Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.3, June 2026, pp. 2588~2597
e-ISSN: 2548-6861

OCR and NLP for Consumer Product Label Analysis: A Systematic Literature Review

Musthofa Dzikry Pamungkas[1]*, Noor Falih[2]*, Muhammad Panji Muslim[3]*, Nanang Nasrulloh[4]**
* Fakultas Ilmu Komputer Universitas Pembangunan Nasional "Veteran" Jakarta, Indonesia
** Fakultas Ilmu Kesehatan, Universitas Pembangunan Nasional "Veteran" Jakarta, Indonesia
2110511040@mahasiswa.upnvj.ac.id [1], falih@upnvj.ac.id [2], muhammadpanji@upnvj.ac.id [3], nasrullah@upnvj.ac.id [4]

Article Info

Article history:

Received 2026-04-24
Revised 2026-05-18
Accepted 2026-05-25

Keyword:

BERT-based Natural Language Processing (NLP),
Diabetes Risk,
Food Composition Label,
Optical Character Recognition (OCR),
Ultra-Processed Food (UPF).

ABSTRACT

Composition labels on consumer products serve as an essential source of information for consumers in making purchasing decisions and for producers in ensuring regulatory compliance. However, manual reading of labels is prone to errors due to small font sizes, blurred images, low resolution, and perspective distortion, which reduce the accuracy of information recognition. This issue becomes increasingly critical as many label compositions contain ingredients associated with metabolic diseases, such as added sugar, artificial sweeteners, and carbohydrates, whose risks are well-documented in epidemiological literature. While numerous systematic reviews have addressed OCR and NLP in document intelligence, no prior review has systematically mapped the integration of OCR robustness, BERT-based semantic extraction, and food label-specific challenges in the context of multilingual Indonesian consumer products. To address this gap, this study conducts a Systematic Literature Review (SLR) following PRISMA guidelines, synthesizing recent advances in Transformer-based OCR and BERT-based NLP for consumer product label analysis. The review aims to map OCR performance under real-world conditions, identify best practices in preprocessing and post-correction, and evaluate end-to-end integration with BERT for semantic understanding of label compositions. The expected outcome is a theoretical contribution in the form of an integrative OCR–BERT framework proposal, along with practical design recommendations for future systems aimed at supporting nutritional literacy and informed consumer decision-making. Empirical validation of the proposed framework remains a direction for future research.

I. INTRODUCTION

Consumer products are widely available in the community to meet consumer needs, prompting businesses to create products that align with consumer demands. One of the strategies employed by producers is the use of labels to attract consumers in purchasing their products [1], [2]. Labels provide essential information regarding product content, brand, and expiration date, which are regulated by the Indonesian National Agency of Drug and Food Control (BPOM). These regulations must be complied with by businesses to meet the established standards [3]. Labels appear in various forms, such as nutrition facts and ingredient composition. This information is required by consumers when making purchasing decisions [4]. Nutrition facts play an important role in providing product information and raising consumer awareness about nutritional intake, which in turn may influence consumer behavior [5], [6]. Consequently, health-conscious consumers have emerged who carefully consider product composition in their consumption patterns. Nutrition facts label also contain detailed information on product composition in accordance with regulations set by BPOM RI [7]. Composition refers to the ingredients used in a product, including the names of substances and their quantities per package. As consumers become increasingly aware of their health, composition has a direct impact on consumer behavior toward a product [8], [9], [10]. Thus,

composition labels hold particular importance for consumers when evaluating products.

The selection of appropriate composition benefits not only consumers but also producers. Consumers become more health-aware, while producers can develop products that meet consumer needs and maintain product quality. Appropriate ingredient selection directly affects both product quality and competitiveness [11], [12]. Therefore, composition influences not only consumer health but also the overall quality and market performance of a product. Increased awareness of ingredient selection encourages more prudent use of chemical substances. The chemical composition of a product can significantly influence its nutritional value [5], [6], [13], [14]. Since health is a crucial factor in society, greater awareness of food composition supports the creation of a healthier population in Indonesia.

Ultra-processed foods (UPF) circulating in the Indonesian market are identified through their composition labels. This issue is particularly relevant in the Indonesian context, where the prevalence of diabetes among adults has increased from 6.9% in 2013 to 10.9% in 2018 [15], with more recent national health surveys confirming a continued upward trend to 11.3% in 2023, driven by aging, obesity, and risky dietary patterns [16]. Epidemiological evidence further suggests that high consumption of ultra-processed foods (UPF) is associated with increased risk of type 2 diabetes, with each 10% increase in UPF consumption raising the risk by approximately 15% [17]. Consumption of UPF exceeding 10% of total dietary intake is associated with a higher incidence of diabetes, including gestational diabetes [18]. Biological mechanisms highlight that additives such as monosodium glutamate (MSG) and phosphate compounds exacerbate oxidative stress, lipogenesis, insulin resistance, and vascular complications related to diabetes [19]. Individuals with type 2 diabetes tend to restrict their consumption of UPF compared to non-diabetic individuals, reinforcing the association between reduced processed food intake and better disease management [20]. Additionally, inadequate fruit and vegetable consumption further increases the incidence of type 2 diabetes in the population [15]. These findings underscore the importance of monitoring composition labels on consumer products, as unhealthy dietary patterns and high UPF consumption are directly associated with the rising burden of diabetes in Indonesia.

Composition labels provide crucial information; however, manual reading at scale is prone to errors due to small font size, blurred text, low resolution, and perspective distortion, all of which hinder consistent character recognition. These conditions reduce recognition accuracy and demand preprocessing strategies as well as recognition architectures that are robust to noise and text variability [21], [22]. The limitations of manual recognition make it difficult for consumers to identify added sugar, sweeteners, and sodium that are associated with metabolic diseases such as diabetes. Therefore, automated text extraction from labels using OCR designed for non-ideal conditions is necessary to ensure reliable nutritional analysis [21], [22], [23].

Advances in OCR offer a solution through encoder–decoder models that integrate image transformers and text transformers, achieving strong performance on printed text, handwriting, and scene text without complex pre- or post-processing, while being relatively easy to extend to multilingual contexts. The architecture converts images into patches and generates wordpieces that align with text analytics, reducing mismatches between OCR outputs and NLP tokenizers. This design also provides greater robustness against visual variations [24], [25], [26]. OCR thus not only improves the accuracy of label reading but also enables automated screening of food compositions that may contribute to diabetes risk.

Text extraction further introduces challenges in semantic understanding, as chemical names, abbreviations, synonyms, and additive functions often appear in short but contextual phrases, such as “sodium benzoate – preservative”. Classical statistical methods often fail to capture such meaning. BERT is highly relevant due to its bidirectional contextual representation, which makes models more sensitive to word relationships within short descriptions. BERT is not only applicable for ingredient extraction or additive categorization but also for identifying terms associated with sugar, artificial sweeteners, or saturated fats, that are factors contributing to diabetes risk [21], [27], [28].

Modern OCR is evolving from Convolutional Recurrent Neural Network (CRNN) pipelines toward end-to-end Transformer-based architectures that combine image encoders with text decoders, reducing the need for hand-crafted processing and external language models. In the context of composition labels, encoder–decoder models perform well on small text, varied fonts, and complex backgrounds, supported by global self-attention to overcome the limitations of local receptive fields in earlier approaches. Lightweight preprocessing (contrast enhancement, deblurring, adaptive resizing) and data augmentation strategies enhance robustness under real-world conditions, such as blur, low contrast, and perspective distortion. Transformer-based OCR thus provides a strong foundation for extracting composition text from label images of suboptimal quality [23], [24], [25], [26].

Once text is extracted, the challenge shifts to semantic understanding, namely identifying ingredient names, additive functions, allergens, and risk indicators from short, contextual phrases. BERT provides bidirectional contextual embeddings that capture word meaning from both left and right contexts, offering higher accuracy for named entity recognition and functional classification than traditional statistical approaches. In practice, BERT is often combined with sequential layers to model token dependencies, consistently improving precision in real-world text tasks. This approach is particularly relevant for chemical composition analysis since synonyms, abbreviations, and alternative names (e.g., *natrium bikarbonat* and *sodium bicarbonate*) represent the same

substance, but product labels may vary in terminology [21], [27], [28].

Although access to composition labels has become easier, no comprehensive review has systematically mapped OCR robustness in recognizing small, blurred, and low-contrast label text, along with preprocessing practices and BERT based NLP strategies for chemical terminology mapping, entity extraction, and functional classification in the Indonesian consumer product context, which involves multilingual and diverse packaging [22], [23], [28]. This gap is critical given that consumer product labels often contain compositions related to metabolic diseases, particularly diabetes, including total sugar, artificial sweeteners, and sodium. Therefore, this study aims to review and analyze findings on the performance and limitations of OCR under real-world conditions, as well as evaluate end-to-end integration schemes with BERT based NLP. The expected outcome is not only a mapping of the most relevant methods and datasets but also design recommendations for an OCR- and BERT based chemical composition analysis system capable of supporting automated screening of diabetes-related ingredients through a web application.

II. METHOD

This paper adopts a Systematic Literature Review (SLR) methodology to identify, categorize, and synthesize relevant research on the use of OCR and NLP for consumer product label analysis, following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines to ensure transparency and methodological rigor, as illustrated in Figure 1. The methodology comprises three key steps: defining research questions, conducting a systematic database search, and applying predetermined inclusion and exclusion criteria to select the final studies.

A. Research Questions

The Research Questions (RQs) provide the necessary focus and direction for this systematic review. They are formulated to directly address the research gaps identified in the Introduction and guide the entire process of searching, filtering, and synthesizing the literature. The three RQs for this study are presented in Table I.

TABEL I
RESEARCH QUESTIONS AND MOTIVATIONS

Research Question	**Motivation**
RQ1-How effective are modern OCR methods, specifically Transformer-based architectures, in terms of performance and limitations when extracting text from consumer product labels under real-world conditions (i.e., small, blurry, low-contrast text, and variable perspective)?	Extracting text from non-ideal label images (small, distorted, blurred) is the primary technical bottleneck in automating consumer product analysis. Investigating Transformer-based OCR is crucial as it represents the current *state-of-the-art* solution for robust scene text recognition, which is necessary to ensure the reliability of the subsequent NLP analysis.
RQ2-What is the role of BERT-based Natural Language Processing (NLP) in enhancing semantic understanding, particularly for the extraction of chemical entities, synonyms, and the functional classification of ingredients within consumer product composition texts?	Chemical and ingredient names on labels contain high semantic ambiguity (synonyms, abbreviations, multilingual variations). BERT's capability for bidirectional contextual embedding is essential to accurately map these entities and their functions, which traditional statistical NLP methods often fail to do, thus providing the necessary nutritional clarity.
RQ3-How can an integrated OCR and BERT-based NLP system be applied to consumer product labels to identify ingredients linked to metabolic diseases (e.g., diabetes), thereby supporting informed nutritional choices?	There is a critical gap in translating raw label data into actionable health intelligence. This question justifies the entire research, demonstrating the practical utility of the integrated OCR-BERT pipeline in automatically screening ingredients (like sugar, sodium), directly connecting product consumption to a major public health concern (diabetes risk).

To ensure the relevance and quality of the selected articles, a set of inclusion and exclusion criteria was established prior to the screening process. These criteria guided the assessment of all retrieved articles across the four screening stages and are summarized in Table II.

TABEL II
INCLUSION AND EXCLUSION CRITERIA

Domain	**Inclusion Criteria**	**Exclusion Criteria**
Technology	Studies applying Transformer-based or Encoder-Decoder OCR architectures, and BERT-based NLP for semantic understanding, NER, or text classification	Studies using only traditional OCR without deep learning advancement, or general NLP unrelated to label or chemical entity extraction
Application	Studies focusing on product labels, food/drug packaging, or scene text with similar characteristics (small, blurred, or distorted text)	Studies on unrelated domains such as license plate recognition or historical handwriting
Document Type	Primary research published in peer-reviewed journals or conference proceedings	Editorials, white papers, books, blog posts, or non-primary sources
Publication Year	Articles published between 2019-2026	Articles published before 2019

Language	Articles written in English	Non-English articles
Accessibility	Full text accessible	Abstract-only or inaccessible full text
Duplication	Unique articles per database	Duplicate entries across databases

B. Searching Process

The literature search was conducted in October 2025 across three prominent academic databases to ensure a comprehensive and systematic collection of relevant articles. These databases, namely IEEE Xplore, ScienceDirect, and Scopus, were selected for their high coverage of computer science, engineering, and medical technology literature. To capture the latest advancements in deep learning and AI, the search was strictly limited to articles published between 2019 and 2026.

To ensure that the search results are highly relevant to the integrated research questions (OCR, BERT-NLP, and health risk), a modular and highly specific search strategy using Boolean operators was employed. The search was structured to capture both the core technological and the application domain.

The search queries were executed in two main modules to maximize capture of relevant studies:

i. ("Optical Character Recognition" OR OCR) AND ("Natural Language Processing" OR NLP) AND (BERT OR Transformer)
ii. ("Chemical Composition" OR "Ingredient Analysis" OR "Food Label") AND (OCR OR NLP) AND (Health OR Risk OR Diabetes)

Those queries aim to ensure that articles focusing on either the *state-of-the-art* architecture or the specific application of extracting health-related ingredients are captured for the subsequent screening process.

C. Result Finding

The final set of articles for review was selected through a rigorous, multi-stage filtering process based on the predetermined inclusion and exclusion criteria. This structured process ensures transparency and consistency, progressing from broad search results to a specific. The screening process was executed in four distinct stages, aligning with the principles of the PRISMA framework:

1. Stage 1: The initial systematic search across the three databases (ScienceDirect, IEEE Xplore, and Scopus) yielded a cumulative total of 1,081 articles (822 from Query 1 and 259 from Query 2) as shown in Table III. This stage applied database-level filters to select only Research Articles (Journal and Conference Papers) and articles published between 2019 and 2026. All 1,081 retrieved articles were then consolidated for the subsequent duplication check.

TABEL III
NUMBER OF RETRIEVED STUDIES

No	Search Database	Results
1	ScienceDirect	702
2	IEEE Xplore	233
3	Scopus	146
Total		**1,081**

2. Stage 2: All 1,081 articles retrieved from the search executions were consolidated and imported into a reference management tool. Automated and subsequent manual procedures were executed to identify and remove all cross-database and intra-database duplicate entries. This critical step resulted in a consolidated list of 939 unique articles ready for relevance screening.
3. Stage 3: The remaining 939 unique articles were critically assessed based on their titles and abstracts against the full set of Inclusion and Exclusion Criteria. This process focused on excluding studies not clearly relevant to the core technologies (OCR/NLP/BERT) or the specific application domain (chemical entity extraction/health risk). A total of 798 articles were excluded during this relevance screening round.
4. Stage 4: The remaining 141 articles proceeded to the full-text screening stage. Each article was read in-depth to confirm its final eligibility, ensuring the study presented relevant primary data and directly addressed the research questions. A total of 113 articles were excluded during this final stage.

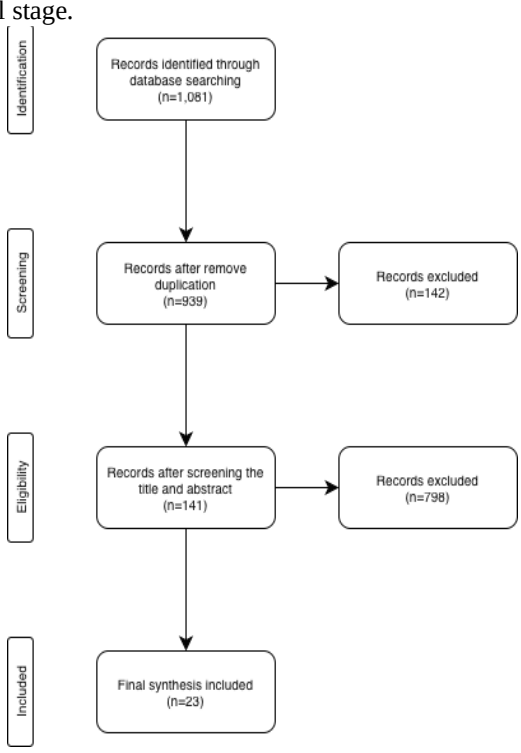

Figure 1. Article Selection Process (PRISMA Flow Diagram)

The remaining 23 articles proceeded to the full-text screening stage and were selected for the comprehensive data

synthesis and review. These articles form the basis for answering the research questions in the following chapter. Table IV shows the result of the filtering process. A comprehensive flowchart illustrating the article selection process is provided in Figure 1 (PRISMA Flow Diagram).

TABEL IV
RESULT OF FILTERING PROCESS

No	Publication	# of Article
1	Q1 Journal	10
2	Q2 Journal	3
3	Q3 Journal	2
4	Conference Proceeding	8
Total		**23**

III. RESULT AND DISCUSSION

A. OCR Robustness and Noise Mitigation

This section synthesizes the necessary solutions for robustly extracting text from low-quality, real-world images of consumer product labels, directly addressing the performance challenges and noise mitigation required by RQ1. The findings confirm that reliable extraction demands a multi-stage approach, integrating specialized architectural design with contextual refinement.

The fundamental shift in OCR state-of-the-art is defined by the move toward highly efficient, non-recurrent Transformer architectures, marking a critical departure from traditional CRNN pipelines that rely on hand-crafted feature engineering and struggle with geometric text variations. Foundational models such as TokenOCR overcome model weight limitations while outperforming established benchmarks including Tesseract and TrOCR [29], and DTrOCR further simplifies the pipeline through a single-stage decoder-only design that surpasses current SOTA across printed, handwritten, and scene text [30], demonstrating consistently lower CER and WER while eliminating the need for separate language models. To achieve robust detection under irregular conditions, multimodal information fusion and multi-task optimization further suppress feature misalignment and information redundancy [31]. This quantitative superiority and reduced architectural complexity collectively justify the adoption of Transformer-based OCR as the foundation for consumer product label analysis systems.

Empirical assessment confirms that raw OCR output is a major bottleneck and that a post-correction phase is mandatory for extraction accuracy. Research explicitly shows that OCR noise poses a significant obstacle to language modelling [32], causing models to sharply diverge from their noiseless targets. This degradation is mitigated by utilizing advanced contextual models. The two-step pipeline of applying transformer models (such as BERT) for post-processing is crucial for text post-processing [33], effectively correcting mistakes and improving the overall quality of the text. This strategy is validated by approaches like the MLM-BERT for Hindi [34], which addresses errors from the inflectional nature of language, achieving a measurable 3.58%-word accuracy improvement. Furthermore, highly generalized frameworks like the Learning with Noisy and Pseudo Label (LNPL) framework are explicitly designed to manage uncertainty and noisy data through joint learning on clean and noisy samples, proving to be robust and consistently outperforming alternatives [35].

To ensure the highest quality input for the OCR system, effective pre-processing is an essential first stage. The challenge of images captured in non-ideal conditions (e.g., extremely low-light text images) is addressed by Supervised Deep Curve Estimation models, which notably outperform state-of-the-art metrics on low-light datasets [36]. Even in standard environments, OCR often struggles with imprecise stroke edge extraction and the limitations of conventional convolutions. This issue is overcome by advanced techniques like Gated Convolutions-based Document Binarization (GDB), which successfully outperforms SOTA across all metrics on average by refining the visual features at the stroke level [37].

In the specific context of consumer product labels, OCR accuracy is further influenced by a combination of real-world factors that compound recognition difficulty. Multilingual label text, particularly the Indonesian-English mix common in local markets, introduces additional variability in character sets and tokenization. Label design factors such as small font sizes in nutrition tables, low-contrast printing on reflective packaging, and irregular text layouts create conditions that consistently degrade traditional OCR performance. Environmental factors during image capture, including varying store lighting, perspective distortion from handheld photography, and motion blur, further reduce recognition accuracy. These compounding challenges reinforce the necessity of the multi-stage approach synthesized in this review, where robust pre-processing, Transformer-based recognition, and BERT-based post-correction collectively address the full spectrum of real-world degradation encountered in consumer product label analysis.

Collectively, the synthesis confirms that the challenges inherent to real-world label images are mitigated through a three-pronged technological strategy: adopting transformer-based foundational architectures for core recognition, leveraging BERT/MLM models for linguistic post-correction, and utilizing specialized deep learning for visual pre-processing. This multi-stage approach successfully validates the requirements set forth in RQ1, establishing a highly robust output text necessary for the subsequent phase. The effectiveness of this initial pipeline, however, is merely a prerequisite; the true utility of the extracted text hinges on the capacity of Natural Language Processing (NLP) models to accurately interpret its semantic meaning and extract value-added entities, which is addressed in detail in the following section. Table V shows the synthesis of key methodologies for OCR robustness.

TABEL V
SYNTHESIS OF KEY METHODOLOGIES FOR OCR ROBUSTNESS

No	Methodology Contribution	Key Metric	Article
1	Transformer-Based Foundational OCR	Outperforms Tesseract and TrOCR on standard benchmarks; DTrOCR surpasses SOTA across printed, handwritten, and scene text	[29], [30], [31]
2	Contextual Post-Correction	3.58% word accuracy improvement (MLM-BERT); significant reduction in downstream model degradation	[32], [34], [35]
3	Vision-Based Pre-processing	Outperforms SOTA on low-light and binarization benchmarks	[36], [37]

B. Semantic Entity Extraction and Contextual Refinement

This section synthesizes the empirical evidence demonstrating the function and measured efficacy of BERT and Large Language Models (LLMs) in enhancing precision by transforming unreliable OCR output into structured, semantically coherent entities, which directly addresses RQ2. The analysis confirms that Transformer-based architectures are essential for achieving the semantic understanding required for accurate entity extraction, particularly when dealing with domain-specific terminology and complex document layouts.

The superiority of Transformer-based models is evident in the challenging clinical domain. Advanced architectures integrating BERT embeddings with sequential models like Bi-LSTM and CRF [38], [39] are the standard for high-performance Semantic Entity Extraction. For Chinese biomedical texts, this combination achieved a state-of-the-art F1-score of 82.1% [39], significantly outperforming previous methods. Even for complex tasks like classifying medical chief complaints based on hierarchical descriptions, models employing BERT and Hierarchical Relational Networks (HRNA) demonstrate strong reasoning capabilities, achieving a Macro-F1 score of 62.35% [38]. These results underscore the necessity of contextual embeddings provided by BERT for interpreting nuanced medical language.

Beyond clinical text, the models demonstrate high efficacy on semi-structured documents analogous to product labels. For extracting information from research articles, Layout-MetaBERT achieved the highest F1-score at 93.6% [40] compared to other BERT variants, highlighting the benefit of layout awareness. While simpler models like Logistic Regression can achieve high F1-scores (up to 98%) on basic invoice extraction tasks [41], complex layouts necessitate more advanced techniques. Multimodal frameworks like VS-MRC [42], which combine visual and semantic guidance, outperform traditional models by achieving an F1 score of 84.58% on social media NER tasks. For tabular data, crucial for nutritional information, the Split, Embed, and Merge architecture achieves an outstanding average F1-score of 97.11% on structured table recognition [43]. Furthermore, explicitly encoding spatial information using methods like Layout Quadrant Tags (LayoutQT) significantly boosts performance [44]; when applied with BERT, it improved the F1-score from 80.4% to 83.6% on document image datasets. This confirms that integrating layout information is critical for differentiating entities based on position. General-purpose BERT-based frameworks such as BERT-CRF have also demonstrated consistent outperformance over non-contextual baselines in biomedical NER tasks, further solidifying the architectural choice for entity extraction [45].

In the specific context of Indonesian consumer product labels, BERT-based entity extraction faces additional challenges related to multilingual synonym variation and local ingredient terminology. Ingredient names commonly appear in multiple forms across labels, for instance, 'sukrosa', 'gula pasir', and 'added sugar' are semantically equivalent but linguistically distinct, as are 'natrium' and 'sodium', or 'sirup glukosa-fruktosa' and 'high-fructose corn syrup (HFCS)'. Classical rule-based approaches such as regex or keyword matching fail to capture these variations reliably, particularly when terms appear in short, contextual phrases embedded within semi-structured nutrition tables. BERT's bidirectional contextual embeddings address this directly, by encoding each token in the context of its surrounding text, the model can distinguish that 'gula' in 'kandungan gula 10g' refers to a nutritional entity, while the same word in a product name carries a different semantic role. Combined with a bilingual synonym dictionary (Indonesian–English), fine-tuned BERT can systematically map diverse ingredient aliases to standardized entity categories, enabling consistent classification regardless of the linguistic variation present on the label.

The synthesized evidence consistently validates that the optimal strategy for precise entity extraction (RQ2) relies on Transformer-based models (BERT/LLM), particularly those enhanced with layout-aware or multimodal capabilities. This methodology effectively transforms potentially noisy OCR output into high-precision chemical entities, synonyms, and functional classifications necessary for downstream health risk assessment. Table VI shows the synthesis of key methodologies for extraction and refinement.

TABEL VI
SYNTHETIS OF KEY METHODOLOGIES FOR EXTRACTION AND REFINEMENT

No	Methodology Contribution	Key Metric	Article
1	Quantitative Extraction (Table)	F1-score 97.11% on structured table recognition	[43]
2	Layout-awareness	F1-score 93.6% (Layout-MetaBERT); F1 improved from 80.4% to 83.6% (LayoutQT)	[40], [44]
3	NER Clinical & Biomedical	F1-score 82.1% (Chinese biomedical); Macro-F1 62.35% (medical complaints)	[38], [39], [45]
4	Semi-structured	F1-score up to 98% (invoice); F1-score 84.58% (social media NER)	[41], [42]

C. End-to-End System Design for Health Risk Assessment

The design of a functional, real-time system for automated ingredient screening is validated by empirical evidence demonstrating its architectural requirements, efficiency, and successful domain transferability, thus fulfilling RQ3.

The system's feasibility is grounded in the proven capability of Multimodal Transformer Pipelines to convert semi-structured documents into structured data. OCR combined with Generative Transformer architectures (e.g., GPT-4o) achieves an F1-score of 98.2% and accuracy of 96.6% in converting unstructured surgical records [46], while simpler OCR + classification pipelines (SVM and Decision Trees) achieve 90% classification accuracy for nutritional claim verification [47]. Crucially, system integrity under noise is also established: Transformer Sequence-to-Sequence models maintain reliable entity extraction from noisy OCR output, with performance degradation limited to only 3.49%–5.23% [48], directly mitigating the primary risk of erroneous ingredient identification.

The methodology's applicability to health-related domains is further validated by dedicated systems such as NutriGuard [49] and an AI-Powered Ingredient Detector [50], which leverage LLMs to convert complex label data into actionable health intelligence for chronic disease management. Quantitative efficiency metrics further confirm deployment viability: a Zero-shot NLP Tool (GPT-3.5) reduces clinical report abstraction time from 93 seconds to 15 seconds per report [51], demonstrating the scalability needed for real-time consumer-facing applications. Table VII shows the synthesis of key methodologies for end-to-end system design.

TABEL VII
SYNTHESIS OF KEY METHODOLOGIES FOR END-TO-END SYSTEM DESIGN

No	Methodology Contribution	Key Metric	Article
1	Feasibility & Accuracy	F1-score 98.2%, accuracy 96.6% (GPT-4o); classification accuracy 90% (SVM)	[46], [47], [50]
2	Noise Robustness	Performance degradation ≤5.23% under significant OCR noise	[49], [51]
3	Domain Applicability	Health risk assessment for chronic disease management	[48], [49]
4	Efficiency & Scalability	Data abstraction time reduced from 93s to 15s per report	[51]

Figure 2 illustrates the proposed end-to-end framework for automated consumer product label analysis, integrating OCR and BERT-based NLP into a modular pipeline. The process begins when a consumer application, accessible via mobile or web interface, initiates the capture of a real-world label image. The image is first passed through the Vision Module, which consists of three sequential stages: pre-processing (contrast enhancement, deblurring, and adaptive resizing), a Transformer-based OCR engine for raw text extraction, and a BERT/MLM-based post-correction stage to mitigate residual linguistic noise introduced during OCR. The processed text is then forwarded to the Language Module, where a BERT-based Key Information Extraction (KIE) model performs entity spotting and Named Entity Recognition (NER) to identify ingredient-related entities. A normalization stage subsequently maps extracted entities to standardized forms using a bilingual synonym dictionary, informed by a knowledge base containing relevant nutrition regulations and classification thresholds. The final structured output, comprising health risk indicators derived from the extracted and classified ingredient data, is returned to the consumer application for display. This modular design ensures that each stage can be independently evaluated and optimized, supporting the layered evaluation protocol described in RQ1, RQ2, and RQ3.

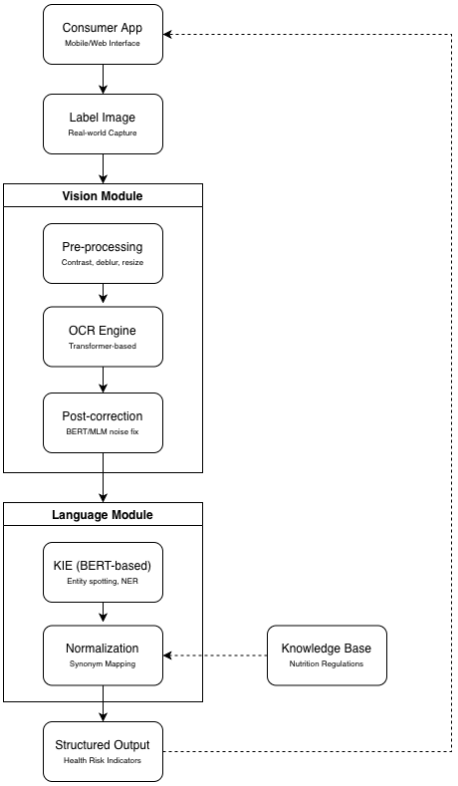

Figure 2. Proposed End-to-end Framework

D. Limitations

This study is subject to several limitations that should be considered when interpreting the findings. First, the literature search was restricted to three academic databases, such as IEEE Xplore, ScienceDirect, and Scopus, which, while comprehensive, may not capture all relevant studies published in other repositories or conference proceedings. Second, the search was conducted exclusively in English, potentially excluding relevant studies published in Indonesian that may be particularly relevant to the local consumer product label context. Third, as with any systematic review, publication bias may have influenced the findings, as studies reporting positive or significant results are more likely to be published than those reporting null or negative outcomes. Fourth, the search cutoff of October 2025 means that studies published after this date are not reflected in the synthesis. Finally, the proposed OCR–BERT framework is derived entirely from secondary literature and has not been empirically validated on real-world Indonesian consumer product labels, which limits the generalizability of the practical recommendations offered.

IV. CONCLUSION

This systematic review synthesizes evidence on the feasibility of an OCR–BERT pipeline for automated consumer product label analysis, directly addressing all three research questions. The findings confirm that robust text extraction (RQ1) requires Transformer-based Scene Text Recognition to mitigate visual noise, which must be followed by layout-aware BERT or LLM processing (RQ2) to achieve high-precision semantic entity extraction from semi-structured documents. Collectively, this integrated approach (RQ3) is supported by empirical evidence from analogous domains demonstrating substantial efficiency gains and high F1-scores in clinical and nutritional entity extraction tasks, suggesting its potential applicability to consumer product label analysis.

It is important to note that the proposed OCR–BERT framework represents a theoretical synthesis derived from existing literature, not an empirically validated system. The efficiency metrics cited (e.g., reduction of data abstraction time from 93 to 15 seconds) originate from studies in related domains and serve as indicative benchmarks rather than direct evidence of the framework's performance on Indonesian consumer product labels. Empirical validation, including real-world testing on multilingual food labels under varied imaging conditions, remains an essential direction for future research. Additionally, ethical and regulatory considerations, such as the accuracy of label interpretation and the potential for misinformation if the system misidentifies ingredient content, must be addressed prior to any public-facing deployment. Furthermore, while this framework is designed specifically for processed food and beverage labels in the Indonesian market, its underlying architecture, combining Transformer-based OCR with BERT-based KIE, may be adaptable to other consumer product domains such as cosmetics, supplements, or pharmaceuticals, provided that domain-specific synonym dictionaries, entity schemas, and regulatory rule-sets are developed accordingly. Such extensions represent promising directions for future research.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to Universitas Pembangunan Nasional "Veteran" Jakarta and to the supervising lecturers for providing the opportunity and guidance to conduct this preliminary research prior to the completion of the undergraduate thesis. Their continuous support and constructive feedback have been instrumental in shaping the direction and quality of this study.

REFERENCES

[1] Amran Pandjaitan, Muhammad Rizki, and Orlando Tenang Saputra Rumahorbo, "Pengaruh Label Dan Harga Terhadap Keputusan Pembelian Rokok," *JURNAL EKONOMI BISNIS DAN MANAJEMEN*, vol. 3, no. 1, pp. 271–280, Jan. 2025, doi: 10.59024/jise.v3i1.1106.

[2] U. Hasanah and B. Pambudi, "Pengaruh Kemasan dan Label terhadap Keputusan Pembelian," *Co-Value Jurnal Ekonomi Koperasi dan kewirausahaan*, vol. 14, Nov. 2023, doi: 10.59188/covalue.v14i6.3903.

[3] S. Albasyira, A. Rasyid, and A. I. Fahrika, "Pengaruh Label BPOM dan E-review Terhadap Keputusan Pembelian Cosmetics Care (Studi pada Generasi Z Muslim di Watampone)," 2025.

[4] D. Rahmah and I. Hasbi, "Pengaruh Citra Merek dan Label terhadap Minat Beli (Studi Kasus Mie Gacoan di Kota Bandung)," *Jurnal Samudra Ekonomi dan Bisnis*, vol. 14, pp. 544–554, Sep. 2023, doi: 10.33059/jseb.v14i3.8287.

[5] S. Aliffia and Kurniawati, "Pengaruh Tingkat Kepedulian Masyarakat Terhadap Label Nutrisi Pada Frozen Food Yang Berhubungan Kepada

Kesehatan Konsumen," *Jurnal Ekonomi Trisakti*, 2023, [Online]. Available: https://api.semanticscholar.org/CorpusID:258155171
[6] P. A. Gunawan and Y. S. Kunto, "Pengaruh Brand Image Dan Nutrition Label Terhadap Keputusan Pembelian Mie Instan Lemonilo: Efek Moderasi Orientasi Makanan Sehat," Apr. 2022.
[7] BPOM RI, *Pedoman Label Pangan Olahan-2020*. 2020.
[8] T. Lamont and M. McSweeney, "Consumer acceptability and chemical composition of whole-wheat breads incorporated with brown seaweed (Ascophyllum nodosum) or red seaweed (Chondrus crispus)," *J. Sci. Food Agric.*, vol. 101, no. 4, pp. 1507–1514, Mar. 2021, doi: https://doi.org/10.1002/jsfa.10765.
[9] A. Ndunge Charles, M. Mburu, D. Njoroge, and V. Zettel, "Chemical composition and consumer acceptability of oyster mushroom and sorghum-pearl millet based composite flours," *Discover Food*, vol. 4, no. 1, Dec. 2024, doi: 10.1007/s44187-024-00219-z.
[10] O. Olatunji, "Chemical Composition, Nutrient Bioavailability And Consumer Acceptability Of Cirina Forda (WESTWOOD) Larva-Enriched Vegetable Soups," 2021.
[11] F. Kusnandar, H. Danniswara, and A. Sutriyono, "Pengaruh Komposisi Kimia dan Sifat Reologi Tepung Terigu terhadap Mutu Roti Manis," *Jurnal Mutu Pangan : Indonesian Journal of Food Quality*, vol. 9, no. 2, pp. 67–75, Oct. 2022, doi: 10.29244/jmpi.2022.9.2.67.
[12] M. S. D. Taula'bi, Y. Y. Oessoe, and M. F. Sumual, "Kajian Komposisi Kimia Snack Bars Dari Berbagai Bahan Baku Lokal: Systematic Review," Jan. 2021.
[13] D. C. Arukwe, J. N. Okoli, C. A. Nwachukwu, and A. L. Kenechukwu, "Chemical Composition, Functional Properties and Consumer Acceptability of Cookies Produced from Blends of Wheat, Orange Fleshed Sweet Potato and Pawpaw Flours," *Sahel Journal of Life Sciences FUDMA*, vol. 3, no. 2, pp. 71–83, Jun. 2025, doi: 10.33003/sajols-2025-0302-09.
[14] Herviana Herviana, Haqqelni Nur Rosyidah, Amalina Rizma, Siska Pratiwi, Citra Dewi Anggraini, and Made Tantra Wirakesuma, "Analisis Pengaruh Sikap terhadap Kesehatan dan Label Kepada Kepatuhan Membaca Label Produk Pangan pada Mahasiswa Gizi Provinsi Kepulauan Riau," *Jurnal Riset Ilmu Kesehatan Umum dan Farmasi (JRIKUF)*, vol. 2, no. 1, pp. 137–143, Jan. 2024, doi: 10.57213/jrikuf.v2i1.159.
[15] F. Milita, S. Handayani, and B. Setiaji, "Kejadian Diabetes Mellitus Tipe II pada Lanjut Usia di Indonesia (Analisis Riskesdas 2018)," *Jurnal Kedokteran dan Kesehatan*, vol. 17, pp. 9–20, Feb. 2021, doi: 10.24853/jkk.17.1.9-20.
[16] F. R. Muharram, J. B. Swannjo, R. R. Melbiarta, and S. Martini, "Trends of diabetes and pre-diabetes in Indonesia 2013–2023: a serial analysis of national health surveys," *BMJ Open* , vol. 15, no. 9, Sep. 2025, doi: 10.1136/bmjopen-2024-098575.
[17] S. Moradi *et al.*, "Ultra-processed food consumption and adult diabetes risk: A systematic review and dose-response meta-analysis," Dec. 01, 2021, *MDPI*. doi: 10.3390/nu13124410.
[18] M. I. Almarshad, R. Algonaiman, H. F. Alharbi, M. S. Almujaydil, and H. Barakat, "Relationship between Ultra-Processed Food Consumption and Risk of Diabetes Mellitus: A Mini-Review," *Nutrients*, vol. 14, no. 12, 2022, doi: 10.3390/nu14122366.
[19] S. Sinha and M. Haque, "Obesity, Diabetes Mellitus, and Vascular Impediment as Consequences of Excess Processed Food Consumption," *Cureus*, Sep. 2022, doi: 10.7759/cureus.28762.
[20] A. Mahajan, A. Deshmane, and A. Muley, "A Comparative Study on the Consumption Patterns of Processed Food Among Individuals With and Without Type 2 Diabetes," *International Journal of Public Health* , vol. 70, 2025, doi: 10.3389/ijph.2025.1607931.
[21] H. Wang, C. Pan, X. Guo, C. Ji, and K. Deng, "From object detection to text detection and recognition: A brief evolution history of optical character recognition," *WIREs Computational Statistics*, vol. 13, no. 5, p. e1547, Sep. 2021, doi: https://doi.org/10.1002/wics.1547.
[22] M. A. M. Salehudin *et al.*, "Analysis of Optical Character Recognition using EasyOCR under Image Degradation," in *Journal of Physics: Conference Series*, Institute of Physics, 2023. doi: 10.1088/1742-6596/2641/1/012001.
[23] U. Salimah, V. Maharani, and R. Nursyanti, "Automatic License Plate Recognition Using Optical Character Recognition," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1115, no. 1, p. 012023, Mar. 2021, doi: 10.1088/1757-899x/1115/1/012023.
[24] X. Wang *et al.*, "Intelligent Micron Optical Character Recognition of DFB Chip Using Deep Convolutional Neural Network," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–9, 2022, doi: 10.1109/TIM.2022.3154831.
[25] N. Sarika, N. Sirisala, and M. S. Velpuru, "CNN based Optical Character Recognition and Applications," in *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, 2021, pp. 666–672. doi: 10.1109/ICICT50816.2021.9358735.
[26] M. S. Kasem, M. Mahmoud, and H.-S. Kang, "Advancements and Challenges in Arabic Optical Character Recognition: A Comprehensive Survey," Dec. 2023, [Online]. Available: http://arxiv.org/abs/2312.11812
[27] Hubert, P. Phoenix, R. Sudaryono, and D. Suhartono, "Classifying Promotion Images Using Optical Character Recognition and Naïve Bayes Classifier," *Procedia Comput. Sci.*, vol. 179, pp. 498–506, 2021, doi: https://doi.org/10.1016/j.procs.2021.01.033.
[28] R. Hemnath, "Integrating Natural Language Processing with BERT and LSTM for Employee Sentiment Analysis in HRM," 2025. [Online]. Available: www.ijahss.com
[29] C. Gunasekara, Z. Hamel, F. Du, and C. Baillie, "TokenOCR: An Attention Based Foundational Model for Intelligent Optical Character Recognition," in *International Conference on Pattern Recognition Applications and Methods*, M. Castrillon-Santana, M. De Marsico, and A. Fred, Eds., Science and Technology Publications, Lda, 2025, pp. 151–158. doi: 10.5220/0013340100003905.
[30] M. Fujitake, "DTrOCR: Decoder-only Transformer for Optical Character Recognition," Institute of Electrical and Electronics Engineers Inc., 2024, pp. 8010–8020. doi: 10.1109/WACV57701.2024.00784.
[31] C. Wang, Y. Yang, and M. Hu, "Scene Text Detection Method Based on Multimodal and Multi-Task Optimization," in *2025 7th International Conference on Natural Language Processing (ICNLP)*, 2025, pp. 650–653. doi: 10.1109/ICNLP65360.2025.11108565.
[32] K. Todorov and G. Colavizza, "An Assessment of the Impact of OCR Noise on Language Models," in *International Conference on Agents and Artificial Intelligence*, A. Rocha, L. Steels, and J. van den Herik, Eds., Science and Technology Publications, Lda, 2022, pp. 674–683. doi: 10.5220/0010945100003116.
[33] U. Kumaran, D. Biswas, B. Sneha, S. Nadipalli, and S. Raja, "Text Post-processing on Optical Character Recognition output using Natural Language Processing Methods," Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/MysuruCon59703.2023.10396964.
[34] T. Kundaikar, S. Fadte, R. Karmali, and J. D. Pawar, "Automatic Hindi OCR Error Correction Using MLM-BERT," *Ingenierie des Systemes d'Information*, vol. 29, no. 2, pp. 619–626, 2024, doi: 10.18280/isi.290223.
[35] M. Ahmed *et al.*, "Towards Robust Learning with Noisy and Pseudo Labels for Text Classification," *Inf. Sci. (N. Y).*, vol. 661, p. 120160, 2024, doi: https://doi.org/10.1016/j.ins.2024.120160.
[36] C.-T. Lin *et al.*, "Text in the dark: Extremely low-light text image enhancement," *Signal Process. Image Commun.*, vol. 130, p. 117222, 2025, doi: https://doi.org/10.1016/j.image.2024.117222.
[37] Z. Yang, B. Liu, Y. Xiong, and G. Wu, "GDB: Gated Convolutions-based Document Binarization," *Pattern Recognit.*, vol. 146, p. 109989, 2024, doi: https://doi.org/10.1016/j.patcog.2023.109989.
[38] Z. Zhang, Z. Lu, J. Liu, and R. Bai, "Medical chief complaint classification with hierarchical structure of label descriptions," *Expert Syst. Appl.*, vol. 252, p. 123938, 2024, doi: https://doi.org/10.1016/j.eswa.2024.123938.
[39] M. Zhou, J. Tan, S. Yang, H. Wang, L. Wang, and Z. Xiao, "Ensemble Transfer Learning on Augmented Domain Resources for Oncological Named Entity Recognition in Chinese Clinical Records," *IEEE Access*, vol. 11, pp. 80416–80428, 2023, doi: 10.1109/ACCESS.2023.3299824.
[40] J. Choi, H. Kong, H. Yoon, H. Oh, and Y. Jung, "LAME: Layout-Aware Metadata Extraction Approach for Research Articles," *Computers, Materials and Continua*, vol. 72, no. 2, pp. 4019–4037, 2022, doi: https://doi.org/10.32604/cmc.2022.025711.

[41] H. T. Ha and A. Horák, "Information extraction from scanned invoice images using text analysis and layout features," *Signal Process. Image Commun.*, vol. 102, p. 116601, 2022, doi: https://doi.org/10.1016/j.image.2021.116601.

[42] J. Hu, Y. Lyu, and Y. Xue, "VS-MRC: A visual semantics-guided machine reading comprehension framework for multimodal named entity recognition with multiple images," *Knowl. Based. Syst.*, vol. 326, p. 114024, 2025, doi: https://doi.org/10.1016/j.knosys.2025.114024.

[43] Z. Zhang, J. Zhang, J. Du, and F. Wang, "Split, Embed and Merge: An accurate table structure recognizer," *Pattern Recognit.*, vol. 126, p. 108565, 2022, doi: https://doi.org/10.1016/j.patcog.2022.108565.

[44] P. M. L. de Lucena Drumond, L. P. Leite, T. E. de Campos, and F. A. Braz, "LayoutQT—Layout Quadrant Tags to embed visual features for document analysis," *Eng. Appl. Artif. Intell.*, vol. 122, p. 106091, 2023, doi: https://doi.org/10.1016/j.engappai.2023.106091.

[45] X. Ye, T. Shi, D. Huang, and T. Sakurai, "Multi-Omics clustering by integrating clinical features from large language model," *Methods*, vol. 239, pp. 64–71, 2025, doi: https://doi.org/10.1016/j.ymeth.2025.03.017.

[46] X. Yang *et al.*, "Cross language transformation of free text into structured lobectomy surgical records from a multi center study," *Sci. Rep.*, vol. 15, no. 1, 2025, doi: 10.1038/s41598-025-97500-7.

[47] S. I. S. P, K. L, P. Bhoomika, K. Tejaswi, L. S. Khande, and N. S. Vemishetti, "Automating Nutritional Claim Verification: The Role of OCR and Machine Learning in Enhancing Food Label Transparency," in *2024 International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS)*, 2024, pp. 1164–1171. doi: 10.1109/ICICNIS64247.2024.10823177.

[48] P. A. Villa-García, R. Alonso-Calvo, and M. García-Remesal, "End-to-end entity extraction from OCRed texts using summarization models," *Neural Comput. Appl.*, vol. 36, no. 35, pp. 22347–22363, 2024, doi: 10.1007/s00521-024-10422-9.

[49] M. A. Hakim, R. A. Ifty, K. E. Delowar, S. H. Chowdhury, I. Rashid, and M. Shakib, "Nutriguard: LLM-Driven Nutritional Assessment for Chronic Disease Prevention," in *2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)*, 2025, pp. 1–6. doi: 10.1109/QPAIN66474.2025.11171750.

[50] S. K, A. R, C. R, and N. R, "AI-Powered Ingredient Detector for Allergies: Enhancing Food Safety Through Natural Language Processing and Computer Vision," in *2025 International Conference on Computing and Communication Technologies (ICCCT)*, 2025, pp. 1–5. doi: 10.1109/ICCCT63501.2025.11019718.

[51] B. Kaufmann *et al.*, "Validation of a Zero-shot Learning Natural Language Processing Tool to Facilitate Data Abstraction for Urologic Research," *Eur. Urol. Focus*, vol. 10, no. 2, pp. 279–287, 2024, doi: https://doi.org/10.1016/j.euf.2024.01.009.

.