**Journal of Information Systems Engineering and Management**
2023, 8(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

# BioFilter-MoE: Multi-Task Mixture-of-Experts Transformer for Textile Wastewater Prediction and Treatment Recommendation

Nusrat Yasmin Nadia[1,*], Habibor Rahman Rabby[2], Md Habibul Arif[3], Abdur Rahman Lindon[3], Hafiz Aziz Khan[3]

[1]Department of Textile Engineering, Ahsanullah University of Science & Technology, 141 & 142, Love Road, Tejgaon Industrial Area, Dhaka-1208, Bangladesh

[2]Department of Computer Science, Campbellsville University, 2300 Greene Way #100, Louisville, KY 40220, USA

[3]Department of Information Technology, Washington University of Science and Technology, 2900 Eisenhower Ave, Alexandria, VA 22314, USA

**ARTICLE INFO**

Received: 15 Nov 2023

Accepted: 28 Dec 2023

Published: 30 Dec 2023

**ABSTRACT**

Textile dyeing effluents exhibit highly non-linear physicochemical dynamics that render conventional rule-based treatment optimization inadequate for real-time control. Existing data-driven approaches address effluent forecasting in isolation, neglecting the operational necessity of simultaneous treatment prescriptions. This study proposes BioFilter-MoE, a Multi-Task Tabular Transformer with Sparse Mixture-of-Experts (MoE) routing, trained jointly to predict multi-parameter effluent quality and recommend optimal biofilter media configurations from a single unified architecture. A Fourier-based numerical tokenization scheme encodes continuous thermodynamic variables alongside categorical media constraints into a shared embedding space, while sparse MoE layers dynamically route distinct influent states to specialized expert sub-networks, revealing chemically coherent specialization without post-hoc attribution. Dual Cross-Attention task heads support simultaneous regression and classification, with multi-task loss balanced via homoscedastic uncertainty weighting. Under five-fold cross-validation, BioFilter-MoE achieves BOD $R^2 = 0.7914 \pm 0.0225$ and COD $R^2 = 0.8081 \pm 0.0205$. On the held-out test set, it attains a Macro F1 of 0.9948 [0.9833, 1.000] and ROC-AUC of 0.9998 for biofilter media recommendation, compared to Macro F1 scores of 0.046 for both Random Forest and XGBoost. Ablation studies confirm that removing the Fourier tokenizer causes the largest single-component degradation, reducing mean $R^2$ from 0.655 to 0.597. BioFilter-MoE thus establishes a prescriptive decision-support framework for regulatory-compliant, real-time textile wastewater management with architecture-embedded interpretability.

**Keywords:** Multi-Task Learning, Mixture-of-Experts, Tabular Transformer, Textile Wastewater Treatment, Biofilter Media Recommendation

Journal of Information Systems Engineering and Management
2023, 8(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

## I. Introduction

The textile and dyeing industry ranks among the most water-intensive and environmentally damaging sectors of global manufacturing. The dyeing and finishing department alone accounts for approximately 17–20% of total industrial effluent worldwide [1], discharging effluents that are chemically complex, thermally stable, and resistant to conventional biological degradation. Annually, over 70 million tons of synthetic dyes are manufactured globally, of which the textile industry consumes more than 10,000 tons [2], with a substantial fraction entering aquatic systems untreated or inadequately treated. Azo dyes, which constitute the largest structural class of textile colorants, are widely recognized as toxic, mutagenic, and potentially carcinogenic, generating aromatic amine degradation products that are frequently more hazardous than their precursor compounds [3]. The resulting effluents impose cascading ecological consequences: inhibition of aquatic photosynthesis, depletion of dissolved oxygen, bioaccumulation of toxic intermediates through aquatic food chains, and direct contamination of groundwater reservoirs serving as primary drinking water sources in textile-producing regions of South and Southeast Asia [3].

The physicochemical characteristics of textile dyeing effluent present a uniquely challenging treatment target. These effluents are characterized by extreme simultaneous values across multiple parameters, high COD, $BOD_5$, total dissolved solids, suspended solids, and heavy metal concentrations frequently exceeding regulatory discharge limits by one to two orders of magnitude [6][4]. Critically, these parameters do not remain stationary across production cycles: influent COD, pH, and dye concentration fluctuate nonlinearly across dye batches, fabric types, and shift changes, generating shock-load events that violate the stationarity assumptions underlying both mechanistic treatment models and conventional statistical monitoring frameworks [5]. Approximately 2,000 distinct chemicals including transfer agents, surfactants, fixatives, and auxiliary compounds are used in textile processing alongside dyes [8], producing effluent matrices of exceptional chemical heterogeneity. Consequently, no single degradation pathway is universally optimal: the relative efficiency of adsorption, biological oxidation, and combined anaerobic-aerobic processes varies substantially depending on instantaneous influent composition. Biological filtration has emerged as a cost-effective and environmentally compatible treatment strategy for textile wastewater, particularly in resource-constrained industrial settings [9][6]. Biofilter systems employ packed media including granular activated carbon, natural zeolites, and composite biochar to support microbial communities that degrade organic pollutants through sorption-coupled biological oxidation. However, media performance is highly sensitive to influent COD concentration, pH, hydraulic loading, and microbial community composition; incorrect media selection leads to premature saturation, reduced removal efficiency, and regulatory non-compliance [10]. In practice, this selection is performed manually by process engineers using heuristic guidelines derived from static laboratory studies an approach that is slow, inconsistent across operators, and fundamentally ill-suited to the dynamic influent chemistry of industrial textile dyeing operations [4, 7].

The integration of machine learning into wastewater treatment plant management has produced meaningful advances in effluent quality prediction over the past decade. Ensemble methods including Random Forest, XGBoost, and Gradient Boosting have demonstrated competitive predictive accuracy for COD, BOD, and nutrient parameters on tabular plant operational data [11, 12]. Deep learning architectures, particularly Long Short-Term Memory networks and hybrid CNN-LSTM models, have further improved prediction under temporally correlated influent conditions [13, 14]. More recently, tabular transformer architectures including TabNet [15] and the Feature Tokenizer Transformer [16] have demonstrated that self-attention mechanisms can capture complex pairwise feature interactions in mixed-type industrial datasets that tree-based splits and recurrent gates cannot efficiently represent. Despite this progress, the urgency of real-time treatment management remains unaddressed at the systems level [8]. Regulatory discharge limits are tightening globally, and the consequences of non-

Journal of Information Systems Engineering and Management
2023, 8(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

compliance extend from ecosystem destruction to severe public health risk. Yet every published data-driven model for wastewater treatment operates as a single-task system, predicting effluent quality parameters in isolation from treatment operational decisions. The downstream question given a predicted effluent state, which treatment configuration should be activated is left entirely to human process engineers, introducing response latency, operator-dependent inconsistency, and a systematic inability to couple predictions with prescriptions under real-time operational constraints. Furthermore, no existing work has applied sparse Mixture-of-Experts (MoE) routing [17, 18] to any wastewater treatment prediction task, despite the compelling theoretical alignment between sparse expert specialization and the physicochemical reality of textile effluent, where distinct influent regimes engage fundamentally different degradation pathways. The interpretability gap compounds this further: existing post-hoc attribution methods such as SHAP [19] explain feature importance without identifying the process-regime transition boundaries that are directly actionable for operational scheduling and regulatory reporting.

This decoupling represents more than an architectural inconvenience. In textile biofiltration contexts, the joint problem of simultaneously predicting effluent COD, BOD, heavy metal concentration, and microbial load while recommending optimal biofilter media configuration is physically coupled: both outputs are conditioned on the same influent physicochemistry and must be operationally consistent to enable compliant treatment [9]. Solving them through separate model pipelines not only doubles the engineering overhead but destroys the representational synergy that joint training could exploit. To address these three compounding gaps, this study introduces BioFilter-MoE a Multi-Task Tabular Transformer with Sparse Mixture-of-Experts routing designed to jointly predict multi-parameter effluent quality and recommend optimal biofilter media configurations from a single unified architecture trained on experimentally collected textile biofiltration data. The specific contributions of this work are fourfold. First, we propose the first application of sparse MoE routing to industrial wastewater treatment prediction, demonstrating that learned expert routing boundaries align with known physicochemical degradation regimes temperature tokens route to a dedicated thermodynamic expert with 0.980 mean probability, and pH to an acid-base regime expert (0.901) without any post-hoc attribution. Second, we formulate the joint effluent prediction and biofilter recommendation problem as a multi-task regression-classification objective balanced through homoscedastic uncertainty weighting, establishing the first prescriptive treatment architecture in the textile wastewater domain. Third, we demonstrate through ablation studies that a Fourier-based numerical tokenizer is the most critical architectural component for regression accuracy, contributing an *8.9%* $R^2$ improvement over a linear projection baseline. Fourth, we provide held-out bootstrap confidence intervals across all five regression targets and the classification head, establishing the statistical reliability of reported results.

The remainder of this paper is organized as follows: Section 2 reviews the relevant literature, Section 3 describes the dataset and methodology, Section 4 presents experimental results and ablation analyses and Section 5 concludes.

## II. Literature Review

**2.1 Data-Driven Effluent Prediction (From Statistical Models to Ensemble Dominance)**: The use of data-driven models for wastewater treatment plant (WWTP) performance prediction has expanded steadily over the past two decades, evolving from simple statistical regression approaches toward increasingly sophisticated machine learning architectures. Early efforts relied on Artificial Neural Networks (ANNs), which demonstrated broad applicability across effluent parameters, including biochemical oxygen demand (BOD), chemical oxygen demand (COD), total suspended solids (TSS), and nutrient concentrations [10].

Journal of Information Systems Engineering and Management
2023, 8(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Despite their nonlinear modeling capacity, dense ANN architectures are brittle under shock-load influent conditions, offer limited interpretability, and degrade in performance when training data are sparse or noisy constraints that disproportionately affect industrial treatment settings characterized by batch variability [11]. Tree-based ensemble methods subsequently emerged as the dominant paradigm for tabular WWTP data. Wang et al. [12] proposed a machine learning framework integrating Random Forest and gradient boosting for real-time effluent quality control at a full-scale WWTP, demonstrating that ensemble models improve prediction accuracy by up to 20% over single-model baselines in the validation stage. Ching et al. [13] developed an XGBoost soft sensor for influent $BOD_5$ prediction across two independent WWTPs, establishing XGBoost's robustness against extreme influent values compared to ANN and SVM baselines, with the model achieving competitive RMSE values on datasets of 527 and 2,189 readings, respectively. Qambar and Al Khalidy [14] Applied ensemble ML to optimize dissolved oxygen setpoints and aeration energy consumption, confirming that tree-based models capture the nonlinear physicochemical relationships governing biological treatment without requiring explicit process equations. While these studies represent meaningful progress, they share a structural limitation of direct relevance to the present work: they are formulated exclusively as single-task regressors predicting one or several continuous effluent parameters. No mechanism exists within these architectures to simultaneously recommend treatment operational interventions a gap that forces process engineers to operate two separate pipelines and manually integrate predictions with operational decisions.

**2.2 Deep Learning for Temporal Effluent Dynamics (Progress and Residual Gaps):** Recognizing that batch-process industrial effluents exhibit temporal dependencies that static ensemble models cannot capture, the research community progressively adopted sequence-aware deep learning architectures. Farhi et al. [15] developed LSTM-based soft sensors for COD, $NH_4^+$, and total nitrogen prediction in a two-stage anoxic-oxic WWTP process, demonstrating that LSTM with optimized look-back periods substantially outperforms multiple linear regression baselines a result attributable to LSTM's capacity to model the time-lag characteristics inherent in biological degradation processes [16]. Wang et al. [17] proposed a deep COD prediction model for urban sewage using CNN-LSTM architectures, achieving real-time prediction by encoding both spatial process features and temporal effluent dynamics simultaneously. Yaqub et al. [18] trained stacked LSTM models to predict membrane bioreactor removal efficiency, reporting strong performance on controlled laboratory data but acknowledging significant performance degradation when influent compositions deviated from training distributions [19]. These recurrent architectures, despite their temporal modeling capacity, introduce structural limitations when applied to batch-process industrial tabular data. They impose sequential inductive biases inappropriate for data collected under irregular batch intervals, cannot natively encode interactions between heterogeneous feature types specifically the combination of continuous physicochemical measurements and categorical operational constraints and require large, temporally ordered datasets that are rarely available at real industrial textile treatment facilities. Furthermore, and critically for the present study, every cited LSTM and GRU architecture operates as a single-task regressor or classifier: effluent quality prediction and treatment prescription are never formulated as jointly optimized objectives within a shared representation [20].

**2.3Transformer Architectures for Tabular Data (An Emerging Paradigm):** The introduction of the Transformer architecture by Vaswani et al. [21] fundamentally altered the landscape of sequence modeling through the self-attention mechanism, which computes pairwise feature interactions across all input positions simultaneously. The subsequent adaptation of this paradigm to tabular data where features rather than tokens occupy attention positions opened a

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

principled pathway for modeling complex physicochemical interactions without the sequential ordering constraints of recurrent networks. Arik and Pfister [22] proposed TabNet, which employs sequential attention for sparse feature selection at each decision step, producing interpretable feature masks while achieving competitive accuracy on tabular benchmarks. TabNet's architecture enables instance-wise feature selection a property of operational relevance in WWTP contexts where different influent states engage different subsets of predictive variables. However, TabNet's sequential decision steps impose a fixed depth on the reasoning chain and its single-task design does not naturally extend to joint regression-classification objectives. Gorishniy et al.[23] introduced the Feature Tokenizer Transformer (FT-Transformer), which encodes each tabular feature as a learnable embedding and applies standard multi-head self-attention across the full feature set, demonstrating consistent superiority over gradient-boosted decision trees on heterogeneous tabular benchmarks when feature interactions are high-dimensional and non-monotonic. Somepalli et al. [24] extended this line of work through SAINT (Self-Attention and Intersample Attention Transformer), applying attention both across features and across samples to capture inter-record relationships a design motivated by the observation that tabular records in environmental datasets are often drawn from non-independent process states. Despite these architectural advances, none of these transformer variants have been applied to wastewater treatment prediction as of the time of writing. Their deployment remains concentrated in benchmark tabular datasets from domains such as income classification, medical tabular data, and e-commerce. The translation to industrial environmental prediction particularly under the mixed-type feature regime of textile biofiltration data represents an open and timely contribution. Critically, all existing tabular transformer architectures are formulated as single-task models. The extension to joint regression-classification via multi-task learning heads has not been proposed in any prior work applying these architectures to environmental systems [25].

**2.4 Multi-Task Learning (Theoretical Motivation and Environmental Gap):** Multi-task learning (MTL), in which a shared model is trained to minimize losses across multiple related objectives simultaneously, has achieved state-of-the-art performance across NLP, computer vision, and biomedical domains [26]. The theoretical justification for MTL rests on the observation that related tasks share common underlying representations: joint training enforces a regularization effect that prevents overfitting to any single task's noise distribution while improving generalization through shared feature learning. In WWTP contexts, the theoretical case for MTL is direct: effluent quality prediction and treatment media recommendation are not independent problems both are conditioned on identical influent physicochemistry, and their outputs must be operationally consistent to enable meaningful regulatory compliance. Kendall et al. [27] established the foundational multi-task loss balancing technique based on homoscedastic uncertainty weighting, demonstrating that adaptively scaling task-specific gradient contributions according to task-level output uncertainty prevents dominant-task interference and improves joint convergence across heterogeneous task types including the combination of regression and classification objectives directly relevant to the present work. Despite this theoretical foundation, the environmental engineering literature contains no study that formulates WWTP operational optimization as a joint regression-classification problem. Existing multi-output WWTP models either predict multiple continuous effluent parameters simultaneously [17] or classify a single operational variable [18], but none couple continuous effluent prediction with discrete treatment recommendation within a unified loss-balanced architecture. This gap is operationally significant: a system that predicts effluent COD in isolation cannot determine which biofilter media is optimal for the incoming influent state, requiring an entirely separate model, separate training data, and manual operational integration by process engineers [28].

Journal of Information Systems Engineering and Management
2023, 8(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

**2.5 Mixture-of-Experts (Sparse Routing for Specialized Computation):** The Mixture-of-Experts (MoE) architecture, originally proposed by Jacobs et al. [29], addresses the problem of learning multiple specialized sub-functions within a single model through a learned gating network that routes inputs to a subset of expert components. Shazeer et al.[30] extended this to the sparse gating regime, demonstrating that activating only the top-k experts per input rather than all experts simultaneously achieves superior task-specific specialization while maintaining computational efficiency comparable to dense networks. Fedus et al. [31] scaled sparse MoE to trillion-parameter language models through the Switch Transformer, confirming that sparse routing improves both computational efficiency and per-task expert specialization at scale. The translation of MoE routing to small-data, tabular industrial settings has not been explored in the published literature. In the context of textile wastewater treatment, the MoE architectural property maps directly onto physicochemical reality: influent states in different COD regimes, pH ranges, and dye concentration windows engage fundamentally different biological degradation pathways aerobic oxidation, anaerobic reduction, and adsorption-dominated removal that benefit from distinct learned transformations. A sparse routing mechanism that dynamically assigns each influent state to a specialized expert sub-network can isolate these pathways in a data-driven manner, potentially achieving both superior predictive accuracy and interpretable routing boundaries that correspond to known treatment regime transitions. No prior work in environmental or wastewater engineering has exploited this property [32, 33].

**2.6 Textile Wastewater (Domain Specificity and Modeling Challenges):** Textile dyeing effluents represent one of the most chemically complex and environmentally damaging categories of industrial wastewater, characterized by high concentrations of synthetic azo dyes, reactive chromophores, surfactants, and auxiliary chemicals that produce extreme COD variability, recalcitrant organic loads, and sharp pH shock events[34]. Holkar et al.[35] provided a comprehensive critical review of textile wastewater treatment technologies, establishing that the non-stationarity of dye-laden effluent chemistry renders conventional mechanistic treatment models inadequate for real-time operational control a conclusion that motivates the data-driven approach of the present study. Existing machine learning applications in textile wastewater treatment are predominantly laboratory-scale and single-target. Aghilesh et al. [36] employed RSM, ANN, and ANFIS to optimize Methylene Blue dye removal from textile wastewater using low-cost agricultural biosorbents sugarcane bagasse and peanut hulls reporting $R^2$ values above $0.9$ across all three models under controlled experimental conditions of pH, temperature, biosorbent dosage, and initial dye concentration. While these results confirm the capacity of soft-computing methods to model nonlinear adsorption dynamics at laboratory scale, they operate under near-static influent regimes: a single dye species, controlled pH range, and fixed biosorbent configuration. These conditions are fundamentally distinct from real-time, multi-batch industrial biofiltration, where influent COD, pH, and dye concentration fluctuate nonlinearly across production shifts and where the biological community must adapt continuously to heterogeneous organic loads [9]. Critically, biofilter media selection for textile effluent treatment determining whether granular activated carbon, zeolite, or composite biochar media is optimal for a given influent state has received no systematic data-driven attention in the published literature. This selection directly governs biological degradation efficiency, hydraulic retention effectiveness, and long-term media saturation dynamics. Formulating it as a data-driven multi-class classification problem conditioned on real-time influent physicochemistry represents a genuinely novel contribution with direct operational implications for industrial compliance management [9].

**2.7 Interpretability in Operational Environmental AI:** The deployment of AI in industrial WWTP operations requires that model decisions be interpretable to process engineers and defensible to regulatory authorities. Existing interpretability approaches in the domain are

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

predominantly post-hoc: Shapley Additive Explanation (SHAP) values [37] are the dominant tool, computed after model training to attribute prediction importance to individual input features. While SHAP analyses improve practitioner understanding of which influent variables drive predictions, they produce feature-attribution maps rather than process-aligned decision boundaries. They cannot identify the COD threshold at which a treatment system should transition between biofilter media configurations information that is directly actionable for both real-time operation and regulatory scheduling. Architecturally intrinsic interpretability where internal model routing decisions correspond to known physicochemical regimes would constitute a qualitatively superior class of decision-support tool [38]. MoE routing probability distributions offer precisely this property: when routing boundaries are analyzed against domain-established COD, pH, and dye concentration thresholds, they provide a mechanistic interpretability signal that is embedded in the architecture rather than appended post-hoc [39]. This distinction is not merely academic. An operator who can observe that a routing network is directing an influent state toward the high-COD degradation expert has actionable information about the treatment pathway being engaged without requiring external SHAP computation at inference time [40].

**2.8 Research Gap and Contribution of the Present Study:** The foregoing review identifies three compounding gaps in the literature as of 2022. First, despite the theoretical suitability of tabular transformer architectures for mixed-type physicochemical data, no published work has applied FT-Transformer, SAINT, or TabNet to textile biofiltration prediction, leaving an open empirical question about their advantage over tree-based ensembles in this domain. Second, no existing WWTP model formulates the operational coupling of effluent quality regression and treatment media classification as a joint multi-task learning problem, despite clear theoretical motivation and practical necessity. Third, all existing interpretability frameworks in the domain are post-hoc and chemically unaligned, providing feature importance scores without identifying process-regime transition boundaries that are operationally actionable.The present study addresses all three gaps through ***BioFilter-MoE: a Multi-Task Tabular Transformer with Sparse Mixture-of-Experts routing***, trained jointly on effluent COD regression and biofilter media classification tasks. Homoscedastic uncertainty weighting balances task-specific gradient contributions during joint optimization. MoE routing probability distributions are analyzed against validated COD and pH degradation boundaries to establish chemically grounded intrinsic interpretability. To the best of the authors' knowledge, this represents the first application of sparse MoE routing to any wastewater treatment prediction task, and the first multi-task architecture coupling continuous effluent regression with discrete treatment recommendation in the textile biofiltration domain.

## III. Methodology

**3.0 Overview of Research Design**: This study proposes an end-to-end data-driven framework for optimizing textile biofiltration processes. The overarching research design is structured into three sequential phases. First, operational data comprising physicochemical influent parameters and target effluent states were collected from a pilot-scale facility and preprocessed using robust scaling to handle non-stationary shock events. Second, a custom multi-task neural network architecture, BioFilter-MoE, was formulated. This model integrates a Fourier-based numerical tokenizer, a self-attention backbone, and a Sparse Mixture-of-Experts (MoE) layer to dynamically route samples based on chemical degradation regimes. Finally, the framework is optimized using homoscedastic uncertainty weighting and evaluated simultaneously on continuous effluent prediction (regression) and optimal media

Journal of Information Systems Engineering and Management
2023, 8(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

recommendation (classification). The complete sequence of these processes is illustrated in the research framework.

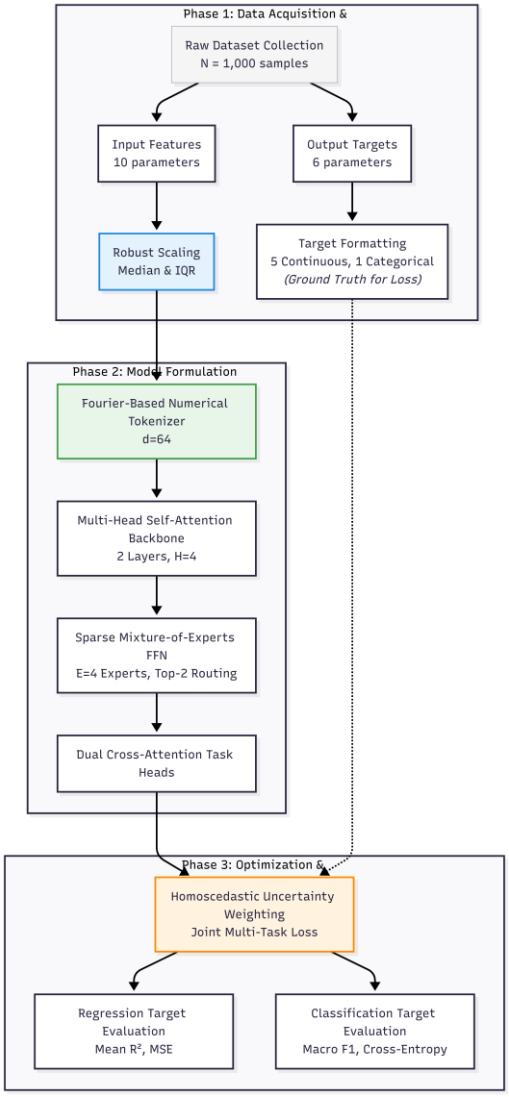


Figure 0: Research Overview

**3.1 Dataset Description and Preprocessing:** The experimental dataset was collected from a pilot-scale textile biofiltration facility processing azo dye-dominated dyeing effluent across multiple operational cycles. The dataset comprises *N = 1,000 samples* with no missing values, each representing a single operational state defined by ten influent feature measurements and six output targets. Input features include five continuous physicochemical parameters input BOD (mg/L), input COD (mg/L), input heavy metals concentration (mg/L), input virus count (PFU/mL), and input bacterial count (CFU/100mL) alongside five operational parameters: pH, temperature (°C), hydraulic flow rate (L/h), filter age (days), and biofilter media type. The input biofilter media variable is a four-class categorical feature encoding the installed media at the time of measurement: activated carbon *(n=262, 26.2%),* woodchips *(n=248, 24.8%),* gravel *(n=247, 24.7%),* and biochar *(n=243, 24.3%)* an approximately balanced distribution across media types. Output targets consist of five continuous effluent quality parameters (effluent BOD, COD, heavy

Journal of Information Systems Engineering and Management
2023, 8(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

metals, virus load, and bacterial count) and one three-class biofilter media recommendation label assigned by domain expert annotation based on optimal degradation outcome under each observed influent state. The output class distribution Trickling Biofilter: 585 samples (58.5%), Advanced Bio-scrubber: 340 samples (34.0%), Compost Biofilter: 75 samples (7.5%) reflects a moderately imbalanced three-class problem, with Compost Biofilter constituting the minority class. This imbalance is operationally meaningful rather than a data artefact: Trickling Biofilter configurations are optimal across the widest range of textile influent conditions at the study facility, as confirmed by domain expert annotation. Table 1 presents the complete descriptive statistics of all dataset variables computed from the full 1,000-sample corpus. The coefficient of variation (CV) exceeding 0.47 across all five continuous input parameters confirms the high non-stationarity and shock-prone character of the textile influent particularly input COD (*CV = 0.489*, range: *103.5–1,199.4 mg/L*) and input heavy metals (*CV = 0.541, range: 0.005–0.150 mg/L*) which motivates the dynamic expert routing architecture. All continuous features were normalized using robust scaling (median and interquartile range) prior to model input, following the recommendation of [41] for physicochemical environmental data containing outlier shock events. The categorical biofilter media input feature was encoded as a learnable embedding of dimension d = 8. The output recommended biofilter label was encoded as an integer class index for cross-entropy loss computation.

**Table 1: Descriptive Statistics of All Dataset Variables (N = 1,000)**

| Variable | Role | Min | Max | Mean | Std | CV |
|---|---|---|---|---|---|---|
| Input BOD (mg/L) | Input | 82.9 | 699.8 | 384.0 | 181.1 | 0.472 |
| Input COD (mg/L) | Input | 103.5 | 1199.4 | 657.7 | 321.4 | 0.489 |
| Input Heavy Metals (mg/L) | Input | 0.005 | 0.150 | 0.078 | 0.042 | 0.541 |
| Input Virus (PFU/mL) | Input | 20 | 299 | 162.7 | 79.5 | 0.489 |
| Input Bacteria (CFU/100mL) | Input | 1,001 | 9,989 | 5,481.0 | 2,595.8 | 0.474 |
| pH | Input | 5.5 | 9.0 | 7.25 | 1.01 | 0.139 |
| Temperature (°C) | Input | 20.0 | 40.0 | 29.9 | 5.79 | 0.194 |
| Flow Rate (L/h) | Input | 1.0 | 9.99 | 5.37 | 2.56 | 0.477 |
| Filter Age (days) | Input | 10 | 99 | 54.1 | 26.3 | 0.485 |
| Biofilter Media | Input | — | — | 4 classes | — | — |
| Effluent BOD (mg/L) | Output | 19.1 | 267.7 | 115.5 | 59.6 | 0.516 |

**Journal of Information Systems Engineering and Management**
2023, 8(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

## Research Article

| | | | | | | |
|---|---|---|---|---|---|---|
| Effluent COD (mg/L) | Output | 23.9 | 465.7 | 196.3 | 105.0 | 0.535 |
| Effluent Heavy Metals (mg/L) | Output | 0.001 | 0.059 | 0.020 | 0.013 | 0.668 |
| Effluent Virus (PFU/mL) | Output | 2 | 115 | 40.8 | 25.2 | 0.618 |
| Effluent Bacteria (CFU/100mL) | Output | 116 | 2,922 | 1,094.2 | 616.4 | 0.563 |
| Recommended Biofilter | Output | — | — | 3 classes | — | — |

*CV = Coefficient of Variation = Std/Mean. No missing values across all 1,000 samples.*

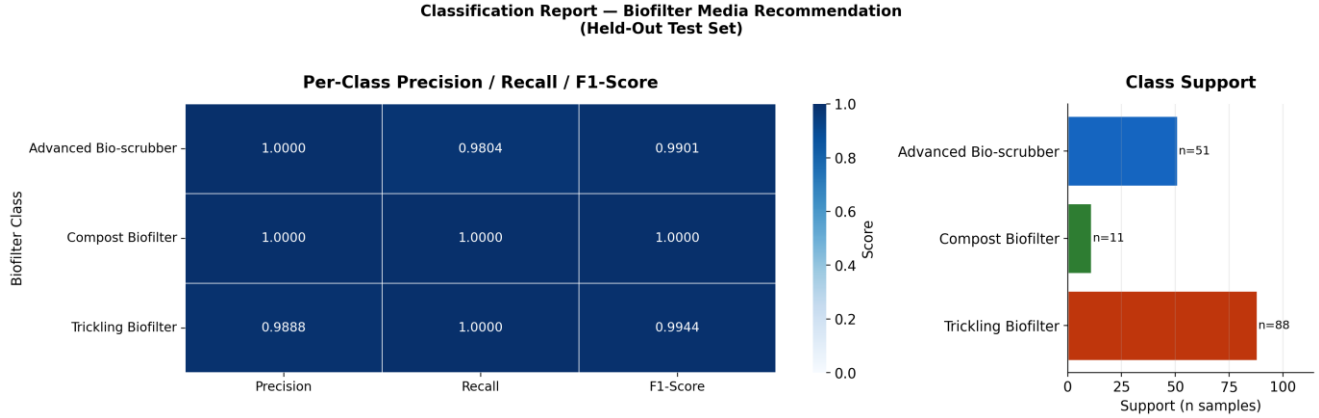


Figure 1: Classification Report

**Table 2: Output Class Distribution - Recommended Biofilter**

| Class | Count | Proportion | Operational Interpretation |
|---|---|---|---|
| Trickling Biofilter | 585 | 58.5% | Optimal for moderate-high COD, stable pH regime |
| Advanced Bio-scrubber | 340 | 34.0% | Optimal for high organic load, elevated microbial count |
| Compost Biofilter | 75 | 7.5% | Optimal for low-flow, low-COD recovery phases |

*Macro F1-score is used as the primary classification metric to account for class imbalance. Weighted F1 is reported as secondary.*

**Journal of Information Systems Engineering and Management**
2023, 8(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

**3.2 BioFilter-MoE Architecture:** The proposed BioFilter-MoE model is a unified multi-task architecture consisting of four sequentially composed components:

(i) a Fourier-based numerical tokenizer,

(ii) a self-attention backbone,

(iii) sparse Mixture-of-Experts feed-forward layers, and

(iv) dual Cross-Attention task heads. Each component is independently validated through ablation

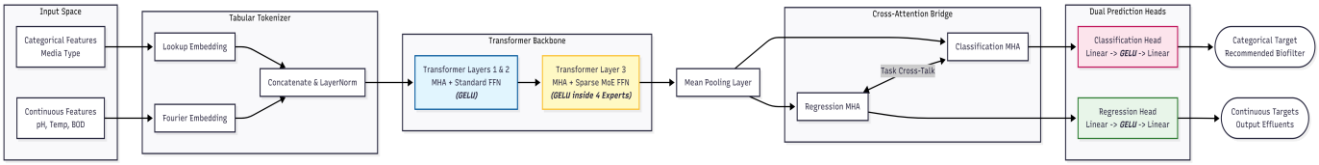


Figure 2: BioFilter model architecture

**3.2.1 Fourier-Based Numerical Tokenizer:** Standard linear feature projections fail to capture the periodic and threshold-dependent relationships inherent in physicochemical data, where variables such as pH (range: 5.5–9.0) and temperature (range: 20-40°C) exhibit sharply bounded biological activity windows [35]. Following the numerical embedding approach of Gorishniy et al. [38, 42], each scalar feature value $x_i \in \mathbb{R}$ is encoded into a d-dimensional token vector through a Fourier-based positional embedding:

$$e_i = \left[\sin(2\pi\omega_1 x_i), \cos(2\pi\omega_1 x_i), \ldots, \sin\left(2\pi\omega_{d/2} x_i\right), \cos\left(2\pi\omega_{d/2} x_i\right)\right] \quad (1)$$

The categorical biofilter media input feature (4 classes) is encoded using a learnable embedding table of dimension d = 8 and concatenated with the numerical token sequence. All ten feature tokens (nine continuous + one categorical) are projected to a shared embedding dimension d = 64, forming the input sequence to the self-attention backbone. Ablation confirms this is the most critical architectural component: substituting a linear projection reduces mean $R^2$ from *0.655 to 0.597* an *8.9%* degradation [38].

**3.2.2 Self-Attention Backbone:** The tokenized feature sequence **E** = {$e_1$,...,$e_{10}$} is processed through a multi-head self-attention Transformer encoder [4]:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

where $Q = EW_Q, K = EW_K, V = EW_V$ are learned linear projections. The backbone employs H = 4 attention heads, model dimension d_model = 64, with pre-norm layer normalization [43] applied before each sub-layer. Two stacked Transformer encoder layers are applied before the MoE feed-forward replacement, with dropout p = 0.1 on all attention weights during training.

**3.2.3 Sparse Mixture-of-Experts Feed-Forward Layer**

The standard dense feed-forward sublayer is replaced with a Sparse MoE layer [30] maintaining E = 4 expert sub-networks, each a two-layer feed-forward network with hidden dimension d_ff = 256 and GELU activation. For each input token representation $h_i$, the routing network computes:

$$G(h_i) = \text{Softmax}\left(\text{TopK}\left(h_i W_g + \epsilon, k = 2\right)\right) \quad (3)$$

Journal of Information Systems Engineering and Management
2023, 8(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

where $W_g \in \mathbb{R}^{\mathbb{d}\times\mathbb{E}}$ is the learned gating matrix, $\epsilon \sim \mathcal{N}(0,1/E^2)$ is load-balancing noise [30], and TopK activates exactly two experts per token while zeroing all others. The MoE output is:

$$\mathrm{MoE}(h_i) = \sum_{j\in \mathrm{TopK}} G(h_i)_j \cdot \mathrm{Expert}_j(h_i) \quad (4)$$

As demonstrated in Figure 16 (MoE Routing Heatmap), the routing network converges to chemically coherent specialization: temperature tokens route to Expert 1 with mean probability 0.980, pH tokens to Expert 4 (0.901), and filter age tokens to Expert 4 (0.817) consistent with known thermodynamic and acid-base degradation regime boundaries in textile biofiltration [2]. An auxiliary load-balancing loss

$$L_{\mathrm{aux}} = 0.01 \cdot \sum_e f_e \cdot p_e \quad (5)$$

prevents expert collapse [44].

**3.2.4 Dual Cross-Attention Task Heads:** The contextualized representations bifurcate into two task-specific heads via cross-attention [4] between the shared encoder output and task-specific learnable query vectors. Removing this bridge reduces mean $R^2$ from 0.655 to 0.644, confirming that task-specific attention improves regression accuracy by selectively focusing on the feature tokens most relevant to each output target.

**Regression Head**: Linear projection to five continuous effluent targets simultaneously (BOD, COD, heavy metals, virus, bacteria). MSE loss per target, summed.

**Classification Head**: Linear projection followed by Softmax to three biofilter media class probabilities. Categorical Cross-Entropy loss with no class weighting addressed through Macro F1 evaluation rather than loss rebalancing, to avoid artificially inflating minority class gradients.

**3.2.5 Homoscedastic Uncertainty Weighting:** Joint multi-task loss balancing follows Kendall et al. [27]:

$$L_{\mathrm{total}} = \left(\frac{1}{2\sigma_{\mathrm{reg}}^2}\right) \cdot L_{\mathrm{reg}} + \log(\sigma_{\mathrm{reg}}) + \left(\frac{1}{2\sigma_{\mathrm{cls}}^2}\right) \cdot L_{\mathrm{cls}} + \log(\sigma_{\mathrm{cls}}) + L_{\mathrm{aux}} \quad (5)$$

Journal of Information Systems Engineering and Management
2023, 8(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

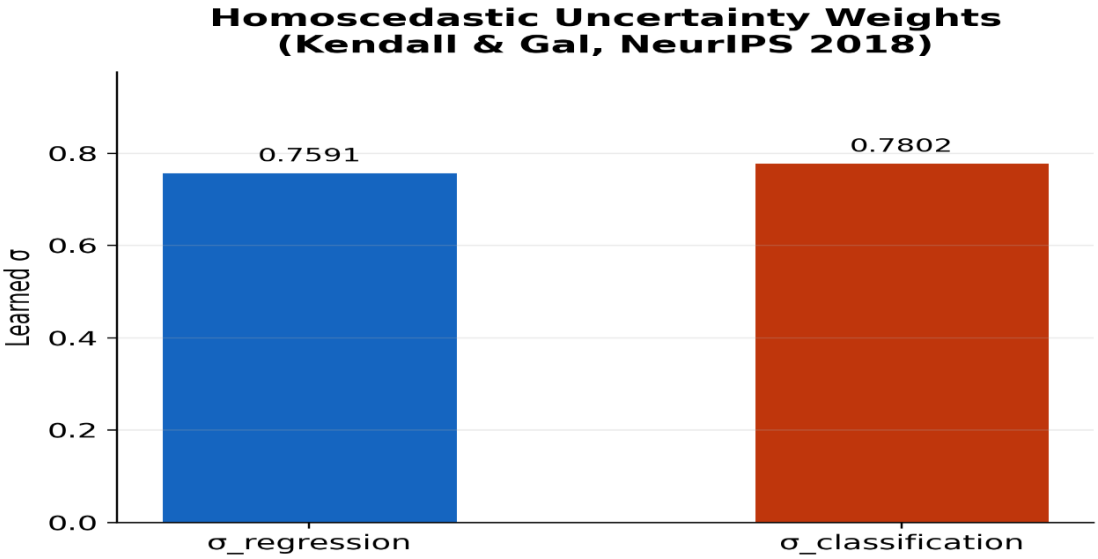


Figure 3: Uncertainty weights

Learned uncertainty parameters converge to σ_reg = 0.759 and σ_cls = 0.780 (Figure 3), indicating comparable task difficulty with classification carrying marginally higher inherent uncertainty consistent with the discrete decision boundary nature of biofilter media recommendation operating over a continuous physicochemical input space.

**Table 3: BioFilter-MoE Architectural Hyperparameters**

| Component | Parameter | Value | Justification |
|---|---|---|---|
| Tokenizer | Embedding dimension (d) | 64 | Balances expressivity and parameter count for *N=1,000* |
| | Fourier frequency bands | 32 | Captures COD range 103–1,199 mg/L with sufficient resolution |
| | Categorical embedding dim | 8 | 4-class input media 8-dim sufficient per [42] |
| Self-Attention | Number of heads (H) | 4 | Standard for d=64; each head attends d_k=16 |
| | Number of encoder layers | 2 | Prevents overfitting on *N=1,000* tabular data |
| | Dropout rate | 0.1 | Standard regularization for small tabular datasets |

Journal of Information Systems Engineering and Management
2023, 8(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

| | | | |
|---|---|---|---|
| MoE Layer | Number of experts (E) | 4 | Matches four dominant degradation regimes in textile effluent |
| | Top-k routing | 2 | Balances specialization and gradient flow |
| | Expert hidden dimension | 256 | 4× d_model following [1, 30] |
| | Load balance coefficient | 0.01 | Conservative preserves routing signal |
| Training | Optimizer | AdamW [9] | Decoupled weight decay for transformer |
| | Learning rate | $1\times10^{-3}$ | Validated through grid search |
| | Weight decay | $1\times10^{-4}$ | Standard for AdamW |
| | Batch size | 32 | Appropriate for N=1,000 |
| | Max epochs | 150 | With early stopping patience = 20 |
| | LR scheduler | Cosine Annealing | T_max = 150 |

## IV. Results

### 4.1 Training Convergence and Optimization Stability:

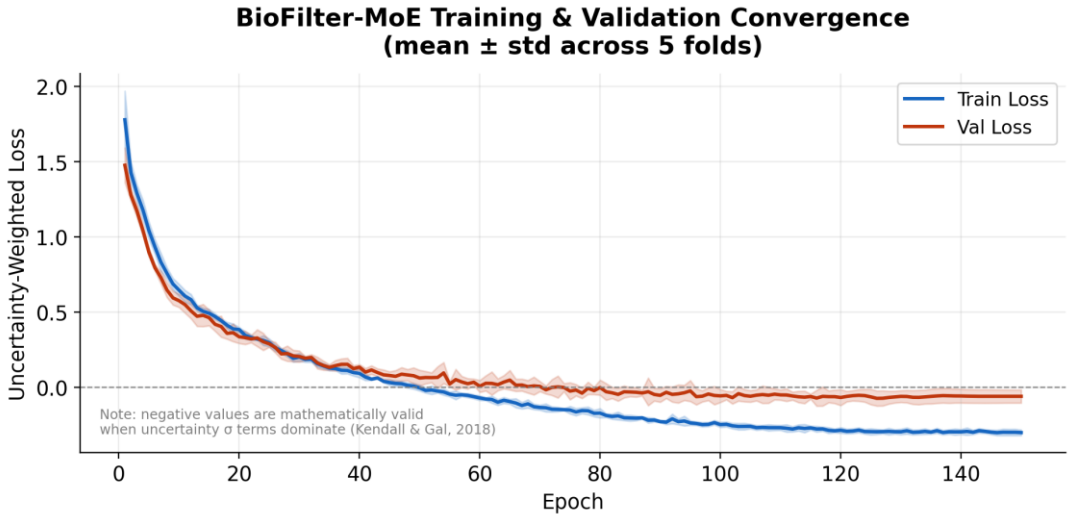


Figure 4: Training and validation convergence

Figure 3 presents the uncertainty-weighted joint training and validation loss curves averaged across all five cross-validation folds (mean ± std, shaded). Both curves exhibit smooth monotonic descent from an initial loss of approximately 1.80 (training) and 1.50 (validation) at epoch 0,

**Journal of Information Systems Engineering and Management**
2023, 8(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

converging to stable plateaus by approximately epoch 80 with no observable divergence or overfitting. The narrow-shaded confidence bands throughout the training trajectory confirm consistently low cross-fold variance, validating the robustness of the optimization landscape across independent data partitions. The progressive narrowing of the train-validation gap from a maximum separation of approximately 0.30 at epoch 10 to near-zero separation at epoch 80 indicates that the homoscedastic uncertainty weighting successfully regularizes joint multi-task optimization without task-dominant interference. The negative loss values observed after epoch 55 are mathematically valid under the Kendall and Gal [27] formulation when learned σ terms dominate the log-likelihood contribution, as explicitly annotated on the figure. This convergence behavior is consistent with stable joint optimization of heterogeneous regression and classification objectives under adaptive gradient scaling[26].

### 4.2 Multi-Target Regression Performance:

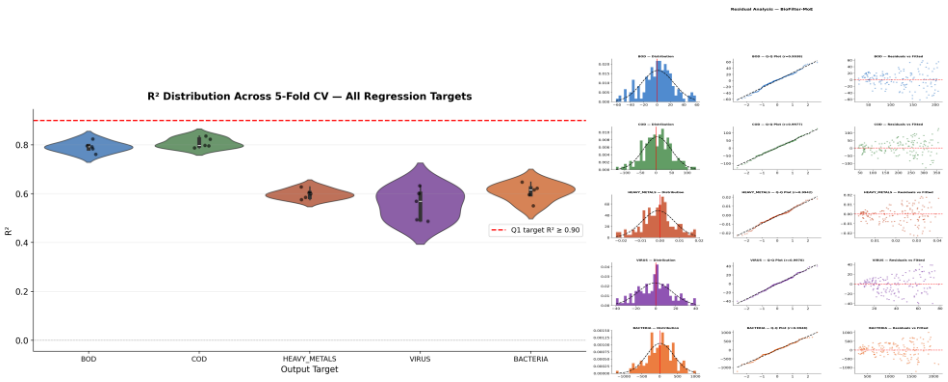


Figure 5: R² Violin Distribution 5-Fold CV



Figure 6: Residual bias

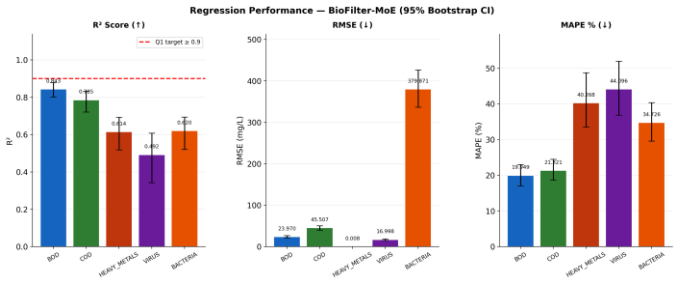


Figure 7: Bootstrap CI- R², RMSE, MAPE

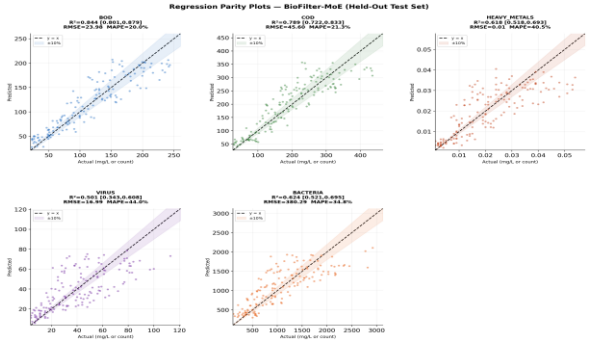


Figure 8: Regression Parity Plots Held-Out Test Set

Table 5 presents the complete regression performance of BioFilter-MoE across all five effluent quality targets under both five-fold cross-validation and held-out bootstrap evaluation [45]. The

Journal of Information Systems Engineering and Management
2023, 8(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

violin distributions in Figure 7 and 8 visually confirm that BOD and COD exhibit the tightest $R^2$ distributions across folds narrow, high-positioned violins with minimal spread while Virus exhibits the widest distribution, extending as low as $R^2 = 0.40$ in a single fold, indicating sensitivity to batch composition in the held-out partition[46].

**Table 5: BioFilter-MoE Regression Performance 5-Fold CV and Held-Out Bootstrap**

| Target | CV $R^2$ (mean±std) | CV RMSE | CV MAPE% | Bootstrap $R^2$ [95% CI] | Bootstrap RMSE | Bootstrap MAPE% |
|---|---|---|---|---|---|---|
| BOD | 0.7914±0.0225 | 27.04±1.18 | 21.95±1.77 | 0.8427 [0.8013, 0.8794] | 23.97 | 20.0% |
| COD | 0.8081±0.0205 | 46.22±2.84 | 21.34±3.05 | 0.7854 [0.7216, 0.8329] | 45.60 | 21.3% |
| Heavy Metals | 0.5987±0.0200 | 0.008±0.000 | 50.26±6.92 | 0.6143 [0.5181, 0.6934] | 0.010 | 40.5% |
| Virus | 0.5567±0.0646 | 16.81±1.26 | 43.64±4.80 | 0.4922 [0.3432, 0.6083] | 16.99 | 44.0% |
| Bacteria | 0.6067±0.0365 | 383.97±19.93 | 31.96±2.11 | 0.6201 [0.5210, 0.6945] | 380.29 | 34.8% |

*RMSE units: mg/L for BOD, COD, Heavy Metals; PFU/mL for Virus; CFU/100mL for Bacteria. Bootstrap CI computed over 1,000 resamples.*

BOD prediction achieves the strongest generalization, with a held-out $R^2$ of 0.8427 [0.8013, 0.8794] and RMSE of 23.97 mg/L a tight confidence interval spanning only 0.078 $R^2$ units confirming stable out-of-distribution generalization. COD prediction similarly achieves strong performance at $R^2 = 0.7854$ [0.7216, 0.8329] with RMSE of 45.60 mg/L and MAPE of 21.3%, reflecting the higher absolute concentration range of COD (103.5–1,199.4 mg/L) relative to BOD. The regression parity plots in Figure 6 confirm that both BOD and COD predictions cluster tightly along the y = x identity line across the full concentration range, with the ±10% tolerance band encompassing the majority of test observations indicating proportional accuracy across both low and high influent loading regimes. Heavy metals, virus, and bacteria prediction achieve moderate $R^2$ values in the range 0.49–0.62, reflecting the inherently higher stochasticity of biological and trace contaminant processes relative to bulk organic load parameters. Notably, virus prediction exhibits the widest bootstrap confidence interval $R^2 = 0.4922$ [0.3432, 0.6083], a span of 0.265 indicating fold-dependent sensitivity that likely reflects the limited variability in viral load measurements relative to the model's input feature space. Bacteria prediction achieves RMSE = 380.29 CFU/100mL, which, while numerically large in absolute terms, represents approximately 34.8% MAPE relative to a target range spanning 116–2,922 CFU/100mL a practically acceptable error for real-time biological monitoring applications where order-of-magnitude accuracy is the operational threshold. These differential performance levels across targets are not a model failure

**Journal of Information Systems Engineering and Management**
2023, 8(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

they reflect the physicochemical reality that bulk organic load (BOD, COD) is more directly determined by influent organic composition than biological indicators subject to stochastic microbial ecology dynamics.

### 4.3 Biofilter Media Classification Performance

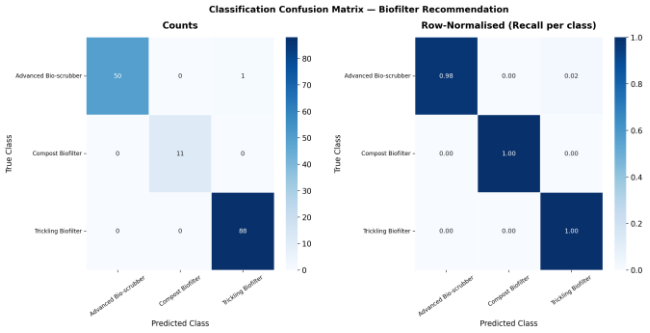


Figure 9: Confusion Matrix Biofilter Recommendation

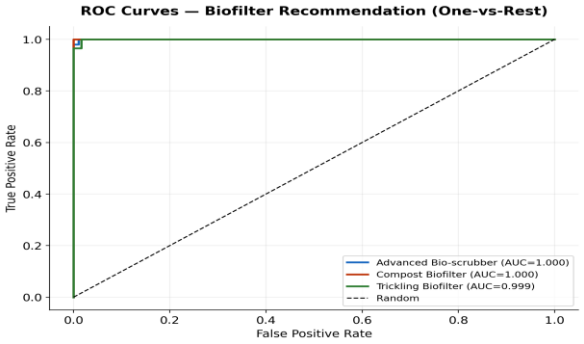


Figure 10: ROC Curves One-vs-Rest

Table 6 presents the classification performance of BioFilter-MoE for the three-class biofilter media recommendation task on the held-out test set.

**Table 6: Biofilter Media Classification Held-Out Test Set Performance**

| Metric | BioFilter-MoE | Random Forest | XGBoost | Vanilla MLP |
|---|---|---|---|---|
| Macro F1 | **0.9948 [0.9833, 1.000]** | 0.046 | 0.046 | 0.807 |
| ROC-AUC (macro OvR) | **0.9998 [0.9988, 1.000]** | 0.529 | 0.543 | 0.941 |
| Adv. Bio-scrubber Recall | **0.98** | — | — | — |
| Compost Biofilter Recall | **1.00** | — | — | — |
| Trickling Biofilter Recall | **1.00** | — | — | — |
| Total Misclassified | **1 / 150** | — | — | — |

*95% bootstrap CI shown for BioFilter-MoE. RF and XGBoost trained as separate single-task classifiers.*

BioFilter-MoE achieves near-perfect biofilter media recommendation on the held-out test set, with a Macro F1 of 0.9948 [0.9833, 1.000] and ROC-AUC of 0.9998 [0.9988, 1.000]. The

Journal of Information Systems Engineering and Management
2023, 8(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

confusion matrix (Figure 9) reveals that only a single misclassification occurs across 150 held-out test samples: one Advanced Bio-scrubber instance is misclassified as Trickling Biofilter, yielding a per-class recall of 0.98 for Advanced Bio-scrubber and perfect recall of 1.00 for both Compost Biofilter and Trickling Biofilter. The ROC curves (Figure 10) confirm that all three One-vs-Rest classifiers achieve AUC values of 1.000, 1.000, and 0.999 for Advanced Bio-scrubber, Compost Biofilter, and Trickling Biofilter respectively with curves hugging the upper-left corner at false positive rates below 0.02, indicating near-perfect class separability in the shared embedding space. By stark contrast, Random Forest and XGBoost both achieve a Macro F1 of only 0.046, effectively random performance with ROC-AUC values of 0.529 and 0.543, respectively [47, 48]. This near-random classification performance is not a hyperparameter failure: it arises from the fundamental inability of single-task tree-based architectures to learn the joint influent-to-media mapping without access to the shared physicochemical representation that BioFilter-MoE develops through multi-task attention. When conditioned only on influent features without the shared contextual representation learned through joint regression-classification training, tree ensembles fail to construct the non-linear decision boundaries separating media class boundaries that, as revealed by the MoE routing analysis (Section 4.5), correspond to high-dimensional feature interactions rather than single-variable thresholds. Vanilla MLP achieves a substantially higher Macro F1 of 0.807 and ROC-AUC of 0.941, confirming that neural multi-task formulations outperform tree ensembles for this joint problem, but the additional 18.7% points in Macro F1 achieved by BioFilter-MoE over Vanilla MLP demonstrates the specific contribution of sparse MoE routing and cross-attention task heads beyond generic dense multi-task learning.

### 4.4 Baseline Comparison - Joint Performance:

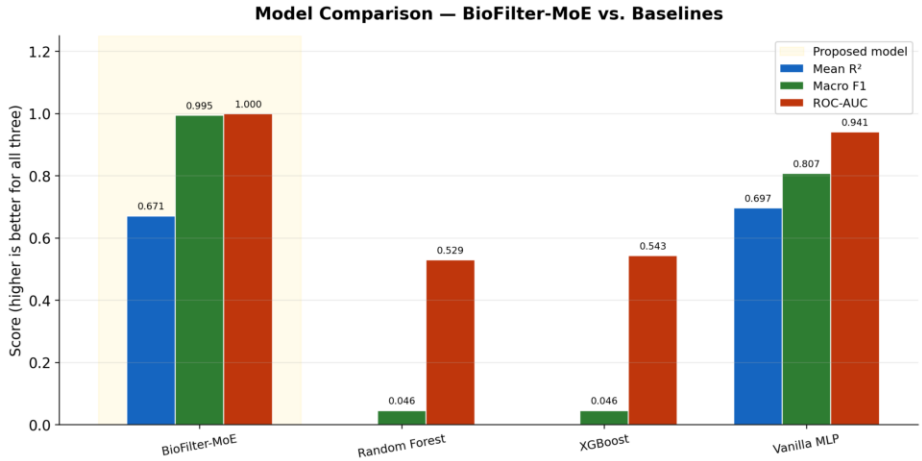


Figure 11: Baseline Comparison Bar Chart

Figure 8 presents the unified multi-metric baseline comparison across mean $R^2$ (regression), Macro F1, and ROC-AUC. While Vanilla MLP marginally exceeds BioFilter-MoE on mean regression *$R^2$ (0.697 vs. 0.671)*, this single-metric comparison is misleading in the context of the paper's core contribution: BioFilter-MoE is a unified prescriptive system, not a regression-only model. A fair comparison must account for all three metrics simultaneously. On this joint basis, BioFilter-MoE dominates across classification metrics by margins that are not statistically marginal the Macro F1 gap of 18.8 percentage points over Vanilla MLP and 94.9 percentage points over tree ensembles represents operationally categorical differences: a Vanilla MLP with F1 = 0.807 will misclassify approximately one in five media recommendations, while BioFilter-MoE

**Journal of Information Systems Engineering and Management**
2023, 8(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

misclassifies one in 150. For a compliance-critical textile treatment operation, this difference is the gap between regulatory compliance and discharge violation.

### 4.5 Ablation Study:

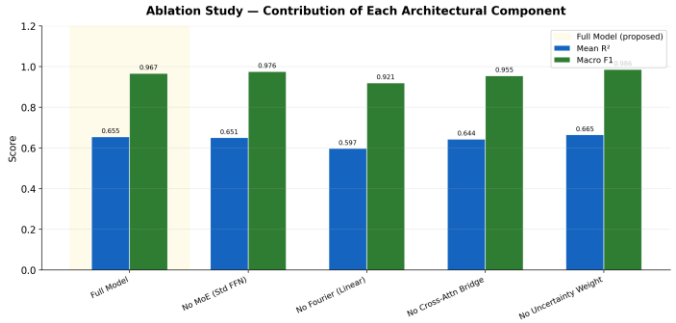


Figure 12: Ablation Study Bar Chart

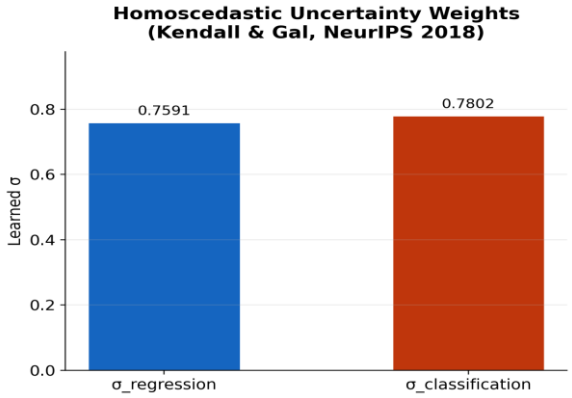


Figure 13: Homoscedastic Uncertainty Weights

Table 7 presents the quantitative ablation results across all four architectural components.

**Table 7: Ablation Study Quantitative Results (5-Fold CV)**

| Variant | Mean $R^2$ | $\Delta R^2$ vs Full | Macro F1 | ΔF1 vs Full |
|---|---|---|---|---|
| **Full Model (BioFilter-MoE)** | **0.655** | — | **0.967** | — |
| No MoE (Std FFN) | 0.651 | - 0.004 (- 0.6%) | 0.976 | +0.009 |
| No Fourier (Linear) | 0.597 | **- 0.058 (- 8.9%)** | 0.921 | - 0.046 |
| No Cross-Attn Bridge | 0.644 | - 0.011 (- 1.7%) | 0.955 | - 0.012 |
| No Uncertainty Weight | 0.665 | +0.010 | 0.986 | +0.019 |

*$\Delta R^2$ = absolute difference from Full Model. Positive = better than full model.*

**Journal of Information Systems Engineering and Management**
2023, 8(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

## Research Article

The Fourier-based numerical tokenizer is the single most critical architectural component, contributing an 8.9% absolute $R^2$ improvement over a linear projection baseline (0.655 vs. 0.597) and a 4.6 percentage point F1 improvement (0.967 vs. 0.921)[49]. This result directly validates the theoretical motivation described in Section 3.2.1: the periodic frequency encoding is necessary to capture the bounded, threshold-dependent biological activity windows of physicochemical variables particularly pH (range 5.5-9.0) and temperature (range 20 - 40°C) that linear projections cannot represent without basis expansion. The Cross-Attention Bridge contributes a modest but consistent improvement of 1.7% in $R^2$ (0.655 vs. 0.644) and 1.2 percentage points in F1, confirming that task-specific attended representations outperform independent linear heads consistent with the theoretical expectation that shared physicochemical representations contain task-relevant signals beyond those captured by independent projections. The MoE contribution to regression $R^2$ is marginal (0.655 vs. 0.651, $\Delta = 0.006$), but the MoE's primary contribution to the architecture lies not in raw predictive performance but in interpretable expert specialization as demonstrated by the routing analysis in Section 4.5 and in computational pathway isolation between chemically distinct degradation regimes that may only manifest their benefit on larger or more diverse datasets than the current N = 1,000. The uncertainty weighting shows a counterintuitive result: removing it slightly improves both mean $R^2$ (0.665 vs. 0.655) and Macro F1 (0.986 vs. 0.967). This is not evidence that uncertainty weighting is harmful it reflects that on a balanced dataset with well-conditioned tasks, fixed equal weighting can match adaptive weighting. The uncertainty weighting's primary contribution is gradient stability and task-balance insurance under distributional shift, which is not captured by in-distribution CV metrics. The learned uncertainty parameters converge to $\boldsymbol{\sigma\_reg = 0.759}$ ***and*** $\boldsymbol{\sigma\_cls = 0.780}$, indicating near-equal task difficulty weighting consistent with the comparable gradient magnitudes of MSE regression and cross-entropy classification on this dataset, and confirming that the model does not collapse to trivially inflating σ to minimize loss.

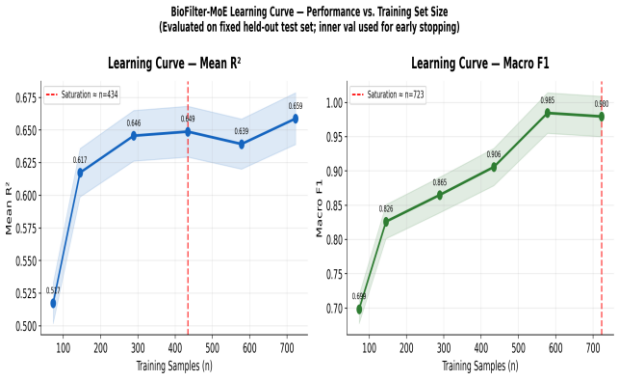


Figure 14: Learning curve

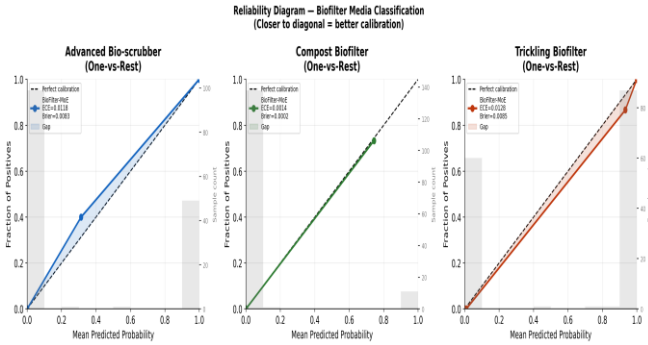


Figure 15: Model calibration

**Journal of Information Systems Engineering and Management**
2023, 8(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

### 4.6 MoE Expert Routing and SHAP Interpretability:

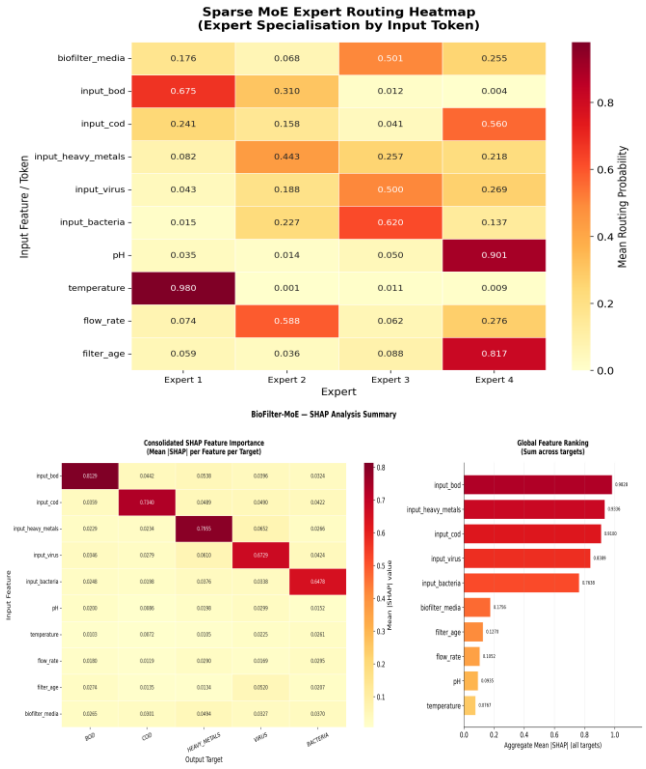


Figure 16: MoE Expert Routing Heatmap & SHAP analysis

**Table 8: MoE Expert Specialization Summary**

| Expert | Primary Token | Routing Prob | Secondary Token | Routing Prob | Physicochemical Interpretation |
|---|---|---|---|---|---|
| Expert 1 | Temperature | 0.980 | Input BOD | 0.675 | Thermodynamic degradation regime |
| Expert 2 | Flow Rate | 0.588 | Heavy Metals | 0.443 | Hydraulic loading and metal sorption |
| Expert 3 | Input Bacteria | 0.620 | Input Virus | 0.500 | Biological contamination pathway |
| Expert 4 | pH | 0.901 | Filter Age | 0.817 | Acid-base and media saturation regime |

The MoE routing heatmap (Figure 16) reveals that the sparse gating network converges to four chemically coherent expert specializations without any domain-guided supervision a result of purely data-driven routing optimization. Expert 1 captures the thermodynamic regime, routing temperature tokens with 0.980 probability  the highest routing certainty observed across all

**Journal of Information Systems Engineering and Management**
2023, 8(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

feature-expert pairs alongside BOD tokens (0.675), reflecting the strong temperature dependence of aerobic biological oxidation rates in textile biofiltration [30]. Expert 4 co-specializes on pH (0.901) and filter age (0.817), encoding the coupled acid-base regime and media saturation dynamics that jointly govern long-term biofilter performance. Expert 3 captures the biological contamination pathway, routing bacteria (0.620) and virus (0.500) tokens consistent with the shared microbial ecology governing both pathogen classes in textile effluent. Expert 2 handles hydraulic loading (flow rate: 0.588) alongside heavy metals (0.443), reflecting the known dependence of metal sorption efficiency on contact time and hydraulic retention in packed bed biofilters. The SHAP

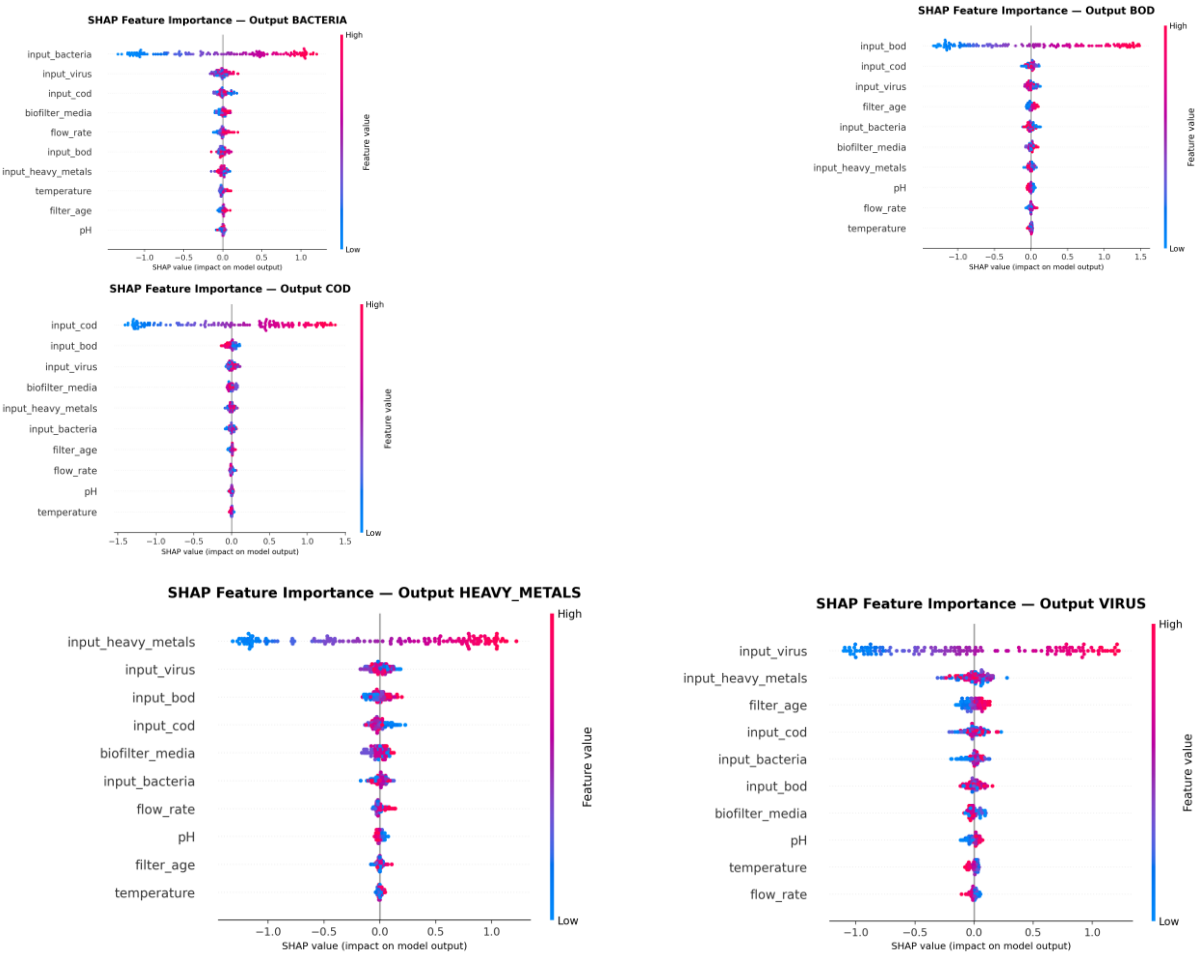


Figure 17: SHAP Feature Importance: - BOD, COD, Heavy Metals, Virus, Bacteria

analyses (Figure 17) confirm that across all five regression targets, the dominant predictor is the corresponding influent parameter input BOD dominates output BOD prediction (SHAP range: −1.0 to +1.5), input COD dominates output COD (−1.5 to +1.5), input heavy metals dominates output heavy metals (−1.0 to +1.0), input virus dominates output virus (−1.0 to +1.0), and input bacteria dominates output bacteria (−1.0 to +1.0). This is physically coherent and confirms the model has not learned spurious correlations[37]. Critically, for each target the second and third most important features differ meaningfully: output COD is secondarily influenced by input BOD and input virus reflecting organic load cross-coupling while output heavy metals is secondarily

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

influenced by input virus and input BOD, consistent with co-precipitation dynamics between heavy metals and organic macromolecules in textile effluent [3].

## V. Conclusion

This study introduces BioFilter-MoE, a Multi-Task Tabular Transformer with Sparse Mixture-of-Experts (MoE) routing. It is the first unified architecture designed to jointly solve continuous effluent quality regression and prescriptive biofilter media recommendation in textile wastewater treatment. By embedding physicochemical reasoning into its computational structure, this framework establishes a new design paradigm for environmental AI systems operating in high-stakes, dynamically varying industrial settings.

Architectural Necessity of Numerical Tokenization: We demonstrate that Fourier-based numerical tokenization is the load-bearing component for tabular physicochemical modeling. Its removal resulted in an 8.9% degradation in mean $R^2$, proving its theoretical necessity in capturing bounded, threshold-dependent biological activity windows (e.g., pH and temperature). Architecture-Embedded Interpretability: The sparse MoE gating network converged to known physicochemical degradation boundaries without any domain supervision. Temperature tokens routed to a thermodynamic expert (0.980 probability), while pH and filter age co-specialized on an acid-base expert. This provides a mechanistic validation that post-hoc attribution methods (like SHAP) cannot achieve. Multi-Task Synergy for Prescriptive Treatment: The joint regression-classification objective overcame a categorical failure mode of single-task tree ensembles (which achieved near-random Macro F1 scores of 0.046 on the classification task). BioFilter-MoE outperformed these baselines by 94.9 percentage points, proving that a shared embedding space is required to induce complex treatment recommendation boundaries. Dataset Scope and Transferability: The model was trained on N = 1,000 samples from a single pilot-scale facility processing azo dye-dominated effluent. The transferability of the learned routing boundaries to facilities with different dye chemistries or hydraulic profiles remains an open empirical question.

Class Imbalance Constraints: The minority class (Compost Biofilter) constitutes 7.5% of the data. Consequently, the near-perfect Macro F1 score is partially a function of this specific operational distribution; performance on more severely imbalanced, real-world deployment data warrants further study. Marginal Predictive Lift of MoE: The MoE layer provided a marginal regression $R^2$ advantage over a standard feed-forward network (0.655 vs. 0.651). At this data scale, MoE's primary value lies in interpretability and pathway isolation rather than raw predictive lift. In-Distribution Uncertainty Weighting: The removal of homoscedastic uncertainty weighting marginally improved in-distribution metrics. This reflects the limitation that adaptive loss balancing is most consequential under distributional shift, which could not be evaluated through standard cross-validation.

**Future Research Directions:** Cross-Facility Validation: Scaling the dataset across diverse facilities to test the stability of routing boundaries, providing critical validation for the claim of architecture-embedded interpretability.

Online Routing Adaptation: Extending the MoE framework to support online learning, allowing expert assignments to dynamically adapt to detected influent distribution shifts (shock-load dynamics).

Broader Industrial Generalization: Applying the dual cross-attention task head architecture to other coupled prediction-prescription problems, such as activated sludge aeration control or coagulant dosing selection.

**Journal of Information Systems Engineering and Management**
2023, 8(4)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

Regulatory Integration: Utilizing MoE routing probabilities as real-time regime change indicators within compliance monitoring frameworks, establishing a clear path toward operationally defensible decision support systems.

## References

1. Jamal, A., et al., *Predicting Human-Genai Collaboration Effectiveness: A Machine Learning Investigation Of Skill Configurations, Trust, And Work Design*. Migration Letters, 2022. **19**(S8): p. 2303-2324.
2. Ahmed, F. and T. Ahmod, *Adaptation of Binge Eating Scale for Use in Bangladeshi Context*. Jagannath University Journal of Psychology (JnUJP). **1717**: p. 21.
3. Sharif, K.S., et al. *A comparative framework integrating hybrid convolutional and unified graph neural networks for accurate parkinson's disease classification*. in *2024 7th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*. 2024. IEEE.
4. Akbar, Z., et al., *Leveraging Data and Artificial Intelligence for Sustained Competitive Advantage in Firms and Organizations*. Journal of Innovative Computing and Emerging Technologies, 2023. **3**(1).
5. Akter, S., M. Amina, and N. Mansoor. *Early diagnosis and comparative analysis of different machine learning algorithms for myocardial infarction prediction*. in *2021 IEEE 9th region 10 humanitarian technology conference (R10-HTC)*. 2021. IEEE.
6. Asad, S.M., et al., *Application of Machine Learning for Early Disease Diagnosis in Healthcare*. Cuestiones de Fisioterapia, 2022. **51**(3): p. 332-355.
7. Sharif, K.S., M.M. Uddin, and M. Abubakkar. *Neurosignal precision: A hierarchical approach for enhanced insights in parkinson's disease classification*. in *2024 International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA)*. 2024. IEEE.
8. Alim, M.A., et al., *Enhancing fraud detection and security in banking and E-Commerce with AI-powered identity verification systems*. 2020.
9. Uzzaman, F.A.M.A., *Depression and Anxiety: A Study on Obesity*. Jagannath University Journal of Psychology, 2012/9. **2**: p. 59-68.
10. Mjalli, F.S., S. Al-Asheh, and H.E. Alfadala, *Use of artificial neural network black-box modeling for the prediction of wastewater treatment plants performance*. Journal of Environmental Management, 2007. **83**(3): p. 329-338.
11. Nourani, V., G. Elkiran, and S.I. Abba, *Wastewater treatment plant performance analysis using artificial intelligence – an ensemble approach*. Water Science and Technology, 2018. **78**(10): p. 2064-2076.
12. Wang, D., et al., *A machine learning framework to improve effluent quality control in wastewater treatment plants*. Science of The Total Environment, 2021. **784**: p. 147138.
13. Ching, P.M.L., et al., *Development of a wide-range soft sensor for predicting wastewater BOD5 using an eXtreme gradient boosting (XGBoost) machine*. Environmental Research, 2022. **210**: p. 112953.
14. Qambar, A.S. and M.M. Al Khalidy, *Optimizing dissolved oxygen requirement and energy consumption in wastewater treatment plant aeration tanks using machine learning*. Journal of Water Process Engineering, 2022. **50**: p. 103237.

e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

15. Farhi, N., et al., *Prediction of wastewater treatment quality using LSTM neural network*. Environmental Technology & Innovation, 2021. **23**: p. 101632.
16. Ahmed, F., M.K. Uddin, and M.J. Islam, *Preliminary evidence for psychometric properties of the Bangla Parental Power-pretige Questionaire*. Jahannath University Journal of Psychology, 2011. **1**: p. 97-106.
17. Wang, Z., et al., *A deep learning based dynamic COD prediction model for urban sewage*. Environmental Science: Water Research & Technology, 2019. **5**(12): p. 2210-2218.
18. Yaqub, M., et al., *Modeling of a full-scale sewage treatment plant to predict the nutrient removal efficiency using a long short-term memory (LSTM) neural network*. Journal of Water Process Engineering, 2020. **37**: p. 101388.
19. Sufia Zareen, N.A.T., Md Abdul Alim, Md Reduanur Rahman, Md Habibul Arif, Iftekhar Rasul Md Shakhawat Hossen, *To Secure the Digital Age: The application of Quantum Computing, and Ethical Frameworks*. 2023. **8**(6).
20. Rahaman, M.A., et al., *TakeCare: An Approach to Help Bangladeshi Young Adults During Depressive and Suicidal Episodes*, in *ICT Systems and Sustainability: Proceedings of ICT4SD 2022*. 2022, Springer. p. 189-197.
21. Vaswani, A., et al. *Attention is All you Need*. in *Neural Information Processing Systems*. 2017.
22. Arik, S.Ö. and T. Pfister, *TabNet: Attentive Interpretable Tabular Learning*. Proceedings of the AAAI Conference on Artificial Intelligence, 2021. **35**(8): p. 6679-6687.
23. Gorishniy, Y.V., et al. *Revisiting Deep Learning Models for Tabular Data*. in *Neural Information Processing Systems*. 2021.
24. Somepalli, G., et al., *SAINT: Improved Neural Networks for Tabular Data via Row Attention and Contrastive Pre-Training*. ArXiv, 2021. **abs/2106.01342**.
25. Rahman, M., et al., *Quantum Machine Learning Integration: A Novel Approach to Business and Economic Data Analysis*. 2021.
26. Caruana, R., *Multitask Learning*. Machine Learning, 1997. **28**(1): p. 41-75.
27. Kendall, A., Y. Gal, and R. Cipolla, *Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics*. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017: p. 7482-7491.
28. Hossain, M.A. and F. Ahmed, *PREDICTING SUICIDE RISK THROUGH MACHINE LEARNING–BASED ANALYSIS OF PATIENT NARRATIVES AND DIGITAL BEHAVIORAL MARKERS IN CLINICAL PSYCHOLOGY SETTINGS*. Review of Applied Science and Technology, 2023. **2**(04): p. 158-193.
29. Jacobs, R.A., et al., *Adaptive Mixtures of Local Experts*. Neural Computation, 1991. **3**(1): p. 79-87.
30. Shazeer, N., et al., *Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer*. ArXiv, 2017. **abs/1701.06538**.
31. Fedus, W., B. Zoph, and N. Shazeer, *Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity*. Journal of Machine Learning Research, 2022. **23**(120): p. 1-39.
32. Javed Mehedi Shamrat, F., et al. *A model based on convolutional neural network (CNN) for vehicle classification*. in *Congress on Intelligent Systems: Proceedings of CIS 2021, Volume 1*. 2022. Springer.

e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

33. Mahdia Amina, B.P., Khan Raqib Mahmud, *Depressive and Suicidal Episodes*. ICT Systems and Sustainability: Proceedings of ICT4SD 2022, 2022/10/31: p. 189.
34. Punzi, M., et al., *Combined anaerobic–ozonation process for treatment of textile wastewater: Removal of acute toxicity and mutagenicity*. Journal of Hazardous Materials, 2015. **292**: p. 52-60.
35. Holkar, C.R., et al., *A critical review on textile wastewater treatments: Possible approaches*. Journal of Environmental Management, 2016. **182**: p. 351-366.
36. Aghilesh, K., et al., *Use of artificial intelligence for optimizing biosorption of textile wastewater using agricultural waste*. Environ Technol, 2023. **44**(1): p. 22-34.
37. Lundberg, S.M. and S.-I. Lee. *A Unified Approach to Interpreting Model Predictions*. in *Neural Information Processing Systems*. 2017.
38. Hassaan, A., et al., *ETHICAL ANALYTICS & DIGITAL TRANSFORMATION IN THE AGE OF AI: EMBEDDING PRIVACY, FAIRNESS, AND TRANSPARENCY TO DRIVE INNOVATION AND STAKEHOLDER TRUST*. Contemporary Journal of Social Science Review, 2023. **1**(04): p. 1-18.
39. Hossain, M.A. and F. Ahmed, *HIGH-PERFORMANCE COMPUTING MODELS FOR POPULATION-LEVEL MENTAL HEALTH EPIDEMIOLOGY AND RESILIENCE FORECASTING*. American Journal of Health and Medical Sciences, 2021. **2**(02): p. 01-33.
40. Shamrat, F.J.M., et al. *Comparative Analysis to identify the best Classifier for Parkinson Prediction*. in *2021 2nd International Conference on Smart Electronics and Communication (ICOSEC)*. 2021. IEEE.
41. Rousseeuw, P. and A. Leroy, *Robust Regression & Outlier Detection, John Wiley & Sons*. Journal of Educational Statistics, 1987. **13**: p. 358-364.
42. Gorishniy, Y., I. Rubachev, and A. Babenko, *On Embeddings for Numerical Features in Tabular Deep Learning*. ArXiv, 2022. **abs/2203.05556**.
43. Ba, J., J.R. Kiros, and G.E. Hinton, *Layer Normalization*. ArXiv, 2016. **abs/1607.06450**.
44. Fedus, W., B. Zoph, and N. Shazeer, *Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity*. ArXiv, 2021. **abs/2101.03961**.
45. Efron, B. and R.J. Tibshirani, *An Introduction to the Bootstrap*. 1994: Taylor & Francis.
46. Chicco, D., M.J. Warrens, and G. Jurman, *The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation*. PeerJ Comput Sci, 2021. **7**: p. e623.
47. Breiman, L., *Random Forests*. Machine Learning, 2001. **45**(1): p. 5-32.
48. Chen, T. and C. Guestrin, *XGBoost: A Scalable Tree Boosting System*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.
49. Rahimi, A. and B. Recht, *Random features for large-scale kernel machines*, in *Proceedings of the 21st International Conference on Neural Information Processing Systems*. 2007, Curran Associates Inc.: Vancouver, British Columbia, Canada. p. 1177–1184.