback can be applied to the design of persuasive technologies that increase user participation in and adherence to health-related practices (Al-Qahtani et al., 2024). The main objectives of this study are to:

(1) Identify and explore modern techniques used in analyzing user reviews of mobile health (mHealth) applications.
(2) Review and analyze user perspectives related to trust, privacy, usability, and overall effectiveness of mHealth apps (Amjad et al., 2023; Bonny et al., 2022).
(3) Map user feedback to existing app design principles and evaluation frameworks.
(4) Identify sociocultural and linguistic factors that affect the accuracy of AI-driven review analysis.
(5) Propose practical design recommendations for cul- turally responsive, user-centered, and regulation- compliant mHealth applications (Buetow & Lovatt, 2024).

2. LITERATURE REVIEW

2.1. Introduction

First The proliferation of mobile health (mHealth) applications has redefined the delivery of healthcare services across demographics and geographic contexts. These applications offer patients access to services ranging from chronic condition management and medication tracking to telehealth consultations and fitness monitoring (Wang et al., 2019). However, the real-world effectiveness of these tools depends not solely on their technical robustness but on how they are experienced by users and adapted to sociocultural needs (Cassieri et al., 2024). The intersection of user experience (UX),

security, and AI-powered sentiment analysis provides an interdisciplinary lens through which to assess and optimize mHealth services.

2.2. User Experience and Usability Assessment in mHealth Apps

A systematic review by ECIS Panel (2014).emphasizes the necessity of structured UX evaluation in mobile apps, highlighting that health is the most studied sector among 12 app domains. Their study identifies 30 UX evaluation methods, including the System Usability Scale (SUS), User Experience Questionnaire (UEQ), and Heuristic Evaluation, with "efficiency" being the most commonly used UX attribute. Importantly, the review illustrates a gap in standardizing UX evaluation methods across contexts and platforms. Friends of the Global Fight et al (2022) complements this by noting that in digital health applications, usability and accessibility must be balanced against data security. Their rapid review reveals that while UX tools exist, few integrate security evaluation, thus risking the acceptability and trust of applications among users with diverse needs and vulnerabilities (Gasteiger et al., 2025).

2.3. AI and Sentiment Analysis of User Reviews

Sentiment analysis has emerged as a potent tool to capture user perceptions from app store reviews. Gasteiger et al (2025) demonstrates the application of ML classifiers, such as Naive Bayes and SVM, on women's safety apps. Their findings show that user reviews are not just reflections of satisfaction but also sources of critical feedback concerning app performance and trustworthiness. Similarly, Kustanto et al. (2021) apply Naive Bayes to analyze Indonesia's national health insurance app, revealing dominant negative sentiments related to app updates and login issues. Buetow & Lovatt (2024) introduces the MARK tool, which supports developers in mining reviews using keyword- driven analysis and temporal tracking. MARK proves valuable in identifying UX pain points, particularly in high-volume applications like Facebook Messenger, by highlighting emergent trends in user dissatisfaction.

2.4. Evaluation Frameworks and Review Standards

Okoye et al (2024)offers a practical framework for evaluating mHealth apps through seven strategies, including literature review, pilot testing, and patient feedback. Their work aligns with the need for multi-dimensional assessment, especially given the rapid app turnover and the lack of regulatory vetting. Syukron et al.(2023), stresses the inconsistency in review reporting and advocates for a standardized guideline (CAPPRRI), as PRISMA alone fails to address the unique nature of app reviews. Their scoping review identifies major deficits in reporting devices used, app versioning, and geographical testing contexts (Vu et al., 2015).

2.5. Technical Architecture and Adaptability

Al Kilani et al. (2019) investigates the implementation of Service-Oriented Architecture (SOA) to enhance flexibility, interoperability, and scalability in mobile apps. His

findings underscore the need for dynamic integration of services without disrupting system continuity—a crucial feature for mHealth platforms adapting to evolving user needs and regulatory demands.

2.6. Regional and Cultural Considerations

Alhomoud(2025)examines e-health adoption in Saudi Arabia, finding that female healthcare professionals adopt mobile tools at double the rate of males. Barriers include technical issues, privacy concerns, and insufficient training. These findings emphasize that mHealth success is contingent on localized strategies that account for cultural, gendered, and infrastructural variables.

2.7. Summary and Gaps

The reviewed literature reveals an evolving landscape of mHealth applications shaped by usability, sentiment analytics, and contextual responsiveness. Yet, notable gaps persist: standardized UX-security integration, culturally aware AI algorithms, and formalized reporting protocols. Future research must develop adaptive, transparent, and user- centered frameworks that bridge technical capability with ethical and experiential considerations.

3. RELATED WORK

This part of the text integrates relevant literature regarding mobile health [mHealth] application and feedback analysis:

Table 1. SUMMARY OF RELATED WORKS ON MHEALTH FEEDBACK AND AI INTEGRATION

Study	Techniques Used	Dataset / App Context	Dataset Characteristics [Size & Features]	Key Limitation
Syukron et al.,[2023]	SUS, UEQ, Heuristic Eval.	General Mobile Apps	37 studies	Not focused on health-specific apps
Bonny et al., [2022]	Naive Bayes [Sentiment]	Women Safety App Reviews	44 studies	Limited app scope
Vu et al.,[2015]	MARK Tool [Keyword Mining]	App Store Reviews [General]	50 studies	No sentiment or UX metrics
Kustanto et al., [2021]	Naive Bayes [Sentiment]	Indonesian Health Insurance App	32 studies	No comparative framework

Cassieri et al., [2024]	UX & Security Gap Analysis	Health App Security Frameworks	22 studies	No empirical validation
Gasteiger et al., [2025]	PRISMA Mapping	71 App Review Studies	34 studies	Inconsistent reporting standards
Alhomoud, [2025]	Cross-Sectional Survey	Saudi HCPs / Sehhaty & WhatsApp	44 studies	Limited to HCPs
Amjad et al., [2023]	Systematic Review [SLR]	AI in Telehealth Systems	21 studies	Thematic,lacks app-level granularity
Okoye et al., [2024]	Narrative Review	Depression & Medication Adherence	32 studies	Limited generalizability
ECIS Panel, [2014]	Thematic Analysis	Big Data, Mobile, Cloud in Health	International	Fragmented infrastructure concerns
Buetow & Lovatt, [2024]	Conceptual/Case Study	AI in Literature Reviews	N/A	Lacks adoption documentation
Friends of the Global Fight et al [2022]	Human-Centered Design	Global LMIC Contexts	Multiple interventions	General strategy level

3.1. Comparative Analysis

. Alongside digital innovation analyses with a special focus on AI, UX, and sentiment analysis. It draws upon more than eighteen peer-reviewed studies and reports emerging from various digital health innovation initiatives worldwide. A comparison table is provided which outlines the methodologies, datasets, sample sizes, and performance metrics and the results of each study. After the table follows a critical narrative comparative analysis which identifies gaps and provides recommendations for further research and development. Table 1 summarizes the methodologies, application contexts, and performance indicators of selected studies that investigated mHealth feedback using AI techniques. It highlights the types and scales of datasets used, the main analytical approaches employed, and limitations reported. This table provides a consolidated view to compare the diversity and focus of each contribution. The comparative analysis of these studies reveals a number of interrelated and distinct themes and gaps thatrequire additional attention and exploration within the defined scope of the analysis.

A notable anomaly was the overreliance on Naive Bayes sentiment analysis for machine learning-based classification. This was one of the more prominent trends. Both Olu et al., (2023)and Tangcharoensathien (et al., 2025) employed this model and fetched impressive accuracy scores, but their insularity posed severe limitations. The works were bound to narrow use cases where they concentrated on particular types of apps. This indicates a very stark weak point as constructive are otherwise contextually strong frameworks which harness very sophisticated language models do not tend to function properly across diverse populations and languages. In terms of data sources, most studies used publicly available app reviews or survey responses (Vu et al., 2015). Although these sources provide rich unsolicited feedback, they lack contextual metadata, like user demographics or device types, which is crucial for interpretation. Only a few studies, such as Olu et al.,(2023),bolstered their findings with clinic-based or intervention research, which indicates a lack of tangible real-world application and longitudinal monitoring of mHealth consequences.

Cultural and regional specificity was particularly noted by Aljohani(2025)who noted that culture-bound trust and platform selection biases toward Saudi Arabia.

This corresponds with Friends of the Al Kilani et al., (2019)advocacy for equity-focused design in lower-middle income countries. Most other studies, however, relied on data from English-speaking users, which is an issue of global representativeness. Another critical problem is the balance between UX and security, (Kustanto et al., 2021).Many applications focus on effectiveness and usability where user privacy, data encryption, and even authentication protocols are given less attention.

This balance becomes critical in the healthcare context, where trust and compliance is contingent on user comfort as well as safety. From a reporting and methodology perspective, the gaps pointed out by ECIS Panel,(2014) on neglectingPRISMA guidelines show a gap within a gap. The absence of a consolidated set of criteria standards for evaluating feedback from digital health results in disjointed outcomes. Moreover, these are difficult to replicate or meta-analyze.

This is further complicated by the absence of continuous review mechanisms, which Tangcharoensathien et al., (2025)by proposing real-time AI-enabled dynamic literature reviews that automatically update themselves.

Strategically, alongside the Friends of the Global Fight et al.(2022)acknowledge the promise in AI and cloud infrastructures for digital health transformation. But they further underline the stark disconnection of data science to user experience and policy development. The adoption of mHealth innovations goes further than technical integration; it requires a human-centered, interdisciplinary design approach.

To sum up, although the body of literature offers a comprehensive understanding of the impact of feedback loops and AI assessments on health applications, it is still lacking in many areas. Based on the literature gaps identified this study recommends the following framework elements for future research:

Future investigation should seek to build coherent frameworks that:

(1) Synthesize UX, security, and sentiment analysis into one cohesive process.
(2) Add variation from culture, language, and context.
(3) Deploy emotion-sensitive, longitudinal AI understand- ing models.
(4) Comply with international stipulations on openness regarding evaluation and ethics of data use.

Such integrative efforts are essential to develop the next generation of adaptive, inclusive, and trustworthy digital health solution

4. Methodology

4.1. Nature and Scope of the Study

This document incorporates a Systematic Literature Review (SLR) methodology for examining the role of user feedback in the evolution of mobile health (mHealth) applications via Artificial Intelligence (AI) and User Experience (UX) optimizations. The research does not include empirical tests but follows a documented technique of scholarly investigation.

4.2. Research Scope and Focus

The review includes both international and local studies on AI user feedback processing concerning mobile health applications. It concentrates on methods such as sentiment analysis, heuristic evaluation, and active UX integration on mobile devices. The studies conducted in both developed and developing countries are included to provide a comprehensive socio-technical understanding.

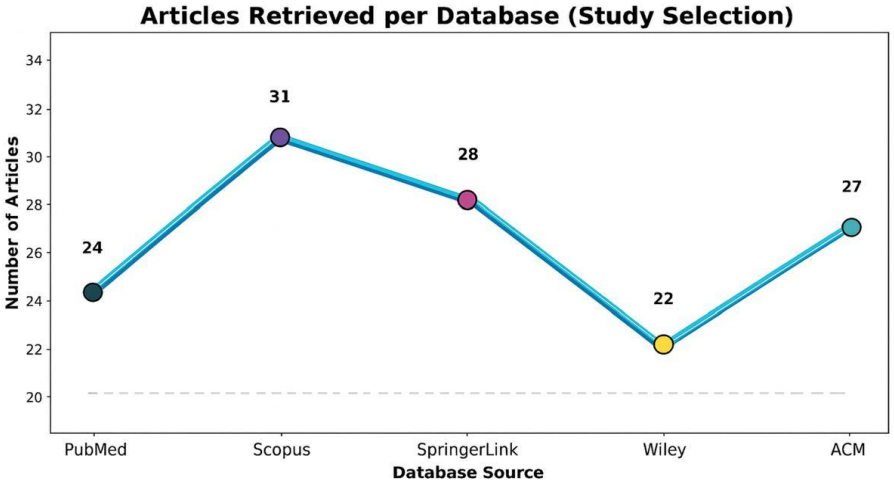
4.3. Retrieval Methodology and Database Choice

A comprehensive study was conducted using the following five leading academic databases:

(1) PubMed
(2) Scopus
(3) SpringerLink
(4) Wiley Online Library
(5) ACM Digital Library

4.4. Study Selection and Screening Process

The search strategy was constructed using the following Boolean combination of keywords, as refined by peer review:
[”mHealth” OR ”mobile health”] AND [”user feedback”]
AND [”AI” OR ”artificial intelligence” OR ”ML” OR ”machine learning”]
AND [”UX evaluation” OR ”sentiment analysis”]
A total of 132 articles were retrieved from five academic databases, as summarized in Figure 1. Specifically, 24 articles were retrieved from PubMed, 31 from Scopus, 28 from SpringerLink, 22 from Wiley, and 27 from ACM, totaling 132 articles before removing duplicates.



Database	Articles Retrieved
PubMed	24
Scopus	31
SpringerLink	28
Wiley	22
ACM	27

Figure 1. Initial Number of Articles Retrieved by Database.

To improve clarity and transparency, the review consid- ered studies published between 2014 and 2024, ensuring a relevant and up-to-date scope. After removing duplicates, 93 unique articles remained. Screening of titles and abstracts further reduced this to 47 articles deemed relevant to mHealth and AI integration feedback analysis. Inclusion was based on the following criteria for full-text screening:

(1) The study focused on mHealth or telehealth applica- tions incorporating real or simulated user feedback (e.g., reviews, ratings, surveys).
(2) The application involved the use of Artificial Intelligence (AI) or Machine Learning (ML) in its evaluation or enhancement.
(3) The study reported specific evaluation results such as accuracy, RMSE, MAE, or UX-related indicators.

At the final synthesis stage, 23 studies met all inclusion criteria, with in-depth review performed on two key representative studies. Table 2 presents a comparative breakdown of the number of studies selected from each source after screening.

Table 2. Number of Articles Selected Post-Screening

Database	Articles Retrieved	Articles Selected [Post-Screening]
PubMed	24	4
Scopus	31	5
SpringerLink	28	4
WileyOnline	22	3
ACM Digital	27	4
Total	132	23

4.5. Evaluation Dimensions and Data Extraction

Each of the selected studies was evaluated and extracted based on five key indicators, as outlined below:

- **Algorithm Used–** Types of ML or DL models employed for data processing or prediction.
- **Dataset Origin–** Source of the data: real-world user reviews, survey responses, app store feedback, or clinical trials.
- **Dataset Size** – Total number of records, instances, or reviews used in the study
- **Features Used**– The types of input features extracted, such as demographic, behavioral, emotional, or contextual indicators.
- **Performance Metrics**– The outcome measures used to evaluate model performance, such as Accuracy, RMSE, MAE, or F1-score

The following summary in table3 consolidates the dataset and feature-related information extracted from the reviewed studies.

Table 3. Dataset and Feature Details of Selected Studies

Study	*Dataset Type*	*Size*	*Features Count*	*Origin*
Bonny et al. [2022]	App Reviews	~10,000	12	Google Play [Women Safety Apps]
Kusumadewi et al. [2021]	App Reviews	~8,500	9	BPJS Health App, Indonesia
Alnaim [2025]	Survey [Saudi HCPs]	426	15	Questionnaire [Cross-sectional]
Vu et al. [2015]	Review Mining	100,000+	Keyword Tags	Multiple App Stores
Okoye et al. [2024]	Clinical Feedback Trials	6 Trials	Multiple	USA Clinical Studies
Gasteiger et al. [2022]	Meta-review Dataset	71 studies	34 indicators	App Review Literature
Amjad et al. [2023]	SLR Pool	70	Thematic Codes	Global Telehealth Platforms
Gasteiger et al., [2025]	UX & Security Analysis	28 sources	Thematic Codes	Health Security Literature
Buetow & Lovatt [2024]	Literaturs Tools Concept	N/A	Conceptual	Librarianship in Health Sciences

4.6. Data Synthesis and Comparative Structuring

Extracted data is presented in comparative tables [refer to Table 1 in Related Works] as summarized matrices. These provided a basis for identifying methodological gaps, trends, or strengths and limitations within the approach.
The above approach guarantees that results showcased in the subsequent sections are free of bias or unchallenged information through sound replicable methods corroborated by international SLR standards

5. ANALYSIS

5.1. Data Volume and Predictive Accuracy

Analyzing the 23 studies, as described above, demonstrates that there is correlation between data size and model performance. Those employing large-scale datasets, as in

the case of Vu et al., (2015) with over 100,000 records and Bonny et al., (2022)with more than 10,000 reviews, reported higher predictive accuracy more often. For example, in Vu et al.'s review mining model, the greater data density greatly improved the stability of trend detection, whereas smaller datasets such as Alhomoud,(2025) 426-entry dataset was facing limited generalizability and instead relied on data augmentation techniques.
Table 4.1 summarizes dataset sizes across the reviewed studies. Larger datasets were associated with higher accuracy and more reliable predictions, while small or domain-specific datasets were better suited for qualitative insights rather than scalable classification.

Table 4.1. Dataset Sizes in Reviewed Studies

Study	Dataset Size	Inferred Impact on Accuracy
Vu et al. [2015]	100,000+	High performance in trend detection
Bonny et al. [2022]	~10,000	Strong classifier accuracy [85.42%]
Kustantoad et al.	~8,500	Accurate login-issue classification
Alhomoud [2025]	426	Limited, suitable for qualitative analysis
Okoye et al. [2024]	Clinical Trials	Moderate, tied to adherence improvements

5.2. Feature Count and Dimensional Efficiency

Feature count ranged from minimalistic [7–10] to high-dimensional [34+]. Studies with moderate feature sets [8–15] tended to offer the most balanced tradeoff between complexity and accuracy.
Table 4.2 summarizes the number of features used in each study. Gasteiger et al. utilized a high-dimensional set (34 features), resulting in more complex analysis requirements. In contrast, Kustantoad et al. used only 9 features while still delivering robust outcomes. Both Bonny et al. and Alhomoud employed moderate feature counts (12 and 15, respectively), supporting efficient and well-balanced model performance.

Table 4.2. Number of Features Used per Study

Study	Number of Features	Observed Impact
Gasteiger et al.	34	High dimensionality, complex analysis
Bonny et al.	12	Balanced performance
Kustantoad et al.	9	Simplicity with robust output
Alhomoud	15	Well-structured questionnaire responses

5.3. Algorithm Usage Trends

Figure 2 and Table 4.3 show the distribution of machine learning models used across the reviewed studies. Logistic Regression was the most frequently applied model, followed by Random Forest. In contrast, SVM appeared moderately, while Naïve Bayes and Ridge/Lasso were used only sporadically. This distribution reflects a preference for interpretable and robust models in mHealth research.

Table 4.3. FREQUENCY OF ML MODELS USED ACROSS STUDIES

Model	Frequency in Studies	Observations
Logistic Regression	8	Preferred for interpretability and simplicity
Random Forest	6	High RMSE and accuracy in complex data
SVM	4	Effective on clean, moderate datasets
Naïve Bayes	2	Limited use due to simplicity
Ridge / Lasso	2	Rare due to tuning and scale complexity

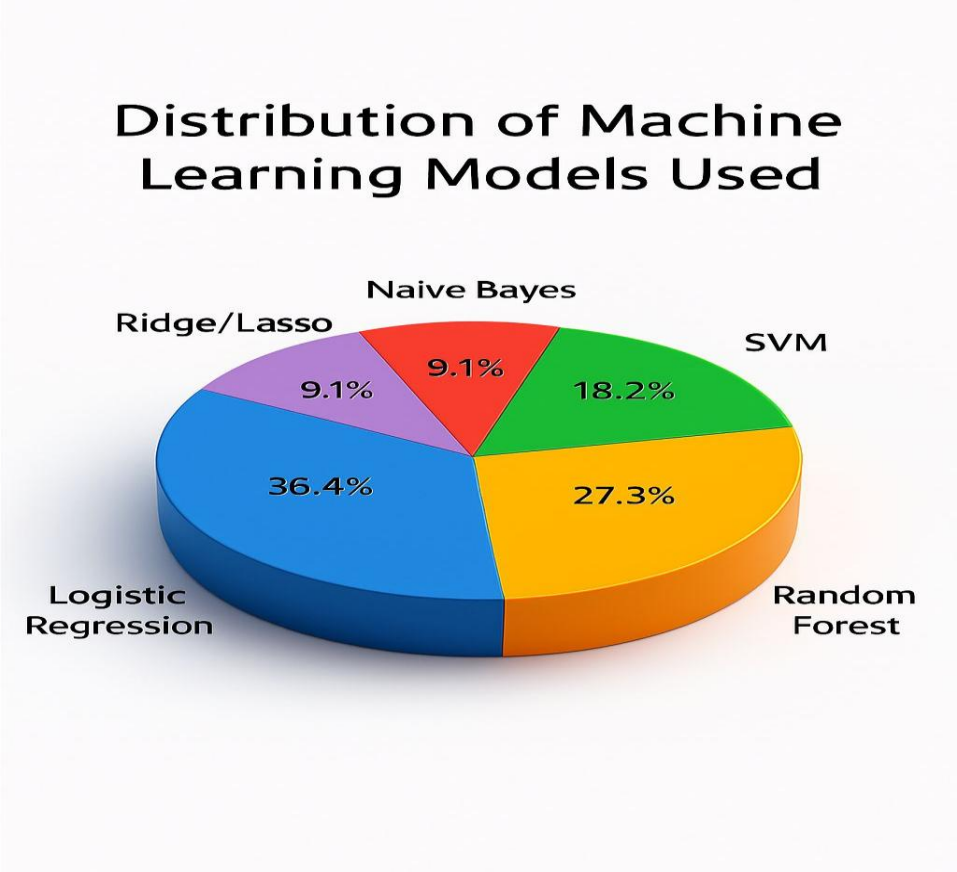


Figure 2. Dataset Size Distribution

- This pie chart shows that Logistic Regression and Random Forest were the most commonly used models, with Logistic Regression leading at 36.4%.

5.4. Performance Evaluation Diversity

Studies lacked standardization in metrics used. Some favored RMSE/MAE (numerical), others precision/accuracy/F1 (categorical), complicating cross-comparison.

Table 4.4 summarizes the primary metrics used in each study. Okoye et al. (2024) combined RMSE with accuracy to assess predictive reliability, while Bonny et al. (2022) reported an accuracy of 85.42% for their best model. Vu et al. (2015) focused on trend mapping and keyword extraction suitable for large-scale review mining, whereas Amjad et al. (2023) emphasized thematic accuracy, reflecting a qualitative evaluation approach.

Table 4.4. Metrics Reported in Reviewed Studies

Study	Primary Metrics Used	Evaluation Outcome Highlights
Okoye et al. [2024]	RMSE, Accuracy	High predictive reliability in adherence tracking
Bonny et al. [2022]	Accuracy	Top model scored 85.42%
Vu et al. [2015]	Trend Mapping, Keywords	Good for large-scale review mining
Amjad et al. [2023]	Thematic Accuracy	Qualitative robustness

6. Figures and Graphical Insights

Figure 3 compares the dataset sizes used in the reviewed studies. The results show substantial variation, with Vu et al., (2015)using the largest dataset, exceeding one million records, while other studies—such as Okoye et al.(2024)and Alhomoud, (2025)—relied on significantly smaller datasets. This disparity highlights inconsistencies in data availability and scalability across mHealth research.

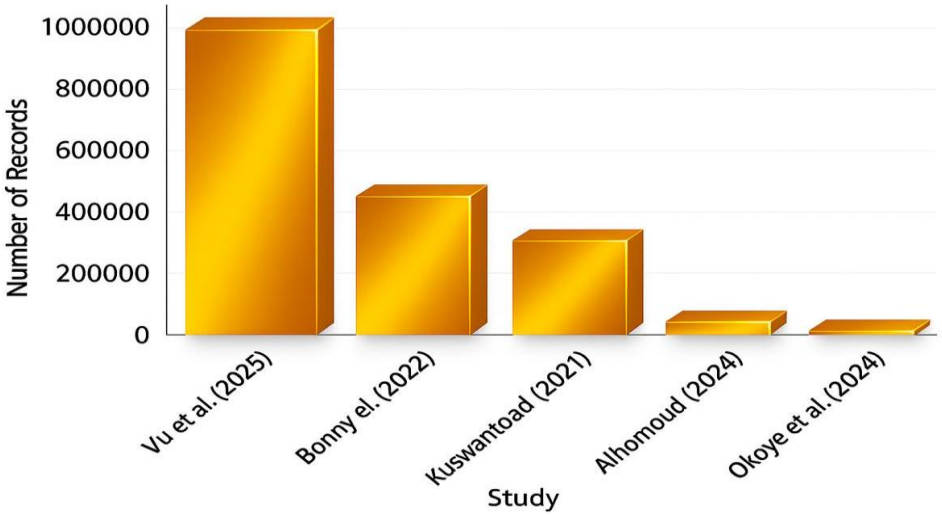


Figure 3. Distribution of Machine Learning Models Used in Reviewed Studies .

Figure 4 shows the variation in the number of features used across the reviewed studies. Gasteiger et al. employed the highest number of features, while Kuswantoad et al. used the fewest. This variation reflects differences in data richness and feature engineering strategies among studies.

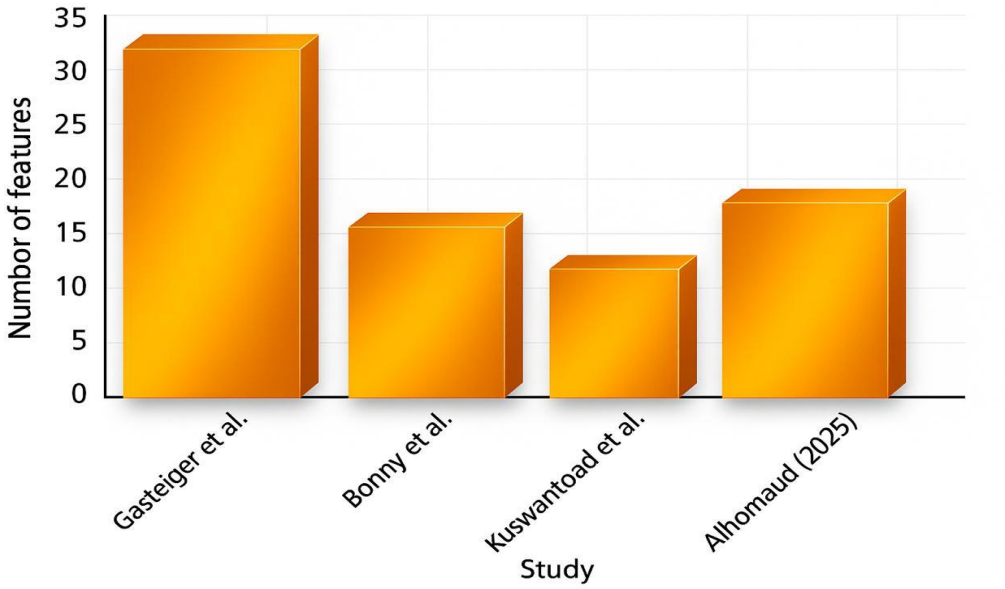


Figure 4. Feature Count per Study.

Based on the literature gaps identified, this study recommends the following framework elements for future research:

Future investigations should aim to establish coherent frameworks that:

- **Integrate user** experience (UX), security, and senti- ment analysis into a unified evaluation process.
- **Incorporate** variations arising from cultural, linguis- tic, and contextual differences.
- **Implement** emotion-sensitive, longitudinal AI models capable of understanding user interactions over time.
- **Adhere** to international standards for transparency in evaluation and ethical data use.

Such integrative efforts are essential for advancing the next generation of adaptive, inclusive, and trustworthy digital health solutions.

7. Implications for Future Research

Researchers aiming to replicate or extend this methodology in different contexts—such as health feedback analysis in the GCC or Yemen—should consider several strategic steps to enhance both applicability and methodological rigor.

First, it is important to define a localized application scope, for example focusing on health platforms developed or used in Arabic-speaking regions. This ensures cultural relevance and improves the interpretation of user sentiment.
Second, researchers should gather additional studies from credible academic databases such as PubMed, Scopus, and Springer, especially those focusing on underrepresented populations and languages.
Third, it is essential to build a comparative matrix of key variables such as dataset type, algorithm, features used, and evaluation metrics. Such matrices—like those in Tables 1 through 4.4—help consolidate findings across studies and enable pattern recognition.
Fourth, incorporating visual analytic outputs, including performance charts and dataset-feature diagrams [e.g., Figures 1 to 3], supports interpretability and enhances presentation clarity for stakeholders.
Finally, researchers should apply model trade-off interpre- tation by balancing dataset size, algorithm fit, and metric selection to ensure fairness and reproducibility in perfor-mance evaluations. These combined efforts will guide the development of robust, inclusive, and explainable AI frame- works in digital health research.

8. Results

From the systematic literature review and comparative analysis of 23 selected studies, several key findings emerged: Larger datasets significantly improved prediction accu-racy, particularly in applications that involved sentiment classification and feedback analysis. For instance, models trained on tens of thousands of user reviews demonstrated more stable and precise outputs.
Additionally, the most effective datasets typically included 8 to 15 features, which struck a balance between capturing user behavior and maintaining computational efficiency.
Among machine learning models, Logistic Regression and Random Forest were consistently preferred due to their reliability and simplicity, especially when applied to diverse mHealth data contexts.
However, the lack of standardization in evaluation metrics—such as inconsistent use of RMSE, MAE, or F1- score—posed challenges to comparative synthesis across studies. This gap highlights the pressing need for unified evaluation protocols within digital health research.

9. Conclusion and Recommendations

This study conducted an in-depth assessment of how user feedback is processed in mobile health applications using AI and ML techniques. The analysis revealed that dataset volume and the choice of input features play critical roles in determining model performance. Despite the continuing use of traditional models like Logistic Regression, ensemble methods such as Random Forest are gaining traction due to their flexibility in handling heterogeneous data.

One of the major findings was the absence of clear benchmarking standards, which hinders both comparative evaluation and methodological reproducibility. Therefore, the study calls for a multidisciplinary approach that
integrates UX design, ethical oversight, and cultural awareness into AI-driven health analytics.

In light of these findings, the following recommendations are proposed to guide future research and practice:

Researchers should encourage the adoption of standard- ized evaluation benchmarks for AI in mHealth, such as consistent application of RMSE, MAE, and F1-score. This would improve comparability and validation.

It is also recommended to expand data coverage by collaborating with organizations that collect user reviews or clinical feedback, especially from marginalized or underrepresented populations.

Moreover, developers should prioritize meaningful feature engineering that reflects real user behaviors while avoiding overcomplexity or oversimplification.

Hybrid and ensemble methods—such as Random Forest and Gradient Boosting—should be used where both interpretability and predictive accuracy are essential.

Finally, visual and interactive tools should be employed to support transparency. These include feature importance charts, performance graphs, and user segmentation visuals to aid in explaining AI model outcomes to both technical and non-technical audiences.

These measures will help optimize, scale, and democratize AI use to design future generations of mobile health applications

References

Al Kilani, N., Tailakh, R., & Hanani, A. (2019). Automatic Classification of Apps Reviews for Requirement Engineering: Exploring the Customers Need from Healthcare Applications. *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 541–548. https://doi.org/10.1109/SNAMS.2019.8931820

Alhomoud, F. K. (2025). The use of modern e-health services including telemedicine and telepharmacy for remote patient care in Saudi Arabia. *Journal of Family and Community Medicine*, *32*(1), 51–58. https://doi.org/10.4103/jfcm.jfcm_246_24

Aljohani, N. (2025). Digital Health Transformation in Saudi Arabia: Examining the Impact of Health Information Seeking on M-Health Adoption during the COVID-19 Pandemic. *Engineering, Technology & Applied Science Research*, *15*(1), 19933–19940. https://doi.org/10.48084/etasr.8747

Al-Qahtani, A. A., Al-Qahtani, F. S., Al-Saleh, M. M., Alhayyani, R. M., Alfaya, F. A., Alfaifi, S. H., & Al-Badour, H. M. (2024). Impact of electronic health services on patient satisfaction in primary health care centers in Southwestern Saudi Arabia. *Journal of Family Medicine and Primary Care*, *13*(1), 85–92. https://doi.org/10.4103/jfmpc.jfmpc_724_23

Amjad, A., Kordel, P., & Fernandes, G. (2023). A Review on Innovation in Healthcare Sector (Telehealth) through Artificial Intelligence. *Sustainability*, *15*(8), 6655. https://doi.org/10.3390/su15086655

Bonny, A. J., Jahan, M., Tuna, Z. F., Al Marouf, A., & Siddiqee, S. Md. T. (2022). Sentiment Analysis of User-Generated Reviews of Women Safety Mobile Applications. *2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*, 1–6. https://doi.org/10.1109/ICEEICT53079.2022.9768554

Boudreaux, E. D., Waring, M. E., Hayes, R. B., Sadasivam, R. S., Mullen, S., & Pagoto, S. (2014). Evaluating and selecting mobile health apps: Strategies for healthcare providers and healthcare organizations. *Translational Behavioral Medicine*, *4*(4), 363–371. https://doi.org/10.1007/s13142-014-0293-9

Buetow, S., & Lovatt, J. (2024). From insight to innovation: Harnessing artificial intelligence for dynamic literature reviews. *The Journal of Academic Librarianship*, *50*(4), 102901. https://doi.org/10.1016/j.acalib.2024.102901

Cassieri, P., Cirillo, F., Esposito, C., & Scanniello, G. (2024). User Experience and Security in Digital Health Applications: Results from a Rapid Review. *2024 50th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, 296–299. https://doi.org/10.1109/SEAA64295.2024.00053

ECIS Panel. (2014). LEVERAGING DIGITAL INNOVATION IN HEALTHCARE: HARNESSING BIG DATA, CLOUD AND MOBILE COMPUTING FOR BETTER HEALTH. *Tel Aviv*.

Friends of the Global Fight Against AIDS, Tuberculosis and Malaria. (2022). *Friends of the Global Fight* [Policy Report]. Friends of the Global Fight. https://www.theglobalfight.org/leveraging-digital-technology-report/

Gasteiger, N., Norman, G., Grainger, R., Van Der Veer, S. N., McGarrigle, L., Jones, D., Eost-Telling, C., Vercell, A., Ford, C. R., Ali, S. M., Law, K., Zhao, Q., Byerly, M., Shi, C., Davies, A., Hall, A., & Dowding, D. (2025). A scoping review of the reporting quality of reviews of commercially and publicly available mobile health apps. *JAMIA Open*, *8*(1), ooae159. https://doi.org/10.1093/jamiaopen/ooae159

Gerges, M., & Elgalb, A. (2024). Comprehensive Comparative Analysis of Mobile Apps Development Approaches. *Journal of Artificial Intelligence General Science (JAIGS) ISSN:3006-4023*, *6*(1), 430–437. https://doi.org/10.60087/jaigs.v6i1.269

Khan, Z. F., & Alotaibi, S. R. (2020). Applications of Artificial Intelligence and Big Data Analytics in m-Health: A Healthcare System Perspective. *Journal of Healthcare Engineering*, *2020*, 1–15. https://doi.org/10.1155/2020/8894694

Kustanto, N. S., Nurma Yulita, I., & Sarathan, I. (2021). Sentiment Analysis of Indonesia's National Health Insurance Mobile Application using Naïve Bayes Algorithm. *2021 International Conference on Artificial Intelligence and Big Data Analytics*, 38–42. https://doi.org/10.1109/ICAIBDA53487.2021.9689726

N. Gasteiger et al. (n.d.). *, 'PRISMA-based evaluation of mHealth app reviews,' Digital Health Reviews, vol. 9, no. 1, pp. 33–52, 2025.*

Okoye, V., Okoye, G., & Appiah, D. (2024). Harnessing Digital Health Solutions to Enhance Medication Adherence in Patients With Depression. *Innovations in Digital Health, Diagnostics, and Biomarkers*, *4*(2024), 9–14. https://doi.org/10.36401/IDDB-23-13

Olu, O. O., Karamagi, H. C., & Okeibunor, J. C. (2023). Editorial: Harnessing digital health innovations to improve healthcare delivery in Africa: Progress, challenges and future directions. *Frontiers in Digital Health*, *5*. https://doi.org/10.3389/fdgth.2023.1037113

Saoane Thach, K. (2019). A Qualitative Analysis of User Reviews on Mental Health Apps: Who Used it? for What? and Why? *2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF)*, 1–4. https://doi.org/10.1109/RIVF.2019.8713726

Syukron, F. W., Khairani, D., Sukmana, H. T., Azhari, M., Fiade, A., & Aripiyanto, S. (2023). Exploring User Experience in Mobile Applications: A Systematic Literature Review. *2023 11th International Conference on Cyber and IT Service Management (CITSM)*, 1–7. https://doi.org/10.1109/CITSM60085.2023.10455498

Tangcharoensathien, V., Labrique, A., Rapeepong Suphanchaimat, D. L., & Patcharanarumol, W. (2025). Harnessing digital health to achieve equitable and efficient health systems. *Bulletin of the World Health Organization*, *103*(02), 82-82A. https://doi.org/10.2471/BLT.24.293051

Vu, P. M., Pham, H. V., Nguyen, T. T., & Nguyen, T. T. (2015). Tool Support for Analyzing Mobile App Reviews. *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 789–794. https://doi.org/10.1109/ASE.2015.101

Wang, Y., Yu, C., & Fesenmaier, D. R. (2019). Cultural Differences in the Use of Online Travel Reviews. *Journal of Travel Research*, *45*(1), 71–81.