Urban Waterbody Conservation Through Knowledge Guided Machine Learning : An Integrated Approach Using Experts Insights

Abstract

Urban waterbodies are increasingly under stress due to unplanned development, ecological neglect, and fragmented policy implementation. Traditional conservation planning often suffers from delays and inconsistent decision-making, especially when domain knowledge is not systematically integrated into the process. This study presents a novel framework that leverages Knowledge-Guided Machine Learning (KGML) to support timely, context-sensitive urban waterbody conservation efforts. The method integrates expert insights with natural language processing to analyze policy documents and match them with issues identified through field assessment. Using a case study from Delhi, the framework demonstrates how policy-aligned recommendations—such as “buffer zone demarcation” and “community participation”—can be extracted efficiently, improving both the speed and relevance of decision-making. A second case study highlights the real-world steps involved in conservation planning, offering a practical perspective to complement the model-driven approach. Together, these cases illustrate the potential of KGML to enhance urban waterbody management by bridging data, policy, and expert knowledge in a scalable and adaptable manner.

Keywords : *Urban Waterbody Conservation, Knowledge-Guided Machine Learning (KGML), Policy Text Mining, Decision-Support Framework, Expert-Informed Modelling, Textual Analysis for Environmental Planning, Sustainable Urban Governance.*

1. Introduction

Urban waterbodies such as lakes, ponds, and wetlands are critical ecological assets within growing cities. Yet across India and many developing regions, these systems are rapidly deteriorating due to land encroachment, pollution, and lack of coordinated governance. Despite the presence of policy documents and conservation schemes, there remains a disconnect between what’s prescribed and what’s practiced. One key reason is the absence of a structured, data-informed system to interpret these policies in light of local conditions and site-specific challenges.

Conservation planning typically involves manual review of extensive government documents, expert consultations, and piecemeal field assessments. This not only slows the process but can result in fragmented or poorly matched interventions. Recent advancements in machine learning, especially those guided by domain expertise, offer promising opportunities to address this gap. Knowledge-Guided Machine Learning (KGML) frameworks have shown success in environmental prediction, such as large-scale surface water change detection and lake

temperature modeling. However, their application in the policy-text domain, especially in the context of urban waterbody conservation, remains underexplored.

This study responds to that gap by presenting an integrated KGML approach that combines expert insights, natural language processing, and policy document analysis to support conservation planning. Unlike black-box ML tools, this method emphasizes interpretability, allowing recommendations to be grounded in both regulatory frameworks and field realities. The research is built around two complementary case studies in Delhi, one that demonstrates the model's application for issue-policy mapping, and another that documents the procedural steps involved in actual conservation implementation. Through these studies, the paper aims to contribute a scalable yet grounded method for bridging knowledge, data, and governance in urban waterbody management.

2. Literature Review

Urban waterbody conservation has been a growing area of concern across environmental planning, especially in rapidly urbanizing regions such as India. Numerous studies have identified degradation due to encroachment, declining water quality, and ecological loss, prompting government bodies to formulate policies aimed at sustainable restoration. However, implementation often falls short due to a lack of structured, data-driven support mechanisms. Most conservation efforts remain fragmented, heavily reliant on expert discretion, and are disconnected from the vast body of policy and guideline documents that already exist.

In recent years, machine learning techniques have begun to find a place in environmental monitoring, particularly in modeling land use change, water quality parameters, and hydrological forecasting. Yet, these applications often prioritize predictive accuracy over interpretability, limiting their relevance in decision-making processes that require transparency and accountability. To address this, a shift toward Knowledge-Guided Machine Learning (KGML) has emerged. KGML integrates domain expertise into machine learning workflows, allowing models to respect known relationships and constraints while still learning from data. This is especially useful in complex systems like urban waterbodies, where both environmental and institutional dynamics are at play.

Two recent studies serve as important foundations for this work. The first, conducted across the contiguous United States, used KGML to classify over 100,000 lakes and reservoirs based on surface area change patterns. By embedding ecological knowledge into the model, the study demonstrated how KGML could enhance the interpretability of large-scale remote sensing analysis. A second study focused on lake water temperature modeling, applying a process-guided deep learning framework to incorporate physical laws into the prediction process. Both studies

underscore the growing viability of KGML in environmental contexts but neither explore its use in textual analysis or policy-driven decision-making.

In the Indian context, the use of KGML remains nascent, particularly for urban governance applications. Conservation plans often remain isolated from analytical models, with little attempt to match field issues to existing policy frameworks. At the same time, the potential of natural language processing (NLP) tools like spaCy and BERT to interpret textual documents remains largely untapped in the waterbody conservation domain. This gap presents a unique opportunity: to build a KGML-based framework that not only analyzes field data but also interprets policy texts, bridging the divide between environmental data and regulatory action.

3. Methodology

This study adopts a structured, multi-stage methodology that combines domain knowledge, expert input, machine learning, and policy document analysis to guide urban waterbody conservation planning. The aim is to develop a proof-of-concept framework that enables faster and more context-aware decision-making using a Knowledge-Guided Machine Learning (KGML) approach. The methodology is tested through two case studies situated in Delhi, India.

3.1 Study Design and Scope

The study focuses on urban waterbodies within the National Capital Territory (NCT) of Delhi. Given the exploratory nature of the research and limitations related to data availability and scale, a local-level application was chosen to validate the approach. The selected site, Sanjay Lake, represents a typical urban waterbody facing ecological stress, governance fragmentation, and encroachment issues.

3.2 Data Sources

The methodology relies on three main categories of data:

1. Field Assessment Reports: Existing assessments of the lake condition, including catchment analysis, pollution sources, and land use changes.
2. Policy and Guideline Documents: National and state-level conservation frameworks, such as the Wetlands Rules (2017), Jal Shakti Abhiyan materials, and Delhi Master Plans (2001 and 2021).
3. Expert Inputs: Semi-structured interviews were conducted with professionals in urban planning, hydrology, and environmental governance to identify critical factors and validate model outputs.

3.3 Framework Process

The methodology was designed in the following steps:

1. Problem Identification & Factor Finalization
   Domain experts helped shortlist key issues relevant to urban waterbody conservation (e.g., catchment encroachment, buffer zone loss, untreated input water). These were used to create a base "issue vocabulary" for the model.
2. Textual Corpus Preparation
   Policy documents were converted into machine-readable formats and preprocessed using spaCy NLP pipeline. Texts were cleaned, tokenized, and lemmatized for consistent comparison.
3. Keyword Matching & KGML Model Training
   The KGML approach was applied to find overlaps between identified issues and their corresponding treatment or policy solutions within the corpus. The model was guided by the expert-derived issue list and used rule-based weighting to emphasize conservation-relevant keywords.
4. Recommendation Matrix Generation
   Based on matched phrases, a matrix of Issue–Policy–Recommendation was generated. This matrix highlights the policy-backed actions for specific challenges observed in the field.
5. Validation Through Case Studies
   a. Case Study 1 applied the model to identify relevant keywords and matched them with conservation policy actions.
   b. Case Study 2 looked at a live conservation effort to map actual tasks and assess the alignment with model outputs.

3.4 Tools and Technologies

i. spaCy NLP for text processing
ii. Python (custom scripts) for KGML implementation
iii. Manual scoring + rule-based filters for enhanced interpretability
iv. QGIS and field maps for spatial referencing (in Case Study 2)

3.5 Limitations

The study is limited by:

1) A relatively small, localized dataset
2) Reliance on available English-language policy documents
3) The early-stage, exploratory nature of the KGML model
   However, these constraints also reinforce the adaptability of the framework, which can be scaled with improved data and domain coverage in future studies.

4. Case Studies & Application

To demonstrate the real-world applicability and performance of the proposed KGML framework, two complementary case studies were conducted. These cases serve to validate both the model's analytical capability and its relevance to on-ground conservation processes.

4.1 Case Study 1: KGML Model Demonstration on Sanjay Lake

The first case study was designed as a proof-of-concept to test the KGML framework's ability to identify and align conservation challenges with actionable policy responses. Using the Sanjay Lake area in East Delhi as the study site, textual assessment reports were analyzed to extract site-specific issues such as encroachment of the catchment area, obstruction in natural flow, and lack of buffer delineation.

These issues were processed through the trained KGML model, which referenced national and local conservation policies to suggest context-specific interventions. For instance, the model identified terms like "buffer zone demarcation" and "community participation" as highly relevant but previously underutilized policy-aligned responses. The recommendation matrix developed through this model enabled a quicker, more structured mapping between observed problems and governance-backed solutions.

The results demonstrated that the KGML method can surface targeted conservation strategies that are both evidence-informed and regulatory-compliant—saving considerable time compared to manual document reviews.

4.2 Case Study 2: Observing Conservation in Practice

While the first case study showcased the model's analytical power, the second case study aimed to understand the practical sequence of steps involved in conserving an urban waterbody. A different lake undergoing active restoration (also located in the Delhi NCT region) was selected to document on-ground efforts, from preliminary assessment to intervention.

Field visits and document analysis revealed a structured process involving:

a) Mapping of physical characteristics and threats
b) Stakeholder consultations
c) Prioritization of interventions (e.g., wetland creation, treated water input)
d) Biodiversity recovery actions (such as removal of invasive species)
e) Community outreach and engagement initiatives

This descriptive mapping of real-world conservation workflows helped contextualize the model's outputs from the first case study. It confirmed that many of the policy-aligned recommendations generated by the model—such as **afforestation**, **wetland polishing**, and **citizen training**—were not only theoretically appropriate but also practically relevant.

4.3 Linking Model and Practice

Together, these two case studies illustrate the dual utility of the proposed framework:

- Case Study 1 validates its potential as a **rapid decision-support tool** grounded in both policy and expert logic.
- Case Study 2 connects the model's outputs to **tangible conservation tasks**, showing how analytical insights can complement real-world planning.

The combination provides both **technical validation** and **field-level relevance**, establishing a strong basis for the framework's broader adoption and further refinement.

5. Results & Interpretation

The proposed KGML framework produced a set of structured outputs that align conservation challenges with specific policy-backed recommendations. The results from Case Study 1 (Sanjay Lake) and insights from Case Study 2 are interpreted below in terms of issue extraction, policy matching, expert validation, and system-level observations.

5.1 Identified Issues and Model-Matched Responses

Using the KGML model, key challenges at the Sanjay Lake site were extracted from field reports and mapped against relevant policy text. The following table (summarized below) captures the relationship between issue categories, supporting policy sections, and model-suggested recommendations:

Identified Issue	Policy Basis	Model-Suggested Recommendation
Catchment Encroachment	Sec. 2.1 - Catchment Conservation	Afforestation, zoning enforcement
Flow Obstruction	Sec. 1.2 - Buffer Delineation	Restore buffer zones
Treated Water Input	Sec. 2.1 - Pollution Management	Construct wetland/polishing ponds
Water Pollution	CPCB Toolkit Guidelines	Monitor DO, BOD, salinity, hardness
Ecosystem Loss	Biodiversity Action Plan	Remove invasives, reintroduce native species
Public Role & Participation	Community Governance Framework	Train citizen groups, conduct awareness drives

This matrix reflects how the KGML model successfully bridges textual field data and policy documents. Recommendations such as **buffer zone restoration** and **community participation**, which were underemphasized in initial manual reviews, were effectively highlighted by the model.

5.2 Expert Validation and Usefulness

Semi-structured interviews with domain experts validated the relevance of the model-generated outputs. Experts confirmed that the matched responses were not only aligned with best practices but also covered several overlooked areas. In particular, they appreciated the system's ability to structure conservation recommendations in a manner that facilitates **faster decision-making** for planners and practitioners.

Experts also highlighted the usefulness of the model in:

1) Shortlisting actionable interventions
2) Clarifying policy ambiguity
3) **Saving manual review time**, particularly across large volumes of policy text

5.3 Real-World Alignment from Case Study 2

Case Study 2—focused on observing actual conservation work—revealed that several of the model-generated recommendations (e.g., afforestation, treated water input, stakeholder engagement) were already part of the on-ground action plan. This overlap reinforces the validity of the model and demonstrates its **potential to support ongoing or future interventions**.

5.4 Interpretability and Flexibility

One of the strengths of the KGML approach was its **transparency**. By relying on rule-based matching, expert weighting, and document structuring, the model maintained interpretability—making it suitable for governance contexts that demand traceable logic.

Moreover, the approach is **flexible**. The methodology can be extended to:

1) Other cities or regions with similar conservation contexts
2) Additional thematic domains such as wetland protection, river restoration, or green space planning
3) Larger corpora, including multilingual or GIS-tagged documents

6. Discussion

The results of this study offer valuable insights into both the **capabilities and limitations** of using a Knowledge-Guided Machine Learning (KGML) framework in urban waterbody conservation. This section discusses the implications of the findings, positioning the research within current practice while highlighting future potential.

6.1 Addressing Fragmentation in Conservation Planning

Urban waterbody conservation in India is often hindered by fragmented institutional responsibilities, lack of integrated planning, and the underuse of existing policy resources. By integrating policy text into a structured, machine-readable format, this study bridges the **information gap** between conservation field assessments and regulatory frameworks. The model aids in **reclaiming underutilized knowledge** embedded in policy documents and turning it into actionable insights. This can accelerate workflows and improve coordination among stakeholders.

6.2 Bridging Expertise and Automation

One of the unique contributions of this study is its ability to merge **human expertise with automation**. While ML models often operate as black boxes, the KGML approach embeds expert judgment into the system design, especially in selecting key issues and assigning weights to thematic areas. This not only improves model accuracy but ensures the output remains **interpretable and trustworthy**, which is essential in governance-related applications.

6.3 Relevance Beyond the Case Study

Although the model was tested on a localized case (Sanjay Lake), the approach is inherently scalable. The rule-based, document-guided structure allows for replication across other waterbodies, cities, or even thematic domains beyond urban lakes. Additionally, its focus on **modular tasks**—such as issue identification, policy alignment, and recommendation generation—makes it adaptable to varied conservation challenges.

The approach also complements recent global KGML efforts (e.g., the US-wide ReaLSAT study or temperature modeling with physics-guided networks) by adding a **textual-policy layer** not previously addressed.

6.4 Limitations and Critical Reflections

Despite its strengths, this study has a few notable limitations:

1) The model relies on English-language documents, excluding regional or multilingual policies.
2) A smaller data corpus and a focused case limit the generalizability at this stage.

3) Manual preprocessing (e.g., expert curation of issue lists) can introduce bias if not reviewed systematically.

These limitations highlight the need for further technical refinement and expanded datasets, particularly as the framework evolves toward broader, real-time applications.

6.5 Theoretical and Practical Contributions

Theoretically, the study advances the application of KGML into **governance and planning documents**, which have traditionally been outside the reach of environmental ML models. Practically, it contributes a tested, flexible method that can reduce turnaround time for conservation decisions and help planners align site-level issues with national strategies.

7. Conclusion and Future Work

This study introduced a Knowledge-Guided Machine Learning (KGML) framework designed to support urban waterbody conservation by aligning site-level issues with policy-guided solutions. By combining expert insights, field data, and natural language processing of policy documents, the model offers a decision-support system that is both interpretable and adaptable.

The framework was applied to a focused case study—Sanjay Lake in Delhi—demonstrating how conservation challenges such as catchment encroachment, pollution, and ecosystem loss can be linked to targeted, policy-aligned recommendations. A second case study illustrated the sequence of real-world conservation activities, validating the practical relevance of the model's outputs. Together, these case studies underscore the dual role of the proposed method: as a tool for rapid analytical insight and as a guide for on-ground action.

The findings highlight several key contributions:

1) A novel application of KGML in textual, policy-oriented environmental decision-making.
2) A structured way to interpret and act upon field observations using existing guidelines.
3) A replicable and modular approach for other urban waterbodies and planning domains.

Future Work

There are several promising directions for extending this research:

Scaling the Framework: Applying the model across multiple urban lakes or water systems to validate its robustness at a regional or national scale.

Multilingual Capability: Incorporating regional-language policy texts to broaden applicability, especially in linguistically diverse contexts like India.

Automated Policy Corpus Expansion: Using web scraping and AI-driven document curation to update the policy database dynamically.

Integration with Geospatial Tools: Linking outputs to GIS-based visualizations for more intuitive planning and stakeholder engagement.

User-Centered Interfaces: Developing a front-end tool that allows urban planners or local authorities to input site-specific issues and receive model-driven recommendations in real-time.

In a policy environment where delays and mismatches between plans and practice are common, this study shows how machine learning—when guided by domain knowledge—can bridge the gap between data and decisions. With further refinement and wider deployment, the proposed framework has the potential to enhance transparency, responsiveness, and effectiveness in the conservation of urban water resources.