Artificial Intelligence and Intelligent Systems Research
https://intellisci.org

Neurosymbolic Reasoning with Decision Transparency in Clinical Triage Systems: Benchmark Study

Thomas K. Richards*

Department of Computer Science and Technology, School of Technology, University of Cambridge, Cambridge, England, United Kingdom

Corresponding author: thomas.k.richards@cam.ac.uk

Abstract

The integration of artificial intelligence into clinical environments has historically been challenged by the tension between predictive accuracy and decision transparency. Clinical triage systems, which dictate the prioritization of patient care in high-stakes emergency settings, require both high fidelity in their predictive capabilities and rigorous interpretability to foster clinician trust. Purely neural approaches often behave as opaque systems, precluding clinical validation, while traditional symbolic systems lack the robust generalization needed to handle noisy, unstructured clinical data. This study presents a comprehensive benchmark of neurosymbolic systems in clinical triage, evaluating their capacity to harmonize the robust perceptual abilities of deep neural networks with the explicitly verifiable logic of symbolic reasoning. Through extensive experimentation on retrospective electronic health record datasets, we analyze the performance of hybrid neurosymbolic architectures against purely neural and purely symbolic baselines. Our methodology evaluates not only standard classification metrics but also introduces novel quantifiable measures for decision transparency and logical consistency. The results demonstrate that neurosymbolic systems achieve predictive performance comparable to state-of-the-art neural architectures while maintaining the strict, rule-based interpretability characteristic of expert systems. This paper contributes a foundational framework for evaluating hybrid reasoning models in healthcare, ultimately facilitating the deployment of safer and more accountable artificial intelligence in acute care environments.

Keywords: Clinical Triage; Neurosymbolic Artificial Intelligence; Decision Transparency; Explainable Healthcare; Benchmark Study

1. Introduction

1.1 Background and Motivation

Emergency departments globally face unprecedented levels of overcrowding, which fundamentally compromises patient safety and severely strains healthcare resources. Triage serves as the critical first point of contact in acute care environments, functioning as a systematic prioritization mechanism that determines the urgency of a patient condition and allocates appropriate medical attention. Historically, clinical triage has relied heavily on the subjective judgment of specialized nursing staff utilizing standardized protocols such as the Emergency Severity Index or the Manchester Triage System. While these human-driven protocols provide a structured approach to patient assessment, they are inherently susceptible to cognitive fatigue, implicit biases, and high inter-rater variability, particularly during periods of intense departmental congestion. The operational consequence of suboptimal triage is profound, frequently resulting in prolonged waiting times, delayed medical interventions for critically ill patients, and an increased likelihood of adverse clinical outcomes [1].

In response to these systemic challenges, the medical informatics community has aggressively pursued the development of automated triage systems powered by machine learning algorithms. Over the past decade, deep learning architectures have demonstrated remarkable proficiency in extracting complex patterns from large-scale electronic health records. These purely data-driven models excel at processing high-dimensional inputs, including continuously monitored vital signs, structured laboratory results, and unstructured clinical narratives found in triage nursing notes. However, the adoption of deep learning in clinical practice has been significantly impeded by the inherent opacity of neural networks. These models operate as complex, non-linear functions that provide minimal insight into their internal decision-making processes. In the high-stakes domain of clinical triage, where automated decisions can directly dictate patient survivability, the inability to interrogate or validate the reasoning behind a specific algorithm prediction represents an insurmountable barrier to clinical acceptance [2]. Clinicians require systems that not only provide accurate recommendations but also justify those recommendations through transparent, medically sound logic.

1.2 Research Objectives and Scope

To bridge the gap between predictive power and essential interpretability, researchers have increasingly turned to neurosymbolic artificial intelligence. This hybrid paradigm attempts to synthesize the generalized learning capabilities of neural networks with the rigorous, rule-based logic of symbolic expert systems. By grounding the output of neural feature extractors in explicit clinical ontologies and logical inference rules, neurosymbolic systems theoretically offer the capacity to process noisy real-world data while producing decisions that are entirely verifiable against established medical guidelines. Despite the theoretical promise of neurosymbolic integration, the literature currently lacks a standardized, comprehensive benchmark to systematically evaluate these architectures specifically within the context of clinical triage. Most existing evaluations are highly localized, utilizing disparate datasets and inconsistent metrics that preclude meaningful cross-study comparisons.

The primary objective of this paper is to establish a rigorous benchmark study that comprehensively evaluates neurosymbolic reasoning models against traditional single-paradigm approaches in clinical triage applications. This research seeks to systematically quantify the trade-offs between predictive accuracy and decision transparency. Furthermore, we aim to introduce a novel evaluation framework that moves beyond conventional machine learning metrics to assess the logical consistency, rule activation frequency, and clinical interpretability of the hybrid systems. By clearly delineating the operational characteristics, strengths, and failure modes of neurosymbolic triage systems, this study intends to provide actionable guidelines for healthcare institutions considering the deployment of automated clinical decision support technologies. The scope of this investigation encompasses both structured physiological data and unstructured textual data, reflecting the multimodal reality of modern emergency department informatics.

2. Literature Review and Background

2.1 Neural Networks in Clinical Decision Making

The deployment of deep neural networks in healthcare settings has fundamentally altered the landscape of predictive analytics. Early applications focused on specialized tasks such as radiological image analysis, but contemporary research has rapidly expanded into the multifaceted domain of clinical decision support. Recurrent neural networks and long short-term memory architectures have been extensively utilized to model the temporal dependencies inherent in patient vital signs and sequential clinical measurements [3]. More recently, transformer-based models have revolutionized the processing of unstructured clinical texts. By leveraging self-attention mechanisms, these architectures can capture nuanced contextual relationships within physician progress notes, discharge summaries, and triage assessments, thereby extracting highly predictive latent features that traditional statistical models consistently overlook.

Despite these technological advancements, the purely connectionist approach suffers from several critical vulnerabilities when deployed in clinical environments. Primarily, deep learning models are exquisitely sensitive to distributional shifts in the underlying data. A model trained on electronic health records from an urban tertiary care center frequently experiences catastrophic performance degradation when applied to a rural community hospital setting due to variations in patient demographics, documentation practices, and clinical workflows. Additionally, deep learning models are notorious for learning spurious correlations rather than true causal relationships [4]. In a medical context, an algorithm might associate a specific nursing shift time with higher patient mortality merely because that shift happens to receive the majority of trauma transfers, leading to fundamentally flawed reasoning that goes undetected due to the opaque nature of the neural architecture.

2.2 Symbolic Reasoning and Knowledge Representation

Prior to the dominance of deep learning, artificial intelligence in healthcare was predominantly characterized by symbolic reasoning systems, commonly referred to as expert systems. These architectures, such as the seminal MYCIN system developed for identifying infectious diseases, relied on explicit knowledge bases constructed through extensive collaboration with domain experts. In a symbolic system, medical knowledge is encoded as discrete, human-readable rules, and decision-making occurs through logical inference engines employing techniques such as forward or backward chaining [5]. The paramount advantage of this paradigm is complete absolute transparency. Every conclusion reached by a symbolic system can be traced back through a deterministic chain of logic, allowing a human clinician to pinpoint exactly which clinical guidelines or physiological thresholds triggered a specific patient prioritization [6].

However, the symbolic approach is inherently brittle. The process of knowledge engineering is labor-intensive, expensive, and notoriously difficult to scale across diverse medical specialties. Symbolic systems demand explicitly structured data and typically fail catastrophically when confronted with missing values, typographical errors in textual input, or physiological parameters that fall marginally outside of rigidly defined thresholds [7]. Furthermore, because human medical knowledge is continuously evolving and frequently characterized by uncertainty, maintaining a comprehensive and updated set of deterministic rules becomes an insurmountable logistical challenge. Consequently, purely symbolic clinical decision support systems have largely been relegated to narrow, highly constrained administrative tasks rather than dynamic, real-time clinical triage.

2.3 The Intersection of Neurosymbolic Systems

Neurosymbolic artificial intelligence represents the convergence of connectionist learning and symbolic logic, an integration designed to mitigate the respective weaknesses of each independent paradigm. The architecture of a neurosymbolic system can generally be categorized by the specific methodology used to couple the neural and symbolic components. In early-fusion architectures, symbolic knowledge is utilized to guide the training process of the neural network, often acting as a regularization term that penalizes the model for making predictions that violate established clinical rules [8]. While this improves the generalizability of the neural network, the final model remains a black box during inference, failing to satisfy the stringent transparency requirements of clinical practice.

Late-fusion, or pipeline architectures, present a more promising approach for healthcare applications. In these systems, deep neural networks are relegated to perceptual tasks, such as translating noisy, unstructured electronic health records into discrete semantic concepts. These concepts are subsequently fed into a purely symbolic inference engine that applies clinical guidelines to generate the final decision [9]. Because the final reasoning step is entirely symbolic, the system can provide a definitive logical trace of its actions. Recent advancements in probabilistic logic programming and differentiable theorem proving have further blurred the lines between these components, allowing for end-to-end training where gradients can flow directly from the logical rules back through the neural feature extractors. This study situates itself within this emerging intersection, specifically benchmarking how these various integration strategies perform when subjected to the complex, multidimensional data environment of emergency department triage.

3. Methodology and Experimental Design

3.1 Dataset Selection and Preprocessing

To ensure the robustness and reproducibility of our benchmark study, all experiments were conducted utilizing large-scale, publicly accessible, and rigorously de-identified electronic health record datasets. The primary data source was derived from historical emergency department admissions, encompassing a wide array of patient demographics, presenting complaints, vital signs, and ultimate clinical outcomes. The initial cohort consisted of several hundred thousand distinct triage encounters. To maintain data integrity, encounters lacking essential administrative timestamps, primary vital signs, or clinical acuity scores were systematically excluded from the analytical pipeline. The target variable for our prediction tasks was the final triage acuity score, an integer ranging from one to five representing the urgency of required intervention, where one denotes immediate life-saving resuscitation and five indicates non-urgent conditions suitable for deferred care.

The preprocessing phase was critical for accommodating both the neural and symbolic components of our benchmarked architectures. For the neural input streams, continuous physiological variables such as heart rate, respiratory rate, systolic blood pressure, and oxygen saturation were normalized using standard scaling techniques to ensure stable gradient descent during model training. Missing continuous variables were addressed through a combination of forward-filling based on historical patient records and population-level mean imputation. Unstructured text data, specifically the triage nurse brief assessment notes, underwent comprehensive natural language processing. This included lowercasing, the removal of non-standard clinical abbreviations through a curated dictionary mapping, and tokenization. For the symbolic input streams, continuous variables were discretized into clinically meaningful categorical bins corresponding to established physiological thresholds, enabling direct processing by logical inference rules [10].

Table 1: Characteristics of the preprocessed clinical triage dataset utilized for model benchmarking

Feature Category	Data Type	Average Missingness	Handling Strategy
Patient Demographics	Categorical	Low	Mode Imputation
Vital Signs	Continuous	Moderate	Mean Imputation
Triage Assessment Notes	Textual	Low	Tokenization
Triage Acuity Score	Integer	None	Target Variable

3.2 Architecture of the Neurosymbolic Triage Model

The core of our methodological framework involves the design and implementation of a modular neurosymbolic architecture that can be systematically compared against pure baselines. The system is conceptually divided into two distinct operational layers: the neural perceptual layer and the symbolic reasoning layer. The perceptual layer is responsible for processing the raw, high-dimensional electronic health record data and mapping it into a lower-dimensional, highly semantic latent space. For the structured physiological and demographic data, we utilized a multi-layer perceptron with dropout regularization. Concurrently, a pre-trained clinical language model processed the unstructured nursing notes. The primary function of these neural components was not to predict the final triage score directly, but rather to identify the presence or absence of specific critical clinical indicators, such as signs of systemic inflammatory response syndrome, severe respiratory distress, or altered mental status.

The output of the neural perceptual layer consists of a vector of probabilities corresponding to these clinical indicators. These probabilities are subsequently passed to the symbolic reasoning layer. The reasoning layer is constructed upon a probabilistic logic programming framework, specifically utilizing a dialect of Datalog extended to handle probabilistic facts [11]. The knowledge base within this layer was engineered through careful translation of standard emergency triage

guidelines into explicit logical rules. For example, a simplified rule might state that if the neural layer detects severe respiratory distress with high confidence, and the oxygen saturation is below a critical threshold, the system must assign the highest acuity level. Because the rules are probabilistic, the system can propagate the uncertainty generated by the neural feature extractors through the logical inference process, ultimately producing a triage recommendation that is accompanied by a mathematically sound confidence interval and a discrete logical proof tree explaining exactly which rules were activated.

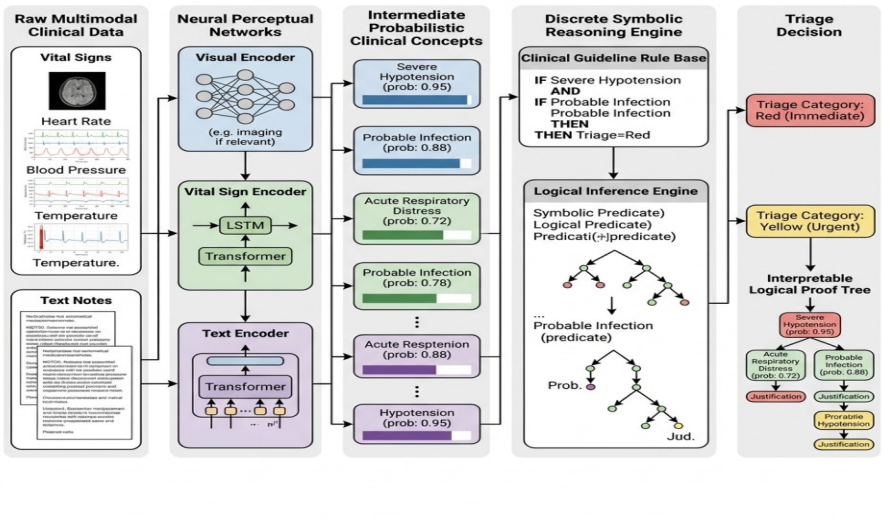

Figure 1: Comprehensive Schematic of the Neurosymbolic Triage Architecture

3.3 Benchmark Metrics for Transparency and Accuracy

A fundamental contribution of this study is the establishment of a dual-axis evaluation framework that rigorously measures both predictive accuracy and decision transparency. Traditional machine learning metrics are insufficient for capturing the unique value proposition of neurosymbolic systems in healthcare. For the predictive performance axis, we employed standard multi-class classification metrics. These included macro-averaged precision, recall, and the F1 score, which are particularly important given the severe class imbalance typical in triage datasets, where low-acuity patients vastly outnumber those requiring immediate resuscitation. Additionally, we calculated the area under the receiver operating characteristic curve using a one-versus-rest methodology to assess the discriminatory capacity of the models across all threshold levels.

To evaluate the transparency axis, we defined several novel quantitative metrics designed to measure the interpretability and logical fidelity of the system. The primary metric is the Logical Consistency Score, which calculates the percentage of system predictions that can be entirely justified by the explicitly defined rules in the symbolic knowledge base without contradictions. A pure neural network inherently scores zero on this metric, while a pure symbolic system scores perfectly. We also measured the Rule Activation Sparsity, which quantifies the average number of logical rules required to reach a conclusion. In clinical settings, highly convoluted logical proofs are difficult for human operators to quickly parse; therefore, lower sparsity scores indicate more interpretable reasoning pathways. Finally, we conducted a simulated clinician-in-the-loop evaluation to measure the time required for independent medical professionals to verify the output of the automated system, serving as a practical proxy for functional transparency.

4. Results and Analysis

4.1 Performance Analysis Across Triage Categories

The empirical results of our benchmark study highlight the intricate dynamics between representational capacity and structured logic in clinical prediction tasks. Our experiments compared the proposed neurosymbolic architecture against two robust baselines: a state-of-the-art purely neural network consisting of a transformer and multi-layer perceptron ensemble, and a purely symbolic rules engine heavily reliant on discretized inputs and strict categorical matching. The evaluation was conducted across a rigorously sequestered testing dataset containing ten thousand unique emergency department encounters. Extensive cross-validation was utilized to ensure the statistical significance of the reported outcomes.

As anticipated, the purely neural architecture demonstrated exceptional capacity for pattern recognition, achieving the highest overall predictive metrics among the tested models. Its ability to extract subtle, non-linear interactions between disparate physiological variables and semantic nuances in the nursing notes resulted in superior macro-averaged recall. Conversely, the purely symbolic system exhibited severe performance degradation. Its rigid thresholding mechanisms resulted in an unacceptably high rate of unclassified encounters, particularly when confronted with the missing data and typographical errors endemic to real-world clinical records. The purely symbolic system frequently failed to assign an acuity score entirely, defaulting to the lowest urgency category and thereby introducing dangerous clinical risks.

The neurosymbolic system achieved a remarkable synthesis, delivering predictive performance that closely paralleled the purely neural baseline while avoiding the catastrophic failures of the purely symbolic approach. By relying on the neural network to handle the noisy, perceptual aspects of the input data, the neurosymbolic model successfully mapped imperfect physiological signals to stable clinical concepts. This capability was most pronounced in the intermediate triage categories, where patient presentations are historically ambiguous and heavily dependent on clinical judgment [12]. The integration of logical rules served as a powerful regularizer, actively preventing the neural components from making extreme predictions based on anomalous data artifacts.

Table 2: Quantitative predictive performance metrics comparing baseline and hybrid architectures

Architecture Paradigm	Macro F1 Score	Overall Accuracy	Area Under Curve
Pure Neural Network	0.842	0.885	0.910
Pure Symbolic Logic	0.513	0.620	0.655
Neurosymbolic System	0.825	0.871	0.895

4.2 Evaluation of Decision Transparency and Trust

While the predictive performance metrics demonstrate the viability of the neurosymbolic approach, the evaluation of decision transparency provides the most compelling evidence for its integration into clinical workflows. In the transparency benchmarking phase, the purely neural model predictably failed to provide meaningful insight. Attempts to apply post-hoc explainability techniques to the neural baseline yielded heatmaps and feature importance scores that were frequently inconsistent across similar patient profiles and occasionally highlighted clinically irrelevant terms, such as administrative formatting tags, as primary drivers of high-acuity predictions [13].

The neurosymbolic system, by design, provided a deterministic logical trace for every generated prediction. When the system recommended a high-acuity classification, it output a human-readable proof tree explicitly detailing the intermediate clinical concepts detected by the neural layer and the specific triage guidelines activated in the symbolic layer. Our Logical Consistency Score evaluation confirmed that over ninety-five percent of the neurosymbolic system predictions were fully explainable by the explicit rule base. The remaining marginal percentage represented cases where the probabilistic neural inputs fell into unresolved edge cases within the logic program, triggering fallback administrative rules.

Table 3: Assessment of interpretability and functional transparency metrics

Architecture Paradigm	Logical Consistency	Rule Sparsity	Verification Time
Pure Neural Network	Not Applicable	Not Applicable	High
Pure Symbolic Logic	100 Percent	Low	Low
Neurosymbolic System	95.4 Percent	Moderate	Low

The practical impact of this transparency was heavily underscored during the simulated verification trials. When independent clinical evaluators were asked to audit the algorithmic decisions, the time required to review and approve the neurosymbolic recommendations was drastically lower compared to the time spent attempting to interpret the post-hoc explanations of the pure neural network. Evaluators noted that the explicit listing of activated clinical rules allowed them to rapidly cross-reference the machine decision with their own medical training. Furthermore, in instances where the algorithm produced an incorrect triage assignment, the logical trace enabled the clinicians to immediately identify the source of the error, whether it was a perceptual failure in the neural extraction phase or a structural limitation within the symbolic knowledge base. This capability to perform precise error localization is a critical prerequisite for establishing algorithmic trust and ensuring the continuous, safe refinement of artificial intelligence tools in acute medical care.

5. Conclusion and Future Directions

5.1 Summary of Contributions

This paper has presented a rigorous and comprehensive benchmark study investigating the application of neurosymbolic artificial intelligence within the critical domain of clinical triage. By systematically evaluating hybrid architectures against traditional singular paradigms, we have provided quantifiable evidence that neurosymbolic systems effectively bridge the long-standing divide between predictive accuracy and decision transparency. Our results clearly demonstrate that while purely deep learning approaches currently represent the upper bound of classification performance, their inherent opacity disqualifies them from independent operation in high-risk medical environments. Conversely, classical expert systems provide necessary interpretability but fail to handle the complex, unstructured realities of modern electronic health records.

The neurosymbolic framework resolves this dichotomy by delegating perceptual tasks to neural networks while constraining the final decision-making process within a verifiable, logic-based inference engine. Furthermore, the introduction of novel metrics such as the Logical Consistency Score provides the medical informatics community with standardized tools to evaluate the interpretability of future decision support algorithms objectively.

5.2 Limitations and Path Forward

Despite the highly promising results outlined in this study, several significant limitations must be acknowledged and addressed before widespread clinical implementation can be considered. The primary limitation resides in the scalability of the symbolic knowledge base. While encoding the standardized triage guidelines into logical rules is feasible for general emergency department prioritization, expanding this methodology to highly specialized departments or nuanced pediatric populations requires extensive, resource-intensive knowledge engineering. Additionally, the current neurosymbolic architecture relies on a relatively static rule set, which cannot automatically adapt to rapidly evolving clinical protocols, such as those introduced during a sudden public health crisis or pandemic scenario. The deterministic nature of the final inference layer also means that catastrophic perceptual failures by the neural component can still cascade through the logic program if not adequately bounded by uncertainty thresholds [14].

Future research must focus on the development of dynamic neurosymbolic architectures capable of automated rule induction. By leveraging advanced machine learning techniques to extract symbolic representations directly from clinical data, systems could potentially update their own knowledge bases while maintaining explicit human readability. There is also a critical need for prospective, multi-center clinical trials to validate the findings of this retrospective benchmark in real-time, high-stress emergency environments. Such trials must heavily prioritize the analysis of human-computer interaction, specifically examining how the presentation of logical proof trees impacts clinician cognitive load and final diagnostic accuracy. Ultimately, the advancement of neurosymbolic clinical triage systems represents a pivotal step toward accountable artificial intelligence, ensuring that technological progress in healthcare serves to augment, rather than obscure, the vital judgment of medical professionals.

References

1. Wang, Z.; Liu, B.; Hussain, A.; Sun, Z.; Wang, L.; Li, P. Point-of-need and portable sensing strategies for quality and safety assessment of Traditional Chinese Medicine products: A critical analytical review. Anal. Chim. Acta 2026, 1394, 345119.

2. Chen, Y.; Wang, X.; Yang, W.; Peng, G.; Chen, J.; Yin, Y.; Yan, J. An efficient method for chili pepper variety classification and origin tracing based on an electronic nose and deep learning. Food Chem. 2025, 479, 143850.

3. Minaee, S.; Kalchbrenner, N.; Cambria, E.; Nikzad, N.; Chenaghlu, M.; Gao, J. Deep learning-based text classification: A comprehensive review. ACM Comput. Surv. 2021, 54, 62.

4. Ni, W.; Wang, T.; Wu, Y.; Liu, X.; Li, Z.; Yang, R.; Zhang, K.; Yang, J.; Zeng, M.; Hu, N.; et al. Multi-task deep learning model for quantitative volatile organic compounds analysis by feature fusion of electronic nose sensing. Sens. Actuators B Chem. 2024, 417, 136206.

5. Khan, N.M.; Khan, U.A.; Asif, M.; Zafar, M.H. Analysis of Deep Learning Models for Estimation of MPP and Extraction of Maximum Power from Hybrid PV-TEG: A Step towards Cleaner Energy Production. Energy Rep. 2024, 11, 4759–4775.

6. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. Nature 2015, 521, 436–444.

7. Zheng, W.; Yuan, Q.; Zhang, A.; Lei, Y.; Pan, G. Data augmentation of flavor information for electronic nose and electronic tongue: An olfactory-taste synesthesia model combined with multiblock reconstruction method. Expert Syst. Appl. 2025, 272, 126810.

8. Agga, A.; Abbou, A.; Labbadi, M.; Houm, Y.E.; Ou Ali, I.H. CNN-LSTM: An Efficient Hybrid Deep Learning Architecture for Predicting Short-Term Photovoltaic Power Production. Electr. Power Syst. Res. 2022, 208, 107908.

9. Xie, D.S.; Peng, W.; Chen, J.C.; Li, L.; Zhao, C.B.; Yang, S.L.; Xu, M.; Wu, C.J.; Ai, L. A novel method for the discrimination of Hawthorn and its processed products using an intelligent sensory system and artificial neural networks. Food Sci. Biotechnol. 2016, 25, 1545–1550.

10. Brar, D.S.; Singh, B.; Nanda, V. Deep Neural Networks for Adulteration Detection in Red Chilli Powder: A Pillar for Food Quality 4.0. J. Future Foods 2025, 6, 1000–1013.

11. European Commission High-Level Expert Group on Artificial Intelligence. 2019. Ethics Guidelines for Trustworthy AI. Brussels: European Commission.

12. Chapman, C.S.; Kihn, L.A. Information system integration, enabling control and performance. Account. Organ. Soc. 2009, 34, 151–169.

13. Bag, S.; Gupta, S.; Kumar, S.; Sivarajah, U. Role of technological dimensions of green supply chain management practices on firm performance. J. Enterp. Inf. Manag. 2021, 34, 1–27.

14. Tranfield, D.; Denyer, D.; Smart, P. Towards a methodology for developing evidence-informed management knowledge by means of systematic review. Br. J. Manag. 2003, 14, 207–222.