Milky Way Scientific Journal (MWSJ)
Doi: https://doi.org/10.65021/mwsj.v2.i1.35
ISSN: 3069-6399
www.mwsJournal.com

Foundations of Machine Learning and Its Applications in Healthcare: A Review of Methods, Clinical Impact, and Open Challenges

Abdulazeez Mousa

Affiliation: Computer Department, Nawroz University, Duhok, Iraq

Corresponding Author: abdulazizmoosa93@gmail.com

ABSTRACT

Healthcare systems generate large volumes of imaging, laboratory, genomic, medication, sensor, and narrative clinical data. Machine learning provides methods for learning patterns from these data and converting them into predictions that may support diagnosis, prognosis, workflow prioritization, and biomedical discovery. This narrative review explains the principles of machine learning for healthcare readers and connects them to major clinical applications. It begins with an introduction to supervised, unsupervised, semi-supervised, and reinforcement learning, followed by common algorithms, deep-learning architectures, model training, generalization, evaluation metrics, and calibration. The review then examines medical imaging, electronic health records analytics, risk prediction, decision support, genomics, drug discovery, drug-drug interaction prediction, and clinical natural language processing. Evidence from landmark studies is discussed alongside recent guidance on external validation, reporting, foundation models, federated learning, explainability, and implementation. The review concludes that machine learning has achieved clinically meaningful performance in selected tasks, but safe adoption requires prospective testing, fairness assessment, privacy protection, workflow integration, and human oversight.

Keywords: clinical decision support; deep learning; electronic health records; healthcare; machine learning

1. INTRODUCTION

Machine learning is the study of algorithms that improve their performance on a task by using data rather than by following hand-written rules. In healthcare, this means that a model can learn statistical relationships between clinical inputs, such as images, laboratory values, medication histories, clinical notes, or genomic markers, and clinically meaningful outputs, such as diagnosis, prognosis, treatment response, or adverse-event risk. This basic definition separates machine learning from ordinary software: instead of being fully programmed by explicit rules, the system is trained from examples and then evaluated on new cases.

Medicine generates data at a scale and density that no individual clinician can fully process. A single hospital admission can produce laboratory panels, imaging studies, continuous vital-sign streams, medication records, and free-text notes. Machine learning is attractive in this setting because it offers a way to learn regularities directly from these data and turn them into predictions that support, rather than replace, clinical judgment. Over the past decade, the field has matured from proof-of-concept demonstrations to systems validated on large patient populations and, in some cases, cleared by regulators for use in care.

Milky Way Scientific Journal (MWSJ)
Doi: https://doi.org/10.65021/mwsj.v2.i1.35
ISSN: 3069-6399
www.mwsJournal.com

The momentum is real, but so is the risk of overstatement. A model that achieves high accuracy on a retrospective dataset from one hospital may degrade when applied to a different population, scanner, workflow, or documentation practice. Predictions that are accurate on average may still be systematically wrong for under-represented groups. A model that cannot explain its reasoning may be difficult for clinicians to trust and difficult for regulators to evaluate. Any honest review of machine learning in healthcare must therefore hold two ideas together: the methods are genuinely powerful, and the path from validation to safe, equitable, routine clinical use remains demanding.

This paper is written for readers who want a coherent picture of both halves of that story. Unlike highly specialized surveys that focus only on one modality or one algorithmic family, this review links foundational concepts to practical clinical domains. It builds a foundation first by explaining learning paradigms, workhorse algorithms, and principles of generalization, and then uses that foundation to interpret evidence from medical imaging, electronic health records, risk prediction, genomics, drug discovery, drug-drug interaction modeling, and clinical natural language processing. The treatment of drug-drug interaction prediction is informed by the broader AI literature and by Marqas et al. [24], but it is presented here as one example within a wider healthcare machine-learning landscape rather than as the central basis of the review.

The intended contribution is pedagogical and integrative. The objectives are: (1) to give a non-specialist clinical or scientific reader an accurate mental model of how machine learning works; (2) to map that model onto application areas where machine learning has produced clinically meaningful evidence; (3) to distinguish retrospective model performance from deployable clinical value; and (4) to state plainly, with reference to primary literature and recent reporting guidance, where the limitations and unresolved problems lie.

2. METHODS

This article is a narrative review rather than a systematic review or meta-analysis. Its purpose is to synthesize foundational machine-learning concepts and representative healthcare applications for a broad clinical and scientific readership. The review therefore prioritizes conceptual clarity, landmark studies, and recent methodological guidance rather than exhaustive enumeration of every published model.

Relevant literature was identified through targeted searches of PubMed, Google Scholar, IEEE Xplore, ACM Digital Library, ScienceDirect, Nature Portfolio journals, BMJ, and major biomedical informatics sources. Search terms included combinations of "machine learning in healthcare," "deep learning medical imaging," "electronic health records prediction," "clinical decision support," "foundation models medicine," "federated learning healthcare," "explainable AI medicine," "drug discovery machine learning," and "clinical natural language processing." Priority was given to peer-reviewed landmark studies, review articles, reporting guidelines, validation frameworks, and recent publications from 2020-2026 in rapidly changing areas.

Studies were included when they were peer-reviewed articles, influential conference papers, major methodological statements, or highly cited clinical AI studies with direct relevance to healthcare machine learning. Sources were excluded when they were not related to healthcare applications, lacked sufficient methodological detail, duplicated findings already covered by more

authoritative sources, or presented commercial claims without peer-reviewed evidence. Because this is a narrative review, no pooled effect sizes were calculated.

3. PRINCIPLES OF MACHINE LEARNING

Machine learning in practice begins with a task, data, feedback, and an evaluation criterion. Instead of a programmer specifying exactly how to distinguish a malignant lesion from a benign one, a learning algorithm is shown examples and adjusts its internal parameters until its predictions match the labels in those examples well enough to be useful on data it has not seen. The discipline is conventionally divided by the kind of feedback the algorithm receives.

3.1 Learning Paradigms

Supervised learning is the most common paradigm in clinical applications. Each training example is paired with a known outcome, such as a diagnosis, risk label, survival time, or measured value. Classification predicts a category, while regression predicts a continuous quantity. Most diagnostic and prognostic models are supervised classifiers or regressors.

Unsupervised learning works with data that have no outcome labels. The goal is to discover structure: clustering groups similar patients or images, and dimensionality reduction compresses many correlated measurements into a few informative ones. This is useful for exploration, phenotyping, and visualization.

Semi-supervised learning sits between the two and is practically important in medicine, where unlabeled data are abundant but expert-labeled data are scarce and expensive. It uses a small labeled set together with a large unlabeled set to learn better representations than either alone would allow.

Reinforcement learning trains an agent to make a sequence of decisions by rewarding desirable outcomes. It is less common in routine clinical use but has been explored for treatment-policy optimization, including sequential fluid and vasopressor management in intensive care [18].

3.2 Classical Algorithms

A small set of classical algorithms still accounts for a large fraction of deployed clinical models, partly because they are interpretable, computationally efficient, and robust on structured tabular data. Logistic regression remains a standard baseline for binary clinical prediction and produces probabilities that are easy to calibrate. Decision trees split the feature space with simple rules and are valued for transparency. Random forests average many de-correlated trees built on bootstrap samples and random feature subsets, reducing variance and providing an internal out-of-bag error estimate [3]. Support vector machines find a maximum-margin boundary and, through kernels, handle non-linear problems in high-dimensional spaces [8]. Gradient-boosted trees build an additive ensemble in which each new tree corrects residual errors; XGBoost is widely used for structured clinical and tabular data because it combines high predictive accuracy with regularization and efficient handling of sparse inputs [5].

Milky Way Scientific Journal (MWSJ)
Doi: https://doi.org/10.65021/mwsj.v2.i1.35
ISSN: 3069-6399
www.mwsJournal.com

A critical point for healthcare is that simpler models may outperform deep learning when the dataset is modest, tabular, missing in systematic ways, or collected from a narrow clinical environment. Deep learning is not automatically superior; the appropriate method depends on the data type, sample size, interpretability requirement, deployment environment, and clinical cost of error.

Table 1: Common machine learning algorithm families and their typical clinical use.

Algorithm family	Learning type	Typical clinical use	Key strength
Logistic regression	Supervised	Binary risk scoring	Interpretable, calibrated probabilities
Decision tree	Supervised	Triage rules	Human-readable decision path
Random forest	Supervised	Tabular diagnosis/prognosis	Low variance, robust to noise
Support vector machine	Supervised	Tissue/sample classification	Effective in high dimensions
Gradient boosting	Supervised	Structured EHR prediction	Strong accuracy on tabular data
k-means / hierarchical	Unsupervised	Patient phenotyping	Discovers structure without labels
PCA	Unsupervised	Feature compression	Reduces dimensionality and aids visualization

3.3 Training, Generalization, and Evaluation

Whatever the algorithm, training is an optimization problem. A loss function measures how far the model's predictions are from the truth, and an optimizer, usually a variant of gradient descent for neural networks, adjusts parameters to reduce that loss. The central difficulty is generalization: a model that memorizes training data performs poorly on new patients, while a model that is too simple misses real signal. This is the bias-variance trade-off. Regularization techniques penalize unnecessary complexity, and cross-validation provides a more honest estimate of out-of-sample performance than a single split.

Evaluation deserves particular care in medicine because errors are not symmetric. Accuracy alone is misleading when disease prevalence is low. Sensitivity measures the fraction of true cases the model catches; specificity measures how well it avoids false alarms; precision measures how trustworthy a positive prediction is. ROC-AUC summarizes discrimination across thresholds, but precision-recall curves are often more informative for rare outcomes. Calibration, the agreement between predicted probabilities and observed frequencies, is often as important as discrimination because a poorly calibrated risk score can mislead decisions even if it ranks patients correctly.

Clinical usefulness also depends on whether predictions change decisions in ways that improve patient outcomes. Decision-curve analysis is valuable because it evaluates net benefit across clinically relevant risk thresholds. Reporting frameworks such as TRIPOD+AI and bias-assessment tools such as PROBAST+AI increasingly emphasize transparent reporting, external validation, and risk-of-bias assessment for prediction models that use regression or machine learning [7,26].

4. DEEP LEARNING FOUNDATIONS

Deep learning is the branch of machine learning that uses neural networks with many layers to learn hierarchical representations directly from raw or lightly processed data. A neuron computes a weighted sum of inputs followed by a non-linear activation; stacking many such units allows a network to compose simple features into increasingly abstract ones [20]. This representational power enabled many imaging and language breakthroughs, but it also increases data demand, opacity, and the risk of overfitting.

Convolutional neural networks exploit the spatial structure of images by learning local filters that are reused across the image. Residual connections allowed CNNs to be trained at great depth and produced a step change in image-recognition accuracy [14],

building on the earlier demonstration that GPU-trained CNNs could dominate large-scale image classification [19]. Recurrent neural networks and long short-term memory cells process sequential data by maintaining memory across time steps [15]. Transformers replace recurrence with self-attention, letting every element of a sequence attend to every other element [35], and now underpin most modern language systems and many vision systems.

Encoder-decoder architectures such as U-Net add skip connections between contracting and expanding paths and remain important for biomedical image segmentation [31]. In current healthcare AI, architecture choice is increasingly tied to modality: CNNs and vision transformers dominate imaging, gradient-boosted trees and temporal neural networks are common in structured EHR prediction, graph neural networks are important in molecular and interaction modeling, and transformers dominate clinical and biomedical text processing.

4.1 Transfer Learning, Foundation Models, and Explainability

Two practical ideas make deep learning feasible in data-scarce clinical settings. Transfer learning reuses a network pretrained on a large general dataset and fine-tunes it on a smaller medical dataset. Foundation models extend this idea to very large models pretrained on broad corpora or multi-modal datasets and then adapted to many downstream tasks. In healthcare, foundation models are being investigated for radiology reporting, clinical summarization, triage support, biomedical question answering, and multi-modal reasoning over text, images, and structured data.

However, foundation models introduce risks that classical prediction models do not fully capture. Large language models may produce fluent but unsupported statements, omit clinically important information, or hallucinate citations and facts. Recent clinical safety work has therefore emphasized hallucination measurement and domain-specific safety evaluation for medical summarization and decision support [2,12]. These models also raise governance concerns because their training data, update cycles, and failure modes are often difficult to inspect.

Explainable artificial intelligence aims to make predictions inspectable. LIME approximates a complex model locally with a simple interpretable one [29], while Shapley-value methods estimate how input features contribute to predictions. A careful distinction is needed between intrinsic interpretability, where the model structure itself is understandable, and post hoc explainability, where an explanation is generated after prediction. Explainability is not a luxury in medicine; it is often a precondition for clinical trust, error detection, and regulatory evaluation.

5. MACHINE LEARNING IN MEDICAL IMAGING

Medical imaging is where deep learning has produced its most visible clinical results. Images are high-dimensional, spatially structured data that CNNs and vision transformers can exploit, and curated labeled datasets became available earlier than in many other clinical domains. A broad survey found deep learning applied across modalities and organ systems, with classification and segmentation as dominant tasks [22].

In dermatology, a CNN trained on roughly 129,000 clinical images spanning more than 2,000 disease classes matched the performance of 21 board-certified dermatologists on biopsy-confirmed melanoma and carcinoma classification [10]. In

Milky Way Scientific Journal (MWSJ)
Doi: https://doi.org/10.65021/mwsj.v2.i1.35
ISSN: 3069-6399
www.mwsJournal.com

ophthalmology, a deep network trained on more than 128,000 retinal fundus photographs detected referable diabetic retinopathy with ROC-AUC close to .99 on two validation sets [11], and a transfer-learning system demonstrated accurate diagnosis from retinal OCT and pediatric chest radiographs [17]. In breast-cancer screening, an AI system evaluated on large UK and US datasets reduced false positives and false negatives relative to radiologists and outperformed six readers on the US test set [25]. In thoracic oncology, an end-to-end three-dimensional model for low-dose chest CT achieved an AUC of about .94 for lung-cancer detection [1].

These results should be interpreted carefully. Comparisons between AI systems and clinicians are often sensitive to case mix, reader experience, prevalence, access to clinical context, and whether the study uses retrospective enriched datasets or real screening populations. An algorithm may perform well under controlled validation yet fail when image quality, scanner type, referral patterns, or disease prevalence change. Therefore, the next level of evidence is not simply higher AUC but external validation, prospective testing, workflow integration, and measurement of patient outcomes.

Table 2: Representative landmark imaging studies and reported performance.

Domain	Task	Reported result	Study
Dermatology	Skin cancer classification	Dermatologist-level on biopsy-proven cases	Esteva et al. [10]
Ophthalmology	Diabetic retinopathy detection	ROC-AUC close to .99 on two validation sets	Gulshan et al. [11]
Radiology	Breast cancer screening	Fewer false positives and false negatives than radiologists	McKinney et al. [25]
Radiology	Lung cancer on low-dose CT	ROC-AUC about .94 using end-to-end 3D model	Ardila et al. [1]
Multi-modal	OCT / chest X-ray diagnosis	Transfer-learning, expert-comparable performance	Kermany et al. [17]

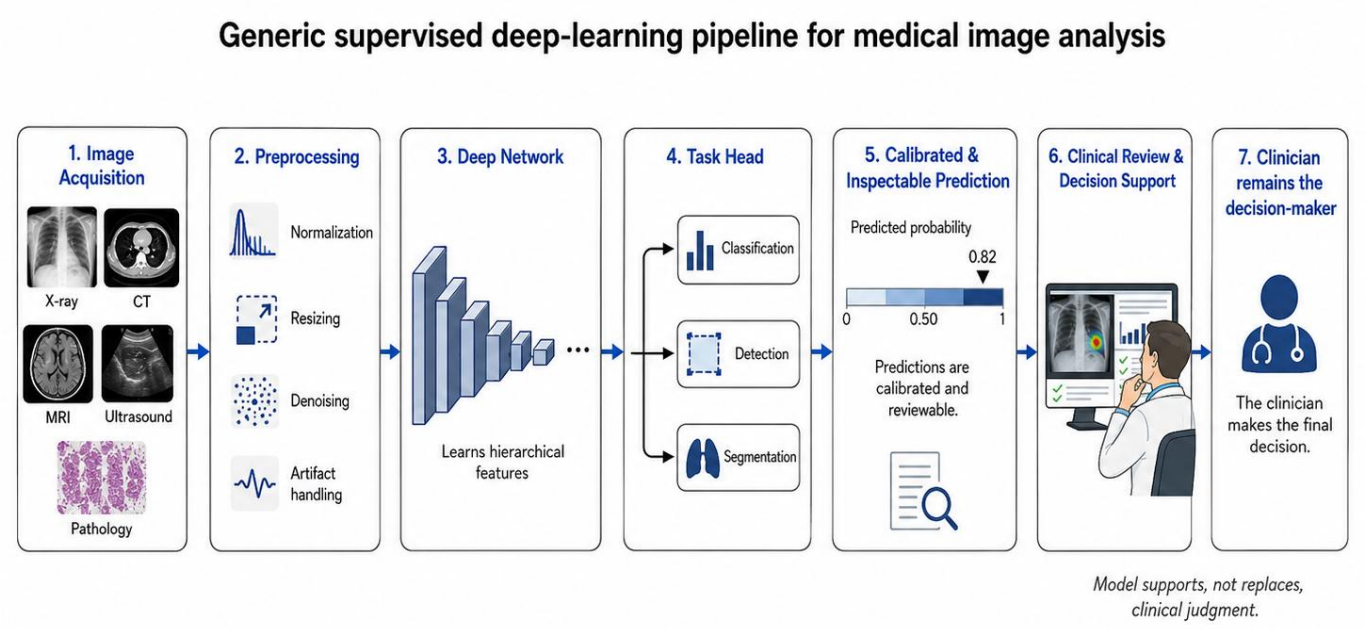

Figure 1: Generic supervised deep-learning pipeline for medical image analysis.

A typical imaging pipeline proceeds from acquisition and preprocessing through a deep network that extracts hierarchical features, a task head that performs classification, detection, or segmentation, and a calibration and clinical-review stage. The clinician remains the decision-maker; the model supplies a calibrated and inspectable prediction.

6. ELECTRONIC HEALTH RECORDS ANALYTICS, RISK PREDICTION, AND DECISION SUPPORT

Electronic health records contain a rich longitudinal picture of each patient, including diagnoses, laboratory values, procedures, medications, flowsheets, clinical notes, imaging reports, and outcomes. Rajkomar et al. [28] showed that deep models trained directly on raw, standardized EHR data from more than 200,000 hospitalized patients could predict in-hospital mortality, unplanned readmission, prolonged length of stay, and discharge diagnoses without the laborious manual feature engineering required by earlier systems. Surveys of the field document a rapid shift from hand-crafted clinical features toward representation learning over heterogeneous record data [32].

EHR analytics differs from imaging because the data are irregular, heterogeneous, and deeply shaped by clinical workflow. A missing laboratory result may mean that a test was unnecessary, unavailable, delayed, or performed elsewhere. Diagnosis codes may reflect billing incentives as much as biological truth. Medication orders may represent intention rather than administration. Clinical notes contain valuable information but also copy-forward artifacts, abbreviations, and institutional language habits. These issues mean that preprocessing, temporal alignment, missingness handling, and outcome definition are not technical details; they determine whether the model learns disease biology, documentation behavior, or hidden shortcuts.

Risk prediction for acute deterioration is especially active. A deep neural network trained on tens of thousands of single-lead electrocardiograms classified a broad range of rhythm abnormalities at a level comparable to cardiologists [13]. For sepsis, reinforcement learning has been explored to learn treatment policies from intensive-care databases, including fluid and vasopressor management [18]. These studies are influential, but they also illustrate the central caveat of clinical prediction: retrospective performance is not the same as prospective clinical benefit. Label leakage, temporal bias, confounding by indication, treatment-selection bias, and changing clinical practice can inflate apparent performance if not carefully controlled.

Decision support systems must also be evaluated as human-machine interventions. A risk score may be accurate but clinically useless if it arrives too late, produces excessive alerts, duplicates existing judgment, or increases workload. Conversely, a modestly accurate model may be valuable if it identifies a high-risk subgroup earlier than usual care and is linked to an actionable pathway. Prospective testing, silent-mode evaluation, randomized or stepped-wedge deployment studies, and post-deployment monitoring are therefore essential. Reporting frameworks such as TRIPOD+AI and trial-reporting guidance such as CONSORT-AI and SPIRIT-AI help move the field from model publication toward accountable clinical implementation [7,23].

7. GENOMICS, DRUG DISCOVERY, AND CLINICAL NLP

7.1 Genomics and Molecular Prediction

Machine learning has reshaped parts of biomedical science upstream of the clinic. Genomic and multi-omics datasets contain thousands to millions of variables, often measured on relatively small patient cohorts. This makes representation learning,

regularization, dimensionality reduction, and careful validation central. In structural biology, a deep-learning system predicted protein three-dimensional structure with near-experimental accuracy at CASP14, addressing a decades-old problem with direct implications for drug target identification [16].

The clinical promise of genomic machine learning lies in linking molecular variation to diagnosis, prognosis, treatment response, and disease mechanism. However, the same high dimensionality that makes these data rich also makes them vulnerable to overfitting and cohort-specific artifacts. Models trained on one ancestry group, sequencing platform, or hospital biobank may not generalize to another. Therefore, multi-cohort validation and explicit fairness assessment are as important in genomics as they are in EHR analytics.

7.2 Drug Discovery and Drug-Drug Interaction Prediction

In antibiotic discovery, a graph neural network trained to predict growth inhibition identified a structurally novel compound with broad-spectrum activity from a chemical library, demonstrating that learned molecular representations can surface candidates that conventional screening might miss [33]. In drug discovery more broadly, machine learning is used for virtual screening, property prediction, toxicity estimation, target identification, binding-affinity prediction, de novo molecule generation, and clinical-trial optimization.

Drug-drug interaction prediction is closely related and has immediate clinical relevance because polypharmacy is common and adverse interactions are preventable sources of harm. Recent work synthesizes approaches spanning classical classifiers, deep networks, graph neural networks, knowledge graphs, transformer architectures, and biologically inspired methods [24]. The strongest future direction is multi-modal modeling that integrates chemical structure, pharmacokinetics, pharmacodynamics, transcriptomic response, clinical co-prescription evidence, and known interaction mechanisms. The main challenge is that interaction severity is not simply a binary label; it depends on dose, route, timing, comorbidity, organ function, and clinical context.

7.3 Clinical and Biomedical Natural Language Processing

Clinical and biomedical NLP converts unstructured text into usable information. Domain-adapted language models such as BioBERT, pretrained on biomedical corpora, substantially improve named-entity recognition, relation extraction, and question answering over general-domain baselines [21], building on bidirectional transformer pretraining introduced by Devlin et al. [9]. In clinical settings, NLP can extract diagnoses, medications, symptoms, adverse events, social determinants of health, and trial-eligibility criteria from notes and reports.

The rise of large language models has expanded the scope of clinical NLP from extraction to generation, summarization, drafting, triage support, and conversational assistance. These applications are promising but require stricter safeguards than ordinary text mining because generated outputs may be fluent, persuasive, and wrong. Recent evidence shows that LLMs can fail to follow diagnostic or treatment guidelines and may not be safe for autonomous decision-making [12]. For this reason, clinical LLMs should be positioned as supervised assistants with audit trails, source grounding, uncertainty communication, and clear accountability rather than independent decision-makers.

Milky Way Scientific Journal (MWSJ)
Doi: https://doi.org/10.65021/mwsj.v2.i1.35
ISSN: 3069-6399
www.mwsJournal.com

8. CHALLENGES, IMPLEMENTATION, AND FUTURE DIRECTIONS

The gap between a strong validation result and safe routine deployment is where most of the field's unresolved problems live. Several challenges recur across applications and should be treated as design requirements rather than afterthoughts.

Data quality and labeling. Models inherit the limitations of their training data. Labels derived from billing codes, clinician behavior, or imperfect proxies introduce systematic error. Expert labels are expensive, and disagreement between experts may reflect genuine clinical ambiguity rather than noise.

- Distribution shift and external validity. A model trained at one institution often degrades elsewhere because of differences in equipment, populations, documentation practices, disease prevalence, and care pathways. External validation on independent cohorts is essential but remains under-reported [4].
- Bias and fairness. Models can encode and amplify existing disparities, performing worse for under-represented groups or recommending different levels of care for clinically similar patients. Bias should be measured through subgroup performance, calibration within groups, error-rate parity, decision-curve differences, and outcome monitoring after deployment. Mitigation may require better sampling, reweighting, fairness-aware objectives, subgroup calibration, model updating, and governance involving affected communities.
- Interpretability and explainability. Intrinsically interpretable models can be easier to audit, but they may not capture complex patterns. Post hoc explanation tools can help but may be unstable or misleading. In clinical use, explanations should be evaluated according to whether they help clinicians detect error, calibrate trust, and make better decisions.
- Privacy, security, and regulation. Clinical data are sensitive and legally protected. Centralized training can conflict with privacy obligations, and adaptive models challenge traditional regulatory pathways. Governance must address consent, data minimization, cybersecurity, auditability, responsibility for error, and post-market monitoring.
- Prospective evidence and workflow integration. Many AI studies are retrospective. Real-world value requires prospective testing, assessment of alert fatigue, clinician behavior, patient outcomes, health economics, and unintended consequences such as automation bias or deskilling.

8.1 Federated Learning and Privacy-Preserving AI

Federated learning trains a shared model across institutions without moving raw patient data off-site, directly addressing one major privacy barrier while enabling larger and more diverse effective datasets [30]. In principle, it can support multi-center learning where legal, ethical, or logistical constraints prevent data pooling. Recent reviews emphasize that federated learning can improve privacy and collaboration but adds complexity in distributed optimization, aggregation, governance, and infrastructure [37].

Federated learning is not a complete privacy solution. Model updates can still leak information unless combined with techniques such as differential privacy, secure aggregation, or encryption. Client heterogeneity, non-independent and non-identically distributed data, communication overhead, site dropout, and inconsistent data standards can reduce performance. Therefore, federated learning should be understood as a promising governance and engineering strategy, not as a guarantee of safe or unbiased clinical AI.

8.2 Research Priorities

- Move from retrospective benchmarking to prospective and pragmatic clinical evaluation, including silent trials, randomized evaluations, and post-deployment monitoring.
- Develop standards for external validation, calibration, and subgroup fairness that are routinely reported rather than treated as optional analyses.
- Strengthen multimodal models that integrate imaging, EHR, genomics, molecular data, and clinical text while preserving interpretability and data governance.
- Create safer foundation-model workflows with retrieval grounding, structured outputs, hallucination testing, audit trails, and explicit human supervision.
- Improve privacy-preserving collaboration through federated learning, secure computation, common data models, and shared evaluation protocols.
- Align technical performance with clinical utility by measuring whether models improve decisions, outcomes, equity, and cost-effectiveness in real care settings.

9. CONCLUSIONS

Machine learning has crossed from promise to clinical relevance in several domains, most clearly in medical imaging and increasingly in EHR-based risk prediction, genomics, drug discovery, drug-drug interaction prediction, and clinical language processing. The methods, from regularized linear models to deep transformers, rest on a coherent set of principles: learn from data, control complexity to generalize, and evaluate with metrics that reflect clinical cost. Understanding those principles allows readers to judge new results rather than simply trust headline performance.

At the same time, strong retrospective performance is necessary but not sufficient. Distribution shift, bias, opacity, privacy, label quality, missingness, temporal leakage, and limited prospective evidence remain major constraints on safe deployment. The most productive posture is therefore disciplined optimism: treat machine learning as a powerful instrument that augments clinical reasoning, hold it to the same standards of validation and equity as any other medical technology, and develop privacy-preserving and explainable methods that make trustworthy implementation possible.

The next phase of healthcare machine learning should be judged less by isolated leaderboard performance and more by reproducible external validation, transparent reporting, prospective clinical impact, fairness across patient groups, privacy-preserving collaboration, and carefully supervised foundation-model use.

ETHICS STATEMENT

This article is a review of previously published literature and did not involve new studies on human participants or animals by the author. No ethical approval was required.

FUNDING ACKNOWLEDGMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

DATA AVAILABILITY STATEMENT

No new data were generated or analyzed in support of this review. All sources discussed are available through their cited publishers and digital object identifiers.

AUTHOR CONTRIBUTIONS

Abdulazeez Mousa conceived the scope of the review, conducted the literature synthesis, and wrote and approved the final manuscript.

CONFLICT OF INTEREST DISCLOSURE

The author declares no conflict of interest.

REFERENCES

1. Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., Naidich, D. P., & Shetty, S. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature Medicine, 25(6), 954-961. DOI: 10.1038/s41591-019-0447-x
2. Asgari, E., Chatterjee, S., Culpin, R., et al. (2025). A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. npj Digital Medicine. DOI: 10.1038/s41746-025-01670-7
3. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. DOI: 10.1023/A:1010933404324
4. Chen, J. H., & Asch, S. M. (2017). Machine learning and prediction in medicine - beyond the peak of inflated expectations. New England Journal of Medicine, 376(26), 2507-2509. DOI: 10.1056/NEJMp1702071
5. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). ACM. DOI: 10.1145/2939672.2939785
6. Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M. M., Xie, W., Rosen, G. L., Lengerich, B. J., Israeli, J., Lanchantin, J., Woloszynek, S., Carpenter, A. E., Shrikumar, A., Xu, J., ... Greene, C. S. (2018). Opportunities and obstacles for deep learning in biology and medicine. Journal of the Royal Society Interface, 15(141), 20170387. DOI: 10.1098/rsif.2017.0387
7. Collins, G. S., Dhiman, P., Ma, J., Schlussel, M. M., Archer, L., Van Calster, B., ... Moons, K. G. M. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, 385, e078378. DOI: 10.1136/bmj-2023-078378
8. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297. DOI: 10.1007/BF00994018
9. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT 2019 (pp. 4171-4186). Association for Computational Linguistics. DOI: 10.18653/v1/N19-1423
10. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118. DOI: 10.1038/nature21056
11. Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA, 316(22), 2402-2410. DOI: 10.1001/jama.2016.17216
12. Hager, P., Jungmann, F., Holland, R., Bhagat, K., Hubrecht, I., Knauer, M., Vielhauer, J., Makowski, M., Braren, R., Kaissis, G., Rueckert, D., & Kather, J. N. (2024). Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nature Medicine. DOI: 10.1038/s41591-024-03097-1
13. Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature Medicine, 25(1), 65-69. DOI: 10.1038/s41591-018-0268-3

14. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770-778). IEEE. DOI: 10.1109/CVPR.2016.90
15. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780. DOI: 10.1162/neco.1997.9.8.1735
16. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zidek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583-589. DOI: 10.1038/s41586-021-03819-2
17. Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M. K., Pei, J., Ting, M. Y. L., Zhu, J., Li, C., Hewett, S., Dong, J., Ziyar, I., ... Zhang, K. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell, 172(5), 1122-1131.e9. DOI: 10.1016/j.cell.2018.02.010
18. Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., & Faisal, A. A. (2018). The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nature Medicine, 24(11), 1716-1720. DOI: 10.1038/s41591-018-0213-5
19. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84-90. DOI: 10.1145/3065386
20. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444. DOI: 10.1038/nature14539
21. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234-1240. DOI: 10.1093/bioinformatics/btz682
22. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sanchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60-88. DOI: 10.1016/j.media.2017.07.005
23. Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., Denniston, A. K., & SPIRIT-AI and CONSORT-AI Working Group. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. Nature Medicine, 26, 1364-1374. DOI: 10.1038/s41591-020-1034-x
24. Marqas, R. B., Simo, Z., Mousa, A., Ozyurt, F., & Iantovics, L. B. (2026). Advancing drug-drug interaction prediction with biomimetic improvements: Leveraging the latest artificial intelligence techniques to guide researchers in the field. Biomimetics, 11(1), 39. DOI: 10.3390/biomimetics11010039
25. McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., ... Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. Nature, 577(7788), 89-94. DOI: 10.1038/s41586-019-1799-6
26. Moons, K. G. M., Collins, G. S., Ma, J., Dhiman, P., Reitsma, J. B., Riley, R. D., ... Wolff, R. F. (2025). PROBAST+AI: An updated quality, risk of bias, and applicability tool for prediction models that use regression or machine learning methods. BMJ, 388, e082505. DOI: 10.1136/bmj-2024-082505
27. Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future - big data, machine learning, and clinical medicine. New England Journal of Medicine, 375(13), 1216-1219. DOI: 10.1056/NEJMp1606181
28. Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G. E., Irvine, J., Le, Q., Litsch, K., ... Dean, J. (2018). Scalable and accurate deep learning with electronic health records. npj Digital Medicine, 1, 18. DOI: 10.1038/s41746-018-0029-1
29. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144). ACM. DOI: 10.1145/2939672.2939778
30. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., & Cardoso, M. J. (2020). The future of digital health with federated learning. npj Digital Medicine, 3, 119. DOI: 10.1038/s41746-020-00323-1
31. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 (Lecture Notes in Computer Science, Vol. 9351, pp. 234-241). Springer. DOI: 10.1007/978-3-319-24574-4_28
32. Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record analysis. IEEE Journal of Biomedical and Health Informatics, 22(5), 1589-1604. DOI: 10.1109/JBHI.2017.2767063

33. Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., MacNair, C. R., French, S., Carfrae, L. A., Bloom-Ackermann, Z., Tran, V. M., Chiappino-Pepe, A., Badran, A. H., Andrews, I. W., Chory, E. J., Church, G. M., Brown, E. D., Jaakkola, T. S., Barzilay, R., & Collins, J. J. (2020). A deep learning approach to antibiotic discovery. Cell, 180(4), 688-702.e13. DOI: 10.1016/j.cell.2020.01.021
34. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56. DOI: 10.1038/s41591-018-0300-7
35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems 30 (pp. 5998-6008). Curran Associates. DOI: 10.5555/3295222.3295349
36. Yu, K.-H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. Nature Biomedical Engineering, 2(10), 719-731. DOI: 10.1038/s41551-018-0305-z
37. Yurdem, B., Kuzlu, M., Gullu, M. K., Catak, F. O., & Tabassum, M. (2024). Federated learning: Overview, strategies, applications, tools and future directions. Heliyon, 10, e38137. DOI: 10.1016/j.heliyon.2024.e38137