safety

MDPI

Review

AI-Driven Approaches for Adverse Event Detection: A Systematic Review of Current Evidence

Francesco De Micco [1,2], Gianmarco Di Palma [1,2,*], Greta Seveso [3,4], Flavia Giacomobono [2], Roberto Scendoni [5,6] and Vittoradolfo Tambone [1]

[1] Bioethics and Humanities Research Unit, Campus Bio-Medico University of Rome, 00128 Rome, Italy; f.demicco@policlinicocampus.it (F.D.M.); v.tambone@unicampus.it (V.T.)
[2] Operative Research Unit of Clinical Affairs, Campus Bio-Medico University Hospital Foundation, 00128 Rome, Italy; flavia.giacomobono@alcampus.it
[3] Department of Public Health, Experimental and Forensic Medicine, University of Pavia, 27100 Pavia, Italy; g.seveso@smatteo.pv.it
[4] Unit of Legal Medicine, IRCCS Fondazione Policlinico San Matteo, 27100 Pavia, Italy
[5] Department of Law, Institute of Legal Medicine, University of Macerata, 62100 Macerata, Italy; r.scendoni@unimc.it
[6] Italian Network for Safety in Healthcare (INSH), Coordination of Marche Region, 60126 Ancona, Italy
* Correspondence: g.dipalma@policlinicocampus.it

Abstract

Introduction: Hospital adverse events are a global patient safety problem that account for avoidable death, long-term disability, extended length of stay, and increased healthcare costs. Underreporting, wherein fewer than 10% of events are indeed recorded, is prevalent and is characterized primarily by cultural and organizational determinants. Artificial intelligence, in the form of machine learning and natural language processing, has been described as a potential tool for enhancing adverse events detection and prediction with the use of large-scale clinical data. **Materials and Methods:** PRISMA-DTA guidelines were followed in this systematic review. Scopus, PubMed, and Web of Science were searched employing keywords associated with adverse events, artificial intelligence methodologies (e.g., machine learning, deep learning, natural language processing), and healthcare settings. Inclusion criteria included original research on artificial intelligence-based solutions for the detection or prediction of adverse events such as medication errors, hospital-acquired infections, and complications during surgery. Reviews, meta-analyses, and non-artificial intelligence studies were excluded. Following screening, 15 studies were found to meet inclusion criteria. **Results:** The referenced studies show a shift from rule-based natural language processing models to advanced deep learning and Bidirectional Encoder Representations from Transformers models. Early approaches, i.e., Support Vector Machine classifiers, achieved AUC scores as high as 0.92, while later models (Random Forest, LightGBM, XGBoost) mirrored AUCs of over 0.93. Large language models achieved F1-scores of 0.84 for named entity recognition. Artificial intelligence models even identified unreported incidents. **Discussion:** Artificial intelligence-powered methods are transforming adverse events detection from retrospective to predictive, proactive monitoring. There remain some challenges, however, including limited external validation, class imbalance, and interpretability of complex models. Future studies must address explainable artificial intelligence, multicenter trials, and high-quality well-annotated datasets to offer secure clinical integration.

Keywords: artificial intelligence; patient safety; adverse events; healthcare risk management; intelligent systems

Academic Editor: Raphael Grzebieta

Received: 27 November 2025
Revised: 27 March 2026
Accepted: 7 April 2026
Published: 14 April 2026

1. Introduction

Hospital adverse events (AEs) are among the greatest threats to the safety and quality of patient care internationally. Not only do they undermine the efficacy of healthcare provision but also the confidence and trust of patients and their families, with adverse effects on the reputation, effectiveness, and financial viability of healthcare centers [1]. A landmark report by the Institute of Medicine was utilized to estimate that medical errors are responsible for 45,000 to 98,000 deaths annually in the United States alone and are therefore among the most frequent causes of preventable death [2]. Subsequent studies have highlighted that the burden of AEs extends beyond mortality, such as long-term disability, extended hospital stay, extra healthcare costs, and extreme emotional and psychological distress in patients and providers alike [3]. Despite using formal reporting systems as well as quality assurance measures, proper reporting of adverse events remains extremely low, with less than 10% of medical mistakes actually being formally documented [4]. Even when they are reported, just 15% of hospital reactions are associated with preventive measures that actually reduce the likelihood of recurrence [5]. Underreporting is generally attributed to an interplay of many structural, organizational, and cultural determinants, such as fear of blame or litigation, lack of good feedback systems, and a lack of a non-punitive safety culture [6]. Eliminating these obstacles is vital to individual and system resilience and improving patient safety as a necessary part of health care quality.

Medical mistakes are broadly defined as avoidable omissions or commission leading to unplanned harm, clinical harm, or failure to achieve the intended therapeutic outcome [7]. These events can be caused by different sources and may touch different stages of the process of care, ranging from medication prescribing and drug administration to diagnosis, surgery, post-surgical care, and observation of patients [8]. Some of the most important categories include adverse drug events, misdiagnosis, hospital-acquired infections, procedural complications, and patient falls, each of which poses different detection and prevention challenges [9].

Healthcare risk management has emerged as a multidisciplinary process whose goal is to identify, examine, and manage risks associated with clinical care in a systematic manner. In the past, it has relied on reactive mechanisms, i.e., incident reporting and root cause analysis, as well as proactive methodologies, i.e., Failure Mode and Effects Analysis (FMECA) and clinical audits, whose aim is to anticipate probable errors prior to their occurrence [10]. These efforts are increasingly complemented by advanced data-driven techniques exploiting the expanding use of electronic health records (EHRs), clinical databases, and real-time monitoring technologies. In this context, artificial intelligence (AI) is a paradigm-shifting technological innovation with the potential to significantly improve adverse event detection and overall clinical safety. AI techniques, particularly machine learning (ML) and natural language processing (NLP), can analyze vast volumes of structured and unstructured clinical data to identify subtle patterns and outliers that can precede the occurrence of adverse events [11]. Predictive models, for example, can highlight early warning signs of sepsis, drug errors, or post-surgical complications so that clinicians can intervene early and reduce harm [12]. Furthermore, AI-powered clinical decision support systems (CDSS) are increasingly being integrated into healthcare workflows in order to enhance the accuracy of diagnosis, reduce care variation, and support evidence-based decision-making.

However, the use of AI technologies in clinical risk management is not without some challenges. Some of these include algorithmic bias, data quality, transparency, explainability, and the need for robust clinical validation, which are both ethical and operational chal-

lenges [13]. Additionally, harmonious integration of AI demands enormous cultural and organizational change, creation of an appropriate regulatory framework and governance models that ensure patient safety as well as data privacy [14].

This systematic review aims to provide a comprehensive and critical summary of the application of AI for adverse event detection in the health care setting. Specifically, it will explore current AI applications, whether they are effective in improving risk prediction and patient outcomes, how they can be replicated across different clinical settings, and the limitations of implementation. By synthesizing recent evidence, this review seeks to identify potential and limitations of AI-based approaches, and to offer insights valuable for understanding how the technologies might be applied to establish a culture of safety through a data-driven environment in modern healthcare systems.

2. Materials and Methods

This systematic review was conducted following the methodological framework provided by the PRISMA-DTA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy Studies) guidelines [13]. Systematic literature searching was conducted for primary studies that reported on the application of AI techniques to identify adverse events in medicine. For wide coverage and international scientific literature with balanced representation, the databases PubMed, Scopus, and Web of Science were used. They were chosen based on their complementarity: PubMed provides wide coverage of biomedical and clinical literature, Scopus has a large number of multidisciplinary journals and conference proceedings relevant to AI, and Web of Science ensures inclusion of highly cited papers with exact citation tracking. The search strategy was developed using a conceptual block approach, grouping keywords into three main sets: A) adverse events ("adverse event" OR "adverse drug event*" OR "adverse drug reaction*" OR "medical error*" OR "patient safety"), B) artificial intelligence techniques ("machine learning" OR "deep learning" OR "neural network*" OR "natural language processing" OR NLP*), and C) healthcare context ("healthcare" OR "electronic health record" OR EHR OR hospital* OR clinic*).* The final search string is as follows and has been adapted to the syntax of each database, as follows: ("adverse drug event" OR "adverse drug events" OR "adverse drug reaction" OR "adverse drug reactions" OR "medication error" OR "medication errors" OR "hospital acquired infection" OR "surgical complication" OR "surgical complications") AND ("machine learning" OR "deep learning" OR "natural language processing") AND ("electronic health record" OR "electronic health records" OR EHR OR "clinical notes" OR "hospital database"). We chose a 15-year time frame (approximately 2010 to the present) because this period encompasses the critical transition from traditional rule-based or statistical systems to advanced predictive models based on machine learning and natural language processing, culminating in large language models and multimodal architectures. Selecting this interval allows us to capture the evolution of AI technologies that are directly relevant to contemporary clinical applications, reflecting both methodological advances and the increasing availability of high-quality electronic health data.

We incorporated research studies that were original and emphasized the application of machine learning, deep learning, or natural language processing techniques in detecting or predicting adverse events (e.g., medication errors, adverse drug events, hospital-acquired infections, or surgical complications) in healthcare settings such as hospitals, clinics, or EHR systems. Studies were included regardless of whether they addressed general-purpose AE detection or focused on specific categories of AE, provided that AI-based methods were applied for AE detection or prediction.

We excluded reviews, meta-analyses, editorials, letters, case reports, and studies not explicitly based on detecting adverse events or employing only standard statistical analyses

without AI methods. We did not employ any language or time restrictions during the search process. However, in the ultimate selection process, studies published in English were prioritized to ensure comparability and consistency of data.

These records were then exported to reference management software, non-duplicate articles were deleted, and a two-step screening was performed: title and abstract screening followed by full-text review to ascertain eligibility (Figure 1; Table S1).

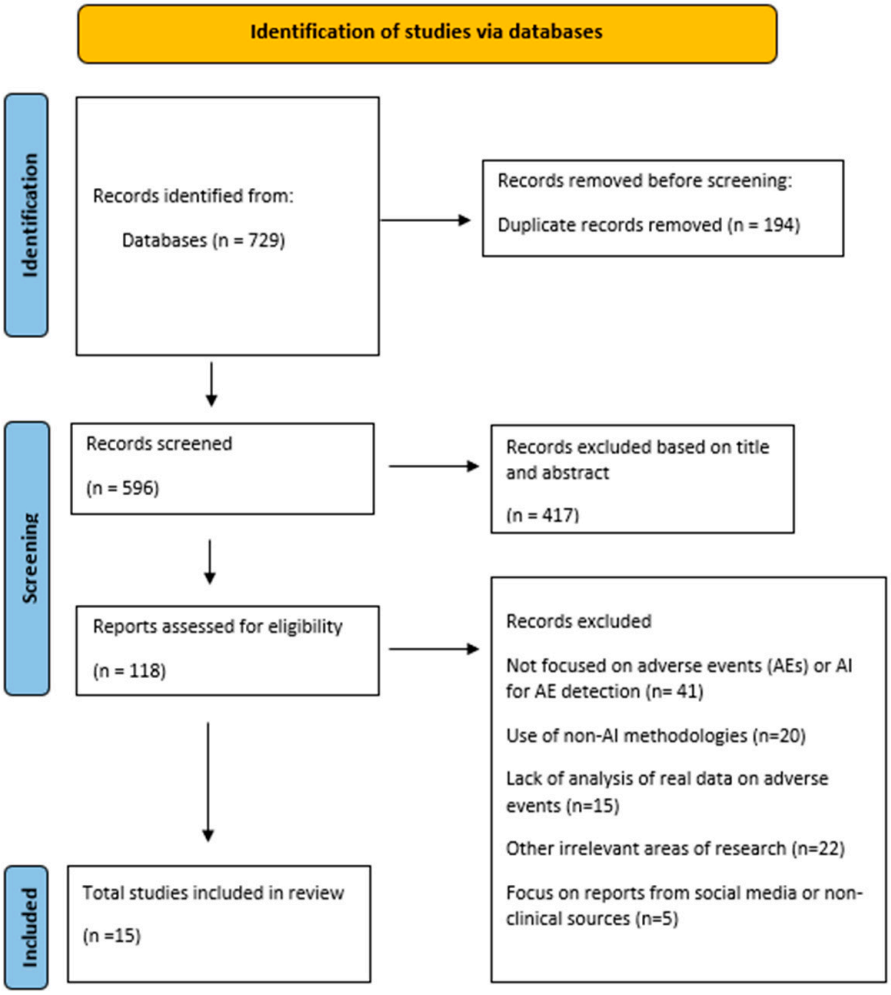

Figure 1. Identification of studies via databases and registers.

3. Results

Synthesis of the 15 studies included showed that the application of AI-driven methods for AE identification has grown more advanced over recent years, moving from basic keyword-based classification models to more sophisticated deep learning techniques and large language models (LLMs). Together, the studies demonstrate that these techniques are not just capable of accelerating clinical data analysis but also detecting unreported or difficult-to-detect events using traditional systems (Table 1).

Table 1. Descriptive summary of the included studies.

Study (Year)	Primary Focus	AI Approach	Key Findings	Data Source	Sample Size (*n*)
Ong (2011) [15]	Automated classification of high-risk clinical incident reports	NLP + SVM	Linear SVM achieved an AUC of 0.92, successfully identifying SAC1 events overlooked by manual review	Incident reports	8083
Gerdes (2013) [16]	Detection of pressure ulcers	Text mining with SAS	Achieved 70% sensitivity and 95% specificity compared to the Global Trigger Tool	EHR	498
Fujita (2012) [17]	Semantic analysis of large-scale incident reports	NLP + clustering	Revealed thematic clusters (e.g., falls, medication management) and highlighted the need for improved incident taxonomies	Incident reports	18,353
Gupta (2015) [18]	Classification of Clinical Incident Types	Naïve Bayes	Accuracy of 80%, AUC of 0.91 using expert-labeled data	Incident reports	5448
Gupta (2019) [19]	Classification of Clinical Safety Incidents	Random Forest	AUC up to 0.98 with optimized WHO-I taxonomy	Incident reports	3600
Barmaz (2021) [20]	Detection of underreporting in clinical trials	Bayesian modeling	Successfully identified outlier sites using rate tail area methods	Clinical trial data	468
Menard (2019) [21]	Quality assurance in clinical trials	Gradient Boosting Machines	AUC up to 0.92 in simulated underreporting scenarios	Clinical trial data	288,254
Kopacheva (2025) [22]	Adverse drug event detection from clinical notes	LLM (BERT-based)	Achieved F1 = 0.84 (NER) and F1 = 0.70 (RE), outperforming traditional NLP approaches	EHR	905
Okamoto (2020) [23]	Detection of unreported severe incidents	SVM applied to clinical notes	Identified 34 severe incidents not reported through conventional systems (precision 28%)	EHR	121
Fujita (2013) [24]	Cross-hospital semantic analysis of incident reports	NLP + clustering	Identified common patterns across hospitals, suggesting unified reporting categories	Incident reports	21,261

Table 1. *Cont.*

Study (Year)	Primary Focus	AI Approach	Key Findings	Data Source	Sample Size (*n*)
Liu (2019) [25]	Analysis of hospital falls	Word2Vec + clustering	Identified five distinct fall clusters characterized by severity and context	Incident reports	1683
Hendrickx (2021) [26]	Analysis of patient complaints for safety signals	NLP + Naïve Bayes	Achieved accuracy between 87 and 93%, supporting prioritization of high-risk providers	Patient complaints	22,509
Fong (2017) [27]	Medication safety event analysis	NLP + SVM	ROC-AUC up to 0.96 for specific error categories using brief text data	Medication safety events	774
Benin (2015) [28]	Automation of adverse event report analysis	Rule-based vs FS-LapSVM	FS-LapSVM reached 91% sensitivity for weight-related errors	Incident reports	9164
Fodeh (2015) [29]	Semi-supervised classification of pediatric errors	FS-LapSVM	Accuracy of 88%, reducing the need for manual labeling	Incident reports	9405

Initial applications of AI for AE detection focused on NLP and text mining to convert narrative clinical reports to structured data. For instance, Ong, M et al. [15] showed that an SVM-based classifier accurately distinguished between high-risk events in incident reports (AUC = 0.92) with a significant reduction in misclassification errors compared to manual review by healthcare professionals.

This was the same approach taken by Gerdes, L.U and C. Hardahl [16], who used text mining as an alternative to the manual reading required by the Global Trigger Tool, with the outcome that NLP systems can achieve 70% sensitivity and 95% specificity for the detection of pressure ulcers. Fujita, K., et al. [17] extended this concept by clustering term co-occurrence networks across tens of thousands of reports to identify thematic patterns (e.g., falls, medication management) not distinctly represented in traditional "top-down" taxonomies. Concurrent with these activities, other studies leveraged traditional machine learning methods for the automatic classification of reports [18,19]. Using thousands of incident reports from Australian hospitals, these studies demonstrated that ML algorithms such as Naïve Bayes and Random Forest, in combination with meticulously crafted taxonomies (e.g., WHO-I), were able to achieve AUC scores near 0.98, significantly outperforming manual coding in both accuracy and efficiency.

Other studies have sought to identify anomalies in the safety data of clinical trials, singling out underreporting specifically. A hierarchical Bayesian model developed by Bermaz et al. was able to estimate AE reporting probability by site in multicenter trials and hence flag "abnormal" behavior with minimal requirements for large-scale audits [20]. Similarly, Menard et al. applied a gradient boosting machine approach, and the AUCs became 0.92 for the simulation of underreporting. Although these procedures do not classify events, they are tactical instruments to verify the quality of safety data [21].

The latest developments in this field are defined by employing next-generation language models, including Bidirectional Encoder Representations from Transformers (BERT)-

based models, for the automatic identification of entities and relations in intricate clinical texts. Kopacheva et al. used a Swe-De-Clin-BERT model to detect adverse events within discharge notes and reported that the NER task achieved an F1-score of 0.84, along with an F1-score of 0.70 on the RE task. This approach circumvents the pitfalls of typical feature-engineered methods but also faces substantial challenges, such as the absence of high-quality annotated data [22].

One of the common threads across most of the research is the prospect of uncovering events not seen in official monitoring systems. Okamoto et al. formulated an SVM-driven system that, in reviewing nearly 300,000 clinical notes, uncovered 34 critical incidents never previously reported to hospital administrations, showing the potential of AI in uncovering reporting gaps [23].

In general, the analysis highlights some key trends. First, SVM and Random Forest algorithms remain effective in supervised classification, particularly on balanced datasets. Second, while old NLP approaches are still viable, they are increasingly being outperformed by LLM-based approaches, which better understand clinical language. Finally, reconciling probabilistic modeling and ML for identifying anomalies (e.g., underreporting) looks to expanding the role of AI, now extending beyond basic event recognition to overall control of safety data.

4. Discussion

4.1. Overview of AI Evolution in AE Detection

Overall analysis of the 15 studies suggests a radical shift in AI methods in the detection of AEs, from primarily retrospective, rule-based approaches to predictive, proactive ones more integrated with clinical processes. This change reflects not only technological improvements (from classical NLP to LLMs) but also growing appreciation for the value of unstructured data and the strength of automation in patient safety surveillance.

Earlier work, such as that of Fujita [17,26] and Ong [15], laid the groundwork for automating text analysis, demonstrating that simple techniques such as bag-of-words, TF-IDF, and SVM were already surpassing manual classification in both precision and efficiency. Such methods have several limitations, though: they are unable to pick up on complex semantic relationships, depend on hand-engineered features, and are difficult to apply to heterogeneous clinical domains.

One of the most important developments was the application of semantic embedding models (e.g., Word2Vec in Liu et al. [25] and more recently LLMs such as Swe-De-Clin-BERT [22]), which enable improved understanding of the clinical language context. Kopacheva's results demonstrate a decisive advantage for Named Entity Recognition (F1 = 0.84), suggesting that the trend towards advanced language models will be a primary force behind future progress [22].

4.2. Comparison with Previous Reviews

Compared to other review studies in the literature, such as those by Hoekstra et al. (2022) and Deimazar and Sheikhtaheri (2023), the analysis conducted in this work presents a broader and more transversal orientation on artificial intelligence approaches applied to the identification and prediction of adverse events in clinical settings [30,31]. Previous reviews focus mainly on specific areas: the first adopts a general perspective on the prediction of events from textual data, including also non-health domains, while the second focuses mainly on ADE and ADR extracted from EHRs and clinical notes. In the comparison, this review extends the field of observation to the different types of hospital adverse events and highlights not only algorithmic performance, but also systemic aspects such as underreporting, integration of models into risk management processes, multicenter

validation needs, data quality and the implications deriving from the use of complex models such as LLMs and explainable AI techniques. This approach allows us to outline a more integrated and operational picture of the potential role of AI in patient safety, overcoming the narrower thematic boundaries of the available reviews.

4.3. Underreporting and Data Governance

The problem of underreporting, so critical in clinical trials, has been addressed innovatively by Barmaz [20] with a hierarchical Bayesian model and by Menard [21] with gradient boosting algorithms. These studies go beyond event detection, quantifying the quality of the data itself by identifying "outlier" sites or anomalous reporting patterns. This is the essential shift in mindset: AI is not a data mining tool but a data governance platform that can support audit and risk management processes.

Most disruptive, possibly, is the discovery that AI is able to identify "hidden" or unreported adverse events. Okamoto et al. [23], for instance, employing an SVM model to trawl through nearly 300,000 clinical notes, identified 34 serious incidents that had not been recorded on official systems. This finding underscores two key points: first, the inherent limitations of human-input-dependent reporting systems; second, the utility of AI as a "second layer" of surveillance that can detect weak signals or occurrences that fall below human radar.

4.4. Generalizability and Data Imbalance

In spite of the great progress observed in the domain, some challenges are still prevalent among the reviewed works. An initial key challenge pertains to the limited external validity of the proposed models. In most cases, assessment is performed on data from one hospital or from very precise clinical environments, which does not support an effective test of their performance across different health care settings. Such a lack of cross-site testing challenges the generalizability and scalability of the findings. From an AI perspective, many studies focus on individual clinical domains, highlighting how adverse event detection still remains fragmented and not fully established as a clinical risk management discipline. This fragmentation limits the ability of models to learn and integrate signals from heterogeneous adverse events, reducing their potential impact on overall clinical risk management. Another major obstacle is that of data imbalance, particularly for rare side effects such as SAC 1 events. By definition, these events produce highly imbalanced datasets and this can considerably weaken the accuracy and stability of predictive models, as highlighted by Hendrickx [26]. Handling such imbalance typically involves applying sophisticated resampling techniques or domain-specific loss functions, but even these may not be sufficient to fully mitigate the problem. Addressing issues such as data imbalance and limitations in external validity is essential not only to confirm and strengthen model performance, but also to build clinical trust in these tools. The adoption of strategies such as multicenter validation, clinician-in-the-loop approaches, and standardized reporting metrics can support safe implementation and promote broader adoption.

4.5. Multicenter Datasets and Methodological Standardization

In order to overcome the limitations currently encountered in terms of external validity and the absence of real cross-site verification, it seems essential that future research is oriented towards the systematic construction of high-quality multicenter datasets, developed according to rigorous and shared methodological criteria. This objective requires a coordinated and multi-level approach, capable of integrating technical, organizational and regulatory aspects. Firstly, it is essential to promote data harmonization processes between institutions, through the adoption of common data models and standardized clinical terminologies (such as ICD, SNOMED CT and MedDRA), so as to reduce semantic

heterogeneity and ensure a consistent representation of adverse events in different care contexts. In the absence of such standardization, the risk is to train models on formally similar but substantially non-comparable data. Secondly, the design of prospective multicenter protocols is desirable, which overcome the traditional dependence on monocentric retrospective datasets. The creation of collaborative networks between healthcare facilities, based on shared inclusion criteria, uniform data extraction pipelines and common annotation guidelines, would significantly improve the representativeness of data and reduce systematic biases linked to the single context. A further crucial element concerns the definition of high-quality annotation frameworks, which involve the involvement of expert clinicians and the use of standardized labeling protocols. In this context, the measurement of the inter-observer agreement should be considered an essential component, in order to ensure consistency and reliability in the different participating fora. At the same time, in light of the growing regulatory restrictions on data protection, the adoption of federated learning approaches and privacy-preserving methodologies, which allow the training and validation of models on distributed data without the need for centralization, is of particular importance. These strategies represent a concrete solution to combine scalability, interinstitutional collaboration and protection of patient confidentiality. From a methodological point of view, it is also necessary that external validation strategies are integrated from the design phase. In particular, the use of schemes such as "leave-one-site-out", in which models are trained on a set of centers and tested on completely independent institutions, allows us to obtain a more realistic estimate of generalizability compared to traditional random subdivisions within the same dataset. Finally, it is essential to improve the transparency and reporting standards of the datasets, through a detailed description of their composition, the specific characteristics of the centers involved, the prevalence of adverse events and the preprocessing procedures adopted. Whenever possible, the sharing of anonymized datasets and common benchmarks could further facilitate the reproducibility of studies and the comparison between different approaches. The systematic integration of these methodological practices would make it possible to overcome the current fragmentation of studies, favoring the transition from models developed in isolated contexts to truly robust, generalizable systems suitable for reliably supporting clinical risk management processes in heterogeneous healthcare environments.

4.6. Interpretability and Explainable AI

Another issue relates to interpretability. Although models like LightGBM and LLMs have strong predictive performance, these models are "black boxes," and clinicians cannot readily see the rationale behind some outputs or recommendations. Without adequate tools of explainability, clinical uptake of such models is problematic, particularly in high-stakes environments where decision transparency is essential. Finally, clinical documentation quality is the underlying limiting factor. Kopacheva reports that only 15% of discharge reports explicitly indicate the adverse drug event for which the patient was hospitalized [22]. Although these models demonstrate high detection capabilities, significant questions remain regarding their actual impact not only on clinical outcomes but also on fundamental aspects such as quality and patient safety. At this stage, a further step is needed, moving from evaluations based on retrospective or experimental datasets to studies conducted in real clinical settings. Future research should assess whether the integration of such AI tools into clinical workflows translates into a measurable and statistically significant reduction in adverse events, thereby improving patient safety. This deficiency in documentation not only disables training a model of AI but also limits the functionality of even the best models, which rely on the availability of correct and full data to deliver reliable results. Under no circumstances can an artificial intelligence system be considered acceptable if

it does not ensure full transparency. Healthcare professionals should not be expected to simply trust the system; rather, design must guarantee maximum interpretability from the outset. Measures should operate on two complementary levels. On one hand, during the design phase, role-based and logged access mechanisms can be implemented to ensure traceability, accountability, and secure interaction with the system, allowing healthcare professionals to inspect model outputs at the individual patient level while preserving data anonymity and integrity. However, such governance measures, while essential, are not sufficient per se to guarantee true interpretability, which instead requires the integration of dedicated explainability techniques capable of making the model's decision-making process transparent and clinically meaningful. In this context, the integration of dedicated explainable AI methods becomes an essential requirement. For example, techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) allow interpretation at the level of the individual patient, quantifying the contribution of individual variables to the specific prediction and thus making the decision-making process more transparent and clinically understandable. At the same time, attention mechanisms in NLP-based models allow the identification of the most relevant textual components within clinical narratives, while feature importance analysis and surrogate models offer a more global view of the system's behavior. The combined adoption of these approaches is therefore essential to bridge the gap between predictive performance and clinical interpretability, fostering trust building by healthcare professionals and facilitating the integration of artificial intelligence systems into daily clinical practice. Moreover, dedicated training programs are essential so that operators can fully understand how the tool works and provide patients with comprehensive information, avoiding uncritical acceptance of system-generated outputs. Only when clinicians are able to accurately interpret and reliably act upon AI-generated outputs can these tools be fully accepted in high-risk settings.

4.7. Limitations of the Review

Given the qualitative and descriptive nature of this review, which did not aim to perform a diagnostic accuracy meta-analysis, the focus was placed on synthesizing methodological approaches and application domains rather than formally evaluating internal validity or systematically assessing sources of bias in the individual studies. A further limitation of this review is the relatively small number of included studies, which reflects the strict focus on clinically relevant AI applications for AE detection and the emerging nature of the field in real-world settings. This should be considered when interpreting the comprehensiveness and generalizability of the findings. Another limitation of this review is that most included studies evaluate AI models using quantitative performance metrics (such as AUC, F1-score, and sensitivity) without reporting clinical utility or demonstrating whether the AI systems effectively reduced AEs or influenced clinician behavior. This reflects the current focus of the literature on methodological performance rather than impact. It should also be noted as a methodological limitation that although PubMed, Scopus and Web of Science were selected for their wide coverage and complementarity, databases more focused on engineering, such as IEEE Xplore and ACM Digital Library, were not included. This choice was made to focus the review on real clinical applications and not purely experimental ones, but it may have resulted in the exclusion of some relevant technical developments in the field of artificial intelligence.

4.8. Future Perspectives

The reviewed literature demonstrates that AI is not merely a technical tool but a strategic resource for patient safety. The integration of ML models with interactive dashboards [27] and automated auditing tools [21] is one step towards more proactive healthcare

systems, where monitoring of adverse events is real-time and continuous [22,29]. In the years to come, it can be revolutionary with the shift towards multimodal models with input of textual, numerical, imaging, and sensor information and LLMs trained on big multilingual clinical databases. Moreover, emerging self-supervised learning approaches for spatio-temporal modeling, such as dual-stream triplet Siamese networks for action recognition, illustrate promising methodological directions. Although these techniques have not yet been applied to clinical adverse event prediction, their ability to capture temporal correlations and structured sequences could inspire future AI models capable of detecting events with strong temporal dynamics. Concurrently, it will be essential to design uniform evaluation metrics and conduct multicenter validations to make these systems reliable and robust. Institutional policies and clinical risk governance within healthcare organizations should not only prioritize the deployment of AI systems, but also promote the establishment of clinical risk management frameworks, ensure staff training programs, and implement continuous auditing mechanisms, so as to guarantee that these tools enhance patient safety and quality of care without introducing new risks. Finally, the focus must be directed towards explainable AI so that models are adopted by clinicians not only for their accuracy but also for the explainability of their decisions. In summary, this review provides a concise overview of the current state of AI applications for AE detection and prediction. By identifying methodological trends, application domains, and key gaps, it offers a foundation for guiding future research and informing the potential clinical implementation of AI systems to improve patient safety.

5. Conclusions

This systematic review has highlighted the significant role of artificial intelligence (AI) methods in transforming adverse event (AE) detection. Current evidence shows that AI, particularly machine learning (ML) and natural language processing (NLP) techniques, is effective at automating the analysis of large volumes of clinical data, reducing the manual burden of identifying and auditing adverse events, and uncovering unreported signals or weak events that might otherwise go unnoticed. Traditional classification models (e.g., SVMs and Random Forest) have demonstrated strong performance on structured data, while more advanced approaches, such as Bayesian models [20], enable enhanced monitoring of data quality and risk signals. Large language models, like Swe-De-Clin-BERT [22], offer further improvements in capturing semantic complexity in clinical text and represent a promising standard for future NLP applications.

Despite these advances, the reviewed studies indicate several challenges that must be addressed before these tools can be widely applied in real-world healthcare settings. Limitations include restricted generalizability—most models are trained and tested on single-hospital or single-country datasets—data imbalance, interpretability issues for complex "black-box" algorithms, and variable quality of clinical documentation [32].

Looking forward, ensuring the transferability of AI tools will require large-scale, multicenter validation and the integration of multimodal data, including text, diagnostic imaging, numerical data, and sensor signals, to provide a comprehensive representation of patient conditions. The development of explainable AI solutions will be essential to build clinicians' trust in automated systems [33], and high-quality annotated datasets will remain critical for training deep learning models and LLMs at scale.

While fully proactive, workflow-integrated prevention systems that reliably avert harm are still largely aspirational, AI is already delivering measurable benefits in automated detection and enhanced surveillance. Embedding these tools into clinical practice through predictive dashboards and early alert systems provides a foundation for future proactive interventions, ultimately supporting healthcare professionals in improving patient safety.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/safety12020052/s1, Table S1: PRISMA Checklist.

Author Contributions: Conceptualization, F.D.M. and G.D.P.; methodology, F.G.; data curation, F.G.; writing—original draft preparation, R.S.; writing—review and editing, G.D.P. and G.S.; visualization, F.D.M.; supervision, V.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

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

References

1. Schwendimann, R.; Blatter, C.; Dhaini, S.; Simon, M.; Ausserhofer, D. The occurrence, types, consequences and preventability of in-hospital adverse events—A scoping review. *BMC Health Serv. Res.* **2018**, *18*, 521. [CrossRef] [PubMed] [PubMed Central]
2. Kohn, L.T.; Corrigan, J.M.; Donaldson, M.S. (Eds.) *To Err is Human: Building a Safer Health System*; National Academy Press: Washington, DC, USA, 1999.
3. Anderson, J.G.; Abrahamson, K. Your Health Care May Kill You: Medical Errors. *Stud. Health Technol. Inform.* **2017**, *234*, 13–17. [CrossRef] [PubMed]
4. Anderson, J.G. Regional patient safety initiatives: The missing element of organizational change. *AMIA Annu. Symp. Proc.* **2006**, *2006*, 1163–1164.
5. Grober, E.D.; Bohnen, J.M. Defining medical error. *Can. J. Surg.* **2005**, *48*, 39–44. [PubMed] [PubMed Central]
6. Ellahham, S. The Domino Effect of Medical Errors. *Am. J. Med. Qual.* **2019**, *34*, 412–413. [CrossRef] [PubMed]
7. Carver, N.; Gupta, V.; Hipskind, J.E. Medical Errors. 7 May 2023. In *StatPearls [Internet]*; StatPearls Publishing: Treasure Island, FL, USA, 2024. [PubMed]
8. Bender, J.A.; Kulju, S.; Soncrant, C. Combined Proactive Risk Assessment: Unifying Proactive and Reactive Risk Assessment Techniques In Health Care. *Jt. Comm. J. Qual. Patient Saf.* **2022**, *48*, 326–334. [CrossRef] [PubMed]
9. Bahl, M.; Barzilay, R.; Yedidia, A.B.; Locascio, N.J.; Yu, L.; Lehman, C.D. High-risk breast lesions: A machine learning model to predict pathologic upgrade and reduce unnecessary surgical excision. *Radiology* **2018**, *286*, 810–818. [CrossRef]
10. Peerally, M.F.; Carr, S.; Waring, J.; Martin, G.; Dixon-Woods, M. Risk Controls Identified in Action Plans Following Serious Incident Investigations in Secondary Care: A Qualitative Study. *J. Patient Saf.* **2024**, *20*, 440–447. [CrossRef] [PubMed]
11. McCarthy, J.; Hayes, P. Some philosophical problems from the standpoint of artificial intelligence. In *Machine Intelligence 4*; Meltzer, B., Michie, D., Eds.; Edinburgh University Press: Edinburgh, UK, 1969; pp. 463–502.
12. Macrae, C. Governing the safety of artificial intelligence in healthcare. *BMJ Qual. Saf.* **2019**, *28*, 495–498. [CrossRef] [PubMed]
13. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef]
14. Joint Commission International. *Accreditation Standards for Hospitals*, 7th ed.; Joint Commission Resources, Inc.: Oak Brook, IL, USA, 2021.
15. Ong, M.-S.; Magrabi, F.; Coiera, E. Automated Identification of Extreme-Risk Events in Clinical Incident Reports. *J. Am. Med. Inform. Assoc.* **2012**, *19*, e110–e118. [CrossRef]
16. Gerdes, L.U.; Hardahl, C. *Text Mining Electronic Health Records to Identify Hospital Adverse Events*; Studies in Health Technology and Informatics; IOS Press: Amsterdam, The Netherlands, 2013; Volume 192, p. 1145. [CrossRef]
17. Fujita, K.; Akiyama, M.; Park, K.; Yamaguchi, E.; Furukawa, H. Linguistic Analysis of Large-Scale Medical Incident Reports for Patient Safety. In *Quality of Life Through Quality of Information: Proceedings of MIE2012*; Studies in Health Technology and Informatics; IOS Press: Amsterdam, The Netherlands, 2012. [CrossRef]
18. Gupta, J.; Koprinska, I.; Patrick, J. Automated Classification of Clinical Incident Types. In *Driving Reform: Digital Health is Everyone's Business*; Studies in Health Technology and Informatics; IOS Press: Amsterdam, The Netherlands, 2015; Volume 214, pp. 87–93. [CrossRef]
19. Gupta, J.; Patrick, J.; Poon, S. Clinical Safety Incident Taxonomy Performance on C4.5 Decision Tree and Random Forest. In *Digital Health: Changing the Way Healthcare is Conceptualised and Delivered*; Studies in Health Technology and Informatics; IOS Press: Amsterdam, The Netherlands, 2019; Volume 266, pp. 83–88. [CrossRef]

20. Barmaz, Y.; Ménard, T. Bayesian Modeling for the Detection of Adverse Events Underreporting in Clinical Trials. *Drug Saf.* **2021**, *44*, 949–955. [CrossRef] [PubMed]
21. Ménard, T.; Barmaz, Y.; Koneswarakantha, B.; Bowling, R.; Popko, L. Enabling Data-Driven Clinical Quality Assurance: Predicting Adverse Event Reporting in Clinical Trials Using Machine Learning. *Drug Saf.* **2019**, *42*, 1045–1053. [CrossRef] [PubMed]
22. Kopacheva, E.; Lincke, A.; Henriksson, A.; Dalianis, H.; Hammar, T. Identifying Adverse Drug Events in Clinical Text Using Fine-Tuned Clinical Language Models: Machine Learning Study. *JMIR Form. Res.* **2025**, *9*, e71949. [CrossRef] [PubMed]
23. Okamoto, K.; Yamamoto, T.; Hiragi, S.; Ohtera, S.; Sugiyama, O.; Yamamoto, G.; Hirose, M.; Kuroda, T. Detecting severe incidents from electronic medical records using machine learning methods. In *Digital Personalized Health and Medicine*; Pape-Haugaard, L.B., Lovis, C., Cort, M., Eds.; IOS Press: Amsterdam, The Netherlands, 2020; pp. 1221–1222. [CrossRef]
24. Fujita, K.; Akiyama, M.; Toyama, N.; Kamemori, Y. Detecting effective classes of medical incident reports based on linguistic analysis for common reporting system in Japan. In *MEDINFO 2013*; Lehmann, C.U., Ed.; IOS Press: Amsterdam, The Netherlands, 2013; pp. 137–141. [CrossRef]
25. Liu, J.; Wong, Z.S.Y.; Tsui, K.L.; So, H.Y.; Kwok, A. *Exploring Hidden In-Hospital Fall Clusters from Incident Reports Using Text Analytics*; Studies in Health Technology and Informatics; IOS Press: Amsterdam, The Netherlands, 2019; Volume 264, pp. 1526–1527. [CrossRef]
26. Hendrickx, I.; Voets, T.; van Dyk, P.; Kool, R.B. Using Text Mining Techniques to Identify Health Care Providers With Patient Safety Problems: Exploratory Study. *J. Med. Internet Res.* **2021**, *23*, e19064. [CrossRef]
27. Fong, A.; Harriott, N.; Walters, D.M.; Foley, H.; Morrissey, R.; Ratwani, R.R. Integrating Natural Language Processing Expertise with Patient Safety Event Review Committees to Improve the Analysis of Medication Events. *Int. J. Med. Inform.* **2017**, *104*, 120–125. [CrossRef]
28. Benin, A.L.; Fodeh, S.J.; Lee, K.; Koss, M.; Miller, P.; Brandt, C. Electronic approaches to making sense of the text in the adverse event reporting system. *J. Healthc. Risk Manag.* **2016**, *36*, 10–20. [CrossRef]
29. Fodeh, S.J.; Miller, P.; Brandt, C.; Benin, A.L.; Lee, K.; Koss, M. Laplacian SVM based feature selection improves medical event reports classification. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*; IEEE: Washington, DC, USA, 2015; pp. 449–454. [CrossRef]
30. Hoekstra, O.; Hurst, W.; Tummers, J. Healthcare related event prediction from textual data with machine learning: A Systematic Literature Review. *Healthc. Anal.* **2022**, *2*, 100107. [CrossRef]
31. Deimazar, G.; Sheikhtaheri, A. Machine learning models to detect and predict patient safety events using electronic health records: A systematic review. *Int. J. Med. Inform.* **2023**, *180*, 105246. [CrossRef] [PubMed]
32. Di Palma, G.; Scendoni, R.; Tambone, V.; Alloni, R.; De Micco, F. Integrating enterprise risk management to address AI-related risks in healthcare: Strategies for effective risk mitigation and implementation. *J. Healthc. Risk Manag.* **2025**, *44*, 25–33. [CrossRef] [PubMed] [PubMed Central]
33. De Micco, F.; Di Palma, G.; Ferorelli, D.; De Benedictis, A.; Tomassini, L.; Tambone, V.; Cingolani, M.; Scendoni, R. Artificial intelligence in healthcare: Transforming patient safety with intelligent systems—A systematic review. *Front. Med.* **2025**, *11*, 1522554. [CrossRef] [PubMed] [PubMed Central]