AI In Disease Diagnosis: A Review of Intelligent Diagnostic Workflows, Applications, And Clinical Challenges

A.M. Suryawanshi[1], Aarshan Allauddin Tamboli[2]
[1,2]*Department of Information Technology, Dattakala Group of Institute Pune, India*

***Abstract*—Artificial Intelligence (AI) is increasingly investigated as a decision-support technology for disease diagnosis. Machine Learning (ML), Deep Learning (DL), Computer Vision, and Natural Language Processing (NLP) can analyse different forms of clinical evidence, including laboratory measurements, electronic health records, medical images, and clinical notes. This paper presents a structured review based on the supplied seminar synopsis and selected literature cited in it. It describes a multi-stage diagnostic workflow, compares commonly used AI methods, and discusses applications in medical imaging, risk prediction, symptom assessment, and clinical decision support. Two conceptual diagrams and summary tables are included to make the workflow and method selection explicit. The paper is a literature-based review; it does not report a newly trained model, a clinical dataset, or original diagnostic-performance measurements. The analysis highlights that sensitivity, specificity, calibration, subgroup performance, external validation, and human oversight are essential alongside conventional accuracy. AI may improve the speed and consistency of selected diagnostic tasks, but its reliability depends on data quality, representative evaluation, privacy protection, and integration into clinical practice. Accordingly, AI should be treated as an assistive component of a clinician-led diagnostic process rather than an autonomous replacement for medical judgement.**

***Index Terms*—Artificial Intelligence, Disease Diagnosis, Machine Learning, Deep Learning, Medical Imaging, Clinical Decision Support, Explainable AI.**

I. INTRODUCTION

Disease diagnosis is a process of combining symptoms, patient history, physical findings, laboratory tests, and imaging evidence to identify a condition and guide further care. Traditional diagnostic workflows rely heavily on professional expertise and can be demanding when the volume of clinical information is large. AI offers computational methods for learning patterns from examples and supporting the review of complex data [1], [2]. ML algorithms can model relationships between structured variables such as age, glucose, blood pressure, and cholesterol. DL methods, especially Convolutional Neural Networks (CNNs), can learn image representations from radiographs, CT scans, MRI scans, and retinal photographs. NLP can extract relevant facts from clinical notes and patient narratives. When used appropriately, these methods may help prioritise cases, flag suspicious findings, or estimate disease risk [3], [4].

However, a prediction is not automatically a diagnosis. Model outputs can be affected by incomplete records, data leakage, population differences, and changes in clinical practice. A system that performs well in a retrospective dataset may fail when used in a different hospital or patient group. Responsible evaluation therefore needs independent test data, clinically meaningful metrics, transparent limitations, and professional review. This paper follows the topic and literature-survey organisation of the supplied synopsis while adopting the compact two-column, numbered-section format of the provided reference publication. Its contribution is a structured synthesis of methods and implementation considerations, not a claim of new clinical experimentation.

II. RELATED WORK

Mirbabaie, Stieglitz, and Frick classify research on AI in disease diagnostics and discuss its current state, challenges, and future directions [1]. Ahsan, Luna, and Siddique review machine-learning approaches for

disease diagnosis, including diabetes, heart disease, kidney disease, and COVID-19 [2]. These reviews indicate that diagnostic tasks differ substantially in their data sources, target outcomes, and validation needs. Aggarwal and colleagues examine diagnostic accuracy in deep-learning medical imaging and highlight the variation between studies and clinical settings [3]. CNN-based image analysis is promising for detecting patterns in X-rays, CT, MRI, and retinal images, but image-level performance alone does not establish clinical benefit. Kumar and colleagues provide a systematic literature review and future research agenda spanning multiple disease areas and evaluation approaches [4].

The supplied synopsis also cites a 2023 article titled "Evaluation of Artificial Intelligence Techniques in Disease Diagnosis and Prediction." Its full bibliographic details were not included in the synopsis, so it is treated here as a synopsis-reported source and should be verified before formal publication [5]. Across the reviewed material, recurring themes are the use of ML/DL for pattern recognition, the importance of dataset quality, and the need for clinical validation, privacy, fairness, and explainability.

TABLE I. SUMMARY OF LITERATURE REVIEW

Ref.	Study / authors	Focus and relevance
[1]	Mirbabaie et al. (2021)	Critical review of AI in disease diagnostics; classification of research themes and implementation challenges.
[2]	Ahsan, Luna & Siddique (2022)	ML-based diagnosis across diabetes, heart, kidney disease and COVID-19.
[3]	Aggarwal et al. (2021)	Systematic review of deep learning in medical imaging; variation in study methods and validation.
[4]	Kumar et al. (2023)	Systematic literature review, disease areas, evaluation metrics and future research agenda.
[5]	Synopsis-cited 2023 article	"Evaluation of AI Techniques in Disease Diagnosis and Prediction"; full bibliographic record needs verification.

III. DIAGNOSTIC WORKFLOW

Fig. 1 presents the conceptual workflow used to organise AI-assisted diagnosis. First, patient information is acquired from authorised sources. Second, data are checked for missing values, inconsistent units, image artefacts, duplicates, and possible label errors. Third, the input is represented in a format suitable for the chosen model. Structured clinical records may be processed by statistical or tree-based ML; images may be analysed by CNNs or vision transformers; and free-text notes may be processed with NLP.

The trained model produces a class, risk estimate, or region of interest. Where clinically appropriate, the output should include confidence or uncertainty information and an explanation of relevant evidence. A qualified healthcare professional then considers the output alongside the full clinical context.

The final stage is post-deployment monitoring: performance, errors, alert burden, and changes in the input population should be reviewed. Privacy and governance are not a single stage; they apply throughout collection, storage, model development, evaluation, and use.

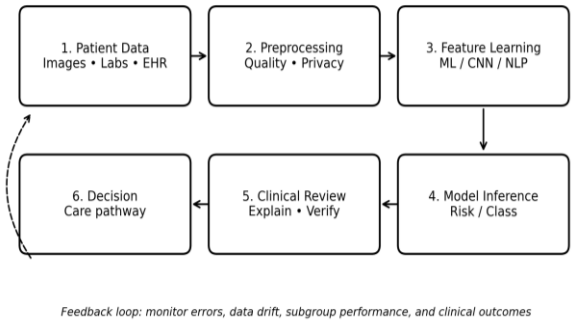


Fig. 1. Conceptual workflow for AI-assisted disease diagnosis.

IV. MULTI-MODAL AI ARCHITECTURE

Fig. 2 illustrates a conceptual architecture with image, structured clinical, and text-processing branches. Outputs may be combined only where the data and evaluation design support multimodal use.

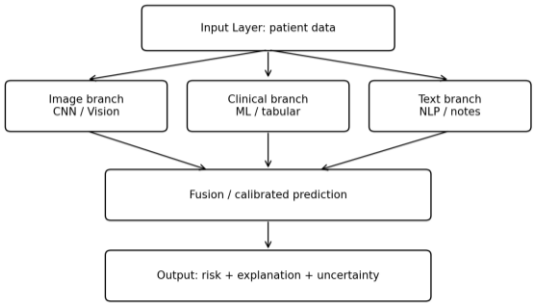


Fig. 2. Conceptual multi-modal architecture for diagnostic support.

V. AI METHODS AND APPLICATIONS

The model family should be selected according to the data type, task, available sample size, interpretability needs, and consequences of errors. Table I summarises common choices. Logistic regression and decision trees can provide useful baselines for structured data. Random Forests and gradient-boosted trees capture non-linear relationships in tabular records. CNNs are widely used for image classification and segmentation, while NLP models process text-based clinical information. No algorithm is universally best: comparisons should use the same data split, a suitable reference standard, and metrics that reflect the clinical task.

TABLE II. COMPARISON OF COMMON AI METHODS

Method	Typical input	Strength	Main caution
Logistic regression	Structured clinical variables	Simple baseline; probability interpretation is possible after calibration.	May miss complex non-linear relationships.
Decision tree / Random Forest	Tabular records	Captures non-linear interactions; tree ensembles are robust baselines.	Can inherit dataset bias; explanations may be approximate.
CNN / vision model	X-ray, CT, MRI, retinal images	Learns spatial features for classification or segmentation.	Needs representative image data and external validation.
NLP / language models	Clinical notes, reports, symptom text	Extracts concepts and contextual information from text.	May omit context or generate unsupported outputs.
Multimodal model	Images + records + text	Can combine complementary evidence.	More complex missing-data, privacy and validation requirements.

Medical imaging: AI may flag suspicious areas in radiographs, CT, MRI, ultrasound, and retinal photographs. The output can help prioritise review but must be checked by qualified clinicians. Structured-data prediction: ML can estimate risk from measurements and history, for example glucose, blood pressure, age, and cholesterol; risk scores should not be interpreted as definitive diagnoses. Symptom assessment and NLP: systems can organise patient-reported symptoms and extract findings from clinical notes. Such tools should communicate uncertainty and provide safe escalation guidance. Clinical decision support: AI can combine evidence from several sources, retrieve relevant information, or identify cases needing further investigation. Monitoring: repeated measurements from devices may reveal changes over time, but alerts require validation to avoid alarm fatigue. Table III maps application areas to inputs, typical methods, and key cautions. These are representative uses described in the literature, not a claim that every listed method is approved or suitable for every clinical setting.

TABLE III. APPLICATIONS, INPUTS AND CLINICAL CAUTIONS

Application	Inputs / method	Intended support	Caution
Image diagnosis	X-ray, CT, MRI, retinal images; CNN / vision models	Flag suspicious regions; prioritise review.	Image findings require clinician interpretation.
Risk prediction	Vitals, lab values, history; ML models	Identify people who may need further assessment.	Risk estimate is not a confirmed diagnosis.
Symptom assessment	Patient narrative; NLP / knowledge-based tools	Organise symptoms and suggest care pathways.	Must communicate uncertainty and escalation advice.
Clinical decision support	Records, reports, guidelines; ML / NLP	Summarise evidence and highlight relevant cases.	Workflow fit and alert burden must be tested.

VI. EVALUATION FRAMEWORK

Accuracy alone may be misleading when disease prevalence is low or classes are imbalanced. Sensitivity (recall) measures the proportion of affected cases detected; specificity measures the proportion of unaffected cases correctly identified. Precision measures the proportion of positive predictions that are correct. The F1-score combines precision and recall, while ROC-AUC summarises discrimination across thresholds. Calibration indicates whether predicted probabilities correspond to observed event rates. Table II summarises these measures and their interpretation. Evaluation data should be separated from training data at the patient level to prevent leakage, and external validation should test the model on data from a different site or period where feasible. Reports should include confidence intervals, missing-data handling, threshold selection, subgroup results, and failure cases. For a screening task, missing a disease may be costly; for a confirmatory pathway, false positives may lead to unnecessary tests. The threshold must therefore be selected with clinical stakeholders rather than chosen only to maximise a single metric.

TABLE IV. CLINICAL INTERPRETATION OF EVALUATION METRICS

Metric	Interpretation	Why it matters
Sensitivity / Recall	Share of actual disease cases detected.	Important when missed cases carry high risk.
Specificity	Share of non-disease cases correctly ruled out.	Helps assess unnecessary follow-up burden.
Precision	Share of positive predictions that are correct.	Important when positive alerts trigger costly or invasive tests.
F1-score	Harmonic mean of precision and recall.	Useful summary when both matter; threshold-dependent.
ROC-AUC	Discrimination across classification thresholds.	Should not replace threshold-specific clinical analysis.
Calibration	Agreement between predicted risk and observed frequency.	Important when outputs are interpreted as probabilities.

VII. CHALLENGES AND LIMITATIONS

The principal limitations are data representativeness, label quality, privacy, fairness, explainability, generalisation, and operational fit. If a dataset under-represents certain age groups or communities, a model may produce unequal error rates. Data collected at one hospital may differ from those at another because of equipment, workflow, patient mix, or coding practices. Such distribution shifts can degrade performance after deployment. Medical records also contain sensitive personal information, so access controls, data minimisation, secure storage, and governance are necessary. Explainability methods can help clinicians inspect model behaviour but do not, by themselves, prove that a model is correct. Human review must be meaningful: users need training, a way to challenge outputs, and clear responsibility for final decisions. Finally, deployment requires interoperability, workflow testing, monitoring, maintenance, and evidence that the system improves care rather than only retrospective scores.

VIII. FUTURE SCOPE

Future work should prioritise prospective and multi-site validation, diverse datasets, reproducible reporting, and measures of clinical utility. Multimodal models that combine imaging, laboratory results, and clinical notes may offer richer context, but they also introduce more complex privacy and integration risks. Explainable AI, calibrated uncertainty, privacy-preserving collaboration, and subgroup monitoring are important research directions. Studies should report whether AI changes diagnostic delay, patient outcomes, workload, and avoidable investigations. Any implementation should follow applicable clinical, legal, and regulatory requirements.

IX. DATA PREPARATION AND MODEL DEVELOPMENT

The quality of an AI diagnostic system depends strongly on how its data are collected, labelled, cleaned, and divided for evaluation. Clinical data may contain missing measurements, inconsistent units, duplicated records, spelling variations, and differences in coding between hospitals. A preparation pipeline should document these issues before model training

begins. Missing values should be handled using methods suitable for the variable and clinical context; replacing every missing value with a single average can hide meaningful patterns. Image datasets may require checks for poor resolution, acquisition artefacts, and annotations that do not match the intended diagnostic target.

Labels require particular attention because they are often treated as the correct answer during supervised learning. A label may come from a specialist report, a pathology result, a follow-up diagnosis, or an administrative code, and these sources do not always have equal reliability. Where possible, the reference standard should be defined before analysis and ambiguous cases should be reviewed. The development record should also describe inclusion and exclusion criteria, patient demographics, data sources, and the period during which data were collected. These details help readers understand the population to which the findings may apply.

Training, validation, and test data must be separated carefully. In medical datasets, multiple images or visits may belong to the same patient. If records from one person appear in both training and test sets, the measured performance can be overly optimistic because the model may recognise patient-specific patterns. A patient-level split is therefore generally preferable. External validation using data from another hospital, device, or time period can reveal whether the model generalises beyond the development environment. These practices support the broader evaluation concerns described in the reviewed literature [1]– [4].

X. CLINICAL APPLICATION SCENARIOS

AI can support different stages of diagnosis, but the expected benefit should be stated for each task. In screening, a model may flag images or measurements that deserve earlier review. In triage, a system may help prioritise patients according to a risk estimate. In diagnostic assistance, a model may identify patterns or present relevant evidence to a clinician. These tasks have different error costs and should not be evaluated as though they were interchangeable. A screening tool may prioritise sensitivity, while a confirmatory tool may need strong specificity and a clear pathway for resolving uncertain results.

For medical imaging, computer-vision models can analyse radiographs, computed tomography, magnetic resonance imaging, ultrasound, and retinal photographs. Potential tasks include classification, segmentation, and detection of suspicious regions. A useful interface should show the finding in context and make clear that highlighted regions are model outputs rather than proof of disease. Image quality, scanner differences, and variations in patient populations can affect results, so a model validated on one dataset should not automatically be assumed to work equally well elsewhere [3].

For structured records, machine-learning models can estimate the probability of an outcome from age, vital signs, laboratory measurements, medical history, and other permitted variables. Such estimates may help identify people who need additional testing or follow-up. They must not be presented as certain diagnoses, and the user should be able to see which inputs were available when the estimate was generated. Clinical notes introduce another challenge: natural-language processing may extract symptoms, medications, family history, and prior findings, but abbreviations, negation, and contextual details can be misunderstood. Human review remains important when the extracted information influences care.

A further scenario is longitudinal monitoring. Repeated measurements from home devices or hospital systems may reveal a change that is not obvious from a single reading. However, alerts should be designed around clinically meaningful thresholds and tested for false alarms. If a system sends too many low-value alerts, staff may begin to ignore them. The implementation plan should define who receives each alert, how quickly it must be reviewed, what action is expected, and what happens when the system is unavailable. These operational details are as important as the underlying algorithm.

XI. IMPLEMENTATION AND HUMAN OVERSIGHT

Before deployment, an institution should define the clinical problem, intended users, eligible patient group, and decision that the tool is expected to support. A model built for research exploration may not be suitable for direct clinical use. Clinicians, data specialists, information-technology staff, administrators, and patient representatives can

contribute different perspectives during planning. Early workflow testing can identify whether the tool adds unnecessary steps, duplicates existing alerts, or delays the consultation. A system should be introduced only when its expected benefit and its limitations are understood by the people who will use it.

The interface should communicate the output, its uncertainty, the time of the latest data, and any important limitations. If the input is incomplete or outside the model's validated range, the system should avoid producing an apparently confident result. Clinicians need a practical way to disagree with the recommendation and document their reasoning. Training should cover appropriate use, known failure cases, privacy requirements, and escalation procedures. Responsibility for the final clinical decision must remain clear rather than being obscured by automated output.

Monitoring must continue after deployment because patient populations, clinical practice, coding systems, and equipment can change. A monitoring plan may track sensitivity, specificity, calibration, subgroup performance, missing inputs, alert frequency, and clinician overrides. A decline in performance should trigger investigation and, where necessary, suspension or recalibration of the system. Updates should be version-controlled and tested before release. The organisation should also maintain a record of incidents and provide a route for users to report unexpected behaviour. This lifecycle approach is consistent with the need for responsible governance described by the World Health Organization [6], [7].

XII. ETHICAL, PRIVACY, AND REGULATORY CONSIDERATIONS

Health information is sensitive, and diagnostic systems should use only the data needed for a defined purpose. Access controls, encryption, secure transfer, retention limits, and audit logs can reduce avoidable exposure. Where data are used for research or system development, the institution should follow applicable consent, ethics-review, and data-protection requirements. Removing names alone may not eliminate re-identification risk, particularly when records contain rare conditions or detailed timelines. Data-sharing arrangements should specify permitted use, storage responsibilities, and procedures for responding to a breach.

Fairness requires more than reporting one overall accuracy value. Performance may vary across age groups, sexes, socioeconomic groups, geographic regions, or populations that are under-represented in the training data. Evaluation should examine clinically relevant subgroups when sample sizes allow, report uncertainty, and avoid making strong claims from very small groups. If a gap is found, the team should investigate whether it is caused by data quality, measurement differences, access to care, or model behaviour. The response may include collecting better data, changing the workflow, or limiting the intended use until adequate evidence is available.

Explainability can help users understand which features or image regions influenced an output, but an explanation is not a guarantee of correctness. Explanations may be incomplete or misleading if they are treated as proof. They should be assessed alongside calibration, error analysis, external validation, and clinical usefulness. Regulatory obligations vary by jurisdiction and by the intended function of a product; teams should verify the requirements that apply before clinical deployment. The ethical goal is not simply to make an algorithm transparent, but to ensure that it is safe, accountable, proportionate, and subject to meaningful human oversight [6], [7].

XIII. RESEARCH GAPS AND RECOMMENDED ROADMAP

The reviewed material points to several continuing gaps. First, many reported results are based on retrospective datasets, while prospective evidence about changes in patient outcomes and clinical workload may be limited. Second, performance may be reported on a narrow population or a single institution, making generalisation uncertain. Third, different studies may use different reference standards, data splits, and metrics, which makes direct comparison difficult. Fourth, technical performance does not necessarily show whether a tool improves access, reduces diagnostic delay, or lowers the burden of unnecessary tests. These gaps indicate the need for evaluation that combines technical quality with evidence of practical benefit.

A staged roadmap can make development more systematic. The first stage is problem definition and

stakeholder consultation. The second is data governance and dataset quality assessment. The third is model development with transparent baselines and patient-level data separation. The fourth is internal and external validation, including calibration and subgroup analysis. The fifth is a limited pilot with clinician training, predefined safety checks, and a clear escalation pathway. The final stage is ongoing monitoring, incident review, and periodic reassessment of whether the tool continues to provide value. At every stage, the project should document limitations and avoid claims that exceed the available evidence.

Future research should also examine human–AI interaction: whether clinicians understand uncertainty, when they rely too heavily on recommendations, and how disagreement is resolved. Multimodal systems that combine images, laboratory values, and text may provide richer context, but they also introduce additional missing-data, privacy, and integration challenges. Research should therefore compare multimodal approaches with simpler alternatives and show that added complexity produces a measurable benefit. Shared reporting practices and independent replication would make findings easier to assess and could help separate promising research results from systems ready for routine care.

XIV. PRACTICAL DESIGN OF AN AI-ASSISTED DIAGNOSTIC SYSTEM

A practical system can be organised into connected components: a secure data interface, a quality-checking module, a prediction service, an explanation and uncertainty layer, a clinician-facing display, and an audit and monitoring component. The data interface should accept only authorised information and record the source and timestamp of each input. The quality-checking module can detect missing fields, unusual units, unsupported file formats, or images that do not meet minimum quality requirements. If an input fails a critical check, the system should request correction or return an “unable to assess” status instead of silently proceeding.

The prediction service should be versioned so that each result can be associated with the model and configuration that generated it. The interface should present the output in a way that matches the clinical task: a probability or risk band for a prediction task, a marked region for an imaging task, or extracted evidence for a text-processing task. Users should be able to inspect relevant source information and recognise when the result is uncertain. The audit component should record appropriate events, such as model version, input status, output, override, and final workflow outcome, while following privacy and access rules.

System architecture should also account for failure. Network interruptions, unavailable servers, incomplete electronic records, and delayed updates can occur in routine healthcare. A fallback process should allow the clinical service to continue safely without the AI component. The system must not prevent access to original images or reports, and it should clearly distinguish a technical failure from a negative diagnostic result. Before rollout, the team should test routine cases, difficult cases, missing-data cases, peak workload, and recovery after service interruption. Such testing reduces the risk that a technically accurate model becomes unreliable in everyday use.

XV. DATA IMBALANCE, UNCERTAINTY, AND ERROR ANALYSIS

Disease datasets often contain far more examples of common or negative cases than rare or positive cases. A model can therefore achieve high overall accuracy by favouring the majority class while missing clinically important cases. Teams should report the class distribution and use metrics that reflect the intended task. Possible development approaches include class-weighted learning, careful sampling, and threshold adjustment, but each choice should be evaluated on data that were not used to fit the model. Synthetic or resampled examples should not replace independent clinical validation.

Uncertainty should be treated as useful information rather than an inconvenience. A prediction may be uncertain because the input is poor quality, the patient differs from the development population, or the available variables do not clearly distinguish competing conditions. A system can support safe use by identifying out-of-range inputs, presenting calibrated probabilities when justified, or directing uncertain cases to a human reviewer. The threshold for escalation should be determined with clinical stakeholders and should consider both the harm of

missed disease and the burden of unnecessary follow-up.

Error analysis should examine more than a single list of incorrect predictions. Reviewers can categorise errors by disease type, severity, data source, image quality, demographic group, and stage of the workflow. They should investigate whether the model learned shortcuts, such as scanner markers or site-specific patterns, instead of clinically relevant evidence. False negatives and false positives may have different consequences, so they should be discussed separately. The findings should lead to documented actions, such as improving data collection, revising the intended use, changing the interface, or stopping deployment until the risk is reduced.

XVI. BENEFITS AND RISKS IN BALANCE

Potential benefits of AI-assisted diagnosis include faster review of large datasets, consistent application of selected rules, prioritisation of urgent cases, and support for clinicians who must combine many sources of information. Automated tools may also help identify patterns that are difficult to notice during a high-volume workflow. These benefits are possibilities, not guaranteed outcomes. They depend on the model's performance, the quality of the input, the way the result is presented, and whether the clinical team can act on it in time.

Risks include automation bias, over-reliance on a recommendation, false reassurance from a negative output, alert fatigue, and unequal performance across populations. A clinician may give excessive weight to a confident-looking result even when the model has not been validated for the patient in front of them. Training and interface design should encourage users to consider the original clinical evidence and question outputs that conflict with the wider picture. A tool should be assessed not only for technical accuracy but also for how it changes human decisions, workload, communication, and patient experience.

Cost and access are also important. High computing requirements, specialist maintenance, integration with hospital software, and staff training may limit adoption in smaller facilities. A lightweight, well-validated model that fits a real clinical need may be more useful than a complex model that is difficult to maintain. Procurement decisions should consider the total lifecycle cost, vendor support, data portability, cybersecurity, update policy, and the ability to discontinue the system safely. These considerations help ensure that adoption is sustainable rather than driven only by novelty.

XVII. EVALUATION PLAN FOR A PILOT STUDY

A pilot evaluation should begin with a written protocol that defines the intended population, clinical workflow, reference standard, primary outcome, and stopping rules. Before the pilot, baseline measures can be collected for diagnostic delay, time spent reviewing cases, number of additional investigations, and workload. During the pilot, the team can compare these measures with the period before implementation or with a suitable comparison group, while accounting for differences in patient mix and practice. Any design should be selected according to the question being asked and the available ethical and operational conditions.

Technical measures should include sensitivity, specificity, precision, calibration, and confidence intervals where appropriate. Workflow measures may include turnaround time, proportion of cases requiring escalation, frequency of overrides, and the number of alerts that do not lead to action. Safety measures should capture missed urgent cases, unexpected outputs, technical failures, and delays caused by the system. Feedback from clinicians and, where appropriate, patients can reveal usability issues that numerical metrics do not capture. Results should be reported transparently, including negative or inconclusive findings.

A pilot should not be considered successful solely because the AI model matches or exceeds a benchmark on a test dataset. Success should be judged against the original clinical purpose, safety requirements, and measurable workflow outcomes. If the tool does not improve the intended process, the team should investigate why before expanding deployment. A documented decision gate can determine whether to continue, modify, restrict, or stop the system. This approach supports responsible evaluation and reduces the chance of scaling a system before its practical value is established.

XVIII. ILLUSTRATIVE END-TO-END WORKFLOW

Consider a hypothetical hospital that wants to use AI to help prioritise chest radiographs for review. The project should begin by specifying that the system is intended to flag selected images for earlier assessment, not to independently confirm or exclude disease. Clinicians and technical staff would agree on the target finding, eligible patients, reference standard, escalation process, and the consequences of an incorrect result. The team would then review whether the available images and reports are sufficiently representative of the hospital's patient population and equipment.

During development, images would be checked for quality and duplicate records, and the reference labels would be reviewed according to a documented process. Training and test partitions would be separated at patient level. A validation set would support model selection and threshold setting, while the final test set would remain separate until the evaluation plan was fixed. The team would compare the model with an appropriate baseline and report sensitivity, specificity, precision, calibration, confidence intervals, and results for relevant patient subgroups. If external data were available, they would provide an additional check of generalisation.

Before clinical use, the system would be tested in a controlled environment without allowing its output to replace normal interpretation. Reviewers could compare the AI flag with the original image and document disagreements. The pilot would examine not only whether the system detects the target finding, but also whether it changes time to review, workload, alert burden, and the number of cases escalated. If the tool produces many low-value alerts or performs inconsistently for a subgroup, the team would revise or stop the pilot rather than assuming that the model is ready for routine use.

If the pilot meets its predefined safety and usefulness criteria, a limited deployment could follow with staff training, a clear fallback procedure, and ongoing monitoring. The system would record its model version and relevant performance indicators, while protecting patient information. Clinicians would retain responsibility for interpreting the complete case. Any major model update would be evaluated before release, and unexpected failures would be reviewed through an established reporting process. This hypothetical example illustrates how data quality, technical validation, clinical workflow, governance, and human oversight must work together.

XIX. SUMMARY OF KEY RECOMMENDATIONS

The main recommendations from this review are as follows: (1) define one clinical task and intended use before selecting an algorithm; (2) use reliable and representative data with documented labels; (3) separate data at patient level and test the system on independent data; (4) report metrics that reflect clinical consequences rather than accuracy alone; (5) assess calibration, uncertainty, and subgroup performance; (6) design interfaces that communicate limitations and support clinician review; (7) protect health information through appropriate governance and security; (8) evaluate the effect on workload and patient outcomes; and (9) monitor the system after deployment, with clear procedures for investigating errors and suspending unsafe use.

These recommendations reinforce that AI-assisted diagnosis is a sociotechnical process, not simply a model-training exercise. The algorithm is one component of a larger system involving data collection, clinical interpretation, workflow design, information security, staff training, and organisational accountability. A technically strong model can still cause harm if its limitations are hidden, if its outputs are misunderstood, or if the institution cannot monitor it effectively.

XX. CONCLUSION

AI techniques can support disease diagnosis by analysing medical images, structured clinical variables, and unstructured text. The supplied literature and synopsis describe applications in early detection, risk prediction, and clinical decision support, while also emphasising data quality and validation challenges. The workflow and comparison tables in this paper show that model choice and evaluation must be tied to the clinical task, not only to algorithmic novelty. Safe adoption requires independent validation, appropriate metrics, privacy safeguards, fairness assessment, transparent communication of uncertainty, and clinician oversight.

AI is best positioned as a carefully monitored aid to healthcare professionals.

REFERENCES

[1] World Health Organization, *Regulatory Considerations on Artificial Intelligence for Health*. Geneva: WHO, 2023.

[2] Y. Kumar, A. Koul, R. Singla, and M. F. Ijaz, "Artificial Intelligence in Disease Diagnosis: A Systematic Literature Review, Synthesizing Framework and Future Research Agenda," 2023.

[3] "Evaluation of Artificial Intelligence Techniques in Disease Diagnosis and Prediction," *Discover Artificial Intelligence*, 2023.

[4] B. Vasey, M. Nagendran, B. Campbell, et al., "Reporting Guideline for the Early-Stage Clinical Evaluation of Decision Support Systems Driven by Artificial Intelligence: DECIDE-AI," *Nature Medicine*, vol. 28, pp. 924–933, 2022, doi: 10.1038/s41591-022-01772-9.

[5] M. M. Ahsan, S. A. Luna, and Z. Siddique, "Machine-Learning-Based Disease Diagnosis: A Comprehensive Review," 2022.

[6] M. Mirbabaie, S. Stieglitz, and N. R. J. Frick, "Artificial Intelligence in Disease Diagnostics: A Critical Review and Classification on the Current State of Research Guiding Future Direction," 2021.

[7] R. Aggarwal et al., "Diagnostic Accuracy of Deep Learning in Medical Imaging: A Systematic Review and Meta-Analysis," 2021.

[8] World Health Organization, *Ethics and Governance of* Artificial *Intelligence for Health: WHO Guidance*. Geneva: WHO, 2021.

[9] U.S. Food and Drug Administration, Health Canada, and Medicines and Healthcare Products Regulatory Agency, "Good Machine Learning Practice for Medical Device Development: Guiding Principles," 2021.

[10] S. M. McKinney, M. Sieniek, V. Godbole, et al., "International Evaluation of an AI System for Breast Cancer Screening," *Nature*, vol. 577, pp. 89–94, 2020, doi: 10.1038/s41586-019-1799-6.

[11] X. Liu, S. Cruz Rivera, D. Moher, M. J. Calvert, and A. K. Denniston, "Reporting Guidelines for Clinical Trial Reports for Interventions Involving Artificial Intelligence: The CONSORT-AI Extension," *Nature Medicine*, vol. 26, pp. 1364–1374, 2020, doi: 10.1038/s41591-020-1034-x.

[12] S. Cruz Rivera, X. Liu, A.-W. Chan, A. K. Denniston, and M. J. Calvert, "Guidelines for Clinical Trial Protocols for Interventions Involving Artificial Intelligence: The SPIRIT-AI Extension," *Nature Medicine*, vol. 26, pp. 1351–1363, 2020, doi: 10.1038/s41591-020-1037-7.

[13] E. J. Topol, "High-Performance Medicine: The Convergence of Human and Artificial Intelligence," *Nature Medicine*, vol. 25, pp. 44–56, 2019, doi: 10.1038/s41591-018-0300-7.

[14] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key Challenges for Delivering Clinical Impact with Artificial Intelligence," *BMC Medicine*, vol. 17, Art. no. 195, 2019, doi: 10.1186/s12916-019-1426-2.

[15] D. Ardila, A. P. Kiraly, S. Bharadwaj, et al., "End-to-End Lung Cancer Screening with Three-Dimensional Deep Learning on Low-Dose Chest Computed Tomography," *Nature Medicine*, vol. 25, pp. 954–961, 2019, doi: 10.1038/s41591-019-0447-x.

[16] A. Rajkomar, E. Oren, K. Chen, et al., "Scalable and Accurate Deep Learning with Electronic Health Records," *npj Digital Medicine*, vol. 1, Art. no. 18, 2018, doi: 10.1038/s41746-018-0029-1.

[17] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks," *Nature*, vol. 542, pp. 115–118, 2017, doi: 10.1038/nature21056.