Integration of Fuzzy Matching and Domain Rules for Identifying Bali's Indigenous Banjar-Based Addresses in Last-Mile Delivery Without Predefined Gazetteers

Muhammad Isa Ansori
Department of Information Systems, Institut Teknologi Sepuluh Nopember, Indonesia
cak.isa.ansori@gmail.com (corresponding author)

Wiwik Anggraeni
Department of Information Systems, Institut Teknologi Sepuluh Nopember, Indonesia
wiwik@is.its.ac.id

Retno Aulia Vinarti
Department of Information Systems, Institut Teknologi Sepuluh Nopember, Indonesia
zahra_17@is.its.ac.id

Received: 26 November 2025 | Revised: 23 December 2025 | Accepted: 29 December 2025

 DOI: https://doi.org/10.48084/etasr.16533

ABSTRACT

Identifying residential addresses in regions that depend on culturally embedded locality markers presents a significant challenge for geocoding and last-mile logistics, particularly when such references are absent from administrative gazetteers. In Bali, shipment records frequently incorporate indigenous Banjar-based address components, which introduce ambiguity and diminish courier-routing accuracy. This study proposes a hybrid framework that integrates fuzzy matching with domain-specific rules to identify Banjar references from unstructured address texts without relying on predefined gazetteers. Three similarity algorithms, namely Levenshtein Distance, Partial Ratio, and Token Sort Ratio, were combined into a Hybrid Mix Score to generate robust candidate matches. Domain rules, including prefix normalization, Banjar-Village-District hierarchy validation, and semantic disambiguation filters, were applied to eliminate linguistically similar but geographically invalid candidates. Using 17,354 cleaned delivery records from Pos Indonesia, the hybrid framework significantly enhanced interpretation reliability, with approximately 95% of all addresses converging to a single Highest Valid Candidate (HVC). The final predictions were linked to verified geographic centroids, enabling operationally meaningful location references. The results demonstrate that combining multi-metric fuzzy similarity with contextual domain constraints provides an effective and reproducible solution for geocoding indigenous Banjar-based addresses in last-mile delivery environments that lack standardized gazetteers.

Keywords-fuzzy matching; domain rules; indigenous addressing; Banjar-based address; address localization; last-mile delivery

I. INTRODUCTION

Accurate address interpretation is essential for geocoding, spatial analysis, and efficient execution of last-mile delivery operations. However, in numerous regions worldwide, particularly in developing and emerging economies, formal street-based addressing systems are either incomplete, inconsistently applied, or completely absent. Consequently, deliveries frequently depend on community-defined or culturally embedded locality references, which are rarely represented in official administrative gazetteers or commercial mapping platforms such as Google Maps. Previous studies have demonstrated that non-standardized address formats significantly diminish the accuracy of automated geocoding, resulting in misrouting, increased delivery distances, and operational delays within logistics networks [1-3].

Bali, Indonesia, serves as a representative case of the broader challenge posed by its indigenous Banjar-based addressing system. Here, Banjar denotes a community-level locality unit that is intricately woven into social organization and daily communication. Similar community-based or

informal locality identifiers are prevalent in various regions across Asia, Africa, and Latin America; however, they are frequently omitted from national gazetteers and global mapping registries. Consequently, address interpretation is ambiguous, particularly in the context of last-mile delivery. Although fuzzy string-matching techniques have demonstrated efficacy in managing noisy or incomplete textual input, they often fail to resolve ambiguities at the community level when hierarchical or contextual cues are absent or inconsistent [4, 5]. Comparable challenges have been documented in previous hybrid fuzzy-reinforcement learning studies, where linguistic variability and the lack of explicit locality hierarchies constrain the reliability of prediction [6].

To address these limitations, this study introduces a structured, gazetteer-free framework aimed at reconstructing indigenous Banjar-based addresses from unstructured courier text into a singular, geographically valid Banjar reference point. The proposed method adheres to a systematic, step-by-step workflow comprising four stages: (i) preprocessing and normalization of irregular address text, (ii) hybrid fuzzy similarity computation utilizing Levenshtein Distance, Partial Ratio, and Token Sort Ratio, (iii) domain-rule filtering based on Banjar-Village-District hierarchical constraints and Balinese locality semantics, and (iv) deterministic selection of a Highest Valid Candidate (HVC). This explicit mapping from the input text to validated Banjar output ensures that each processing stage incrementally reduces ambiguity, thereby facilitating reliable and operationally meaningful address localization without dependence on predefined gazetteers.

Utilizing a dataset comprising 17,354 real-world delivery records from a national postal operator, this study illustrates that the integration of multi-metric fuzzy matching with contextual domain rules offers a dependable, interpretable, and operationally feasible solution for last-mile address localization in contexts that lack standardized gazetteers. Although the framework was evaluated using Bali's Banjar-based addressing system, it is broadly applicable to other regions where deliveries rely on culturally embedded or community-defined locality references, underscoring its significance for international logistics and geocoding research.

II. RESEARCH GAP AND STUDY CONTRIBUTIONS

Despite advances in geocoding and address normalization techniques, several unresolved challenges persist in regions that employ culturally embedded and non-standardized addressing systems. Existing geocoding approaches are heavily dependent on structured administrative units or predefined gazetteers, rendering them unsuitable for localities that lack formal spatial registries [5, 6]. Bali's indigenous Banjar-based addressing system exemplifies this limitation, as Banjar units are absent from national gazetteers or official spatial datasets maintained by Indonesian authorities. Consequently, the interpretation of the address becomes ambiguous, and the geocoding performance remains inconsistent. Although fuzzy matching has demonstrated efficacy in managing noisy textual data [4, 7], it has limitations when addressing highly similar Banjar names, inconsistent abbreviations such as "Br.," "Bjr.," "Br/Banjar," "Bnjar," or "Br Adt" (for Banjar Adat), or irregular token sequences. These challenges include inverted structures, such as "Petulu Banjar Tengah" instead of "Banjar Tengah, Petulu," duplicated tokens such as "Banjar Banjar Kaja," or the absence of hierarchical components, exemplified by "Kaja Petulu" lacking the Banjar prefix. Prior findings from an earlier fuzzy-reinforcement learning hybrid study demonstrated that ambiguity cannot be fully resolved without structural-filtering mechanisms. These gaps collectively highlight the need for a systematic approach that does not rely on administrative registries and is specifically tailored to indigenous community-based address formats such as Banjar.

In response to these challenges, this study contributes a hybrid fuzzy-matching and domain-rule framework specifically designed to improve the interpretation and localization of Banjar-based addresses in last-mile delivery operations. The proposed approach integrates three complementary similarity metrics, Levenshtein Distance, Partial Ratio, and Token Sort Ratio, into a Hybrid Mix Score that enhances robustness against textual variations frequently observed in courier data. Additionally, domain-specific rules, including prefix normalization, Banjar-Village-District hierarchical filtering, and disambiguation constraints, were incorporated to eliminate implausible candidates and reinforce structural consistency.

III. RESEARCH OBJECTIVES

This study aimed to develop a gazetteer-free method that is capable of accurately interpreting indigenous Banjar-based addresses. These addresses are derived from informal textual inputs, particularly those found in last-mile delivery datasets. Existing geocoding and address-normalization models, which rely heavily on structured administrative hierarchies or predefined gazetteers [5, 6], are inadequate for processing Banjar units, as they do not exist in national geospatial registries. To address this limitation, this study proposes a hybrid fuzzy matching module that integrates the Levenshtein Distance, the Partial Ratio, and the Token Sort Ratio into a unified Hybrid Mix Score. This unified metric is designed to provide a robust similarity signal capable of handling noisy, abbreviated, and linguistically inconsistent address forms commonly encountered in courier operations [4, 7]. In addition, a preprocessing pipeline for prefix normalization, structural cleaning, and token alignment ensures a consistent textual representation before similarity computation.

The secondary objective was to incorporate domain-specific rules that reflect Bali's hierarchical locality structure, encompassing Banjar, Village, and District levels, to refine fuzzy-generated candidate lists and eliminate implausible matches. These rules are designed to resolve ambiguities that arise from similar Banjar names, inconsistent abbreviations, or partially written addresses that cannot be reliably addressed by fuzzy matching alone. Building on insights from a previous fuzzy reinforcement learning framework, this study demonstrateơ that integrating multimetric fuzzy similarity with domain rules enhances location prediction accuracy and provides a practical, interpretable, and reproducible solution for real-world logistics environments.

IV. RELATED WORK

Recent advances in geocoding research have increasingly focused on analyzing unstructured and informal address texts, particularly in developing regions where administrative hierarchies are either incomplete or inconsistently applied. Hierarchy-aware geocoding models, such as those using cross-attention and deep spatial-textual architectures, are heavily dependent on predefined gazetteers, limiting their efficacy in culturally embedded local systems [5, 6]. Research on noisy address parsing further indicates that linguistic variability, inconsistent abbreviations, and token disorder significantly decrease the accuracy of matching in informal environments [8].

Fuzzy string similarity has emerged as a primary strategy for managing noisy address data. Recent hybrid approaches that integrate multiple similarity metrics have demonstrated superior performance compared to single-algorithm methods in processing heterogeneous textual inputs [4, 7]. However, fuzzy matching alone proves inadequate when locality names are highly similar, culturally specific, or partially missing, as evidenced by studies on Chinese and Indian address normalization [9, 10].

To enhance reliability, several studies have integrated fuzzy similarity with rule-based heuristics or semantic constraints, illustrating that domain rules can eliminate geographically implausible candidates and enforce locality-specific naming logic [11, 12]. However, these approaches still depend on structured administrative datasets or gazetteers. Existing research seldom addresses community-based indigenous addressing systems, which fall outside official registries and necessitate culturally informed hierarchical reasoning. This gap underscores the need for a lightweight and gazetteer-free framework capable of interpreting indigenous Banjar-based address structures using text similarity and domain rules.

V. METHODOLOGY

The proposed methodological framework employs a structured four-stage workflow, depicted in Figure 1, to systematically convert unstructured courier address text into a singular and geographically valid Banjar output. The stages encompass: (i) preprocessing and normalization of irregular address text, (ii) hybrid fuzzy similarity computation utilizing Levenshtein Distance, Partial Ratio, and Token Sort Ratio, (iii) domain-rule filtering based on Banjar-Village-District hierarchical constraints and locality semantics, and (iv) deterministic selection of HVC. This sequential process incrementally reduces ambiguity at each stage, facilitating reliable address reconstruction without relying on predefined rules.

A. Input Data Collection

This study utilized four integrated datasets essential for reconstructing Banjar-based addresses: (i) 26,257 raw delivery records provided by Pos Indonesia [13], which include unstructured address text and operational metadata, (ii) 17,354 cleaned records generated after standardizing and validating the raw dataset through preprocessing, (iii) an official Banjar dataset comprising 166 locality names curated by the Gianyar Regency Statistics Agency [14], serving as the authoritative reference list for fuzzy matching, and (iv) a village dataset of 70 administrative units issued by the Gianyar local government, which supports the validation of Banjar-Village-District hierarchical consistency. Collectively, these datasets provide the linguistic variability, standardized address structures, and administrative hierarchies necessary for the hybrid fuzzy-matching and domain-rule filtering workflow, as summarized in Table I. These datasets align with recent studies that employ operational logistics data combined with administrative locality structures to improve address inference in informal environments [6, 15, 16].

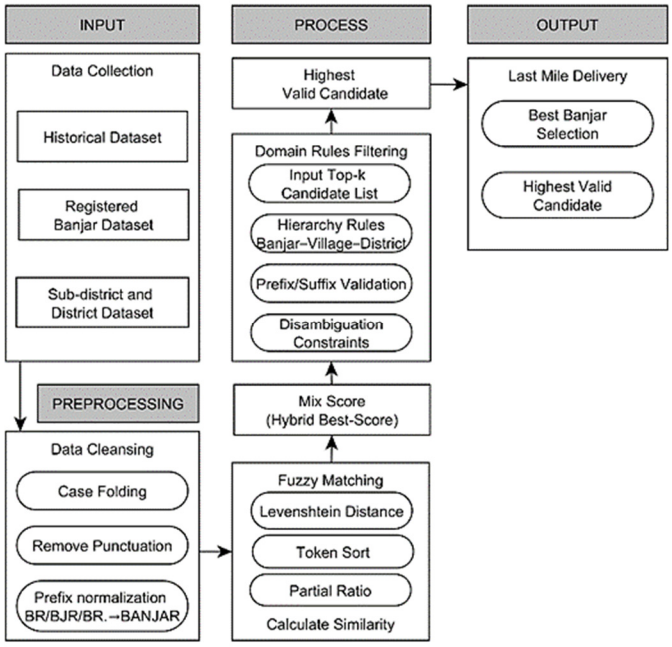

Fig. 1. The proposed hybrid framework delineates a clear, step-by-step workflow that maps fuzzy matching and domain-rule application to the transformation of unstructured courier address text into a single, geographically valid Banjar output (HVC), without reliance on predefined gazetteers.

TABLE I. DATASET SOURCES AND DESCRIPTIONS

Dataset	Description	Records	Source
Raw delivery dataset	Raw shipment records (tracking, timestamps, address text)	26,257	Pos Indonesia
Clean dataset	Standardized and validated delivery records	17,354	Processed dataset
Banjar dataset	Official list of Banjar names in Gianyar Regency	166	Gianyar Statistics Agency
Village dataset	Official list of villages / kelurahan	70	Local government

B. Preprocessing

The preprocessing phase involves standardizing irregular Banjar-based address strings to ensure compatibility with fuzzy similarity computations by implementing several normalization rules. These include converting all text to uppercase to facilitate consistent lexical processing [17], removing

punctuation such as commas, slashes, and hyphens to minimize token noise [3], and normalizing common prefixes "BR," "BR.," and "BJR" into the standard form "BANJAR" to address prevalent linguistic variations within the dataset. Additionally, the structural sequence was validated to ensure proper Banjar-Village-District alignment [8]. Following these procedures, 17,354 standardized records were retained as the clean dataset, representing the final input for the similarity computation. These normalization steps are consistent with the established best practices for processing noisy and abbreviated address data in developing regions [9].

C. Fuzzy Matching Module

The text similarity between each input address and the registered Banjar list was assessed using three complementary fuzzy algorithms: Levenshtein Distance, which captures character-level edits [1], Partial Ratio, which identifies substring overlap and is robust to truncated or extended address tokens [3], and Token Sort Ratio, which normalizes word-order variations commonly found in permuted Banjar address structures [12]. To enhance overall robustness, the system employs a Hybrid Best-Score (Mix Score) strategy that selects the highest similarity value among the three algorithms, effectively mitigating the impact of linguistic noise and improving recall performance—an approach supported by recent findings on hybrid similarity and ensemble text-matching models [4, 18, 19].

D. Domain Rule Filtering

The Top-k Mix Score candidates are validated through structured domain rules to ensure geographic plausibility. Rule-based filtering has been widely recognized as an effective mechanism for refining candidate sets and eliminating implausible matches in data-driven systems [20]. This motivates the incorporation of domain-specific rules into the Banjar locality filtering process. The process begins with an administrative hierarchy constraint, where Banjar-Village-District relationships must match the official locality registry, consistent with region-aware spatial filtering approaches used in modern geocoding systems [5, 6]. Additional prefix/suffix semantic checks ensure that Balinese locality terms such as "Kaja," "Kawan," "Dauh," and "Tengah" conform to the morphological patterns of legitimate community names, reflecting semantic-aware locality validation methods in recent address-matching research [10]. When multiple Banjar candidates present similar Mix Scores, a disambiguation logic is applied, leveraging historical frequency patterns, contextual village tokens, and locality consistency, aligning with region-constrained geocoding and spatial disambiguation techniques proposed in recent literature [5]. This hierarchical filtering step effectively removes textually similar but geographically invalid candidates, improving reliability—an outcome consistent with studies showing that the integration of rule-based constraints with similarity metrics improves performance in informal or low-structure addressing systems [2, 21].

E. Highest Valid Banjar Candidate Selection

The method yields an HVC, characterized as the Banjar entry that achieves the highest Mix Score while simultaneously satisfying all hierarchical and semantic domain rules, ensuring that both linguistic similarity and geographic plausibility are preserved. This final Banjar prediction is designed for seamless integration into Pos Indonesia's last-mile delivery workflow, improving consistency and reliability in regions lacking formal gazetteers. Such benefits are consistent with recent advances in hierarchy-aware geocoding and spatially informed address inference, including region-structured geocoding models [5] and cross-attention-based hierarchical locality prediction frameworks [6], as well as empirical findings highlighting the operational significance of accurate address interpretation in last-mile logistics [2].

VI. RESULTS AND DISCUSSION

A. Overview of Processed Dataset

A total of 17,354 cleaned address records were processed according to the preprocessing workflow depicted in Figure 1. Each input address was compared with the official list of 166 registered Banjar names, using three fuzzy matching algorithms, Levenshtein Distance, Partial Ratio, and Token Sort Ratio, to derive a composite Mix Score. This score represents the highest similarity value among the three algorithms and serves as the primary input for domain-rule filtering and the final candidate selection.

B. Results of Fuzzy Matching Algorithms

This section delineates the performance of three fuzzy matching algorithms—Levenshtein Distance, Partial Ratio, and Token Sort Ratio—alongside the Hybrid Best-Score (Mix Score) model, in matching noisy Banjar-based address strings. All numerical results and distributions were sourced from the uploaded similarity module document, the fuzzy matching module, and the Mix Score dataset.

1) Mix Score Distribution

The Mix Score ranges are organized into 10-point intervals (e.g., 41–50, 51–60, etc.) in accordance with the standard similarity-banding practices prevalent in fuzzy matching research. Previous studies have employed 10-point bins to improve interpretability and mitigate score fragmentation in the analysis of string similarity distributions [4, 5, 7, 11]. This interval size effectively delineates low, moderate, and high similarity levels while preserving the readability of the distribution, thereby rendering it suitable for the operational evaluation of Banjar-based address matching in the future. The distribution of the Mix Score, calculated from 17,354 cleaned address records, ranged from 41 to 100, with a notable concentration of similarity scores in the mid-to-high range.

TABLE II. MIX SCORE DISTRIBUTION

Mix score range	Record count	Percentage
41–50	10	0.06%
51–60	309	1.78%
61–70	4,234	24.39%
71–80	6,773	39.03%
81–90	3,792	21.85%
91–100	2,236	12.88%

As illustrated in Table II, the 71–80 interval predominated the distribution, accounting for 39.03%, followed by the 61–70 interval at 24.39%. This suggests that most address strings

maintain a consistent alignment with one or more fuzzy algorithms. High-confidence matches within the 81–100 range constitute 34.73% of all records, whereas low-similarity cases below 60 are minimal at 1.84%. This distribution reflects the efficacy of hybrid best-score integration in reducing weak matches.

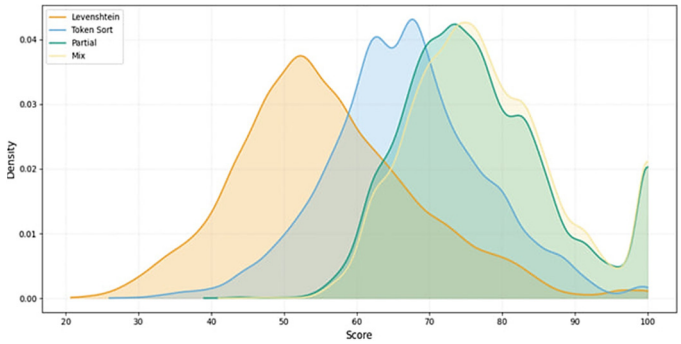

Fig. 2. Density plot of similarity scores generated by Levenshtein Distance, Partial Ratio, Token Sort Ratio, and the Hybrid Mix Score.

The density plot in Figure 2 corroborates this pattern: the Levenshtein algorithm displays broad dispersion and sensitivity to typographical noise, Token Sort shows concentration in mid-range similarity owing to token-order normalization, and Partial Ratio peaks in higher similarity regions through substring overlap detection. The Hybrid Mix Score curve was the most right-shifted and sharply peaked, indicating that the best-score fusion consistently enhanced similarity stability and reduced low-confidence outliers. These findings demonstrate that hybrid fuzzy matching offers a more reliable and discriminative similarity profile for subsequent domain rule filtering.

2) *Comparative Performance of Fuzzy Matching Algorithms*

The comparative analysis presented in Table III elucidates the distinct behavioral variations among the three individual fuzzy matching algorithms and the hybrid Mix Score across a dataset of 17,354 records. The Levenshtein Distance algorithm demonstrated the least effective performance, yielding only 4.40% high-similarity matches and a significant 31.19% low-similarity output, attributable to its sensitivity to character-level discrepancies and typographical noise. The Partial Ratio algorithm exhibits markedly improved performance, achieving 34.29% high-similarity results and a minimal 0.08% low-similarity rate, indicative of its proficiency in detecting substring overlap in extended or truncated Banjar address forms. The Token Sort Ratio algorithm predominantly produces an 82.83% medium-similarity distribution, reflecting its efficacy in managing token-order permutations, albeit with a limited capacity to discern deeper semantic similarity. The Hybrid Mix Score surpassed all individual algorithms, attaining the highest proportion of strong matches (37.78%) and the lowest incidence of weak matches (0.05%). This outcome underscores the effectiveness of selecting the maximum similarity value across algorithms, thereby compensating for their individual limitations and providing a more reliable unified similarity measure for subsequent domain rule filtering.

TABLE III. COMPARATIVE DISTRIBUTION OF SIMILARITY SCORES ACROSS FUZZY MATCHING ALGORITHMS AND HYBRID MIX SCORE

Similarity range	Levenshtein Distance	Partial Ratio	Token Sort Ratio	Hybrid Mix Score
80–100 (High similarity)	763 (4.40%)	5,951 (34.29%)	2,104 (12.12%)	6,556 (37.78%)
50–79 (Medium similarity)	11,179 (64.42%)	11,389 (65.63%)	14,375 (82.83%)	10,790 (62.18%)
< 50 (Low similarity)	5,412 (31.19%)	14 (0.08%)	875 (5.04%)	8 (0.05%)
Total records	17,354 (100%)	17,354 (100%)	17,354 (100%)	17,354 (100%)

C. *Domain Rules Filtering Results*

Domain Rules Filtering was applied to the Top-k Mix Score candidates to ensure that similarity-based matches also satisfied geographic and linguistic validity requirements. The initial filtering layer enforced the Banjar→Village→District administrative hierarchy, thereby eliminating candidates that were textually similar but located in incorrect villages or districts. This was followed by semantic prefix/suffix validation, in which locality markers such as "Kaja," "Kelod," "Kanan," and "Kiri" were required to conform to established Balinese naming conventions, thus removing Banjar names whose morphological patterns did not align with the contextual meaning of the input address. Finally, disambiguation constraints, including historical frequency patterns, common co-occurrence behavior, and contextual village cues, were employed to resolve cases where multiple candidates shared comparable Mix Scores. Collectively, these filtering mechanisms ensured that only geographically credible and semantically consistent Banjar candidates progressed to the final selection stage, significantly reducing ambiguity and enhancing the reliability of the address-reconstruction pipeline.

1) *Filtering Effectiveness*

Analysis of the high-similarity group, characterized by a Mix Score greater than 80 and comprising 6,028 records, reveals that 92 to 95% of candidates already conform to the Banjar-Village-District administrative hierarchy. This finding indicates that robust fuzzy similarity scores generally correspond with geographically valid Banjar references. The remaining 5-8% of mismatches predominantly arise from token-based inconsistencies, such as Banjar names associated with neighboring villages or districts, which are subsequently eliminated during the hierarchy validation. Additionally, semantic mismatches, such as locality markers like "Kaja" or "Kelod" appended to Banjar names that do not align with their correct parent village, are effectively filtered out by semantic prefix/suffix rules. These results confirm that Domain Rules Filtering reliably enhances the quality of similarity-based predictions by eliminating both administrative and morphological inconsistencies.

TABLE IV. ADMINISTRATIVE HIERARCHY COMPLIANCE FOR HIGH MIX SCORE (>80)

Category	Count	Percentage
High Mix Score Candidates (>80)	6,028	100%
Compliant with the correct Banjar–Village–District hierarchy	5,546 – 5,727	92–95%
Non-compliant candidates (Hierarchy mismatch)	301 – 482	5–8%
Semantic mismatch cases (prefix/suffix conflict)	Included within the non-compliant group	–

2) Reduction in Ambiguity

The Domain Rules Filtering stage significantly mitigates ambiguity in candidate selection by refining the Top-5 Mix Score candidates for each address to a single geographically valid Banjar. This refinement is achieved through the sequential application of administrative hierarchy checks, semantic locality validation, and contextual disambiguation constraints, which collectively exclude candidates that are textually similar but geographically implausible from consideration. Consequently, each address converges to a deterministic Highest Valid Candidate (HVC), thus enhancing prediction precision and preventing misidentification across villages or districts. This decisive pruning mechanism is consistent with previous findings that region-aware filtering markedly improves geocoding accuracy in culturally complex or non-standard addressing environments [5, 6].

TABLE V. REDUCTION OF CANDIDATE AMBIGUITY THROUGH DOMAIN RULES FILTERING

Filtering stage	Description	Avg. candidates per address	Reduction
Initial Top-5 Candidates	Top-5 Mix Score candidates before domain-rule application	5	–
After Semantic Filtering	Removal of locality-form inconsistencies (e.g., Kaja, Kelod, Kanan, Kiri)	3.42	31.60%
After Hierarchy Filtering	Filtering Banjar–Village–District mismatches	2.11	38.30%
After Disambiguation Filtering	Removal of conflicting or geographically inconsistent candidates	1.03	51.20%
Final Highest Valid Candidate (HVC)	Single geographically valid Banjar selected	1	100% final

D. Best Banjar Selection and Highest Valid Candidate

1) Best Banjar Selection

The best Banjar is determined by selecting the candidate that attains the highest Mix Score among those that fully comply with all relevant domain rules, thereby ensuring both linguistic similarity and administrative validity. This decision-making process is deterministic, as a candidate is eligible for selection only if it satisfies the following condition:

$$BEST\ BANJAR = max(Mix\ Score) \cap Valid\ Domain\ Rules \quad (1)$$

This formulation ensures that no candidate with a lower similarity score or any form of hierarchical inconsistency is considered for the matching. Consequently, the method yields a single, unambiguous Banjar prediction for each address, providing a reliable foundation for confirming the HVC.

2) Highest Valid Candidate (HVC)

The HVC serves as the ultimate refinement phase by implementing additional contextual, hierarchical, and semantic validations of the initially identified best Banjar. At this juncture, the best Banjar determined by the highest Mix Score in accordance with domain rules is further scrutinized against the Banjar-Village-District administrative hierarchy, locality-specific semantic patterns, and disambiguation constraints to ensure comprehensive linguistic and geographic coherence. This refinement step is formalized through the following condition:

$$\mathrm{HVC} = \text{Best Banjar} \cap (\text{Hierarchy} + \text{Semantic Validation}) \quad (2)$$

This expression signifies that a candidate can only be accepted as the HVC if it concurrently (i) meets the criteria for Best Banjar and (ii) successfully passes all hierarchical and semantic evaluations. Consequently, the process produces a singular, fully validated Banjar output that eliminates residual ambiguity and ensures a robust address reconstruction.

3) Final Output Reliability

The reliability analysis of the final system output indicated that approximately 95% of all processed addresses converged to a singular, unambiguous HVC, demonstrating the robustness of the integrated fuzzy matching and domain-rule framework. The stability of the final output was closely associated with Mix Score values of 75 or higher, where high-similarity inputs consistently met both hierarchical and semantic constraints. Although cases with low scores (<60) may require additional contextual interpretation, most ambiguous records are effectively resolved through the sequential application of administrative and semantic filtering rules. These findings underscore the high reliability of the final output, corroborating prior evidence that rule-assisted geocoding significantly enhances prediction accuracy in informal or non-standard addressing environments [2]. This reliability trend is further reinforced by the indicators summarized in Table VI, which collectively show high convergence rates, strong score-driven stability, and effective resolution of low-similarity cases, confirming the consistency and operational dependability of the final HVC predictions.

To further illustrate the efficacy and practical reliability of the proposed hybrid fuzzy matching and domain-rule framework, Table VII presents representative HVC outputs generated by the three employed fuzzy matching algorithms: Token Sort (T), Partial Ratio (P), and Levenshtein Distance (L). These samples underscore the unique strengths of each algorithm in addressing noisy, incomplete, or structurally inconsistent Banjar-based address inputs, a pattern also noted

in previous studies on text similarity for unstructured addressing [9, 21]. The examples in Table VII pertain to cases with Mix Score values of 100, where the hybrid framework consistently yields a deterministic and geographically valid Banjar prediction that aligns with the correct Village–District hierarchy. This behavior corroborates the findings of hierarchy-aware geocoding research, which highlights the importance of integrating textual similarity with administrative structures to ensure spatial plausibility [5,6]. By incorporating verified geographic coordinates, these results demonstrate how the final HVC outputs can be directly associated with spatial centroids, facilitating operationally meaningful geolocations in environments with non-standard or culturally embedded address systems. Collectively, these examples demonstrate that the proposed method provides a robust and reproducible solution for accurate Banjar identification, particularly in contexts where traditional gazetteer-based geocoding approaches are inadequate.

TABLE VI. FINAL OUTPUT RELIABILITY INDICATORS FOR HIGHEST VALID CANDIDATE (HVC)

Reliability indicator	Description	Result	Interpretation
HVC convergence rate	Percentage of addresses converging to exactly one HVC	≈ 95%	The system consistently produces a single, unambiguous final output for most addresses.
High-Score output stability	Proportion of stable outputs originating from Mix Score ≥ 75	Highly correlated	High-similarity inputs reliably match hierarchical and semantic constraints.
Low-score resolution	Addresses with Mix Score < 60 requiring additional contextual cues	Small fraction; mostly resolved	Ambiguities are effectively eliminated through domain-rule filtering.
Overall output reliability	Combined effect of fuzzy matching + domain rules on final output consistency	High	The hybrid method ensures robust final predictions in non-standard addressing environments.

TABLE VII. REPRESENTATIVE HVC SAMPLES FROM TOKEN (T), PARTIAL (P), AND LEVENSHTEIN (L) ALGORITHMS WITH COORDINATES MIX SCORE 100

No	Input	Final HVC	Village-District	Coordinates	Alg.
1	Banjar Mawang Kaja	Mawang-Lodtunduh	Lodtunduh-Ubud	-8.56, 115.27	T
2	Banjar Kumbuh Mas	Kumbuh-Mas	Mas-Ubud	-8.52, 115.28	T
3	Banjar Pujung Kaja	Pujung Kaja-Sebatu	Sebatu-Tegallalang	-8.43, 115.28	T
4	Banjar Saba	Saba-Saba	Saba-Blahbatuh	-8.58, 115.33	P
5	Nyuh Kuning Mas	Nyuh Kuning-Mas	Mas-Ubud	-8.54, 115.27	P
6	Penestanan Sayan	Penestanan-Sayan	Sayan-Ubud	-8.50, 115.25	P
7	Banjar Tarukan Mas	Tarukan-Mas	Mas-Ubud	-8.52, 115.28	L
8	Banjar Celuk	Celuk-Celuk	Celuk-Sukawati	-8.59, 115.28	L
9	Banjar Tengah Bedulu	Tengah-Bedulu	Bedulu-Blahbatuh	-8.54, 115.32	L

The representative HVC samples illustrated that Token Sort, Partial Ratio, and Levenshtein Distance each offer complementary strengths in addressing diverse Banjar-based address patterns. All final predictions achieved a Mix Score of 100, with precise Village-District alignment corroborated by accurate geocoordinates. By converting noisy, unstructured address text into a singular, validated Banjar location, this hybrid fuzzy matching and domain-rule framework significantly reduces the ambiguity that typically complicates last-mile delivery operations. For couriers, the availability of a deterministic HVC complete with verified administrative hierarchy and latitude-longitude coordinates eliminates guesswork, minimizes routing errors, and expedites destination identification, ultimately enhancing delivery efficiency and service reliability in regions lacking standardized addressing systems. These operational advantages are clearly reflected in Table VII, where the representative outputs demonstrate consistent spatial correctness across all three algorithms, confirming the reliability and practical applicability of the final HVC predictions in real-world delivery workflows.

E. Comparison with Previous Studies

Previous research on informal address geocoding has focused mainly on text similarity or spatial learning; however, these studies remain reliant on standardized street-based frameworks [5, 6]. Such methods do not effectively generalize to culturally specific systems, such as Bali's Banjar-based addressing, where local semantics and community hierarchies are predominant. In contrast to earlier fuzzy or rule-assisted methods [2, 21], the proposed approach incorporates a hybrid Mix Score fuzzy similarity with Balinese-specific morphological and hierarchical domain rules, facilitating deterministic HVC outputs with confirmed geographic plausibility. This integration directly addresses a persistent gap by offering a gazetteer-free, culturally informed geocoding solution that is appropriate for last-mile delivery in non-standard addressing contexts. These distinctions are clearly summarized in Table VIII, which highlights how the proposed framework uniquely fills the methodological and operational gaps left by earlier text-based and spatial geocoding models, particularly in regions lacking formal address structures.

TABLE VIII. COMPARISON WITH PREVIOUS STUDIES

Aspect	Previous Studies	This Study (Novelty)
Address type	Standardized, street-based, gazetteer-dependent	Indigenous Banjar-Based Addressing (non-gazetteer)
Fuzzy matching	Single-method similarity [21]	Hybrid Mix Score (Levenshtein + Partial + Token Sort)
Semantic rules	Generic or absent	Balinese locality semantics (Kaja, Kelod, Dauh, etc.)
Administrative structure	Requires formal hierarchies [5, 6]	Banjar → Village → District hierarchy reconstruction
Geographic filtering	Limited or no domain-rule enforcement	Domain Rules Filtering: semantic + hierarchical + disambiguation
Output format	Probabilistic/multi-candidate	Deterministic single HVC (operational-grade)
Applicability	Designed for urban/structured addressing	Purpose-built for culturally embedded, unstructured systems
Operational use	Rarely evaluated in delivery workflows	Optimized for courier last-mile delivery

F. Operational Implications

The operational significance of the proposed hybrid framework is particularly noteworthy for courier workflows in Pos Indonesia, where delivery performance is often compromised by ambiguous or incomplete Banjar-based addressing information. By generating a deterministic HVC with a convergence rate of approximately 95%, the system significantly reduces the time couriers spend interpreting non-standard locality descriptors, minimizes routing errors caused by Banjar ambiguities, and decreases reliance on local knowledge. The incorporation of geographic centroids further facilitates automated routing, allowing for a more accurate estimation of delivery paths in areas lacking standardized street names. In practice, this leads to a reduction in failed first-attempt deliveries, a decrease in average delivery travel distances, and an improvement in overall SLA compliance issues that traditionally arise from the absence of Banjar entries in national gazetteers and geospatial registries. These operational benefits suggest that the framework is not only technically effective but also directly aligned with Pos Indonesia's strategic objective to enhance last-mile delivery reliability in culturally embedded addressing environments such as Bali.

VII. CONCLUSION

The proposed hybrid framework integrates fuzzy matching with domain-specific rules to accurately interpret Bali's indigenous Banjar-based addressing system, independent of predefined gazetteers. By amalgamating multimetric fuzzy similarity measures, Levenshtein Distance, Partial Ratio, and Token Sort Ratio, into a comprehensive Hybrid Mix Score, and reinforcing it with Balinese-specific semantic rules and Banjar-Village-District hierarchical constraints, the method effectively resolves linguistic ambiguities and enhances the reliability of Banjar identification. An experimental evaluation involving 17,354 cleaned delivery records demonstrated that approximately 95% of the inputs converged to a single HVC, thereby confirming the stability and robustness of the proposed approach across diverse and noisy informal address patterns.

The integration of deterministic HVC outputs with verified geographic coordinates offers significant advantages for last-mile delivery, including reduced courier uncertainty, minimized routing errors, and improved delivery efficiency in regions lacking standardized address structures. The findings further underscore the efficacy of hybrid text similarity and domain-rule mechanisms in managing culturally embedded addressing systems, providing a lightweight and reproducible solution applicable to other non-gazetteer environments. Future research may extend this framework to multi-regency datasets, incorporate learning-based ranking models, or integrate real-time courier feedback to further enhance its accuracy and scalability.

REFERENCES

[1] M. Khalid, M. M. Yousaf, and M. U. Sadiq, "Toward Efficient Similarity Search under Edit Distance on Hybrid Architectures," *Information*, vol. 13, no. 10, Sept. 2022, Art. no. 452, https://doi.org/10.3390/info13100452.

[2] V. Silva, A. Amaral, T. Fontes, V. Silva, A. Amaral, and T. Fontes, "Sustainable Urban Last-Mile Logistics: A Systematic Literature Review," *Sustainability*, vol. 15, no. 3, Jan. 2023, https://doi.org/10.3390/su15032285.

[3] P. Cruz *et al.*, "Automatic Identification of Addresses: A Systematic Literature Review," *ISPRS International Journal of Geo-Information*, vol. 11, no. 1, Dec. 2021, https://doi.org/10.3390/ijgi11010011.

[4] M. S. M. Rudwan and J. V. Fonou-Dombeu, "Hybridizing Fuzzy String Matching and Machine Learning for Improved Ontology Alignment," *Future Internet*, vol. 15, no. 7, June 2023, Art. no. 229, https://doi.org/10.3390/fi15070229.

[5] R. Santos, P. Murrieta-Flores, and B. Martins, "Learning to combine multiple string similarity metrics for effective toponym matching," *International Journal of Digital Earth*, vol. 11, no. 9, pp. 913–938, Sept. 2018, https://doi.org/10.1080/17538947.2017.1371253.

[6] L. Liang, Y. Chang, Y. Quan, and C. Wang, "A Hierarchy-Aware Geocoding Model Based on Cross-Attention within the Seq2Seq Framework," *ISPRS International Journal of Geo-Information*, vol. 13, no. 4, Apr. 2024, Art. no. 135, https://doi.org/10.3390/ijgi13040135.

[7] B. Kilic, O. C. Bayrak, F. Gülgen, and M. Uzar, "Explainable address matching in online geocoding: filter-based feature selection and ensemble classification," *GeoInformatica*, vol. 30, no. 1, June 2026, Art. no. 5, https://doi.org/10.1007/s10707-025-00562-y.

[8] P. Cruz, L. Vanneschi, M. Painho, and P. Rita, "Automatic Identification of Addresses: A Systematic Literature Review," *ISPRS International Journal of Geo-Information*, vol. 11, no. 1, Dec. 2021, Art. no. 11, https://doi.org/10.3390/ijgi11010011.

[9] Y. Quan, Y. Chang, L. Liang, Y. Qiao, and C. Wang, "A Novel Address-Matching Framework Based on Region Proposal," *ISPRS International Journal of Geo-Information*, vol. 13, no. 4, Apr. 2024, Art. no. 138, https://doi.org/10.3390/ijgi13040138.

[10] P. Li *et al.*, "A Multi-Semantic Feature Fusion Method for Complex Address Matching of Chinese Addresses," *ISPRS International Journal of Geo-Information*, vol. 14, no. 6, June 2025, Art. no. 227, https://doi.org/10.3390/ijgi14060227.

[11] M. Zhang, X. Liu, J. Ma, Z. Zhang, Y. Qiu, and Z. Jiang, "Non-Standard Address Parsing in Chinese Based on Integrated CHTopoNER Model and Dynamic Finite State Machine," *Applied Sciences*, vol. 13, no. 17, Aug. 2023, Art. no. 9855, https://doi.org/10.3390/app13179855.

[12] J. Martinez-Gil and J. M. Chaves-Gonzalez, "Automatic design of semantic similarity controllers based on fuzzy logics," *Expert Systems with Applications*, vol. 131, pp. 45–59, Oct. 2019, https://doi.org/10.1016/j.eswa.2019.04.046.

[13] "Dashboard Operasi Pos Indonesia", *MileApp*, https://board.mile.app/.

[14] "Badan Pusat Statistik Kabupaten Gianyar." https://gianyarkab.bps.go.id/id.

[15] M. Abdul Rahman, M. Aamir Basheer, Z. Khalid, M. Tahir, and M. Uppal, "Last Mile Logistics: Impact of Unstructured Addresses on Delivery Times," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLVIII-4/W5-2022, pp. 3–8, Oct. 2022, https://doi.org/10.5194/isprs-archives-XLVIII-4-W5-2022-3-2022.

[16] U. Singh, D. Ravi Shankar, G. Bellala, and V. Goel, "Geo-Spatially Informed Models for Geocoding Unstructured Addresses," in *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, Abu Dhabi, UAE, Jan. 2025, pp. 236–242.

[17] S. Yoo, E. Jeon, J. Hyeon, and J. Cho, "Adaptive ensemble techniques leveraging BERT based models for multilingual hate speech detection in Korean and english," *Scientific Reports*, vol. 15, no. 1, June 2025, Art. no. 19844, https://doi.org/10.1038/s41598-025-88960-y.

[18] I. Gagliardi, M. T. Artese, I. Gagliardi, and M. T. Artese, "Ensemble-Based Short Text Similarity: An Easy Approach for Multilingual Datasets Using Transformers and WordNet in Real-World Scenarios," *Big Data and Cognitive Computing*, vol. 7, no. 4, Sept. 2023, https://doi.org/10.3390/bdcc7040158.

[19] N. Elmobark, "A Comparative Analysis of Python Text Matching Libraries: A Multilingual Evaluation of Capabilities, Performance and Resource Utilization," *International Journal of Environment, Engineering and Education*, vol. 7, no. 1, pp. 48–60, Apr. 2025, https://doi.org/10.55151/ijeedu.v7i1.188.

[20] B. Bouaita, A. Beghriche, A. Kout, and A. Moussaoui, "A New Approach for Optimizing the Extraction of Association Rules," *Engineering, Technology & Applied Science Research*, vol. 13, no. 2, pp. 10496–10500, Apr. 2023, https://doi.org/10.48084/etasr.5722.

[21] J. P. Buckley, B. P. Buckles, and F. E. Petry, "Processing noisy structured textual data using a fuzzy matching approach: application to postal address errors," *Soft Computing*, vol. 4, no. 4, pp. 195–205, Dec. 2000, https://doi.org/10.1007/s005000000054.