

Article

A Multilabel Embedding-Based Framework for Predicting Short-Term Cross-Category Complaint Recurrence

Theng-Jia Law , Choo-Yee Ting * , Hu Ng  and Hui-Ngo Goh 

Centre of Big Data and Blockchain Technologies, CoE for Advanced Cloud, Faculty of Computing & Informatics, Multimedia University, Persiaran Multimedia, Cyberjaya 63100, Selangor, Malaysia; nghu@mmu.edu.my (H.N.); hngoh@mmu.edu.my (H.-N.G.)

* Correspondence: cyting@mmu.edu.my

Abstract

Public complaints exhibit strong spatiotemporal dependencies, where issues often propagate across categories within short timeframes, yet existing studies largely overlook cross-category recurrence and underutilize embedding representations for structured data. To address this gap, this study proposes a multilabel embedding-based framework to predict short-term cross-category complaint recurrence using structured spatiotemporal data. Using 48,103 real-world complaint records, the framework integrates embedding representations with Machine Learning (ML) and Deep Learning (DL) models to predict the likelihood of multiple complaint categories recurring within the next seven days in the same area. The results indicated that both approaches achieved average label-wise F1-scores of 30.5–32.6%, exceeding 81% for highly recurrent categories. The best ML model, Binary Relevance with Logistic Regression using multilingual-e5-large embeddings, achieved the lowest Hamming loss of 0.138 ± 0.126 . Statistical analysis confirmed non-normality with Shapiro–Wilk statistics between 0.796 and 0.819 and p -values below 0.05, and significant differences across models with a Friedman test statistic of 512.531 and p -values below 0.05, although no significant pairwise differences were found with the Nemenyi post hoc test. Furthermore, 95% bootstrap confidence intervals indicate stable performance, with F1 ranging from 74.1% to 74.9% for ML and 73.2% to 74% for DL.

Keywords: public complaints; complaint recurrence; embeddings; machine learning; deep learning; multilabel classification

1. Introduction

Public complaints are formal expressions of dissatisfaction or service requests submitted by individuals to governmental and municipal authorities [1–7]. These complaints are typically classified into few categories such as infrastructure, public safety, environmental concerns, health services, and administrative processes [6–9]. As critical feedback mechanisms, citizen complaints service not only as indicators of service quality but also as early signals of systemic failures in urban governance. In contrast, inefficient complaint management can lead to public frustration, diminish morale among frontline public servants, and erode institutional trust over time [5,7,10,11].

A growing body of research demonstrates that complaint patterns exhibit strong spatiotemporal dependencies, where incidents in one location often precede or co-occur with related issues in adjacent areas or service domains [12–20]. For instance, unresolved infrastructure defects such as potholes or faulty street lighting, may trigger cascading



Received: 22 April 2026

Revised: 26 June 2026

Accepted: 7 July 2026

Published: 8 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

grievances in public safety or traffic management, amplifying citizen dissatisfaction and increasing operational burdens on municipal departments [12,13]. Empirical evidence supporting this phenomenon has been reported in multiple studies. For example, an analysis of 65,680 complaints recorded through the Road Problem Reporting Service (RPRS) application between January 2021 and December 2022 identified pothole complaints on highways, national roads, and residential roads as high-priority issues due to their potential impacts on traffic safety, service quality, and maintenance workloads [12]. Similarly, an analysis of Buffalo's 311 service requests during a severe snowstorm revealed that surges in snow-removal complaints were subsequently followed by increases in garbage collection complaints within the same geographic areas, demonstrating how disruptions in one municipal service can propagate downstream effects across other service domains [19]. This interdependence underscores the limitations of treating complaints as isolated events and highlights the need for predictive models capable of anticipating multiple concurrent risks. Such a challenge extends beyond conventional univariate forecasting or single-label classification frameworks [15], calling for more sophisticated, multimodal approaches.

While prior studies have focused on predicting complaint types [3,7,10,21–34], few approaches address the multifaceted recurrence of complaints across several categories within a short time frame. Cross-category complaint recurrence is defined as this phenomenon, where an initial complaint triggers a sequence of subsequent complaints across different categorical domains, reflecting cascading and interdependent service failures rather than isolated events, and capturing the dynamics of complaint escalation and secondary issue propagation [14,16–20,35]. Moreover, there remains limited research on transforming entire data instances comprising both categorical and temporal features into a unified embedding space that preserves relational and spatial semantics. Existing complaint prediction studies have extensively utilized transformer-based embeddings to represent complaint narratives and improve classification performance [31,36,37]. However, these embeddings are predominantly derived from unstructured textual descriptions and are primarily applied to tasks such as complaint classification, leaving the rich contextual information in structured variables underutilized. Therefore, a key research gap remains in developing a framework that transforms structured spatiotemporal complaint records into embedding representations and leverages them for prediction of short-term cross-category complaint recurrence.

To address this gap, this study makes a two-part contribution. First, this study formalizes short-term complaint recurrence as a multilabel classification problem, enabling the simultaneous prediction of multiple downstream complaint categories within a 7-day horizon. Second, it introduces a hybrid modeling framework that integrates semantic embeddings derived from complaint categories with structured spatiotemporal features including district, state legislative assembly, responsible department, and complaint timestamp to create a more comprehensive representation of municipal service dynamics. By combining sentence-transformer-based embeddings with predictive models, the proposed approach aims to improve the representation of structured spatiotemporal complaint records and facilitate the early identification of potential cross-category complaint propagation. This empowers municipal agencies to anticipate cascading service failures, prioritize preventive interventions, and optimize resource allocation in a proactive, data-driven manner.

The remainder of this paper is structured as follows: Section 2 reviews the relevant literature on spatiotemporal dependencies in public complaints and the use of embedding-based representations for complaint classification. Section 3 describes the methodological framework, detailing the process of data collection, multilabel dataset construction, embedding generation, model construction and evaluation. Section 4 presents and discusses the experimental results, comparing Machine Learning and Deep Learning performances across different embedding models and complaint categories. Finally, Section 5 con-

cludes the study by summarizing key findings, highlighting practical implications for municipal governance.

2. Literature Review

Recent advances in urban analytics have revealed that complaints are deeply interconnected in location and time. Complaints do not occur in isolation but evolve through interrelated processes, where one type of issue can trigger or influence others nearby. The growing recognition of the spatiotemporal and cascading nature of complaints has motivated researchers to explore advanced Machine Learning (ML) and Deep Learning (DL) approaches to move beyond descriptive spatial statistics toward predictive frameworks capable of capturing cross-domain dependencies.

While ML models typically involve simpler approaches like Logistic Regression and Random Forest (RF), DL which is an advanced subset of ML employs deep neural networks with multiple layers, enabling more powerful representations for supervised learning tasks [38,39]. Building upon this, recent studies have explored embedding-based representations that capture relational, temporal, and semantic information within structured complaint data. However, most studies focus on single-category classification or cluster analysis, rather than predicting multiple categories likely to recur within a short temporal window. Addressing this research gap requires transforming structured spatiotemporal features into embedding spaces that preserve contextual relationships among complaint categories. Such an approach forms the foundation for the multilabel cross-category recurrence prediction framework proposed in this study.

2.1. Spatiotemporal Dependencies in Complaints

Complaints do not occur in isolation but are deeply embedded in the spatial and temporal dynamics of urban life [14,16–20]. Existing studies have increasingly emphasized that complaints often cluster geographically and propagate over time, with one issues serving as a precursor to others in adjacent domains. In this context, statistical analysis of citizen-reported complaints have revealed that urban complaints are far from random, but instead exhibit strong spatial autocorrelation and temporal regularities as indicated in Table 1.

The researchers applied Moran's I and hot spot analysis to Beijing's 12,345 hotline data and reported significantly clustered patterns of safety complaints, especially in central districts [16]. Similarly, spatial autocorrelation analysis on New York City's 311 noise complaints revealed positive significant across multiple distance scales and concluded that building density in residential areas had higher complaints during non-working periods, confirming the tendency of complaints to aggregate geographically [17]. Furthermore, the researchers studied San Francisco's 5.94 million 311 service requests from 2009 to 2022 and identified both persistent and emerging hot spots of waste-related complaints, showing how clustering shifts over time and across neighborhoods [18]. In particular, an analysis of Buffalo's 311 service requests during a severe snowstorm demonstrated that spikes in snow removal complaints were quickly followed by surges in garbage collection requests, indicating a clear cause–effect relationship between service failures [19]. Spatial analysis further confirmed that these complaint categories were clustered within the same block groups, reinforcing the idea of temporal propagation embedded within geographic concentration [19].

Table 1. Comparative Analysis of Existing Complaint Studies highlighting Data Sources, Methods, and Problem Types.

Author	Data Source						Method								Problem Type							
	12,345 Hotline	311 Call	Sense City	Baidu Search	RPRS	Patrol-Called	110 Call	Moran's I	ArcGIS	Support Vector Machine	RF	Extreme Gradient Boosting	Geographical RF	SHapley Additive exPlanations	Long Short-Term Memory	Density-based Spatial Clustering of Applications with Noise	Convolutional Neural Networks	Autocorrelation Analysis	Clustering Analysis	Time-Series Analysis	Single Category Prediction	Cross-Category Prediction
[12]					✓											✓			✓			
[16]	✓							✓										✓				✓
[17]		✓						✓										✓				✓
[18]		✓							✓									✓				✓
[19]		✓								✓	✓		✓	✓				✓				✓
[20]			✓										✓							✓		✓
[40]		✓						✓					✓	✓				✓				✓
[41]				✓								✓		✓								✓
[42]						✓											✓				✓	
[43]		✓															✓					✓
[44]											✓	✓										✓

The check mark indicates that the method was included and used by the researcher; blank cells indicate that the method was not included or used.

This growing recognition of the spatiotemporal and cascading nature of complaints has motivated researchers to explore advanced ML and DL models, aiming to move beyond purely statistical descriptions and toward predictive frameworks that can capture complex dependencies across time, space, and complaint categories. Time-series modeling has been one of the earliest approaches to uncover such dependencies. Long Short-Term Memory (LSTM) networks were applied to capture temporal recurrence patterns in Patras, Greece, and the results showed that both seasonal trends and a general increase in complaints could be modeled for prediction [20]. Similarly, the researchers introduced the Airware-Haikou dataset that combines air quality measurements with a complaint corpus and highlighted how citizen reports can provide valuable context for interpreting pollution anomalies when coupled with spatiotemporal modeling [45].

Other studies have focused on spatial heterogeneity and neighborhood effects in complaint patterns. The researchers applied an interpretable spatial ML framework combining Geographically Weighted RF with SHapley Additive exPlanations (SHAP) analysis to urban sanitation complaints in San Francisco [40]. Their results showed that drivers such as restroom density and homelessness rates varied spatially, shaping the clustering of sanitation-related complaints. Other study extended this line of inquiry in China by using an Extreme Gradient Boosting (XGBoost) and SHAP model to examine pollution-related complaints, revealing that socio-economic attributes often had a stronger influence on complaint recurrence than objective environmental measures [41].

Spatiotemporal clustering methods have also been leveraged to understand operational and management implications of complaint data. By geotagging smartphone-

reported road issues, the researchers applied spatiotemporal clustering to link complaint types and locations with their actual processing times [12]. Consequently, the approach showed how certain complaint categories tended to cluster in space and these clusters predicted subsequent response times. Other study integrated road network features into a Convolutional Neural Network (CNN)-based model to predict urban events and achieved high accuracy ($R^2 \approx 0.94$) on patrol-called data [42]. This study demonstrated that combining event and spatial context improves prediction of one event type from others in the network.

Recent advances in graph-based DL have further expanded the modeling of complaint interdependencies. The researchers introduced MuST²-Learn, a spatiotemporal Graph Neural Network (GNN) framework for heterogeneous municipal service requests [43]. This multi-view model jointly encodes spatial, temporal, and inter-type correlations among complaint categories to capture how one service request type may influence the timing or volume of other types, improving predictive accuracy across request categories. Similarly, the researchers proposed a Deep Spatio-Temporal Graph Attention Network (GAT) for street-level police service calls (110 calls) [44]. Treating streets as nodes in a graph, their model uses attention to weight neighboring streets differently, capturing how crime-related calls at one location precede or correlate with calls on adjacent streets.

Together, these works underscore a growing recognition that complaints and service requests exhibit complex spatial and temporal interdependencies. However, most existing studies have either examined recurrence within the same complaint category or have modeled correlations at an aggregate level, focusing on volume forecasting or cluster detection. There remains limited research that explicitly predicts which other complaint categories are likely to occur within a short horizon once an initial complaint is reported in a given area. This gap highlights the need for predictive frameworks capable of capturing short-term, cross-category recurrence of complaints, which is critical for enabling anticipatory governance and proactive resource allocation.

2.2. Complaint Classification with Embedding Representation

The growing availability of high-dimensional tabular data has motivated a transition from conventional feature encoding methods toward embedding-based representations. Conventional approaches such as one-hot encoding often fail to capture latent relationships among categorical variables and may lead to sparse, high-dimensional feature spaces. To address these limitations, embedding techniques have emerged as an effective representation learning paradigm for structured data by mapping discrete feature values into dense continuous vectors that preserve semantic and relational information. Such representations enable ML and DL models to learn more informative patterns while improving generalization and computational efficiency.

A foundational contribution in this area was introduced by [46], who proposed entity embeddings for categorical variables through neural networks trained end-to-end. Their approach demonstrated that semantically related category values can be positioned closer together in the latent space, thereby capturing intrinsic relationships that are inaccessible to one-hot encoding. Building upon this concept, Arik et al. [47] proposed TabNet, a deep tabular learning architecture that generates contextual feature embeddings through sequential attention mechanisms. By dynamically selecting the most relevant features at each decision step, TabNet learns instance-specific representations while maintaining interpretability. More recently, transformer-based approaches have extended representation learning for tabular data. The survey by [48] highlights various strategies for adapting transformers to structured data, including row-wise serialization, column-wise encoding, and hybrid architectures that integrate entity embeddings with positional information. Similarly, Jiang

et al. [49] categorized tabular representation learning methods into specialized, transferable, and foundation-model-based approaches, emphasizing the growing importance of transferable embeddings for downstream predictive tasks.

Recent studies have further explored the integration of LLMs and sentence encoders for tabular representation learning. For example, Cheng et al. [50] transformed tabular records into textual sequences by serializing feature-value pairs and encoding them using DistilBERT, allowing semantic relationships among structured attributes to be captured through pretrained language representations. Likewise, Kasneci et al. [51] demonstrated that contextual embeddings generated by RoBERTa and GPT-2 provide complementary information to conventional numerical and categorical features, consistently improving predictive performance in tabular classification tasks. Extending this direction, Koloski et al. [52] proposed a framework that converts each feature-value pair into natural language descriptions and leverages frozen LLM embeddings before applying lightweight adaptation layers for downstream learning. Furthermore, Hoppe et al. [53] evaluated transferable row embeddings generated by tabular foundation models and reported that although pretrained embeddings can enhance predictive performance, their effectiveness depends on task complexity and data characteristics. Collectively, these studies demonstrate that embedding-based representations have evolved from simple categorical embeddings to sophisticated transformer and LLM-derived representations capable of capturing rich interactions within structured tabular data.

The benefits of embedding representations have also been demonstrated in complaint-related applications as shown in Table 2. In public service informatics, Esperança et al. [54] combined complaint metadata with complaint text using Bidirectional Encoder Representations from Transformers (BERT)-based representations to improve municipal complaint routing and reduce redundancy. Similarly, Jurn et al. [31] integrated KoBERT text embeddings with structured complaint attributes extracted through named entity recognition, achieving improved prediction performance in call center complaint management. Beyond public complaints, multimodal embedding frameworks have been applied in healthcare settings where structured records and complaint narratives are jointly modeled. For instance, Lin et al. [36] combined TabNet embeddings for structured patient information with MacBERT representations of complaint descriptions to predict hospitalization outcomes. Likewise, Zhang et al. [37] demonstrated that integrating structured electronic health record data with BERT-derived complaint embeddings significantly improved emergency department severity prediction compared with using either modality alone. These findings indicate that embedding representations can effectively fuse heterogeneous structured and unstructured information, resulting in more informative predictive models.

Despite these advances, existing studies predominantly employ embeddings for single-complaint classification [3,7,10,21–34,55], clustering [54], and severity estimation [36,37]. More importantly, most studies utilize embeddings primarily to represent textual complaint descriptions, while the potential of embedding-based representations for structured tabular attributes, including complaint categories, temporal characteristics, and spatial information, remains underexplored. Existing research also focuses on predicting outcomes within a single complaint category or domain, overlooking the complex interdependencies that exist among different complaint categories across space and time. Consequently, there is limited understanding of whether embedding representations derived from structured spatiotemporal complaint data can effectively capture latent cross-category relationships and support the prediction of subsequent complaint occurrences. Addressing this gap is particularly important for developing proactive urban governance systems, as it enables early identification of emerging complaint patterns, facilitates anticipatory resource allocation,

and advances complaint analytics beyond reactive classification toward spatiotemporally aware recurrence prediction.

Table 2. Comparative Analysis of Existing Embedding-based Complaint Analytics Studies highlighting Representation Methods and Predictive Models.

Author	Embeddings					ML Model						DL Model		
	Bidirectional Encoder Representations from Transformers	Term Frequency-Inverse Document Frequency	TabNet	BGE-M3	E5	RF	SVM	XGBoost	Adaptive Boosting	Gradient Boosting	Categorical Boosting	Light Gradient Boosting Machine	Neural Network	Bidirectional Encoder Representations from Transformers
[50]	✓													✓
[54]	✓	✓				✓	✓	✓						
[31]	✓										✓			✓
[36]			✓											✓
[37]	✓					✓		✓		✓				
[3]	✓					✓	✓	✓						
[7]		✓				✓	✓		✓					
[10]		✓				✓	✓	✓						
[21]		✓					✓							
[22]		✓				✓	✓		✓					
[23]		✓					✓							
[24]	✓									✓				
[25]	✓	✓				✓	✓	✓					✓	
[26]		✓				✓								
[27]		✓					✓							
[28]		✓				✓		✓	✓					
[29]	✓													✓
[30]	✓													✓
[32]							✓							✓
[33]	✓													✓
[34]	✓								✓			✓		✓
[55]				✓	✓									

The check mark indicates that the method was included and used by the researcher; blank cells indicate that the method was not included or used.

3. Method

This section presents the methodological framework for predicting short-term cross-category complaint recurrence using multilabel classification with embeddings derived from structured spatiotemporal dataset as shown in Figure 1. The proposed approaches investigate five experimental questions to evaluate the effectiveness and practical applicability of the proposed framework, namely whether (i) embedding-based representations improve multilabel complaint recurrence prediction compared with conventional encoding approaches, (ii) performance remains consistent across complaint categories under label

imbalance, (iii) highly recurrent complaint categories can be reliably identified to support proactive intervention planning, (iv) low-frequency complaint categories expose systematic failure modes that limit deployment readiness, and (v) the observed performance differences remain statistically reliable across categories and models. The corresponding hypotheses assume that richer embeddings improve recurrence representation, category imbalance significantly affects prediction behavior, recurrent complaint categories exhibit higher predictive reliability, rare categories remain challenging due to insufficient recurrence signals, and statistical testing confirms that observed differences reflect stable model behavior rather than random variation.

The proposed approach begins with the construction of a multilabel dataset, where each complaint instance represents a unique occurrence along with its subsequent related complaints observed within a predefined prediction horizon of $h = 7$ days. This transformation enables the formulation of the recurrence prediction task as a supervised multilabel classification problem, where each label corresponds to a distinct complaint category that may reappear within the short-term window. Following dataset construction, the embedding representation stage converts each structured complaint record into a dense and high-dimensional vector that captures its semantic and contextual relationships. By embedding the features into a unified representation space, the model is able to learn latent dependencies between complaint categories across time and space. These embeddings provide a rich input format for ML and DL models to effectively recognize cross-category recurrence patterns.

Finally, the model construction and evaluation phase applies a series of ML and DL models, including both existing algorithms and modern architectures optimized for tabular data, to perform multilabel classification. The predictive performance of these models is assessed using established multilabel evaluation metrics such as subset accuracy, hamming loss, precision, recall, and F1 score. Together, these methodological components form an end-to-end framework capable of capturing the spatial, temporal, and categorical dependencies essential for anticipating complaint recurrence within a short-term horizon.

3.1. Data Collection

In this study, a real-world complaints dataset was collected from a local municipal service provider in Seberang Perai, Pulau Pinang, Malaysia. The dataset encompasses a comprehensive record of public complaints logged between 1 January 2024 and 30 June 2025, providing a 1 year and 6 month longitudinal view of civic issues reported by residents. A total of 48,103 complaint instances were initially collected with 6 variables during this period as shown in Table 3, each representing a unique service request or concern submitted to the respective responsible department. The dataset reflects the operational scope of municipal governance, covering various aspects of urban living including infrastructure, public safety, sanitation, and administrative services.

Each complaint is timestamped and geotagged to one of the three administrative districts in Seberang Perai, namely North, Central, and South. As summarized in Table 4, the Central district accounts for the largest proportion of complaints (47.37%), followed by North (23.66%) and South (16.83%). In addition to district information, each complaint is linked to a specific State Legislative Assembly (SLA) constituency. An SLA constituency is an electoral and administrative subdivision represented by a state legislator and serves as an important governance unit for local planning, public service monitoring, and resource allocation. A total of 22 unique SLA constituencies are represented in the dataset, comprising 13 in the North district, 18 in the Central district, and 6 in the South district. Compared with district-level aggregation, SLA constituencies provide a finer spatial granularity that enables the identification of localized complaint recurrence patterns while maintaining

sufficient complaint density for meaningful analysis. This makes SLA a suitable spatial unit for modeling short-term cross-category complaint recurrence.

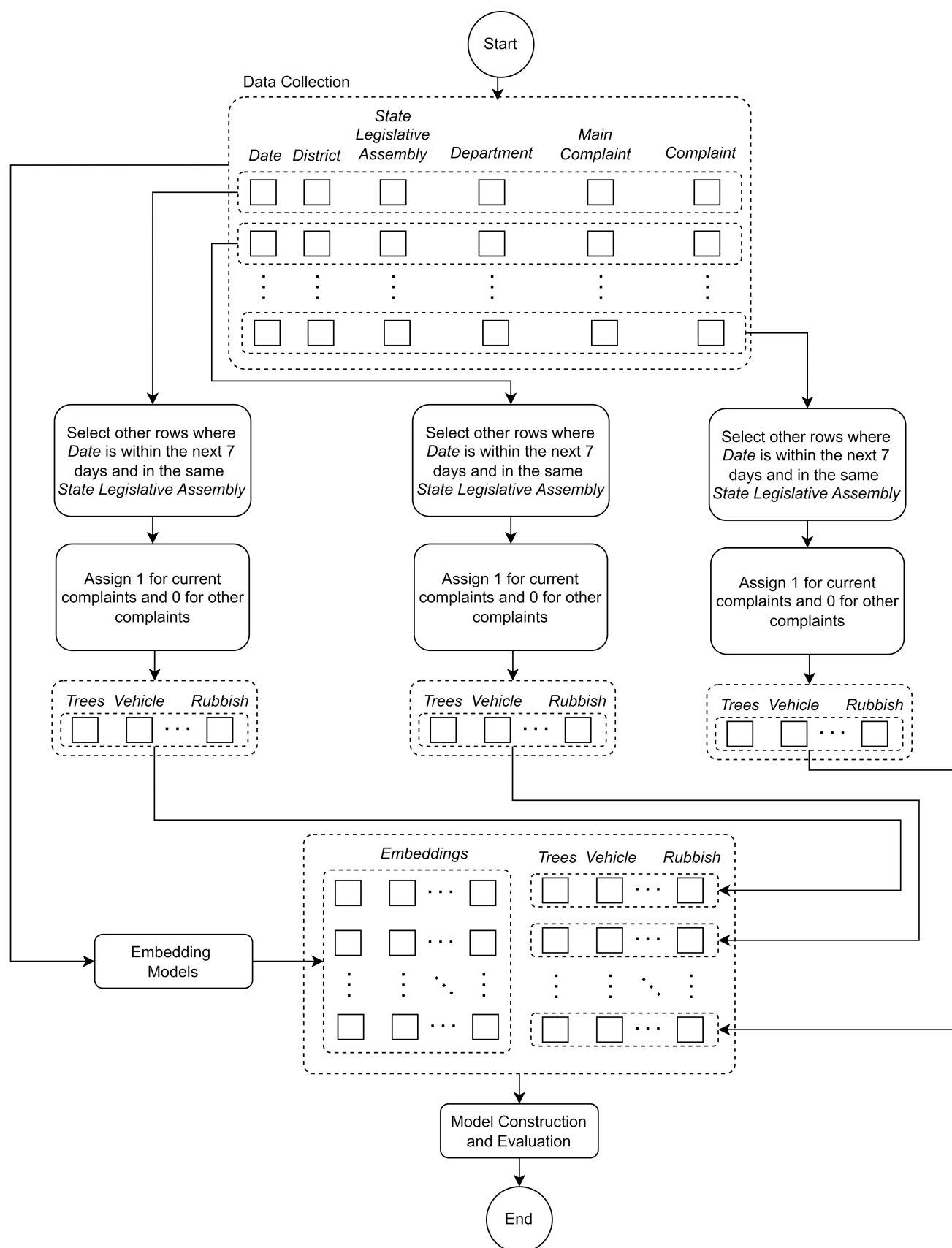


Figure 1. Workflow of the Proposed Multilabel Framework for Predicting Short-Term Cross-Category Complaint Recurrence.

Table 3. Feature Description.

Feature	Type	Description
Date	Datetime	The date the complaint was lodged, formatted as YYYY-MM-DD.
District	Categorical	The administrative district within Seberang Perai where the complaint originated: North, Central, or South.
State Legislative Assembly	Categorical	The legislative assembly constituency within Seberang Perai associated with the complaint.
Department	Categorical	The administrative department responsible for addressing the complaint.
Main Complaint	Categorical	Main category under which the complaint falls.
Complaint	Categorical	Category of a complaint classified under the broader main category.

The responsible department associated with each complaint is also recorded, facilitating insights into inter-departmental workloads and response efficiencies. Overall, the dataset contains complaints managed by 21 unique municipal departments, with 20 departments represented in both the North and Central districts and 19 departments represented in the South district, as shown in Table 4. Complaints are organized under a hierarchical classification system consisting of four main complaint categories, namely Enforcement, Maintenance, Cleaning, and Others. These main categories are further subdivided into 40 specific complaint types, offering more detailed descriptions of the reported issues (e.g., Trees, Vehicle, and Rubbish). This two-tier taxonomy ensures both interpretability at the strategic level and precision at the operational level.

Table 4. Distribution of State Legislative Assembly, Department, and Complaints by District.

District	State Legislative Assembly	Department	Complaints (%)
North	13	20	23.66
Central	18	20	47.37
South	6	19	16.83

Prior to analysis, data quality assessment was conducted to identify incomplete records. Complaint instances containing missing values in any of the 6 variables were removed to ensure consistency during subsequent process. After removing incomplete records, 23,804 complete complaint records remained for subsequent analysis. The inclusion of temporal, geographic, administrative, and categorical dimensions makes this dataset particularly suitable for modeling dynamic civic issue patterns, especially when transformed into a multilabel classification framework to anticipate cascading complaints within a 7-day window. To preserve the temporal integrity of complaint evolution and avoid information leakage, the dataset is partitioned chronologically, with 21,012 complaint records submitted between 1 January 2024 and 30 April 2025 used for model training, and the remaining 2792 records spanning 1 May 2025 to 30 June 2025 reserved as a held-out test set for evaluating predictive performance on future and unseen complaint patterns.

3.2. Formulating Short-Term Cross-Category Complaint Recurrence as Multilabel Solution

To model the problem of predicting short-term cross-category complaint recurrence, this study formulates it as a multilabel classification task. The objective is to determine, for each complaint instance, whether other categories of complaints are likely to recur within a short time window in the same area. The prediction horizon is defined by a parameter h to represent the number of days after the occurrence of a complaint. In this study, $h = 7$ is used to capture week-level recurrence patterns.

Each complaint instance x_i is represented by structured features as shown in Table 3. Formally, a complaint instance can be written as $x_i = (\text{Date} = t_{\text{start}}, \text{District} = r_i,$

State Legislative Assembly = l_i , Department = w_i , Main Complaint = m_i , Complaint = c_i) where $c_i \in C$ is the specific complaint category from the global set C . Within the next h days in the same area l_i , let the observed set of recurrence categories $C_F \subseteq C$. The multilabel output for x_i is then expressed as a binary vector $S_i = [s_{i1}, s_{i2}, \dots, s_{iN}]$ where $s_{ij} = 1$ if category $c_j \in C_F$, and $s_{ij} = 0$ otherwise, with $N = |C|$ being the total number of complaint categories.

Algorithm 1 provides the process to transform the complaint dataset into a multilabel target representation. Each step of the process contributes to identifying all subsequent complaint categories within the defined time window h and encoding them into a binary label vector.

Algorithm 1 Extract Future Complaints Within h days in Same Area

Input

x_i Current complaint
 X Full complaint dataset
 C List of complaint categories
 h Time window size in days

Output

S Binary vector of length N with the presence of each c within h days

```

1: function NEXTCOMPLAINTS( $x_i, X, C, h$ )
2:    $i \leftarrow$  current position of  $x_i$ 
3:    $t_{start} \leftarrow$  current Date of  $x_i$ 
4:    $l \leftarrow$  current State Legislative Assembly of  $x_i$ 
5:    $t_{end} \leftarrow t_{start} + h$ 
6:    $F \leftarrow$  rows from  $X$  where  $t_{start} \leq \text{Date} \leq t_{end}$ , State Legislative Assembly =  $l$ , and
     row index  $\neq i$ 
7:    $C_F \leftarrow$  unique list of complaint categories of  $F$ 
8:    $S \leftarrow \{(k, 1) \mid k \in C_F\}$ 
9:    $S \leftarrow S \cup \{(k, 0) \mid k \notin C_F\}$ 
10:   $S \leftarrow$  reorganize the sequence of  $k$  of  $S$  into an order matching the  $C$ 
11:  return  $S$ 
12: end function

```

The function *nextComplaints*(x_i, X, C, h) begins by locating the current instance x_i within the full dataset X (Algorithm 1, Line 2). The temporal reference t_{start} is then set to the date of the current complaint (Algorithm 1, Line 3), while the spatial reference l is set to the complaint's State Legislative Assembly, ensuring that recurrence is constrained to the same area (Algorithm 1, Line 4). The prediction horizon is then established by adding h days to the starting date t_{start} , yielding t_{end} (Algorithm 1, Line 5). Next, the algorithm extracts all rows F from X when (i) their dates fall within $[t_{start}, t_{end}]$, (ii) they occur within the same State Legislative Assembly l , and (iii) they are not the same row as x_i (Algorithm 1, Line 6). Consequently, this subset F contains all relevant future complaints within the short-term window. From F , the unique set of complaint categories C_F is identified (Algorithm 1, Line 7).

To encode recurrence in a multilabel form, a binary vector S is constructed (Algorithm 1, Lines 7–10). Each element of S corresponds to a category $c \in C$. If a category c appears in C_F , its corresponding entry in S is assigned as 1 (Algorithm 1, Line 8); otherwise, it is assigned as 0 (Algorithm 1, Line 9). The sequence is then reorganized to match the global order of categories in C (Algorithm 1, Line 10) for output consistency. Finally, the binary vector S is returned as the multilabel output for the complaint instance x_i (Algorithm 1, Line 11). This formulation transforms each complaint into a supervised instance: the features consist of the structured spatiotemporal features of x_i , as listed in Table 3, and the label is the multilabel vector S . Each dimension of S represents the

recurrence likelihood of all complaint categories C for a particular complaint category, thereby framing the prediction problem as multilabel classification rather than single-label prediction.

Consequently, the transformed training dataset encompasses 6 features (Table 3) and 21,012 complaint records across 37 categories C . The categories Trees, Vehicle, and Rubbish account for the highest numbers of short-term recurrent complaints, while Earthworks, Open Burning, and Staff represent the least recurrent categories (Table 5). The table summarizes the distribution of recurrence outcomes, where Yes indicates the number of instances in which a complaint category recurred or occurred within the short-term prediction horizon when a complaint is lodged in the same area, while No represents the absence of recurrence. This distribution highlights the importance of framing the task as a multilabel classification problem, since recurrent complaints often arise in relation to one another over short time spans and must therefore be modeled jointly rather than independently.

Table 5. Complaint Categories with the Highest and Lowest Short-Term Recurrence Frequencies.

Group	Category	Yes	No
Highest	Trees	17,452	3560
	Vehicle	16,447	4565
	Rubbish	16,225	4787
Lowest	Earthworks	711	20,301
	Open Burning	691	20,321
	Staff	265	20,747

3.3. Embeddings Representation

After formulating short-term cross-category complaint recurrence as a multilabel problem, the next step is to represent each complaint instance in a numerical space suitable for predictive models. The transformed dataset, consisting of the 6 structured features, is first converted into a textual form by concatenating feature values into a single comma-separated string. The resulting textual representation is then projected into a high-dimensional embeddings space, where semantically similar complaint instances are placed closer together. Existing embedding models are employed for this transformation, each producing vectors of different dimensional sizes depending on the architecture and pretraining corpus. Table 6 summarizes the embedding models considered in this study along with their corresponding output dimensions.

Table 6. Embedding Models and the Corresponding Embedding Dimension Size.

Embedding Model	Dimension Size
all-MiniLM-L6-v2	384
all-mpnet-base-v2	768
jina-embeddings-v4	2048
bge-m3	1024
multilingual-e5-large	1024
Qwen3-Embedding-4B	2560

As shown in Table 6, lightweight models such as all-MiniLM-L6-v2 generate relatively compact embeddings of 384 dimensions, while larger models such as Qwen3-Embedding-4B produce embeddings up to 2560 dimensions. The diversity of embedding sizes provides an opportunity to assess the trade-off between representational richness and computational efficiency when training multilabel classifiers for complaint recurrence prediction. By

embedding complaint records into such vector spaces, structured spatiotemporal features can be effectively aligned with DL architectures, enabling the capture of latent relationships that extend beyond categorical boundaries.

3.4. Model Construction and Evaluation

In this work, a range of ML and DL models were employed, as summarized in Table 7, to perform multilabel classification for predicting short-term cross-category complaint recurrence within $h = 7$ days. Since multilabel classification requires handling multiple outputs simultaneously, common problem transformation methods such as Binary Relevance (BR) and Classifier Chain (CC) were integrated with most of the selected models. BR approaches the task by decomposing it into multiple independent binary classification problems, while CC goes a step further by modeling label dependencies, using the predictions of earlier labels as additional features for subsequent classifiers [56–58].

Table 7. Existing Machine Learning and Deep Learning Models for Multilabel Classification.

Type	Model	Hyperparameter	Description	Values
ML	Logistic Regression (LR)	class_weight	Weights assigned to classes to address class imbalance	None
	Gaussian Naive Bayes (GNB)	var_smoothing	Small variance added for numerical stability	$1 \cdot 10^{-9}$
	Random Forest (RF)	class_weight max_depth	Weights assigned to classes to address class imbalance Maximum depth of a tree	None None
	Adaptive Boosting (AdaBoost)	estimator	Base learner	Decision Tree
	Extreme Gradient Boosting (XGBoost)	max_depth	Maximum depth of a tree	None
	Light Gradient Boosting Machine (LGBM)	class_weight max_depth	Weights assigned to classes to address class imbalance Maximum depth of a tree	None -1
	Categorical Boosting (CatBoost)	class_weights depth	Weights assigned to classes to address class imbalance Depth of a tree	None None
	Multilabel k-Nearest Neighbours (MLkNN)	k	Number of nearest neighbours	10
DL	Modern Neighbor Component Analysis (ModernNCA) [59]	max_epochs lr	Maximum number of training epochs Learning rate controlling parameter updates during training	3 0.001
	Multilayer Perceptron (MLP)	max_epochs lr	Maximum number of training epochs Learning rate controlling parameter updates during training	3 0.001
	Recurrent Neural Network for Tabular Data (TabularRNN) [60]	max_epochs lr	Maximum number of training epochs Learning rate controlling parameter updates during training	3 0.001
	TabM [61]	max_epochs lr	Maximum number of training epochs Learning rate controlling parameter updates during training	3 0.001
	Multi-task TabNet [47]	max_epochs lr	Maximum number of training epochs Learning rate controlling parameter updates during training	3 0.001
	FT-Transformer (FTT) [62]	max_epochs lr	Maximum number of training epochs Learning rate controlling parameter updates during training	3 0.001
	MambaTab (MT) [63]	max_epochs lr	Maximum number of training epochs Learning rate controlling parameter updates during training	3 0.001

While BR and CC provide a flexible means of adapting existing classifiers to multilabel settings, certain models are inherently capable of handling multilabel outputs. MLkNN, for instance, extends the classical K-Nearest Neighbors algorithm by assigning multiple labels to an instance based on the label distribution of its nearest neighbors, effectively capturing relationships between categories in local neighborhoods [57]. Similarly, Multi-

task TabNet is a DL model specifically designed for tabular data. It applies a series of feature transformations and employs an attention mechanism to focus on the most relevant attributes for each task, allowing it to capture nonlinear dependencies and feature interactions [47,64,65]. Because these models are inherently multilabel by design, they do not require BR or CC transformation.

For probabilistic multilabel classifiers, BR and CC models produce label-wise probability estimates that are internally converted into binary predictions using the default decision threshold of 0.5 [56–58]. Similarly, Multi-task TabNet applies a sigmoid activation at each output node and adopts the standard 0.5 cutoff to determine label assignment [47,64,65]. For MLkNN, label assignment follows the algorithm's built-in Bayesian posterior estimation procedure, where a label is predicted as positive when its posterior probability exceeds 0.5 [57]. This default thresholding behavior ensures consistency across models and enables fair comparison across embedding configurations. Overall, the ML models included in this study were selected for their relative simplicity, interpretability, and established performance in classification tasks. Specifically, LR and GNB were included as lightweight probabilistic baselines to assess the effectiveness of semantic embeddings under different modeling assumptions, while RF, AdaBoost, XGBoost, LGBM, and CatBoost were selected to represent tree-based ensemble approaches capable of capturing nonlinear relationships and feature interactions. Moreover, DL models were chosen for their capacity to learn rich representations and capture complex feature interactions in tabular data [66–68]. This complementary set of models provides a balanced foundation for investigating the effectiveness of multilabel learning in predicting complaint recurrence.

To assess the predictive performance of these models, several evaluation metrics were applied namely subset accuracy, hamming loss, precision, recall, and F1 score. Subset accuracy, also known as the exact match ratio, measures the proportion of instances in which the entire predicted label set matches the ground truth exactly, making it the strictest metric for multilabel classification. In contrast, hamming loss evaluates the fraction of misclassified labels, penalizing false positives and false negatives equally across all categories. To provide a more nuanced perspective, precision quantifies the proportion of correctly predicted labels among all predicted labels, while recall measures the proportion of correctly predicted labels among all true labels. Finally, the F1 score which involves the harmonic mean of precision and recall offers a balanced metric that accounts for both false positives and false negatives. For multilabel classification, precision, recall, and F1 score are computed using sample-wise averaging, ensuring that each complaint record contributes equally to the overall evaluation. In addition to overall performance evaluation, label-wise performance analysis is conducted to examine model effectiveness at the individual complaint category level. In this context, label-wise performance refers to the evaluation of prediction effectiveness for each complaint category separately, allowing systematic comparison of how different models handle the heterogeneity and imbalance across complaint types.

To further assess the effectiveness of models under the label-wise setting, statistical significance analyses are performed. First, the Shapiro–Wilk test is applied to the distribution of label-wise performance scores to evaluate the normality assumption. The hypotheses are defined as follows:

H_0 . *The label-wise performances are normally distributed.*

H_a . *The label-wise performances are not normally distributed.*

If the resulting p -value is less than the significance level of 0.05, H_0 is rejected, indicating a violation of normality. In such cases, the non-parametric Friedman test is employed; otherwise, a one-way Analysis of Variance (ANOVA) is conducted to compare multiple

models across labels. The hypotheses for both tests are defined as below. Rejection of H_0 at the 0.05 significance level indicates statistically significant differences among the evaluated models;

H_0 . *There is no significant difference in model performance across models.*

H_a . *At least one model exhibits significantly different performance.*

Following a statistically significant ANOVA or Friedman test result, pairwise post hoc comparisons are performed to determine which models differ significantly. Specifically, Tukey's Honest Significant Difference (HSD) test is employed after ANOVA, whereas the Nemenyi test is applied after the Friedman test. Both post hoc tests evaluate the following hypotheses for each pairwise comparison. Rejection of H_0 at the 0.05 significance level indicates that the performance difference between a pair of models is statistically significant. These post hoc analyses provide a detailed examination of pairwise model differences and identify the specific models contributing to the overall significance detected by the ANOVA or Friedman test;

H_0 . *There is no significant difference between the performances of the two compared models.*

H_a . *There is a significant difference between the performances of the two compared models.*

Furthermore, to complement hypothesis testing, bootstrap confidence intervals are computed for label-wise performance metrics to quantify the uncertainty of and variability in model estimates. Specifically, 95% bootstrap confidence intervals are reported to provide a reliable range within which the true performance is expected to lie. This combination of statistical significance testing and interval estimation ensures a more comprehensive evaluation of model performance beyond point estimates.

4. Results and Discussion

This section presents the experimental findings on short-term complaint recurrence prediction using the proposed framework. The evaluation considers multiple embedding models combined with both ML and DL approaches. Model performance is assessed using subset accuracy, hamming loss, precision, recall, and F1 score. To structure the analysis, five levels of results are reported. First, this study reports the baseline performance of the selected ML and DL models with different embeddings, establishing a benchmark for comparison (Table 8). Next, a statistical summary of performance is provided by computing the mean and standard deviation of each metric across complaint categories, offering a more generalizable view of model effectiveness beyond individual cases (Table 9). The findings narrow the focus to the three most recurrent complaint categories namely Trees, Vehicle, and Rubbish (Table 10) and least most recurrent categories such as Open Burning, Earthworks, and Staff (Table 11), illustrating model behavior on high-frequency and low-frequency labels, respectively. To ensure that the observed performance differences are statistically reliable, statistical significance tests are conducted. Specifically, the Shapiro–Wilk test is first applied to assess whether the distribution of model performance scores satisfies the normality assumption required for parametric testing. Depending on the outcome of the normality test, either a one-way ANOVA or the non-parametric Friedman test is employed to determine whether statistically significant differences exist among the evaluated models. The choice of post hoc procedure depends on the statistical test applied. When a significant result is obtained from the one-way ANOVA, Tukey's HSD test is used to compare all model pairs. Conversely, when the Friedman test indicates a significant

difference, the Nemenyi test is employed to identify the specific pairs of models that exhibit statistically significant performance differences. In addition, bootstrap CI is computed for the top-performing ML and DL models, ranked based on F1 score, to quantify the stability of and variability in their label-wise performance across complaint categories.

Table 8. Comparison of Top Performing Embeddings-Based and Conventional Encoding Models based on F1 Score.

Type	Embeddings	Model	Subset Accuracy	Hamming Loss	Precision	Recall	F1
ML	No	BR + AdaBoost	0.000	0.140	0.684	0.748	0.696
	No	CC + AdaBoost	0.000	0.144	0.672	0.760	0.693
	jina-embeddings-v4	BR + LR	0.001	0.140	0.694	0.674	0.663
	multilingual-e5-large	BR + LR	0.001	0.138	0.695	0.673	0.661
	Qwen3-Embedding-4B	BR + LR	0.002	0.139	0.693	0.670	0.658
DL	bge-m3	BR + TabM	0.000	0.140	0.688	0.695	0.675
	multilingual-e5-large	CC + TabM	0.000	0.140	0.681	0.701	0.674
	bge-m3	CC + FTT	0.001	0.143	0.682	0.698	0.671
	Qwen3-Embedding-4B	CC + FTT	0.000	0.146	0.676	0.708	0.671
	bge-m3	CC + TabM	0.000	0.141	0.686	0.685	0.668

Table 8 compares embeddings-based approaches against conventional encoding base-lines. Among all evaluated configurations, only two non-embedding models using one-hot encoding ranked within the top-performing configurations, namely BR + AdaBoost and CC + AdaBoost. Although these conventional representations remained competitive in overall F1 score ($\geq 69.3\%$), the majority of top-performing positions were occupied by embeddings-based models. Notably, embeddings-based approaches achieved competitive or lower hamming loss while maintaining comparable precision ($\geq 69.3\%$) and recall ($\geq 67\%$). The superior ranking frequency of embedding representations indicates that transforming structured complaint records into dense semantic spaces enables models to capture relationships among complaint categories, temporal dynamics, and administrative attributes more effectively than sparse one-hot vectors. Moreover, A key observation is that subset accuracy values remain extremely low across all models, ranging between 0% and 2%. This indicates that none of the models were able to perfectly predict the entire set of complaint recurrence labels for any given instance. Such a result is expected in multilabel classification tasks with large label spaces, as the exact match ratio becomes overly strict and rarely achieved, especially when multiple labels co-occur with varying dependencies.

In contrast, hamming loss provides a more informative perspective on model performance. Across both ML and DL models, hamming loss values remain relatively consistent, falling between 0.138 and 0.146. These results suggest that, while exact label-set matches were rare, the models were still capable of correctly predicting a substantial proportion of individual labels. Notably, the ML models with BR + LR using multilingual-e5-large embeddings achieved the lowest hamming loss (0.138), indicating fewer misclassifications compared to other configurations. This reinforces the suitability of hamming loss as a practical evaluation metric in multilabel complaint recurrence prediction, as it penalizes errors at the individual label level rather than the entire label set.

When considering F1 score, a clearer separation between model performance emerges in embeddings-based models. Among the embedding-based models, the DL models with TabM consistently outperform ML models with LR, with the top F1 scores reaching 67.5% (BR + TabM with bge-m3 embeddings) compared to 66.3% (BR + LR with jina-embeddings-v4). Therefore, the results suggest that both model families exhibit broadly comparable effectiveness for multilabel complaint recurrence prediction. The primary distinction lies

in their prediction tendencies rather than overall predictive capability. Specifically, ML models generally achieve slightly higher precision, whereas DL models tend to achieve slightly higher recall.

Interestingly, the precision–recall trade-off further explains the observed differences in F1 scores between ML and DL models. For ML models, particularly BR + LR with multilingual-e5-large embeddings, precision values remain slightly higher than recall (69.5% vs. 67.3%), suggesting that these models adopt a more conservative prediction strategy by favoring correct positive predictions at the expense of missing some true labels. In contrast, DL models with TabM, such as CC + TabM with multilingual-e5-large embeddings, tend to yield higher recall than precision (70.1% vs. 68.1%), reflecting a more inclusive prediction approach that captures a broader set of true labels while tolerating additional false positives. This balance contributes to the consistently higher F1 scores of DL models, as the harmonic mean rewards models that effectively balance both metrics rather than excelling in only one.

Table 9. Average Label-Wise Performance of ML and DL Models.

Embeddings	Model	Label-Wise Accuracy	Hamming Loss	Precision	Recall	F1
ML Model						
No	BR + AdaBoost	0.860 ± 0.128	0.140 ± 0.128	0.319 ± 0.364	0.354 ± 0.420	0.326 ± 0.375
No	CC + AdaBoost	0.856 ± 0.136	0.144 ± 0.136	0.315 ± 0.360	0.357 ± 0.425	0.324 ± 0.373
jina-embeddings-v4	BR + LR	0.860 ± 0.127	0.140 ± 0.127	0.375 ± 0.391	0.328 ± 0.402	0.311 ± 0.374
multilingual-e5-large	BR + LR	0.862 ± 0.126	0.138 ± 0.126	0.342 ± 0.371	0.331 ± 0.399	0.318 ± 0.373
Qwen3-Embedding-4B	BR + LR	0.861 ± 0.126	0.139 ± 0.126	0.326 ± 0.373	0.331 ± 0.395	0.319 ± 0.372
DL Model						
bge-m3	BR + TabM	0.860 ± 0.125	0.140 ± 0.125	0.318 ± 0.366	0.328 ± 0.404	0.310 ± 0.375
multilingual-e5-large	CC + TabM	0.860 ± 0.126	0.140 ± 0.126	0.303 ± 0.366	0.337 ± 0.411	0.312 ± 0.377
bge-m3	CC + FTT	0.856 ± 0.133	0.144 ± 0.133	0.335 ± 0.363	0.328 ± 0.405	0.310 ± 0.368
Qwen3-Embedding-4B	CC + FTT	0.854 ± 0.142	0.146 ± 0.142	0.334 ± 0.366	0.338 ± 0.417	0.309 ± 0.374
bge-m3	CC + TabM	0.859 ± 0.128	0.141 ± 0.128	0.322 ± 0.369	0.325 ± 0.405	0.305 ± 0.374

While the previous analysis focused on overall performance aggregated across all labels, Table 9 provides deeper insights into average label-wise behavior by reporting per-category performance. This view is critical because it captures the variability across complaint categories, many of which exhibit imbalanced recurrence frequencies. Although label-wise accuracy remains relatively high across all models (approximately 85.4–86.2%), this metric should be interpreted cautiously. Unlike subset accuracy, which requires every label for a complaint record to be predicted correctly, label-wise accuracy evaluates each category independently. Given the highly sparse multilabel matrix, where most category labels are negative for a given complaint instance, the metric is dominated by correctly predicted negatives. Consequently, high label-wise accuracy does not necessarily imply strong identification of recurrent complaint categories or effective modeling of label dependencies.

Examining hamming loss in this context further supports this observation. Average hamming losses fall consistently between 0.138 and 0.146 across all models, reflecting a relatively stable error rate per label. The closeness of these values to those observed in the overall evaluation highlights the effectiveness of per-label predictions, but the accompanying deviations imply that less frequent categories contribute disproportionately to misclassifications. This confirms that while the models are competent in handling dominant categories, rare complaint types remain more difficult to capture consistently within the 7-day prediction horizon.

A more informative observation is the substantial variability reflected in the standard deviations of precision, recall, and F1 score, which range from approximately $\pm 36\%$ to 42% across categories. These values are larger than the corresponding means, indicating considerable heterogeneity in category-level performance. While certain complaint categories are predicted reliably, others remain extremely difficult to identify. This observation is consistent with the highly imbalanced nature of complaint recurrence, where frequently recurring categories dominate the training data while many rare categories contain very limited positive examples. The one-hot encoding baselines achieved precision values of 31.9% and 31.5% , recall of 35.4% and 35.7% , and F1 scores of 32.6% and 32.4% , indicating strong performance under aggregate evaluation. In comparison, embeddings-based ML models with BR + LR generally achieve higher average precision (up to 37.5% with jina-embeddings-v4) compared to DL models (as low as 30.3% with multilingual-e5-large + CC + TabM). This suggests that ML-based approaches are more conservative, reducing false positives by making fewer positive predictions per label. By contrast, DL models tend to lean toward higher recall, with CC + TabM (multilingual-e5-large) reaching 33.7% , slightly outperforming the ML counterparts in capturing true positives. Among the embeddings-based models, these trade-offs manifest in the F1 scores, where both ML and DL converge around similar ranges ($\approx 30.5\text{--}31.9\%$), despite ML exhibiting marginally more stable precision–recall balance for certain embeddings. The average F1 scores of approximately $30\text{--}33\%$ further demonstrate that category-level prediction remains challenging despite the favorable overall label-wise accuracy. Moreover, the performance differences between ML and DL models are relatively small, suggesting that category frequency and label imbalance exert a stronger influence on predictive performance than the choice of classifier architecture. Therefore, the results indicate that the principal challenge lies not in model selection but in accurately learning recurrence patterns for infrequent complaint categories.

Notably, embeddings play a substantial role in shaping per-label performance. For ML, jina-embeddings-v4 yields the highest average precision and competitive F1, whereas multilingual-e5-large provides slightly stronger recall. For DL models, Qwen3-Embedding-4B embeddings achieve the best balance with F1 scores above 30.9% , outperforming bge-m3 which consistently lags behind. This pattern indicates that embedding dimensionality and semantic richness directly influence how effectively both ML and DL models generalize across complaint categories, particularly those with limited recurrence.

Building upon the label-wise averages, Table 10 zooms into the performance of the three most recurrent complaint categories namely Trees, Vehicle, and Rubbish. These categories dominate the recurrence distribution and therefore provide a strong test case for evaluating the robustness of the proposed models. Across both ML and DL approaches, all three categories achieve notably higher F1 scores than the overall averages reported in Table 9, suggesting that recurrent categories are captured more reliably than less frequent ones. This improvement reflects the greater availability of training data for these categories, which enables the models to better learn recurrence patterns.

For the Trees category, ML and DL models exhibit very close performance. LR with BR and high-capacity embeddings such as Qwen3-Embedding-4B achieves the highest F1 score among ML approaches (89.9%), driven by balanced precision (83.3%) and recall (97.6%). DL models, particularly TabM with bge-m3 and multilingual-e5-large, consistently produce perfect recall (100%) but slightly lower precision (80.8%), yielding F1 scores around 89.4% . Near-perfect recall is likely attributable to the high prevalence of recurring Trees complaints rather than an inherent superiority of a particular modeling approach. Because recurrent Trees complaints constitute a large proportion of positive examples, the models can identify most positive cases successfully. However, the corresponding precision values remain around $80\text{--}83\%$, indicating that the high recall is partly achieved through more

aggressive positive predictions that increase false positives. Similar behavior is observed for Vehicle and Rubbish, where improvements in recall are often accompanied by reductions in precision.

Table 10. Performance of ML and DL Models on the Three Most Recurrent Complaint Categories (Trees, Vehicle, and Rubbish).

Label	Type	Embeddings	Model	Label-Wise Accuracy	Hamming Loss	Precision	Recall	F1
Trees	ML	No	BR + AdaBoost	0.808	0.192	0.808	1.000	0.894
		No	CC + AdaBoost	0.808	0.192	0.808	1.000	0.894
		jina-embeddings-v4	BR + LR	0.811	0.189	0.814	0.993	0.895
		multilingual-e5-large	BR + LR	0.817	0.183	0.831	0.971	0.896
		Qwen3-Embedding-4B	BR + LR	0.823	0.177	0.833	0.976	0.899
	DL	bge-m3	BR + TabM	0.808	0.192	0.808	1.000	0.894
		multilingual-e5-large	CC + TabM	0.808	0.192	0.808	1.000	0.894
		bge-m3	CC + FTT	0.814	0.186	0.835	0.959	0.893
		Qwen3-Embedding-4B	CC + FTT	0.813	0.187	0.816	0.992	0.896
		bge-m3	CC + TabM	0.808	0.192	0.808	1.000	0.894
Vehicle	ML	No	BR + AdaBoost	0.785	0.215	0.874	0.845	0.860
		No	CC + AdaBoost	0.785	0.215	0.874	0.845	0.860
		jina-embeddings-v4	BR + LR	0.797	0.203	0.891	0.841	0.865
		multilingual-e5-large	BR + LR	0.788	0.212	0.870	0.855	0.862
		Qwen3-Embedding-4B	BR + LR	0.796	0.204	0.877	0.857	0.867
	DL	bge-m3	BR + TabM	0.796	0.204	0.858	0.884	0.871
		multilingual-e5-large	CC + TabM	0.788	0.212	0.864	0.863	0.863
		bge-m3	CC + FTT	0.790	0.210	0.796	0.980	0.879
		Qwen3-Embedding-4B	CC + FTT	0.770	0.230	0.820	0.902	0.859
		bge-m3	CC + TabM	0.804	0.196	0.863	0.888	0.875
Rubbish	ML	No	BR + AdaBoost	0.760	0.240	0.763	0.985	0.859
		No	CC + AdaBoost	0.748	0.252	0.753	0.985	0.853
		jina-embeddings-v4	BR + LR	0.782	0.218	0.821	0.905	0.861
		multilingual-e5-large	BR + LR	0.794	0.206	0.829	0.912	0.868
		Qwen3-Embedding-4B	BR + LR	0.785	0.215	0.825	0.903	0.862
	DL	bge-m3	BR + TabM	0.782	0.218	0.808	0.927	0.863
		multilingual-e5-large	CC + TabM	0.765	0.235	0.776	0.963	0.859
		bge-m3	CC + FTT	0.777	0.223	0.814	0.908	0.858
		Qwen3-Embedding-4B	CC + FTT	0.801	0.199	0.828	0.925	0.874
		bge-m3	CC + TabM	0.760	0.240	0.774	0.958	0.856
Label	Type	Embeddings	Model	Subset Accuracy	Hamming Loss	Precision	Recall	F1
Trees, Vechicle, and Rubbish	ML	No	BR + AdaBoost	0.565	0.215	0.790	0.891	0.819
		No	CC + AdaBoost	0.565	0.220	0.789	0.892	0.818
		jina-embeddings-v4	BR + LR	0.574	0.203	0.800	0.849	0.804
		multilingual-e5-large	BR + LR	0.576	0.200	0.784	0.845	0.793
		Qwen3-Embedding-4B	BR + LR	0.581	0.199	0.785	0.844	0.795
	DL	bge-m3	BR + TabM	0.576	0.205	0.795	0.878	0.815
		multilingual-e5-large	CC + TabM	0.560	0.213	0.788	0.884	0.815
		bge-m3	CC + FTT	0.554	0.207	0.782	0.890	0.814
		Qwen3-Embedding-4B	CC + FTT	0.551	0.205	0.794	0.880	0.813
		bge-m3	CC + TabM	0.563	0.210	0.791	0.894	0.820

The Vehicle category presents a more competitive picture. Here, DL models such as CC + FTT with bge-m3 embeddings achieve the highest F1 score (87.9%), surpassing ML

approaches by leveraging stronger recall (98%) while maintaining competitive precision (86.4%). By contrast, ML models with BR + LR and embeddings such as Qwen3-Embedding-4B maintain more balanced precision (87.7%) and recall (85.7%), producing an F1 score of 86.7%. For Rubbish, both ML and DL approaches again achieve strong and comparable performance. ML models using multilingual-e5-large embeddings with BR + LR report the highest ML F1 score (86.8%), combining strong precision (82.9%) with high recall (91.2%). On the other hand, DL models, particularly CC + TabM with bge-m3 embeddings, reach an F1 score of 85.6%, driven by higher recall (95.8%) but slightly lower precision (77.4%). Across different embedding-model pairings, DL approaches consistently show higher recall values, in some cases approaching 96.3%, but their precision is often lower, leading to F1 scores that remain competitive but not consistently superior to ML.

When analyzing these three categories jointly, the contrast between ML and DL approaches becomes clearer. ML models consistently maintain higher subset accuracy (≈ 56 – 58.1%) and slightly lower hamming loss compared to DL models (≈ 55.1 – 57.6% subset accuracy). However, DL approaches show superior F1 scores across the combined categories, with CC + TabM using bge-m3 embeddings reaching the highest F1 score of 82%. This improvement is largely attributed to recall-oriented behavior, where DL models capture a greater proportion of recurrent labels even when precision suffers. Given the marginal performance differences, these findings should not be interpreted as conclusive evidence that DL models generalize better than ML models. Instead, the results suggest that both model families are capable of effectively modeling highly recurrent complaint categories when sufficient training examples are available.

Table 11 presents the category-level performance for the three least recurrent complaint categories, namely Open Burning, Earthworks, and Staff, which represent the most challenging prediction cases due to their extremely limited recurrence frequencies. Across all evaluated configurations, including both non-embedding and embedding-based ML and DL models, performance exhibits a highly uniform pattern. Label-wise accuracy remains exceptionally high, ranging between 98.8% and 99.3%, while hamming loss remains correspondingly low at 0.007–0.012. However, despite these seemingly favorable values, precision, recall, and F1 score are consistently equal to 0 across all models and embeddings. This result indicates that none of the evaluated methods successfully identified positive recurrence instances for these categories. Instead, all models converged toward predicting the negative class for every sample, thereby achieving high apparent correctness due solely to the overwhelming dominance of non-recurrent labels.

This outcome further highlights the limitations of relying on accuracy-oriented metrics in highly sparse multilabel settings. Since label-wise accuracy evaluates each complaint category independently and the majority of observations belong to the negative class, a trivial classifier that predicts no recurrence can still achieve nearly perfect accuracy and low hamming loss. The consistency of this behavior across one-hot encoding, semantic embeddings, ML approaches, and DL architectures suggests that the limitation is not primarily attributable to model selection or representation quality. Rather, the complete absence of positive predictions implies that the underlying data distribution for these categories contains insufficient recurring patterns to support meaningful learning. Even semantically richer embeddings such as multilingual-e5-large, Qwen3-Embedding-4B, jina-embeddings-v4, and bge-m3 failed to produce measurable improvements, indicating that semantic feature enhancement alone cannot compensate for extremely limited positive supervision.

Table 11. Performance of ML and DL Models on the Three Least Recurrent Complaint Categories (Open Burning, Earthworks, and Staff).

Label	Type	Embeddings	Model	Label-Wise Accuracy	Hamming Loss	Precision	Recall	F1
Open Burning	ML	No	BR + AdaBoost	0.993	0.007	0.000	0.000	0.000
		No	CC + AdaBoost	0.993	0.007	0.000	0.000	0.000
		jina-embeddings-v4	BR + LR	0.993	0.007	0.000	0.000	0.000
		multilingual-e5-large	BR + LR	0.993	0.007	0.000	0.000	0.000
		Qwen3-Embedding-4B	BR + LR	0.993	0.007	0.000	0.000	0.000
	DL	bge-m3	BR + TabM	0.993	0.007	0.000	0.000	0.000
		multilingual-e5-large	CC + TabM	0.993	0.007	0.000	0.000	0.000
		bge-m3	CC + FTT	0.993	0.007	0.000	0.000	0.000
		Qwen3-Embedding-4B	CC + FTT	0.993	0.007	0.000	0.000	0.000
		bge-m3	CC + TabM	0.993	0.007	0.000	0.000	0.000
Earthworks	ML	No	BR + AdaBoost	0.988	0.012	0.000	0.000	0.000
		No	CC + AdaBoost	0.988	0.012	0.000	0.000	0.000
		jina-embeddings-v4	BR + LR	0.988	0.012	0.000	0.000	0.000
		multilingual-e5-large	BR + LR	0.988	0.012	0.000	0.000	0.000
		Qwen3-Embedding-4B	BR + LR	0.988	0.012	0.000	0.000	0.000
	DL	bge-m3	BR + TabM	0.988	0.012	0.000	0.000	0.000
		multilingual-e5-large	CC + TabM	0.988	0.012	0.000	0.000	0.000
		bge-m3	CC + FTT	0.988	0.012	0.000	0.000	0.000
		Qwen3-Embedding-4B	CC + FTT	0.988	0.012	0.000	0.000	0.000
		bge-m3	CC + TabM	0.988	0.012	0.000	0.000	0.000
Staff	ML	No	BR + AdaBoost	0.988	0.012	0.000	0.000	0.000
		No	CC + AdaBoost	0.988	0.012	0.000	0.000	0.000
		jina-embeddings-v4	BR + LR	0.988	0.012	0.000	0.000	0.000
		multilingual-e5-large	BR + LR	0.988	0.012	0.000	0.000	0.000
		Qwen3-Embedding-4B	BR + LR	0.988	0.012	0.000	0.000	0.000
	DL	bge-m3	BR + TabM	0.988	0.012	0.000	0.000	0.000
		multilingual-e5-large	CC + TabM	0.988	0.012	0.000	0.000	0.000
		bge-m3	CC + FTT	0.988	0.012	0.000	0.000	0.000
		Qwen3-Embedding-4B	CC + FTT	0.988	0.012	0.000	0.000	0.000
		bge-m3	CC + TabM	0.988	0.012	0.000	0.000	0.000
Label	Type	Embeddings	Model	Subset Accuracy	Hamming Loss	Precision	Recall	F1
Open Burning, Earthworks, and Staff	ML	No	BR + AdaBoost	0.979	0.010	0.000	0.000	0.000
		No	CC + AdaBoost	0.979	0.010	0.000	0.000	0.000
		jina-embeddings-v4	BR + LR	0.979	0.010	0.000	0.000	0.000
		multilingual-e5-large	BR + LR	0.979	0.010	0.000	0.000	0.000
		Qwen3-Embedding-4B	BR + LR	0.979	0.010	0.000	0.000	0.000
	DL	bge-m3	BR + TabM	0.979	0.010	0.000	0.000	0.000
		multilingual-e5-large	CC + TabM	0.979	0.010	0.000	0.000	0.000
		bge-m3	CC + FTT	0.979	0.010	0.000	0.000	0.000
		Qwen3-Embedding-4B	CC + FTT	0.979	0.010	0.000	0.000	0.000
		bge-m3	CC + TabM	0.979	0.010	0.000	0.000	0.000

An additional observation emerges when considering the combined evaluation of Open Burning, Earthworks, and Staff. When aggregated into a joint multilabel setting, subset accuracy remains at 97.9% and hamming loss remains stable at 0.010 across all model configurations, yet precision, recall, and F1 continue to remain at zero. This confirms that the inability to recover positive labels persists even after combining multiple low-frequency

categories and is therefore unlikely to result from isolated category-specific effects. Instead, the results suggest the existence of a lower recurrence threshold below which current predictive architectures cannot reliably learn complaint recurrence behavior within the 7-day forecasting horizon.

The statistical significance analysis further examines the effectiveness of model performance across individual complaint categories in the multilabel classification setting. Prior to conducting comparative statistical tests, the Shapiro–Wilk normality test was applied to determine whether the distribution of label-wise performance scores satisfied the normality assumption required for parametric tests. As shown in Table 12, the Shapiro–Wilk statistics range from 0.796 to 0.819 across the evaluation metrics, including label-wise accuracy (0.819), hamming loss (0.819), precision (0.799), recall (0.799), and F1 score (0.796). However, the corresponding p -values are all equal to 0, which are substantially below the significance level of $\alpha = 0.05$. Therefore, the null hypothesis H_0 which assumes that the model performance scores are normally distributed across complaint categories, is rejected. This result indicates that the label-wise performance values significantly deviate from normality. Consequently, instead of employing a parametric one-way ANOVA, the non-parametric Friedman test was adopted to compare the relative performance of the evaluated models.

Table 12. The Shapiro–Wilk and Friedman Tests of Label-Wise Performance.

Statistical Test	Statistical Measure	Label-Wise Accuracy	Hamming Loss	Precision	Recall	F1
Shapiro–Wilk	Statistics p -value Reject Null Hypothesis	0.819	0.819	0.799 0 Yes	0.799	0.796
Friedman	Statistics p -value Reject Null Hypothesis	3430.777	3430.777	932.61 0 Yes	1535.869	512.531

The Friedman test results, also summarized in Table 12, reveal statistically significant differences in model performance across complaint categories. The Friedman statistics are 3430.777 for label-wise accuracy, 3430.777 for hamming loss, 932.61 for precision, 1535.869 for recall, and 512.531 for F1 score, with all associated p -values equal to 0. Since these values are below the significance threshold of 0.05, the null hypothesis H_0 stating that there is no significant difference in model performance is rejected for all evaluation metrics. These findings confirm that the observed variations in label-wise performance among the evaluated ML and DL models are statistically significant rather than arising from random variation.

In Table 13, the Nemenyi post hoc analysis provides a finer-grained interpretation of the statistical differences identified by the Friedman test by examining pairwise comparisons among model combinations. Unlike the Friedman test, which only indicates that at least one configuration differs significantly from the others, the Nemenyi procedure identifies where those differences occur across the evaluated multilabel configurations. Table 13 shows that several of the largest performance gaps emerge between embedding-based modern tabular architectures and lower-performing baseline configurations. An important observation is that most pairwise comparisons remain close to, but do not consistently cross, significance thresholds ($\alpha = 0.05$). For example, Qwen3-Embedding-4B with BR + MT compared against one-hot encoding + CC + FTT yields a p -value of 0.051, while bge-m3 with CC + MT reaches $p = 0.084$ and bge-m3 with BR + MT yields $p = 0.098$. These values indicate measurable but moderate practical separation between configurations rather than overwhelming superiority of a single architecture. From a modeling perspective, Table 13 suggests that representation quality contributes more strongly than classifier complexity

alone. High-performing comparisons consistently involve richer semantic embeddings such as Qwen3-Embedding-4B, jina-embeddings-v4, and bge-m3 combined with architectures designed to capture structured feature interactions. Conversely, multiple TabNet configurations remain among the weakest despite using embeddings, implying that embedding dimensionality alone does not guarantee improved multilabel prediction. Instead, the interaction between representation learning and model appears to determine category-level effectiveness. Therefore, the Nemenyi results reinforce the earlier conclusion that embedding representations primarily improve reproducibility and cross-category generalization rather than producing universally dominant gains in average F1 performance.

Table 13. The Nemenyi Post Hoc Test of Label-Wise Performance.

Group A			Group B			<i>p</i> -Value
Embeddings	Model	F1	Embeddings	Model	F1	
Qwen3-Embedding-4B	BR + MT	0.360	No	CC + FTT	0.273	0.051
Qwen3-Embedding-4B	CC + MT	0.367	No	CC + FTT	0.273	0.062
bge-m3	CC + MT	0.373	No	CC + FTT	0.273	0.084
Qwen3-Embedding-4B	BR + MT	0.360	all-mpnet-base-v2	TabNet	0.187	0.095
bge-m3	BR + MT	0.382	No	CC + FTT	0.273	0.098
Qwen3-Embedding-4B	CC + MT	0.367	all-mpnet-base-v2	TabNet	0.187	0.113
Qwen3-Embedding-4B	BR + MT	0.360	jina-embeddings-v4	TabNet	0.193	0.119
jina-embeddings-v4	CC + MT	0.376	No	CC + FTT	0.273	0.138
Qwen3-Embedding-4B	CC + MT	0.367	jina-embeddings-v4	TabNet	0.193	0.141
bge-m3	CC + MT	0.373	all-mpnet-base-v2	TabNet	0.187	0.148
Qwen3-Embedding-4B	BR + MT	0.360	Qwen3-Embedding-4B	TabNet	0.197	0.164
Qwen3-Embedding-4B	BR + MT	0.360	No	TabNet	0.187	0.164
bge-m3	TabNet	0.197	Qwen3-Embedding-4B	BR + MT	0.360	0.166
bge-m3	BR + MT	0.382	all-mpnet-base-v2	TabNet	0.187	0.171
bge-m3	CC + MT	0.373	jina-embeddings-v4	TabNet	0.193	0.182
Qwen3-Embedding-4B	BR + MT	0.360	all-MiniLM-L6-v2	TabNet	0.197	0.183
Qwen3-Embedding-4B	CC + MT	0.367	Qwen3-Embedding-4B	TabNet	0.197	0.192
Qwen3-Embedding-4B	CC + MT	0.367	No	TabNet	0.187	0.192
bge-m3	TabNet	0.197	Qwen3-Embedding-4B	CC + MT	0.367	0.194
jina-embeddings-v4	BR + MT	0.382	No	CC + FTT	0.273	0.202
bge-m3	BR + MT	0.382	jina-embeddings-v4	TabNet	0.193	0.208
Qwen3-Embedding-4B	CC + MT	0.367	all-MiniLM-L6-v2	TabNet	0.197	0.213
jina-embeddings-v4	CC + MT	0.376	all-mpnet-base-v2	TabNet	0.187	0.231
bge-m3	CC + MT	0.373	Qwen3-Embedding-4B	TabNet	0.197	0.243
bge-m3	CC + MT	0.373	No	TabNet	0.187	0.243
bge-m3	CC + MT	0.373	bge-m3	TabNet	0.197	0.245
bge-m3	CC + MT	0.373	all-MiniLM-L6-v2	TabNet	0.197	0.267
Qwen3-Embedding-4B	BR + MT	0.360	multilingual-e5-large	TabNet	0.195	0.269
Qwen3-Embedding-4B	BR + MT	0.360	No	BR + FTT	0.288	0.272
bge-m3	BR + MT	0.382	Qwen3-Embedding-4B	TabNet	0.197	0.275
bge-m3	BR + MT	0.382	No	TabNet	0.187	0.276
jina-embeddings-v4	CC + MT	0.376	jina-embeddings-v4	TabNet	0.193	0.276
bge-m3	BR + MT	0.382	bge-m3	TabNet	0.197	0.278
bge-m3	BR + MT	0.382	all-MiniLM-L6-v2	TabNet	0.197	0.302
No	CC + FTT	0.273	all-MiniLM-L6-v2	BR + MT	0.373	0.304
Qwen3-Embedding-4B	CC + MT	0.367	multilingual-e5-large	TabNet	0.195	0.308
No	CC + FTT	0.273	all-mpnet-base-v2	CC + MT	0.377	0.310
Qwen3-Embedding-4B	CC + MT	0.367	No	BR + FTT	0.288	0.311
jina-embeddings-v4	BR + MT	0.382	all-mpnet-base-v2	TabNet	0.187	0.322
Qwen3-Embedding-4B	TabNet	0.197	jina-embeddings-v4	CC + MT	0.376	0.355

To complement the statistical significance analysis, 95% bootstrap confidence intervals were computed for the top-performing configurations selected according to overall F1 score (Table 14). Instead of interpreting narrower intervals as evidence of stability, the confidence intervals were used to examine whether performance differences between competing representations remain meaningful after accounting for sampling variability across complaint categories.

Table 14. Confidence Interval of Label-Wise Performance.

Type	Embeddings	Model	Label-Wise Accuracy	Hamming Loss	Precision	Recall	F1
ML	No	BR + AdaBoost	(0.858, 0.862)	(0.138, 0.142)	(0.712, 0.722)	(0.771, 0.781)	(0.741, 0.749)
	No	CC + AdaBoost	(0.853, 0.858)	(0.142, 0.147)	(0.699, 0.710)	(0.779, 0.788)	(0.738, 0.746)
	jina-embeddings-v4	BR + LR	(0.858, 0.862)	(0.138, 0.142)	(0.735, 0.746)	(0.720, 0.731)	(0.729, 0.737)
	multilingual-e5-large	BR + LR	(0.860, 0.864)	(0.136, 0.140)	(0.740, 0.751)	(0.723, 0.733)	(0.733, 0.741)
	Qwen3-Embedding-4B	BR + LR	(0.859, 0.863)	(0.137, 0.141)	(0.738, 0.749)	(0.722, 0.732)	(0.731, 0.740)
DL	bge-m3	BR + TabM	(0.858, 0.862)	(0.138, 0.142)	(0.732, 0.742)	(0.73, 0.740)	(0.732, 0.740)
	multilingual-e5-large	CC + TabM	(0.857, 0.862)	(0.138, 0.143)	(0.725, 0.736)	(0.740, 0.750)	(0.733, 0.742)
	bge-m3	CC + FTT	(0.854, 0.859)	(0.141, 0.146)	(0.722, 0.732)	(0.729, 0.739)	(0.726, 0.735)
	Qwen3-Embedding-4B	CC + FTT	(0.852, 0.856)	(0.144, 0.148)	(0.710, 0.720)	(0.741, 0.751)	(0.726, 0.734)
	bge-m3	CC + TabM	(0.856, 0.861)	(0.139, 0.144)	(0.729, 0.740)	(0.724, 0.734)	(0.728, 0.736)

The results show substantial overlap among the leading configurations. For example, BR + LR with multilingual-e5-large achieved an F1 confidence interval of (73.3%, 74.1%), while BR + LR with jina-embeddings-v4 produced a closely overlapping interval of (72.9%, 73.7%). Similarly, BR + TabM with bge-m3 yielded an interval of (73.2%, 74.0%), and CC + TabM with multilingual-e5-large reached (73.3%, 74.2%). The overlap suggests that although the Friedman test detected statistically significant differences across models, the practical magnitude of these differences among the top-performing methods is limited.

Importantly, confidence intervals reveal a different pattern from the point estimates alone. While one-hot encoding baselines achieved competitive average F1 values and appeared among the top overall rankings, embedding-based models occupied most of the selected configurations and maintained overlapping intervals across multiple metrics simultaneously, including label-wise accuracy, hamming loss, precision, recall, and F1. This indicates that embedding representations do not necessarily provide large isolated improvements but instead yield more reproducible predictive behavior across heterogeneous complaint categories. Therefore, the confidence interval analysis is interpreted as an assessment of practical equivalence and representation effectiveness rather than stability alone. Combined with the Friedman test results, the findings suggest that embeddings contribute primarily through consistent cross-category generalization rather than statistically large performance gains.

Table 15 presents the label-wise performance of the proposed framework (multilingual-e5-large + BR + LR) under different spatial aggregation levels (State Legislative Assembly (SLA) and District) and prediction horizons ranging from 3 to 28 days. The comparison against three baseline strategies (All 1s, All 0s, and Random prediction) provides important insights into the contribution of spatial granularity and temporal horizon selection for cross-category complaint recurrence prediction. Overall, the proposed framework consistently outperforms the Random and All 0s baselines across all prediction horizons, demonstrating that the learned embedding-based representations capture meaningful spatiotemporal relationships beyond chance-level prediction. Notably, the proposed model achieves substantially lower hamming loss than the Random baseline in both spatial settings, indicating more accurate label-level predictions despite the large and sparse multilabel output space.

For example, under the SLA setting, hamming loss remains between 0.134 and 0.146 across all horizons, compared with approximately 0.498–0.501 for the Random baseline. Similarly, under the District setting, hamming loss ranges from 0.129 to 0.165, remaining considerably lower than the corresponding Random baseline values near 0.5. These results suggest that the proposed framework effectively learns recurrence patterns rather than merely exploiting the dominant class distribution.

Table 15. Spatial Unit Ablation Study of Label-Wise Performance Across Aggregation Levels and Prediction Horizons h .

Spatial Unit	h	Prediction	Label-Wise Accuracy	Hamming Loss	Precision	Recall	F1	Lift
State Legislative Assembly	3	Proposed	0.866 ± 0.135	0.134 ± 0.135	0.254 ± 0.330	0.235 ± 0.321	0.235 ± 0.319	0.673 ± 0.834
		All 1s	0.194 ± 0.236	0.806 ± 0.236	0.194 ± 0.236	0.811 ± 0.397	0.270 ± 0.294	0.811 ± 0.397
		All 0s	0.806 ± 0.236	0.194 ± 0.236	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
		Random	0.501 ± 0.008	0.499 ± 0.008	0.196 ± 0.237	0.412 ± 0.206	0.211 ± 0.213	0.829 ± 0.413
	7	Proposed	0.862 ± 0.126	0.138 ± 0.126	0.342 ± 0.371	0.331 ± 0.399	0.318 ± 0.373	0.779 ± 0.92
		All 1s	0.265 ± 0.293	0.735 ± 0.293	0.265 ± 0.293	0.811 ± 0.397	0.344 ± 0.337	0.811 ± 0.397
		All 0s	0.735 ± 0.293	0.265 ± 0.293	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
		Random	0.500 ± 0.009	0.500 ± 0.009	0.264 ± 0.292	0.397 ± 0.198	0.259 ± 0.238	0.795 ± 0.396
	14	Proposed	0.861 ± 0.124	0.139 ± 0.124	0.389 ± 0.397	0.409 ± 0.434	0.389 ± 0.401	0.805 ± 1.042
		All 1s	0.329 ± 0.335	0.671 ± 0.335	0.329 ± 0.335	0.811 ± 0.397	0.407 ± 0.365	0.811 ± 0.397
		All 0s	0.671 ± 0.335	0.329 ± 0.335	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
		Random	0.502 ± 0.011	0.498 ± 0.011	0.329 ± 0.336	0.403 ± 0.199	0.302 ± 0.252	0.805 ± 0.397
	21	Proposed	0.862 ± 0.126	0.138 ± 0.126	0.444 ± 0.383	0.454 ± 0.438	0.435 ± 0.402	1.02 ± 1.134
		All 1s	0.363 ± 0.349	0.637 ± 0.349	0.363 ± 0.349	0.811 ± 0.397	0.440 ± 0.372	0.811 ± 0.397
		All 0s	0.637 ± 0.349	0.363 ± 0.349	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
		Random	0.500 ± 0.009	0.500 ± 0.009	0.362 ± 0.349	0.409 ± 0.201	0.324 ± 0.254	0.817 ± 0.402
	28	Proposed	0.854 ± 0.130	0.146 ± 0.130	0.448 ± 0.384	0.487 ± 0.434	0.459 ± 0.401	0.943 ± 0.897
		All 1s	0.381 ± 0.355	0.619 ± 0.355	0.381 ± 0.355	0.811 ± 0.397	0.459 ± 0.375	0.811 ± 0.397
		All 0s	0.619 ± 0.355	0.381 ± 0.355	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
		Random	0.499 ± 0.010	0.501 ± 0.010	0.380 ± 0.356	0.402 ± 0.198	0.335 ± 0.254	0.804 ± 0.396
District	3	Proposed	0.856 ± 0.168	0.144 ± 0.168	0.354 ± 0.414	0.419 ± 0.471	0.377 ± 0.431	0.494 ± 0.559
		All 1s	0.377 ± 0.377	0.623 ± 0.377	0.377 ± 0.377	0.811 ± 0.397	0.445 ± 0.388	0.811 ± 0.397
		All 0s	0.623 ± 0.377	0.377 ± 0.377	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
		Random	0.498 ± 0.008	0.502 ± 0.008	0.376 ± 0.376	0.405 ± 0.199	0.324 ± 0.260	0.809 ± 0.397
	7	Proposed	0.871 ± 0.148	0.129 ± 0.148	0.458 ± 0.424	0.483 ± 0.475	0.458 ± 0.444	0.683 ± 0.607
		All 1s	0.454 ± 0.401	0.546 ± 0.401	0.454 ± 0.401	0.811 ± 0.397	0.518 ± 0.401	0.811 ± 0.397
		All 0s	0.546 ± 0.401	0.454 ± 0.401	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
		Random	0.499 ± 0.011	0.501 ± 0.011	0.453 ± 0.401	0.396 ± 0.196	0.368 ± 0.267	0.795 ± 0.393
	14	Proposed	0.860 ± 0.170	0.140 ± 0.170	0.520 ± 0.435	0.551 ± 0.469	0.513 ± 0.448	0.986 ± 2.059
		All 1s	0.509 ± 0.408	0.491 ± 0.408	0.509 ± 0.408	0.811 ± 0.397	0.570 ± 0.402	0.811 ± 0.397
		All 0s	0.491 ± 0.408	0.509 ± 0.408	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
		Random	0.501 ± 0.011	0.499 ± 0.011	0.508 ± 0.407	0.404 ± 0.198	0.403 ± 0.265	0.809 ± 0.397
	21	Proposed	0.842 ± 0.186	0.158 ± 0.186	0.543 ± 0.432	0.585 ± 0.468	0.542 ± 0.443	0.716 ± 0.552
		All 1s	0.538 ± 0.403	0.462 ± 0.403	0.538 ± 0.403	0.811 ± 0.397	0.6 ± 0.398	0.811 ± 0.397
		All 0s	0.462 ± 0.403	0.538 ± 0.403	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
		Random	0.499 ± 0.009	0.501 ± 0.009	0.538 ± 0.404	0.404 ± 0.198	0.421 ± 0.263	0.808 ± 0.396
	28	Proposed	0.835 ± 0.196	0.165 ± 0.196	0.562 ± 0.425	0.639 ± 0.448	0.581 ± 0.427	0.744 ± 0.501
		All 1s	0.558 ± 0.402	0.442 ± 0.402	0.558 ± 0.402	0.811 ± 0.397	0.618 ± 0.397	0.811 ± 0.397
		All 0s	0.442 ± 0.402	0.558 ± 0.402	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
		Random	0.500 ± 0.011	0.500 ± 0.011	0.558 ± 0.401	0.404 ± 0.198	0.433 ± 0.261	0.812 ± 0.398

The most notable finding is the influence of spatial aggregation. Although the methodology initially adopted SLA as the primary spatial unit because it offers finer-grained localization of complaint dynamics, the District-level aggregation consistently achieves higher F1 scores across all prediction horizons. At $h = 7$, the District representation attains an F1 score of 45.8% compared with 31.8% for SLA, and the gap further widens at $h = 28$, where District reaches 58.1% while SLA records 45.9%. Similar improvements are observed in recall, increasing from 33.1% to 48.3% at $h = 7$ and from 48.7% to 63.9% at $h = 28$. These findings suggest that aggregating complaints at the District level reduces spatial

fragmentation and alleviates data sparsity by pooling recurrence events from multiple SLA regions. As a result, the model is exposed to more consistent recurrence patterns and can learn cross-category dependencies more effectively. However, this performance advantage should be interpreted alongside the loss of geographic specificity. While district-level aggregation improves predictive accuracy, SLA-level aggregation retains finer spatial resolution that may be more useful for operational deployment and localized intervention planning. Therefore, the results highlight an important trade-off between predictive performance and spatial interpretability, where coarser aggregation improves model learning while finer aggregation preserves actionable geographic detail.

The comparison with the All 1s prediction provides additional insight into the characteristics of complaint recurrence. As the prediction horizon increases, the All 1s baseline becomes increasingly competitive, with F1 scores rising from 27% to 45.9% under SLA and from 44.5% to 61.8% under District aggregation. This reflects the growing prevalence of positive labels over longer time windows. Nevertheless, unlike the All 1s baseline, which achieves high recall through indiscriminate positive predictions, the proposed framework maintains substantially higher label-wise accuracy and considerably lower hamming loss. This demonstrates that the model is not merely benefiting from increasing recurrence prevalence but is successfully distinguishing which complaint categories are likely to recur.

Table 16 examines the impact of spatial aggregation on the prediction of the Trees, Vehicle, and Rubbish under prediction horizons of 3 and 7 days. Consistent with the findings in Table 15, the results demonstrate that highly recurrent complaint categories are substantially easier to predict than the overall multilabel problem. Across all three categories, the proposed framework achieves F1 scores ranging from 79.3% to 99.7%, considerably higher than the average label-wise F1 values reported in the earlier experiments. This finding is expected because these categories possess a large number of positive recurrence instances, allowing the model to learn stable spatiotemporal recurrence patterns. The low hamming loss values, particularly under the District aggregation level (0.006–0.030), further indicate that the framework can accurately distinguish recurrent and non-recurrent events for these dominant complaint categories. The results therefore reinforce the observation that complaint recurrence frequency is a major determinant of predictive performance, often exerting greater influence than model complexity itself.

A notable pattern emerges when comparing the SLA and District aggregation levels. For all three complaint categories, district-level aggregation consistently produces near-perfect performance. For example, Trees achieves F1 scores of 98.6% and 99.7% under District aggregation for $h = 3$ and $h = 7$, respectively, while Rubbish records F1 scores of 98.5% and 99.5%. Similar behavior is observed for Vehicle, where the performance approaches perfect classification. The lift metric provides further insight into the practical value of the proposed framework. Lift measures the predictive improvement relative to random selection and therefore offers a more meaningful assessment than accuracy alone in highly imbalanced settings. For the SLA aggregation level, the proposed model achieves lift values exceeding 1 across all three categories, reaching 1.364 for Vehicle and 1.307 for Rubbish at $h = 3$. These values indicate that the framework identifies recurrent complaints approximately 30–36% more effectively than random prediction, demonstrating that it successfully captures meaningful recurrence structures rather than relying solely on category prevalence. Interestingly, lift values generally decline as the prediction horizon increases from 3 to 7 days, such as Vehicle decreasing from 1.364 to approximately 1.17 and Rubbish from 1.307 to 1.112. This reduction is expected because recurrence prevalence increases with longer horizons, making the prediction task easier even for baseline methods and thereby reducing the relative advantage over random prediction. At the District level, lift values converge toward 1.0 because complaint recurrence becomes almost ubiquitous

within the aggregated regions. In these circumstances, even simple strategies such as predicting all positives perform nearly optimally. Therefore, the lift analysis highlights that the greatest practical benefit of the proposed framework occurs under finer-grained spatial settings, where meaningful discrimination between recurrent and non-recurrent complaints remains challenging.

Table 16. Spatial Unit Ablation Study of Label-Wise Performance for Trees, Vehicle, and Rubbish.

Label	Spatial Unit	h	Prediction	Label-Wise Accuracy	Hamming Loss	Precision	Recall	F1	Lift
Trees	State Legislative Assembly	3	Proposed	0.709	0.291	0.753	0.845	0.796	1.120
			All 1s	0.672	0.328	0.672	1.000	0.804	1.000
			All 0s	0.328	0.672	0.000	0.000	0.000	0.000
			Random	0.498	0.502	0.672	0.495	0.570	0.999
		7	Proposed	0.817	0.183	0.831	0.971	0.896	1.029
			All 1s	0.808	0.192	0.808	1.000	0.894	1.000
			All 0s	0.192	0.808	0.000	0.000	0.000	0.000
			Random	0.514	0.486	0.813	0.517	0.632	1.006
	District	3	Proposed	0.972	0.028	0.972	1.000	0.986	1.000
			All 1s	0.972	0.028	0.972	1.000	0.986	1.000
			All 0s	0.028	0.972	0.000	0.000	0.000	0.000
			Random	0.491	0.509	0.970	0.491	0.652	0.998
		7	Proposed	0.994	0.006	0.994	1.000	0.997	1.000
			All 1s	0.994	0.006	0.994	1.000	0.997	1.000
			All 0s	0.006	0.994	0.000	0.000	0.000	0.000
			Random	0.514	0.486	0.995	0.513	0.677	1.002
Vehicle	State Legislative Assembly	3	Proposed	0.815	0.185	0.880	0.825	0.852	1.364
			All 1s	0.645	0.355	0.645	1.000	0.785	1.000
			All 0s	0.355	0.645	0.000	0.000	0.000	0.000
			Random	0.503	0.497	0.647	0.504	0.567	1.003
		7	Proposed	0.788	0.212	0.870	0.855	0.862	1.120
			All 1s	0.777	0.223	0.777	1.000	0.874	1.000
			All 0s	0.223	0.777	0.000	0.000	0.000	0.000
			Random	0.497	0.503	0.769	0.504	0.609	0.991
	District	3	Proposed	0.986	0.014	0.986	1.000	0.993	1.000
			All 1s	0.986	0.014	0.986	1.000	0.993	1.000
			All 0s	0.014	0.986	0.000	0.000	0.000	0.000
			Random	0.491	0.509	0.988	0.491	0.656	1.001
		7	Proposed	0.994	0.006	0.994	1.000	0.997	1.000
			All 1s	0.994	0.006	0.994	1.000	0.997	1.000
			All 0s	0.006	0.994	0.000	0.000	0.000	0.000
			Random	0.490	0.510	0.993	0.490	0.656	1.000
Rubbish	State Legislative Assembly	3	Proposed	0.749	0.251	0.791	0.795	0.793	1.307
			All 1s	0.605	0.395	0.605	1.000	0.754	1.000
			All 0s	0.395	0.605	0.000	0.000	0.000	0.000
			Random	0.498	0.502	0.607	0.484	0.539	1.003
		7	Proposed	0.794	0.206	0.829	0.912	0.868	1.112
			All 1s	0.745	0.255	0.745	1.000	0.854	1.000
			All 0s	0.255	0.745	0.000	0.000	0.000	0.000
			Random	0.500	0.500	0.747	0.498	0.597	1.002
	District	3	Proposed	0.970	0.030	0.970	1.000	0.985	1.000
			All 1s	0.970	0.030	0.970	1.000	0.985	1.000
			All 0s	0.030	0.970	0.000	0.000	0.000	0.000
			Random	0.490	0.510	0.971	0.489	0.650	1.001
		7	Proposed	0.990	0.010	0.990	1.000	0.995	1.000
			All 1s	0.990	0.010	0.990	1.000	0.995	1.000
			All 0s	0.010	0.990	0.000	0.000	0.000	0.000
			Random	0.487	0.513	0.989	0.488	0.653	0.999

Finally, the comparison with the All 1s baseline reveals an important characteristic of highly recurrent complaint categories. Although the proposed framework often achieves substantially higher accuracy and lower hamming loss, its F1 scores are occasionally compa-

rable to or slightly below the All 1s strategy, particularly for Trees at the District level. This phenomenon arises because these categories exhibit extremely high recurrence prevalence, causing a naïve strategy that predicts recurrence for every instance to achieve near-perfect recall and competitive F1 scores. However, the proposed framework remains preferable because it maintains significantly lower error rates and provides more selective predictions.

Table 17 presents the opposite end of the recurrence spectrum by evaluating the proposed framework on the Open Burning, Earthworks, and Staff. Unlike the highly recurrent categories analyzed in Table 16, these complaint types contain very few positive recurrence instances in the dataset, with only 691, 711, and 265 positive occurrences, respectively. As a result, the prediction problem becomes highly imbalanced, creating substantial challenges for both learning and evaluation. Across all spatial units and prediction horizons, the proposed framework consistently records F1 scores of 0%, indicating that the model predicts all instances as non-recurrent. At first glance, this may appear to indicate poor predictive performance. However, the simultaneously high label-wise accuracies ranging from approximately 94.4% to 99.6% reveal a different interpretation. The model correctly identifies the overwhelming majority of negative instances, reflecting the extreme rarity of recurrence events within these categories. These findings demonstrate the existing limitation of classifiers when applied to severely imbalanced multilabel datasets, where minimizing overall classification error naturally favors the majority class.

The comparison between spatial aggregation levels further highlights the dominant influence of class imbalance. Unlike the substantial gains observed for Trees, Vehicle, and Rubbish, changing from SLA to District aggregation does not improve predictive performance for Open Burning, Earthworks, and Staff. Regardless of whether complaints are aggregated at the SLA or District level, the proposed framework continues to produce zero recall and zero F1 score. This suggests that the primary limitation is not spatial sparsity but the insufficient number of positive recurrence examples available for learning meaningful recurrence patterns. Even though District aggregation increases complaint density, the absolute number of positive examples remains too small relative to the overwhelming number of negative instances. Consequently, the classifier converges toward a conservative decision boundary that predicts the majority class exclusively. These findings align with the earlier label-wise analysis showing that category frequency exerts a stronger influence on predictive performance than model architecture or representation strategy.

The lift analysis provides additional evidence regarding the difficulty of these categories. For the proposed framework, lift remains 0 across all settings because no positive recurrence events are identified. In contrast, the Random prediction occasionally achieves lift values greater than 1, such as 1.182 for Open Burning, 1.133 for Earthworks, and 1.107 for Staff at $h = 3$. These seemingly favorable lift values should not be interpreted as superior predictive capability. Instead, they arise from stochastic positive predictions generated by random guessing within an extremely sparse label space. Because the prevalence of positive cases is exceptionally low, even a small number of randomly correct predictions can inflate lift values. However, the corresponding F1 scores remain extremely low (typically below 3%), confirming that these gains are not operationally meaningful. Therefore, the lift results highlight an important limitation of evaluating rare-event prediction using isolated metrics and reinforce the need to interpret lift together with precision, recall, and F1 score.

In Table 18, a striking observation is the substantial performance gap between majority and minority complaint categories. For the majority classes, the proposed framework achieves strong predictive performance across all settings, with F1 scores increasing from 60.1% to 79.3% under the SLA setting as the prediction horizon expands from 3 to 7 days. Performance becomes nearly perfect under District aggregation, where F1 reaches 98.5% and 99.5% for $h = 3$ and $h = 7$, respectively. In contrast, the minority categories

consistently record F1 scores of 0 regardless of spatial unit or prediction horizon. This result reinforces the findings from Tables 16 and 17 that complaint recurrence frequency is the dominant factor governing prediction performance. Categories with abundant recurrence examples provide sufficient positive observations for learning meaningful spatiotemporal dependencies, whereas extremely rare categories remain difficult to model due to the overwhelming prevalence of negative instances.

Table 17. Spatial Unit Ablation Study of Label-Wise Performance for Open Burning, Earthworks, and Staff.

Label	Spatial Unit	h	Prediction	Label-Wise Accuracy	Hamming Loss	Precision	Recall	F1	Lift
Open Burning	State Legislative Assembly	3	Proposed	0.996	0.004	0.000	0.000	0.000	0.000
			All 1s	0.004	0.996	0.004	1.000	0.009	1.000
			All 0s	0.996	0.004	0.000	0.000	0.000	0.000
			Random	0.507	0.493	0.005	0.583	0.010	1.182
		7	Proposed	0.993	0.007	0.000	0.000	0.000	0.000
			All 1s	0.007	0.993	0.007	1.000	0.014	1.000
			All 0s	0.993	0.007	0.000	0.000	0.000	0.000
			Random	0.495	0.505	0.005	0.350	0.010	0.697
	District	3	Proposed	0.973	0.027	0.000	0.000	0.000	0.000
			All 1s	0.027	0.973	0.027	1.000	0.052	1.000
			All 0s	0.973	0.027	0.000	0.000	0.000	0.000
			Random	0.492	0.508	0.025	0.473	0.047	0.934
		7	Proposed	0.944	0.056	0.000	0.000	0.000	0.000
			All 1s	0.056	0.944	0.056	1.000	0.106	1.000
			All 0s	0.944	0.056	0.000	0.000	0.000	0.000
			Random	0.507	0.493	0.059	0.516	0.105	1.043
Earthworks	State Legislative Assembly	3	Proposed	0.995	0.005	0.000	0.000	0.000	0.000
			All 1s	0.005	0.995	0.005	1.000	0.010	1.000
			All 0s	0.995	0.005	0.000	0.000	0.000	0.000
			Random	0.496	0.504	0.006	0.571	0.011	1.133
		7	Proposed	0.988	0.012	0.000	0.000	0.000	0.000
			All 1s	0.012	0.988	0.012	1.000	0.023	1.000
			All 0s	0.988	0.012	0.000	0.000	0.000	0.000
			Random	0.499	0.501	0.011	0.485	0.022	0.969
	District	3	Proposed	0.980	0.020	0.000	0.000	0.000	0.000
			All 1s	0.020	0.980	0.020	1.000	0.039	1.000
			All 0s	0.980	0.020	0.000	0.000	0.000	0.000
			Random	0.499	0.501	0.019	0.482	0.037	0.963
		7	Proposed	0.952	0.048	0.000	0.000	0.000	0.000
			All 1s	0.048	0.952	0.048	1.000	0.092	1.000
			All 0s	0.952	0.048	0.000	0.000	0.000	0.000
			Random	0.508	0.492	0.039	0.393	0.072	0.814
Staff	State Legislative Assembly	3	Proposed	0.994	0.006	0.000	0.000	0.000	0.000
			All 1s	0.006	0.994	0.006	1.000	0.013	1.000
			All 0s	0.994	0.006	0.000	0.000	0.000	0.000
			Random	0.499	0.501	0.007	0.556	0.014	1.107
		7	Proposed	0.988	0.012	0.000	0.000	0.000	0.000
			All 1s	0.012	0.988	0.012	1.000	0.024	1.000
			All 0s	0.988	0.012	0.000	0.000	0.000	0.000
			Random	0.507	0.493	0.014	0.588	0.028	1.188
	District	3	Proposed	0.977	0.023	0.000	0.000	0.000	0.000
			All 1s	0.023	0.977	0.023	1.000	0.046	1.000
			All 0s	0.977	0.023	0.000	0.000	0.000	0.000
			Random	0.499	0.501	0.023	0.492	0.044	0.983
		7	Proposed	0.950	0.050	0.000	0.000	0.000	0.000
			All 1s	0.050	0.950	0.050	1.000	0.095	1.000
			All 0s	0.950	0.050	0.000	0.000	0.000	0.000
			Random	0.489	0.511	0.044	0.446	0.080	0.882

The influence of spatial aggregation is particularly evident for the majority categories. Under SLA aggregation, the proposed framework improves from a subset accuracy of 47.3% at $h = 3$ to 57.6% at $h = 7$, accompanied by reductions in hamming loss and improvements in precision, recall, and F1 score. When aggregation is performed at the District level, however, the same categories achieve near-perfect results, with subset accuracy exceeding 92% and hamming loss falling below 0.03. For the minority categories, spatial aggregation provides only limited benefit. Although District aggregation slightly increases positive prevalence and reduces sparsity, the proposed framework continues to predict all instances as non-recurrent, resulting in zero precision, recall, and F1 score. The high subset accuracies observed for these categories (88.9–98.9%) should therefore be interpreted cautiously, as they primarily reflect correct classification of the dominant negative class rather than successful identification of recurrence events.

Table 18. Spatial Unit Ablation Study of Majority and Minority Class Prediction Performance.

Label	Spatial Unit	h	Prediction	Subset Accuracy	Hamming Loss	Precision	Recall	F1	Lift
Trees, Vehicle, and Rubbish	State Legislative Assembly	3	Proposed	0.473	0.242	0.583	0.650	0.601	0.000
			All 1s	0.375	0.359	0.641	0.883	0.716	0.000
			All 0s	0.117	0.641	0.000	0.000	0.000	0.000
			Random	0.121	0.500	0.554	0.434	0.452	0.000
		7	Proposed	0.576	0.200	0.784	0.845	0.793	0.000
			All 1s	0.538	0.223	0.777	0.957	0.836	0.000
			All 0s	0.043	0.777	0.000	0.000	0.000	0.000
			Random	0.127	0.496	0.672	0.482	0.527	0.000
	District	3	Proposed	0.929	0.024	0.976	1.000	0.985	0.000
			All 1s	0.929	0.024	0.976	1.000	0.985	0.000
			All 0s	0.000	0.976	0.000	0.000	0.000	0.000
			Random	0.113	0.509	0.849	0.490	0.597	0.000
		7	Proposed	0.979	0.007	0.993	1.000	0.995	0.000
			All 1s	0.979	0.007	0.993	1.000	0.995	0.000
			All 0s	0.000	0.993	0.000	0.000	0.000	0.000
			Random	0.127	0.503	0.867	0.498	0.607	0.000
Open Burning, Earth-works, and Staff	State Legislative Assembly	3	Proposed	0.989	0.005	0.000	0.000	0.000	0.000
			All 1s	0.000	0.995	0.005	0.011	0.007	0.000
			All 0s	0.989	0.005	0.000	0.000	0.000	0.000
			Random	0.129	0.494	0.005	0.006	0.005	0.000
		7	Proposed	0.979	0.010	0.000	0.000	0.000	0.000
			All 1s	0.000	0.990	0.010	0.021	0.014	0.000
			All 0s	0.979	0.010	0.000	0.000	0.000	0.000
			Random	0.129	0.506	0.008	0.011	0.009	0.000
	District	3	Proposed	0.947	0.023	0.000	0.000	0.000	0.000
			All 1s	0.000	0.977	0.023	0.053	0.032	0.000
			All 0s	0.947	0.023	0.000	0.000	0.000	0.000
			Random	0.133	0.490	0.021	0.026	0.022	0.000
		7	Proposed	0.889	0.051	0.000	0.000	0.000	0.000
			All 1s	0.000	0.949	0.051	0.111	0.069	0.000
			All 0s	0.889	0.051	0.000	0.000	0.000	0.000
			Random	0.126	0.495	0.046	0.058	0.048	0.000

An important observation concerns the lift metric. Unlike Tables 15–17, all lift values in Table 18 are reported as 0 because the evaluation aggregates multiple categories into majority and minority groups rather than assessing individual labels independently. Con-

sequently, the lift metric is no longer informative in its conventional form. Nevertheless, relative improvement over random prediction can still be inferred through F1 score comparisons. For the majority categories under SLA aggregation, the proposed framework achieves F1 scores of 60.1% and 79.3% compared with 45.2% and 52.7% for the Random baseline, representing improvements of approximately 33% and 50%, respectively. Under District aggregation, the improvement becomes even more substantial, with F1 scores of 98.5–99.5% versus only 59.7–60.7% for random prediction. These results demonstrate that the proposed framework captures meaningful spatiotemporal recurrence structures far beyond chance-level behavior. In contrast, for the minority categories, both the proposed framework and the All 0s baseline achieve identical F1 scores of 0%, while the Random baseline occasionally produces small non-zero F1 values due to stochastic positive predictions. However, these values remain extremely low (approximately 0.5–4.8%) and do not represent practically useful predictive performance. Therefore, even though the lift metric itself is unavailable at the aggregated-class level, comparison against random prediction confirms that the proposed framework provides substantial predictive gains for recurrent complaint categories while remaining constrained by extreme class imbalance in rare categories.

Figures 2 and 3 provide a comparison between one-hot encoded features and the proposed embeddings-based representation. Specifically, the same complaint records were projected into two dimensions using t-SNE, allowing the local neighborhood structure of the feature space to be visually examined. For simplicity, the comparison focuses on two complaint categories with substantially different recurrence frequencies: Trees, one of the most recurrent categories, and Staff, one of the least recurrent categories. The color coding indicates the actual recurrence labels where 1 indicates recurrence and 0 indicates non-recurrence of the specific complaint category, while the red cross highlights a representative complaint instance from unseen data used to illustrate the differences between the two feature representations.

In Figure 2, the four label combinations exhibit substantial overlap throughout the projection space. The dominant green points (Trees = 1, Staff = 0) are dispersed across nearly the entire manifold, while the minority purple and blue classes are intermingled without forming clear neighborhoods. More importantly, the highlighted complaint instance (red cross) is located within a region containing mixed label combinations and is incorrectly classified as (Trees = 0, Staff = 0). This observation suggests that one-hot encoding treats each categorical value as an independent dimension without modeling relationships among features. Consequently, complaints sharing similar contextual characteristics, such as occurring within comparable locations, or date, may remain distant in the feature space if their encoded vectors differ in only a few dimensions. The resulting fragmented representation limits the classifier's ability to identify meaningful recurrence patterns.

In contrast, Figure 3 demonstrates that the embeddings produce a more structured representation of complaint instances. Although some overlap remains due to the inherent complexity of multilabel recurrence prediction, complaints associated with the dominant recurrence pattern (Trees = 1, Staff = 0) exhibit denser local neighborhoods and smoother transitions between regions. Notably, the same highlighted complaint instance is now embedded within a neighborhood predominantly occupied by green instances and is correctly classified as (Trees = 1, Staff = 0). This change indicates that the embedding representation preserves contextual similarity more effectively than one-hot encoding. Rather than treating each feature independently, the transformation learns interactions among attributes, allowing complaint records sharing similar combinations of temporal, spatial, and categorical characteristics to be positioned closer together in the latent space.

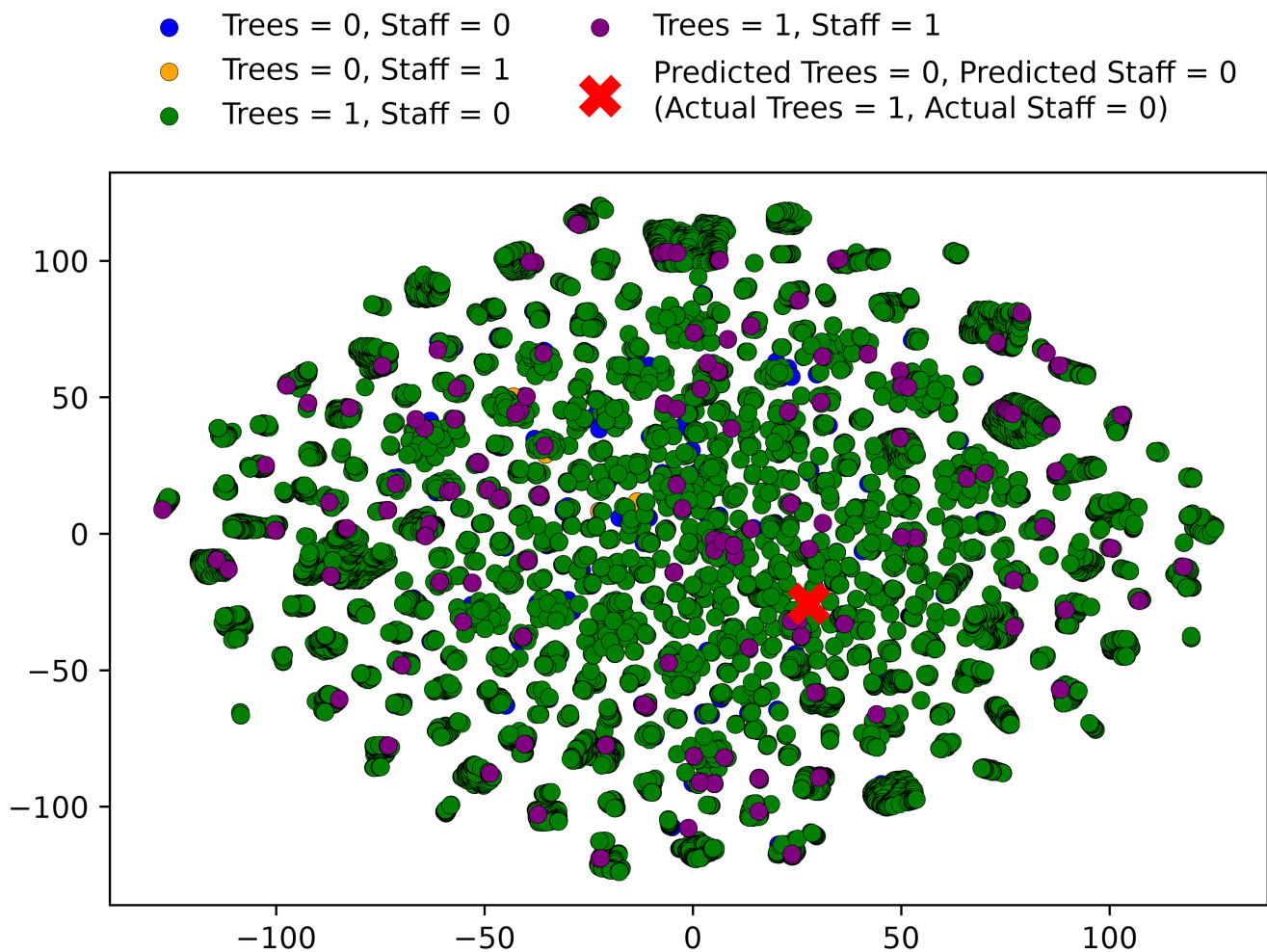


Figure 2. t-SNE Visualization of Complaint Instances represented using One-Hot Encoded Tabular Features.

The results provide evidence supporting the meaningfulness of the generated embeddings. Importantly, the embedding does not attempt to interpret individual attribute values in isolation. Instead, the textual serialization serves as an approach for presenting a structured collection of contextual attributes to a pre-trained embedding model. The embedding model processes the combined attribute sequence and produces a contextualized representation that captures relationships among attributes rather than their literal linguistic meaning. Consequently, an embedding generated from a comma-separated text should be interpreted as a latent representation of a complaint context characterized by a specific temporal occurrence, geographic location, and complaint category. The t-SNE visualization demonstrates that complaints sharing similar contexts are mapped to nearby regions, suggesting that the learned representation contains information relevant to recurrence behavior. Further evidence of meaningful structure can be observed by examining the local neighborhoods surrounding the highlighted complaint instance. Under one-hot encoding, nearby instances belong to multiple label combinations (Figure 2), indicating weak correspondence between feature similarity and recurrence outcomes. In contrast, the embeddings-based representation places the same instance within a more homogeneous neighborhood dominated by the correct recurrence label (Figure 3).

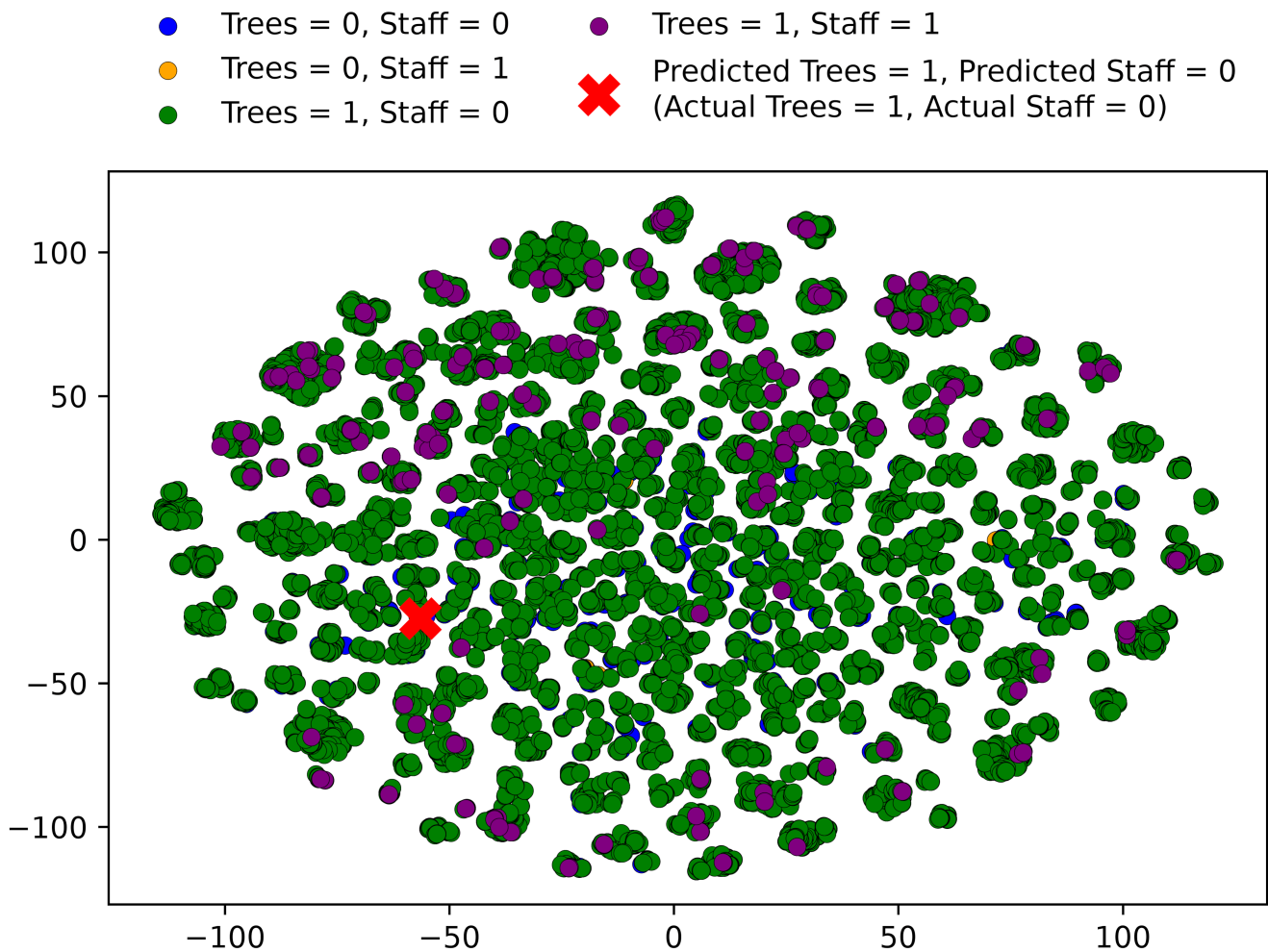


Figure 3. t-SNE Visualization of Complaint Instances represented using Multilingual-E5-Large embeddings generated from Structured Complaint Features Converted into Textual Representations.

In this study, the prediction horizon ($h = 7$) days was chosen to align with the objective of early complaint prediction on a weekly basis, allowing municipal agencies to identify potential recurrence patterns sufficiently early for intervention while maintaining practical planning cycles. A weekly horizon provides a balance between responsiveness and signal availability, since shorter horizons generate fewer observable recurrence events whereas longer horizons increasingly aggregate heterogeneous complaint behaviors and reduce temporal specificity. Under this formulation, complaint instances are constructed using overlapping 7-day windows, which naturally introduce temporal correlation between adjacent observations because nearby samples may share portions of future recurrence periods. Rather than eliminating this dependency, the framework preserves the sequential nature of complaint generation by enforcing chronological separation between training and testing stages to prevent future information leakage while retaining realistic temporal continuity. The observed evaluation outcomes further reflect the implications of this design. Although subset accuracy remains low (0–2%), label-wise accuracy reaches approximately 86%, indicating that the models are considerably more effective at predicting individual complaint categories than reproducing the complete joint recurrence configuration for each instance. Since subset accuracy requires exact prediction of the entire multilabel set, its values become increasingly restrictive under sparse and high-dimensional label spaces. In contrast, label-wise accuracy evaluates categories independently and is largely influenced by correctly predicted negative labels. This discrepancy suggests that the

proposed framework successfully captures dominant category-level recurrence patterns but only partially models higher-order dependencies among complaint categories.

Despite the encouraging predictive performance, several methodological limitations should be recognized. All multilabel predictions were generated using a fixed decision threshold of 0.5 and default hyperparameters. While the decision threshold represents a widely adopted default in multilabel classification, it may not be optimal for highly imbalanced complaint categories. The experimental results reveal evidence of over-prediction behavior in several models. For example, some models achieve near-perfect recall for highly recurrent categories such as Trees while exhibiting substantially lower precision. This pattern suggests that the models increasingly favor positive predictions as category prevalence rises, effectively reducing false negatives at the expense of increased false alarms. From an operational perspective, excessive false-positive predictions may lead to the inefficient allocation of inspection resources and unnecessary preventive interventions. Therefore, categories with extremely low recurrence frequencies may require lower decision thresholds to improve sensitivity, whereas highly recurrent categories may benefit from more conservative thresholds to reduce false positives. Consequently, the reported performance should be interpreted as a baseline performance rather than the optimal achievable configuration. Future work may investigate category-specific threshold optimization, cost-sensitive threshold selection and hyperparameter tuning strategies to better balance precision and recall across complaint categories.

Moreover, the current framework treats complaint categories as independent output labels despite the existence of a two-level complaint taxonomy consisting of 4 main categories and 40 subcategories. Although the multilabel formulation enables simultaneous prediction of multiple recurring complaint categories, it does not explicitly exploit hierarchical relationships between parent and child categories. Incorporating hierarchical multilabel classification mechanisms may further improve predictive performance by enforcing taxonomy-aware consistency and facilitating knowledge transfer between semantically related complaint categories. This direction is particularly relevant for rare complaint categories where limited recurrence observations restrict effective model learning.

From a practical deployment perspective, the proposed framework can be integrated into municipal complaint management systems as an event-driven prediction service. Once a new complaint is submitted, the structured attributes are serialized and transformed into an embedding representation before being passed to the trained multilabel predictor to estimate the likelihood of recurrence across complaint categories within the next seven days. The resulting probabilities can then be aggregated into dashboards, early-warning alerts, or resource scheduling systems to support proactive intervention. Because embedding generation is performed only once per incoming complaint and inference relies on frozen pretrained encoders followed by lightweight prediction models, online prediction remains computationally feasible. In deployment settings, embeddings may additionally be pre-computed and cached for recurring categorical combinations to reduce latency. Although larger embedding models increase representation cost, inference complexity remains substantially lower than model retraining and is compatible with near real-time municipal monitoring workflows. Future optimization may further reduce operational cost through embedding distillation, quantization, or compact encoder architectures while maintaining predictive quality.

5. Conclusions

This study introduced a multilabel embedding-based framework for predicting short-term cross-category complaint recurrence using structured spatiotemporal data. By integrating semantic embeddings with both ML and DL models, the framework effectively

captures inter-category dependencies that existing single-label approaches may overlook. The experimental results demonstrate that both ML and DL models achieved comparable performance, with an average label-wise F1 score ranging from 30.5% to 32.6% and an F1 score exceeding 81% for the top recurrent complaint categories (Trees, Vehicle, and Rubbish). Among all experiments, BR + LR using multilingual-e5-large embeddings achieved the lowest average label-wise hamming loss of 0.138 ± 0.126 , reflecting high per-label accuracy. Meanwhile, CC + TabM using bge-m3 embeddings attained the highest F1 score of 82% for the top recurrent categories, demonstrating its superior capability to balance precision and recall in multilabel classification. Conversely, this study revealed the impact of label imbalance, with models attaining overall accuracies greater than 97% yet producing an F1 score of 0 for specific labels. This outcome suggests that the classifiers were biased toward the dominant negative class, resulting in the complete omission of minority-label predictions despite seemingly strong accuracy performance. Statistical tests confirmed non-normality (Shapiro–Wilk $p < 0.05$) and significant performance differences (Friedman test statistics = 512.531, $p < 0.05$). However, the subsequent Nemenyi post hoc analysis did not identify any statistically significant pairwise differences between individual models ($p > 0.05$), suggesting that while overall ranking differences existed across the evaluated methods, no single model consistently outperformed another at the chosen significance level. Narrow 95% bootstrap confidence intervals (e.g., $F1 \approx 73.3\text{--}74.1\%$) further indicate stable and reliable predictions. These findings indicate that, for the complaint recurrence prediction task studied here, embedding quality may play a more decisive role than model complexity in shaping predictive performance. High-capacity embeddings such as Qwen3-Embedding-4B and multilingual-e5-large substantially enhanced the performance of ML models, enabling lightweight classifiers like LR to achieve results comparable to advanced DL models. This highlights the potential of embedding-driven learning as a scalable and computationally efficient solution in similar multilabel tabular prediction settings, though further validation across domains is warranted.

From a practical perspective, the proposed framework offers an important step toward predictive complaint management. By identifying likely recurrences of different complaint categories within a short temporal window, municipal authorities can anticipate cascading service issues and implement preventive measures before complaints accumulate. Such a proactive approach promotes more efficient resource allocation, reduces operational delays, and strengthens public trust through responsive urban governance. Future work may extend this framework by identifying optimal hyperparameters, conducting hierarchical multilabel classification, and exploring various embedding models across domains; incorporating multimodal inputs such as textual complaint narratives or social media data to adopt adaptive time horizons for recurrence prediction; and leveraging graph-based DL architectures to capture cross-regional dependencies within municipal networks.

Author Contributions: Conceptualization, T.-J.L.; methodology, T.-J.L.; writing—original draft preparation, T.-J.L.; writing—review and editing, C.-Y.T.; supervision, C.-Y.T.; project administration, C.-Y.T.; writing—review and editing, H.N. and H.-N.G.; resources, H.N. and H.-N.G.; validation, H.N. and H.-N.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Telekom Malaysia Research & Development Grant (TM R&D), grant number MMUE/240102.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to reasons of sensitivity.

Acknowledgments: The authors appreciate and express gratitude for the financial support from Telekom Malaysia Research & Development Grant (TM R&D), MMUE/240102.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sarafis, D.; Karamitsios, K. A Technological Survey on Citizen Complaint Management Systems and Future Advances. In Proceedings of the 2024 14th International Conference on Electrical Engineering (ICEENG), Cairo, Egypt, 21–23 May 2024; pp. 265–269.
2. Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K. Think before you classify: The rise of reasoning large language models for consumer complaint detection and classification. *Electronics* **2025**, *14*, 1070.
3. Caldeira, F.; Nunes, L.; Ribeiro, R. Classification of public administration complaints. In *Proceedings of the 11th Symposium on Languages, Applications and Technologies (SLATE 2022)*. Schloss Dagstuhl–Leibniz-Zentrum für Informatik; Schloss Dagstuhl–Leibniz-Zentrum für Informatik: Wadern, Germany, 2022; pp. 9:1–9:12.
4. Wang, Y.; Zhou, Y.; Mei, Y. A joint attention enhancement network for text classification applied to citizen complaint reporting. *Appl. Intell.* **2023**, *53*, 19255–19265.
5. Kim, S.; An, M.; Lee, H.; Kang, J. Materiality-based online complaint classification: An analytical framework for efficient public service using text mining. *ICIC Express Lett* **2024**, *15*, 51–60.
6. Mishra, S.; Das, S. Effective City Planning: A Data Driven Analysis of Infrastructure and Citizen Feedback in Bangalore. *arXiv* **2022**, arXiv:2211.03126.
7. Madyatmadja, E.D.; Sianipar, C.P.; Wijaya, C.; Sembiring, D.J. Classifying crowdsourced citizen complaints through data mining: Accuracy testing of k-nearest neighbors, random forest, support vector machine, and adaboost. *Informatics* **2023**, *10*, 84.
8. Salleh, N.A.; Alauddin, K.; Talib, Y.A.; Ismail, K. A study of complaint trends and strategic improvements: a four-year analysis of six malaysian municipalities. *Plan. Malays.* **2024**, *22*.
9. Torres-Sandoval, A.J.; Tavera-Cortés, M.E.; Acevedo-Ortiz, M.A.; Ortiz-Hernández, Y.D. Transparency, Governance, and Public Service Management: Challenges of Citizen Participation in Ecatepec de Morelos. *Adm. Sci.* **2025**, *15*, 144.
10. Hafidz, I.H.; Sulistya, A.; Lidiawaty, B.R. Sentiment Analysis of Public Complaints: A Machine Learning Comparison of SVM, Naive Bayes, Random Forest, and XGBoost. In Proceedings of the 2025 International Conference on Advancement in Data Science, E-learning and Information System (ICADEIS), Bandung, Indonesia, 3–4 February 2025; pp. 1–6.
11. Theresia, H.T.; Tulusan, F.M.; Londa, V.Y. Complaint Service Management at the Manado City Government Public Service Mall. *J. Dimens. Manag. Public Sect.* **2025**, *6*, 124–135.
12. Jeon, W.; Kim, J.; Lee, J. A Data-Driven Decision-Making Support Method for Priority Determination for an Intelligent Road Problem Reporting System. *Appl. Sci.* **2024**, *14*, 10861.
13. Dong, S.; Yu, T.; Farahmand, H.; Mostafavi, A. Probabilistic modeling of cascading failure risk in interdependent channel and road networks in urban flooding. *Sustain. Cities Soc.* **2020**, *62*, 102398.
14. Wang, G.; Wu, C.; Yao, Y.; Zhang, T. Spatial-temporal analysis of patient complaints in Shanghai from 2015 to 2022. *BMC Health Serv. Res.* **2025**, *25*, 311.
15. Yu, S.; Guo, D.; Fu, Y.; Jin, W. EventFormer: A hierarchical neural point process framework for spatio-temporal clustering events prediction. *J. Big Data* **2025**, *12*, 162.
16. He, J.; Zhang, W.; Yang, M. The spatial and temporal characteristics of urban public safety under the residents' complaints: Evidence from 12345 data in Beijing, China. *J. Urban Manag.* **2024**, *13*, 217–231.
17. Chen, S.; Yu, B.; Shi, G.; Cai, Y.; Wang, Y.; He, P. Scale-Dependent Relationships Between Urban Morphology and Noise Perception: A Multi-Scale Spatiotemporal Analysis in New York City. *Land* **2025**, *14*, 476.
18. Bogulski, C.A.; Watson, W.P.; Young, S.G. Privy by the Bay: Emerging hotspot analysis of 311 reports of human/animal waste near San Francisco Pit Stop locations, 2009–2022. *PLoS ONE* **2025**, *20*, e0327795.
19. Zhou, R.Z.; Hu, Y.; Sun, K.; Muldoon, R.; Clark, S.; Joseph, K. Explainable GeoAI and statistical analysis reveal complementary insights about disparities of 311 help requests during the 2022 Buffalo blizzard. *Int. J. Disaster Risk Reduct.* **2025**, *126*, 105635.
20. Gkontzis, A.F.; Kotsiantis, S.; Feretzakis, G.; Verykios, V.S. Temporal Dynamics of Citizen-Reported Urban Challenges: A Comprehensive Time Series Analysis. *Big Data Cogn. Comput.* **2024**, *8*, 27.
21. Indrahimawan, M.R.; Santosa, P.I.; Adji, T.B. Handling Data Imbalance Using Text Augmentation For Classifying Public Complaints. In Proceedings of the 2023 International Conference on Computer, Control, Informatics and its Applications (IC3INA), Bandung, Indonesia, 4–5 October 2023; pp. 284–289.
22. Obaid, S.; Ilyas, A. Ensemble approach for Complaint Classification. In Proceedings of the International Conference on Biological Research and Applied Science, Karachi, Pakistan, 23–25 January 2024; p. 50.

23. Li, X.; Shu, Q.; Kong, C.; Wang, J.; Li, G.; Fang, X.; Lou, X.; Yu, G. An intelligent system for classifying patient complaints using machine learning and natural language processing: Development and validation study. *J. Med. Internet Res.* **2025**, *27*, e55721.
24. Cai, H.; Shao, X.; Zhou, P.; Li, H. Multi-Label Classification of Complaint Texts: Civil Aviation Service Quality Case Study. *Electronics* **2025**, *14*, 434.
25. Aditama, A.R.; Wicaksono, A.F. Classification of customer complaints on social media for e-commerce in Indonesia. *Int. J. Electr. Comput. Eng. (IJECE)* **2025**, *15*, 2977–2985.
26. Jondhale, R.; Patil, S.; Shinde, A.; Ajalkar, D.; Biradar, S. Predicting Consumer Complaint Disputes in Finance Using Machine Learning (AIOPS). In Proceedings of the 2024 Second International Conference on Advances in Information Technology (ICAIT), Chikkamagaluru, India, 24–27 July 2024; Volume 1, pp. 1–7.
27. Rahman, A.; Zamzami, E.; Sawaluddin, S. Performance Improvement of Support Vector Machine with Firefly Algorithm for Public Complaints Classification. In Proceedings of the 2024 2nd International Conference on Technology Innovation and Its Applications (ICTIIA), Medan, Indonesia, 12–13 September 2024; pp. 1–5.
28. Rani, N.; Singh, D.; Saha, B.; Shukla, S.K. Automated Classification of Cybercrime Complaints using Transformer-based Language Models for Hinglish Texts. *arXiv* **2024**, arXiv:2412.16614.
29. Airani, P.; Pipada, N.; Shah, P. Classification of Complaints Text Data by Ensembling Large Language Models. In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence—Volume 3: ICAART. INSTICC*; SciTePress: Setúbal, Portugal, 2025; pp. 679–686.
30. Mazhar, H.A.; Altaf, U.; Anwar, S.M.; Bhutta, M.S. Auto-Classification of FIA-Cybercrime Wing Complaints Using Bidirectional Encoder Representations from Transformers Model. *UCP J. Eng. Inf. Technol. (HEC Recognized-Y Categ.)* **2024**, *2*, 1–11.
31. Jurn, S.; Kim, W. Improving Text Classification of Imbalanced Call Center Conversations Through Data Cleansing, Augmentation, and NER Metadata. *Electronics* **2025**, *14*, 2259.
32. Oliveira, L.L.; Jiang, X.; Babu, A.N.; Karajagi, P.; Daneshkhah, A. Effective natural language processing algorithms for early alerts of gout flares from chief complaints. *Forecasting* **2024**, *6*, 224–238.
33. Xiangfu, K.; Bo, D.; Ke, X.; Yongliang, T. Text classification model for livelihood issues based on BERT: A study based on hotline compliant data of Zhejiang province. *Beijing Da Xue Xue Bao* **2023**, *59*, 456–466.
34. Laila, K.; Viswanathan, V.; Senoj, S.; Stanley, A. Enhancing Tenant-Owner Communication via BERT-Based Severity Prediction of Housing Complaints. *Int. J. Sci. Eng. Dev. Res.* **2025**, *10*, 298–308.
35. Zheng, C.; Zhang, Y.; Lian, X.; Ke, J.; Chen, H.; Chen, Y. Impact of hospital complaint handling on promoting high-quality development of hospitals via an emotional language analysis model: A case study of a tertiary hospital service center in Quanzhou city, Fujian province. *Front. Health Serv.* **2025**, *5*, 1610004.
36. Lin, Y.T.; Deng, Y.X.; Tsai, C.L.; Huang, C.H.; Fu, L.C. Interpretable deep learning system for identifying critical patients through the prediction of triage level, hospitalization, and length of stay: Prospective study. *JMIR Med. Inform.* **2024**, *12*, e48862.
37. Zhang, X.; Wang, Y.; Jiang, Y.; Pacella, C.B.; Zhang, W. Integrating structured and unstructured data for predicting emergency severity: An association and predictive study using transformer-based natural language processing models. *BMC Med. Inform. Decis. Mak.* **2024**, *24*, 372.
38. Razzaq, K.; Shah, M. Machine learning and deep learning paradigms: From techniques to practical applications and research frontiers. *Computers* **2025**, *14*, 93.
39. Uddin, S.; Yan, S.; Lu, H. Machine learning and deep learning in project analytics: Methods, applications and research trends. *Prod. Plan. Control* **2025**, *36*, 873–892.
40. Yi, S.; Li, X.; Wang, R.; Guo, Z.; Dong, X.; Liu, Y.; Xu, Q. Interpretable spatial machine learning insights into urban sanitation challenges: A case study of human feces distribution in San Francisco. *Sustain. Cities Soc.* **2024**, *113*, 105695.
41. Xu, G.; Liu, H.; Jia, C.; Zhou, T.; Shang, J.; Zhang, X.; Wang, Y.; Wu, M. Spatiotemporal patterns and drivers of public concern about air pollution in China: Leveraging online big data and interpretable machine learning. *Environ. Impact Assess. Rev.* **2025**, *114*, 107897.
42. Jiang, Y.; Zhao, S.; Li, H.; Wu, H.; Zhu, W. Urban Spatiotemporal Event Prediction Using Convolutional Neural Network and Road Feature Fusion Network. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 341.
43. Asif, N.; Hong, Z.; Ren, S.; Zhang, X.; Shang, X.; Yuan, Y. MuST2-Learn: Multi-view Spatial-Temporal-Type Learning for Heterogeneous Municipal Service Time Estimation. *arXiv* **2025**, arXiv:2508.16503.
44. Sui, J.; Chen, P.; Gu, H. Deep Spatio-Temporal Graph Attention Network for Street-Level 110 Call Incident Prediction. *Appl. Sci.* **2024**, *14*, 9334.
45. Liu, X.; Zhi, X.; Zhou, T.; Zhao, L.; Tian, L.; Gao, R.; Luo, J.; Cui, W.; Wang, Q. A holistic air monitoring dataset with complaints and POIs for anomaly detection and interpretability tracing. *Sci. Data* **2025**, *12*, 1288.
46. Guo, C.; Berkhahn, F. Entity embeddings of categorical variables. *arXiv* **2016**, arXiv:1604.06737.
47. Arik, S.Ö.; Pfister, T. Tabnet: Attentive interpretable tabular learning. *Proc. Proc. Aaai Conf. Artif. Intell.* **2021**, *35*, 6679–6687.

48. Badaro, G.; Saeed, M.; Papotti, P. Transformers for tabular data representation: A survey of models and applications. *Trans. Assoc. Comput. Linguist.* **2023**, *11*, 227–249.
49. Jiang, J.P.; Liu, S.Y.; Cai, H.R.; Zhou, Q.L.; Ye, H.J. Representation learning for tabular data: A comprehensive survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2026**, *48*, 6488–6508.
50. Cheng, K.; Li, X.; Wang, Z.; Zhang, C.; Huang, B.; Xu, Y.E.; Dong, X.L.; Sun, Y. Tab-cleaner: Weakly supervised tabular data cleaning via pre-training for E-commerce catalog. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 172–185.
51. Kasneci, G.; Kasneci, E. Enriching tabular data with contextual llm embeddings: A comprehensive ablation study for ensemble classifiers. *arXiv* **2024**, arXiv:2411.01645.
52. Koloski, B.; Margeloiu, A.; Jiang, X.; Škrlj, B.; Simidjievski, N.; Jamnik, M. Llm embeddings for deep learning on tabular data. *arXiv* **2025**, arXiv:2502.11596.
53. Hoppe, F.; Kleinemeier, L.; Franz, A.; Göbel, U. Comparing Task-Agnostic Embedding Models for Tabular Data. *arXiv* **2025**, arXiv:2511.14276.
54. Esperança, M.; Freitas, D.; Paixão, P.V.; Marcos, T.A.; Martins, R.A.; Ferreira, J.C. Proactive Complaint Management in Public Sector Informatics Using AI: A Semantic Pattern Recognition Framework. *Appl. Sci.* **2025**, *15*, 6673.
55. Rakhimzhanov, D.; Belginova, S.; Yedilkhan, D. Automated Classification of Public Transport Complaints via Text Mining Using LLMs and Embeddings. *Information* **2025**, *16*, 644.
56. Tarekegn, A.N.; Ullah, M.; Cheikh, F.A. Deep learning for multi-label learning: A comprehensive survey. *arXiv* **2024**, arXiv:2401.16549.
57. Touw, D.J.; van de Velden, M. Classifier chain networks for multi-label classification. *arXiv* **2024**, arXiv:2411.02638.
58. Yadav, P.S.; Rao, R.S.; Mishra, A. An evaluation of multi-label classification approaches for method-level code smells detection. *IEEE Access* **2024**, *12*, 53664–53676.
59. Ye, H.J.; Yin, H.H.; Zhan, D.C.; Chao, W.L. Revisiting nearest neighbor for tabular data: A deep tabular baseline two decades later. *arXiv* **2024**, arXiv:2407.03257.
60. Thielmann, A.F.; Samiee, S. On the efficiency of nlp-inspired methods for tabular deep learning. *arXiv* **2024**, arXiv:2411.17207.
61. Gorishniy, Y.; Kotelnikov, A.; Babenko, A. Tabm: Advancing tabular deep learning with parameter-efficient ensembling. *arXiv* **2024**, arXiv:2410.24210.
62. Gorishniy, Y.; Rubachev, I.; Khrulkov, V.; Babenko, A. Revisiting deep learning models for tabular data. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 18932–18943.
63. Ahamed, M.A.; Cheng, Q. MambaTab: A plug-and-play model for learning tabular data. In *Proceedings of the 2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*; IEEE: Piscataway, NJ, USA, 2024; pp. 369–375.
64. Liu, Q.; Guo, A.; Qiao, F.; Ma, X.; Liu, Y.A.; Huang, Y.; Wang, R.; Sheng, C. Skillful bias correction of offshore near-surface wind field forecasting based on a multi-task machine learning model. *Atmos. Ocean. Sci. Lett.* **2025**, *18*, 100590.
65. Fan, Y.; Waldmann, P. Tabular deep learning: A comparative study applied to multi-task genome-wide prediction. *BMC Bioinform.* **2024**, *25*, 322.
66. Thielmann, A.F.; Kumar, M.; Weisser, C.; Reuter, A.; Säfken, B.; Samiee, S. Mambular: A sequential model for tabular deep learning. *arXiv* **2024**, arXiv:2408.06291.
67. Somvanshi, S.; Das, S.; Javed, S.A.; Antarksa, G.; Hossain, A. A survey on deep tabular learning. *arXiv* **2024**, arXiv:2410.12034.
68. Borisov, V.; Leemann, T.; Seßler, K.; Haug, J.; Pawelczyk, M.; Kasneci, G. Deep Neural Networks and Tabular Data: A Survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 7499–7519.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.