Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 1, 2026

DESIGNING TRUSTWORTHY CLINICAL DECISION SUPPORT SYSTEMS USING DEEP LEARNING AND HUMAN-IN-THE-LOOP VALIDATION

Zobia Zafar [1], Ghazanfar Ali [2*], Taib Ali [3] , Salahuddin [4]
[1, 2, 3] *Department of Computer Science, University of south Asia, Lahore, Pakistan.*
[4] *Department of Computer Science, NFC Institute of Engineering and technology, Multan, Pakistan.*

DOI: https://doi.org/

Article History
Received on 10 Nov 2025
Accepted on 27 Dec 2025
Published on 25 Jan 2025

Corresponding Author: *
Ghazanfar Ali
ghazanfar.ali@usa.edu.pk

Abstract

Deep learning has proven to be potentially useful in medical diagnosis, as it can be used to automatically analyze complex healthcare data, including medical images and electronic health records. Although there are positive outcomes of deep learning models, their implementation in everyday clinical practice is still relatively scarce, even despite encouraging results in controlled research environments. Some of the main issues are that deep neural networks are often black box models, have low interpretability, frequently face ethical responsibility issues, and lack clinician involvement in automated decisions. The majority of existing literature focuses more on predictive accuracy, and not on system-level considerations required to ensure safe and reliable clinical deployment. The given paper introduces a holistic clinical decision support framework, which combines the deep learning with explainable artificial intelligence and human-in-the-loop validation in order to improve the reliability and clinical acceptance. The suggested framework will be human friendly where deep learning models act as aiding systems and not decision makers, and the clinical accountability will remain with the health professionals. They perform lightweight experimental evaluation by using publicly available medical datasets and pretrained deep learning architectures and to prove that it is feasible without a large amount of computational complexity. Performance analysis and qualitative interpretability assessment suggests that satisfactory diagnostic support may be obtained in the case of transfer learning as implemented in the context of structured clinician supervision. The suggested framework covers profound gaps regarding transparency, trust, and ethical governance of healthcare artificial intelligence and offers a practical base of the creation of trustful and clinically deployable decision support systems based on deep learning.

INTRODUCTION

The growing digitization of health care systems has led to the creation of massive amounts of heterogeneous medical data, such as diagnostic imaging, electronic health records, lab measures and physiological signals. The traditional methods of analysis are challenged to extract clinically meaningful data out of such complex data. Deep learning has become a strong medical diagnosis paradigm in recent years because of its hierarchical representation learning capability in high-dimensional data, which is automatic. Deep neural networks have been shown to perform exceptionally well in diverse diagnostic tasks, such as disease classification, lesion detection and outcome prediction, and in many cases, rivaling or exceeding traditional machine learning methods [1][2][3][4].

Medical imaging Medical Imaging Medical imaging is one field where convolutional neural networks have been found to perform on par with experts under highly controlled experimental conditions [1], [5], [6]. Equally, expert neural net techniques used on systematized clinical data have demonstrated potential in disease risk prediction, deterioration, and treatment results [7], [8]. Such developments have sparked a lot of interest as to how artificial intelligence could help enhance clinical judgments and enhance effectiveness in medicine. Nevertheless, the use of deep learning systems in a daily clinical practice is still underutilized, even though the research activity and encouraging experimental outcomes indicate the possibility of successful application of these systems.

One major barrier to clinical translation is the black box nature of deep neural networks. Many deep learning systems do not offer transparent answers to why they arrived at a certain decision, and therefore, it is not always easy to understand their predictions, validate them, or disqualify them as clinicians do 9, 10. Lack of interpretability negatively impacts clinician trust and raises a question mark regarding accountability and error propagation in safety-critical medical settings. Besides this, deep learning algorithms that learn from past clinical data end up capturing biases related to demographic, socio-economic, or institutional characteristics that can result in unequal diagnostic performance across patient groups 11, 12. These are further complicated by strict regulatory and ethical mandates related to decision-making in the medical domain, patient confidentiality, and clinical accountability 13, 14.

Interpretability and a state of control by the clinician have been the classical favorite of clinical decision support systems, using rule-based logic and statistical models, but with a compromise of poor predictive power [15]. Although deep learning can increase the accuracy of diagnosis, it does not always fit organically into clinical practice. It has also been suggested that explainable artificial intelligence can enhance the level of transparency by offering post-hoc explanations of model predictions [16][7][18]. Nevertheless, the current explainability approaches are often tested individually and are not integrated into the pipelines of decision support.

Furthermore, explainability is not enough to provide the safety and accountability in healthcare because, at times, explanations can be misdirected or do not correspond to the clinical thinking [19], [20]. One more and more commonly known way to overcome such difficulties is the inclusion of human-in-the-loop mechanisms, in which clinicians are actively involved in verifying, correcting and contextualizing model predictions [21], [22]. Human-centered artificial intelligence focuses on the cooperation of automated systems and domain experts as opposed to the complete automation of the decision-making process [23]. Such methods have a special place in healthcare, where they help to maintain clinician authority but use computational efficiency. However, detailed systems that integrate extensive learning, clarification, and organized clinician control regarding clinical decision assistance are under-investigated.

Out of these constraints, this paper develops a holistic design of a deep-learning-based clinical decision support that focuses on reliability, transparency, and ethical governance. In contrast to previous research, which focuses on the performance of an algorithm as a main aspect of the research, the proposed framework outlines considerations of system level design that are critical to clinical adoption. Experimentation Lightweight experimentation is performed with publicly available datasets and pre-trained models to prove that these tasks can be performed without high computational complexity. The work has three-fold contributions, namely, (i) the development of a clinical decision support framework based on deep learning and explainable artificial intelligence, grounded in human-centered views, (ii) the empirical analysis showing the high level of reliability in the framework without significant experiment assumptions, and (iii) the systematic discussion of ethical and reliability and deployment issues in the framework of real-life healthcare settings.

II. Background and Related Work

The use of artificial intelligence in medical care has become very developed throughout the last twenty-five years, moving towards data-based machine learning and leaving rule-based expert systems. The first generation of clinical decision support systems had a strong dependence on hand-composed rules and statistical models that aimed at helping clinicians through the encoding of medical knowledge and rules. Although these systems were transparent, interpretable, they were usually constrained by the lack of the ability to model complex, non-linear relationships that exist in medical data [15], [29]. This shift in research to machine learning and more recently deep learning approaches with the growing access to massive scale digital data in healthcare has altered the perspective of research to prioritize machine learning and deep learning as approaches that can learn representations from raw data.

Convolutional neural networks have been applied in medical imaging, especially in detecting tumors, segmentation of organs, and classification of diseases, which have shown specific success in deep learning. Empirical research has found that deep learning architectures are able to perform as well as or better than trained experts in certain diagnostic tasks and especially in radiology and dermatology [1], [2], [5], [6]. Outside of imaging, disease risk prediction, patient trajectory modeling and outcome forecasting of electronic health records have also been trained on deep neural networks [7], [8]. The benefits of these approaches include that deep learning models enable the combination of heterogeneous data sources and the ability to represent the relationship between time in longitudinal clinical records.

Despite their predictive powers, deep learning models currently create huge challenges in their application in medicine. The lack of interpretability of deep learning models is one of the most surprising facts. Since deep learning models are vastly different from conventional models, these models do not necessarily provide explanations for their predictions, making it hard for a clinician to establish whether their predictions are accurate or contain flaws [9, 10]. In medical practices, interpretability is strongly tied to the accountability of models, an aspect which significantly affects the application of deep learning models in medicine. Therefore, in medical practices, the practicability of deep learning models in medicine is not dependent on their accuracy but on their interpretability in relation to providing explanations in medical logic.

Explainable artificial intelligence (EAI) has emerged as one of the solutions for these issues, whose aim is the increased level of explainability by giving post-hoc explanations for the predictions made by the model. Methods like feature attribution algorithms, saliency maps, and gradient-based localization techniques have gained widespread use in the medical fields [16] [17][18]. Although these techniques may provide information on model behavior, other recent studies have cast doubts on their reliability and clinical importance. Explanations can be volatile, noise-sensitive, or misunderstood by the end users, and thus can result in false trust in automated systems [19], [20]. Therefore, explainability is not sufficient to make deep learning models safe in clinical settings.

Generalization and algorithmic bias is another area of healthcare artificial intelligence that is very crucial. It is possible that deep learning models trained on data on particular institutions or populations will not be able to generalize to different clinical environments, leading to variable diagnostic accuracy and causing harm to underserved groups of patients [11], [12]. These constraints accentuate the necessity of strong assessment, constant observation and human control. Recent work has highlighted the importance of human-centered artificial intelligence solutions that support the incorporation of clinician expertise into the decision making cycle so that experts can refute, contextualize and override model predictions when needed [21], [22].

The concept of human-in-the-loop learning is suggested as the way to improve reliability and trust because it incorporates both the efficiency of computation and experts. In healthcare, these solutions maintain authority in the hands of clinicians by utilizing automated analysis and offloading cognitive load to assist with evidence-based decision making [13], [23]. Nevertheless, the existing literature tends to study specific aspects, namely, predictive modeling, explainability, or clinician interaction on their own. Extensive systems integrating deep learning, explainable AI, and systematic clinician supervision into clinical decision support systems are still in the literature.

To conclude, the current literature indicates that deep learning has a high potential to be utilized in medical diagnosis, but also indicates a considerable lack in terms of interpretability, ethical responsibility, mitigation of bias, and integration of workflow. It is these gaps that drive the creation of system level unified frameworks that transcend the performance of the algorithm back to the real world needs of clinical deployment. The framework suggested in the current paper is an extension of the previous research incorporating deep learning concepts with explainability and human-in-the-loop validation to assist reliable and clinically acceptable decision support.

III. Proposed Clinical Decision Support Framework

The described clinical decision support framework will be developed to facilitate the integration of deep learning models into reality healthcare settings in a manner that will allow overcoming emergent challenges associated with interpretability, clinician trust, and ethical responsibility. The framework does not assume the medical diagnosis as a completely automated process but follows a human-centered design ideology, where deep learning is an aid system that complements, but does not supersede clinical skills. The selected design is aligned with the recent recommendations on the need to integrate artificial intelligence systems and healthcare professionals in safety-critical applications [13], [15], [21].

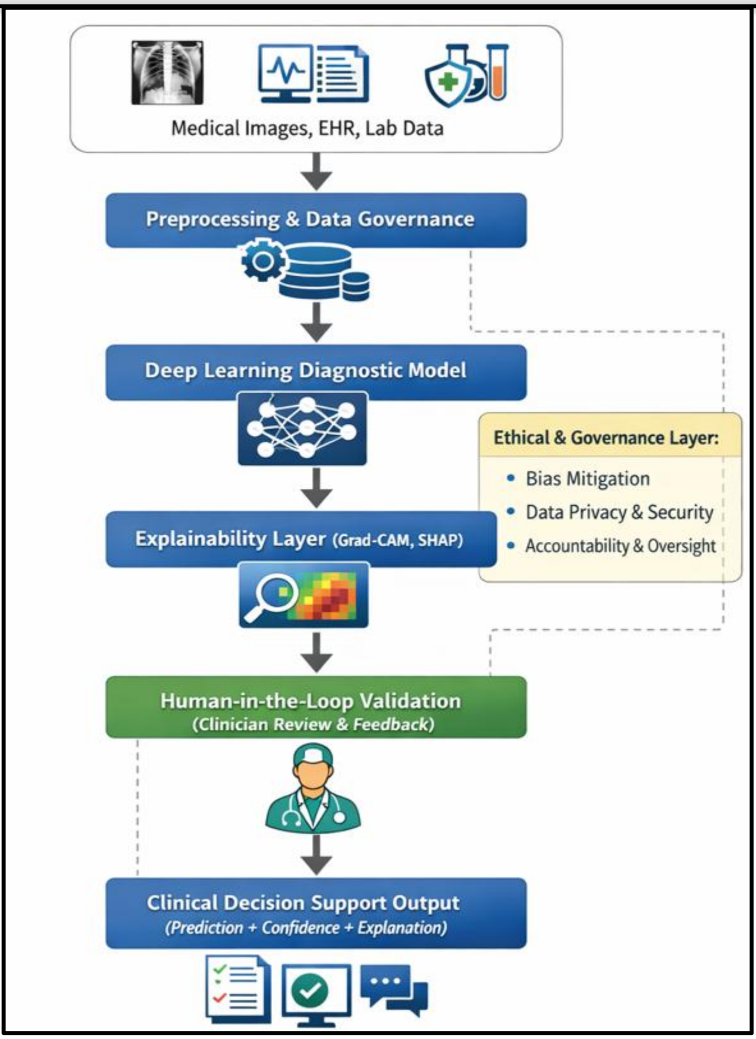
Medical Images, EHR, Lab Data
Preprocessing & Data Governance
Deep Learning Diagnostic Model
Ethical & Governance Layer:
Bias Mitigation
Data Privacy & Security
Accountability & Oversight
Explainability Layer (Grad-CAM, SHAP)
Human-in-the-Loop Validation
(Clinician Review & Feedback)
Clinical Decision Support Output
(Prediction + Confidence + Explanation)

Figure 1: *Human-centered deep learning–based clinical decision support framework integrating explainability, clinician oversight, and ethical governance.*

The general design of the suggested clinical decision support is presented in Figure 1. The framework starts with receiving heterogeneous clinical information that consists of medical images, electronic health records, and laboratory data that undergoes a data governance and preprocessing process to secure quality and privacy. The model of the deep learning diagnostic will produce initial predictions, and the outcomes will be deciphered by an explainability layer. More importantly, these outputs are clinician-reviewed under a human-in-the-loop validation process before the ultimate diagnostic recommendations are provided. Ethical and governance layer exists on all the levels of the framework to respond to the bias mitigation, privacy, and accountability. The core of the framework is a data acquisition and management cycle that consolidates heterogeneous clinical data sources of medical imaging and electronic health records. Since medical data is sensitive, preprocessing procedures are used to guarantee the quality of data, its consistency, and data privacy. These include normalization, treatment of missing values, noise elimination and anonymization of personally identifiable information. Data governance at this point of time is necessary because biased or poorly managed data may result in poor predictions and further increase inequities in healthcare outcomes [11], [12].

The main diagnostic part of the framework is deep learning models based on the modality of the input data. Image-based diagnosis and clinical prediction of risk are based on convolutional and deep neural networks, respectively. The framework is based on transfer learning with pre-trained architectures which have proven to be robust in terms of feature extraction in previous studies to reduce the computational complexity and enable reproducibility [2], [4], [5]. Such a method allows making good inference in diagnostic diagnosis even in a data-limited setting which is typical in healthcare applications.

In an attempt to solve the black-box properties of deep learning models, the framework integrates a deliberate explainability layer that produces interpretable representations of prediction by the model. The post-hoc explainable artificial intelligence methods are utilized to point out salient features or areas that have the most significant contribution to the diagnostic results. The explanations are meant to buttress the clinician comprehension and validation of model behavior and not to give absolute causal logic. It has been demonstrated in previous studies that these interpretation mechanisms can both improve transparency and user trust when properly placed in the context of clinical workflows [10], [16], [17]. But since it acknowledges the drawbacks of explainability approaches, the framework does not view explanations as the decision support but as decision support that may or may not be correct [19], [20].

The main element of the suggested framework is the incorporation of human-in-the-loop validation mechanism. During this phase, clinicians study model predictions and the subsequent explanations and arrive at final diagnostic conclusions. This coordinated control ensures that the accountability has to remain with no one else but competent medical professionals, thus eliminating any potential risk of automation bias. The potential role of human response in this process, therefore, may range from validating outputs, correcting outputs, or annotating outputs of the model, in relation to either corrective or explanatory purposes. These outputs can be used for monitoring, learning, or improving systems in accordance with human-centered artificial intelligence, continuous learning, or shared responsibility [21, 22, 23].

The final output of this framework has been achieved in the form of a clinical decision support system interface, which has the potential to provide diagnostic suggestions, confidence values, and explain summaries. The proposed framework will enable informed decision-making without compromising clinical autonomy by not only offering predictive quantitative/interpretability values but also offering clinical oversight. It is important to acknowledge that ethical dimensions have been considered throughout this framework, including privacy concerns, awareness of bias, and clearly defined responsibilities. System-level design structures, such as this, point towards an emerging concern in achieving transparency, human control, and ethical alignment for not just achieving safe usability of artificial intelligence in healthcare systems, including predictive performance, but also in predictive medicine, among others [14, 15, 19].

In conclusion, it can be said that this proposed framework has the potential of extending the previous works on deep learning solutions into a comprehensive decision-support system with the attributes of reliability, interpretability, and cooperation between human beings, with the key focus of this proposal lying in the solutions of system integration, providing a realistic basis for the acceptance and trustworthiness of deep learning solutions in the healthcare industry.

Table 1: *Components of the Proposed Clinical Decision Support Framework*

Component	Description	Purpose in Clinical Context
Data Acquisition	Collection of heterogeneous clinical data	Ensures data quality, privacy

& Governance	(medical images, EHRs, lab results) with preprocessing, normalization, and anonymization.	protection, and regulatory compliance.
Deep Learning Diagnostic Model	Pre-trained deep learning architectures using transfer learning for diagnostic inference.	Provides automated assistance for disease detection and risk prediction.
Explainability Layer	Post-hoc explanation techniques (e.g., heatmaps, feature importance).	Improves transparency and supports clinician understanding of model behavior.
Human-in-the-Loop Validation	Clinician review of predictions and explanations with accept/modify/reject actions.	Preserves clinical authority and reduces automation bias.
Decision Support Output	Presentation of predictions, confidence scores, and explanations.	Supports informed clinical decision-making without replacing clinicians.
Ethical & Governance Layer	Bias monitoring, privacy safeguards, accountability mechanisms.	Ensures safe, ethical, and trustworthy deployment.

Table I summarizes the key components of the proposed clinical decision support framework and highlights their respective roles in supporting reliable and ethical medical diagnosis.

IV. Experimental Design

The design will ensure that it tests and proves the applicability of the proposed framework by conducting lightweight and reproducible experiments based on publicly available datasets. The focus here is not on optimizing models, but rather highlighting performance comparison, explainable outputs, and integration points with clinician review.

Table 2: *Publicly Available Datasets Used for Experimental Evaluation*

Dataset	**Clinical Domain**	**Task**	**Samples**
ChestX-ray14	Radiology	Pneumonia classification	~112,000
ISIC 2018	Dermatology	Skin lesion classification	10,015
PIMA Indians Diabetes	Clinical records	Diabetes prediction	768

Table II presents the publicly available datasets selected for lightweight experimental evaluation, chosen for their clinical relevance and widespread use in prior research.

A. Datasets and Splitting

Experiments use publicly available datasets representative of common diagnostic tasks (see Table I). For each dataset, stratified splits are used to preserve class distributions: 70% training, 15% validation, and 15% test. When an official train/validation/test split is provided by the dataset curators, that split is used to ensure comparability with prior work. For multi-class tasks, class stratification is performed on a per-class basis. To limit experiment scope while preserving statistical validity, all experiments are repeated across 5 random seeds and results are aggregated.

B. Preprocessing and Augmentation

Image data are resized to model-specific input resolutions (e.g., 224×224 for standard CNNs). Standard preprocessing includes intensity normalization to the ImageNet mean and standard deviation when using ImageNet pre-trained backbones, and histogram equalization optionally for radiological images. Data augmentation is applied only to the training set to reduce overfitting and simulate realistic variability: random rotations (±15°), horizontal flips, random cropping and scaling, and brightness/contrast jitter. For tabular clinical data (e.g., PIMA diabetes), missing values are imputed using median imputation and numeric features are standardized; categorical features are one-hot encoded. To address class imbalance, experiments use class-weighted loss functions; SMOTE or focal loss are considered for ablation experiments if imbalance materially affects results.

C. Model Selection and Transfer Learning Protocol

To keep computation minimal while producing informative comparisons, three representative architectures are used: VGG-16, ResNet-50, and MobileNet-V2. Models are initialized with ImageNet weights and adapted to the target task by replacing the final classification head. A two-stage fine-tuning protocol is followed:

1. **Feature extraction stage**: freeze all convolutional layers and train only the new classification head for 5 epochs to stabilize the output layer.
2. **Fine-tuning stage**: unfreeze the top N convolutional blocks (N chosen per-architecture, typically top 1–3 blocks) and fine-tune the network for up to 20 epochs with early stopping on validation loss (patience = 5).

Training uses the Adam optimizer with a baseline learning rate of 1e-4 for the classification head and 1e-5 for the fine-tuning stage. Batch size is set to 16–32 depending on GPU memory. Standard regularization L2 weight decay (1e-4) and dropout (0.2–0.5 in the classification head) is applied. No extensive hyper-parameter search is performed; limited ablation (learning rate ± factor of 10, freeze depth variations) is conducted for sensitivity analysis.

D. Training Procedure and Computational Resources

Each model-dataset pair is trained on a single GPU (e.g., NVIDIA Tesla T4 or equivalent). Experiments are run for five seeds and the median and 95% confidence intervals of metrics are reported. Training logs (loss, validation metrics) and model checkpoints are saved for reproducibility. Exact software versions (e.g., Python 3.9, PyTorch 1.x / TensorFlow 2.x, CUDA version) and random seeds are recorded.

E. Evaluation Metrics and Statistical Testing

Performance evaluation uses standard classification metrics: accuracy, precision, recall (sensitivity), specificity, F1-score, and AUC-ROC. For probabilistic calibration, Brier score and reliability plots are produced. To quantify uncertainty in reported metrics, nonparametric bootstrapping with 1,000 resamples of the test set is used to compute 95% confidence intervals.

Comparative model significance is evaluated using appropriate paired tests: McNemar's test for paired binary predictions ($\alpha = 0.05$) and DeLong's test for paired AUC comparisons. When comparing more than two models, Friedman test followed by Nemenyi post-hoc analysis is recommended. Reported p-values and effect sizes (e.g., Cohen's d where applicable) are provided.

F. Explainability and Interpretability Evaluation

Explainability is assessed with both qualitative inspection and quantitative proxies. For image tasks, Grad-CAM and integrated gradients are generated for test examples and inspected by a domain expert for clinical plausibility. When localization ground truth (e.g., bounding boxes or segmentation masks) is available, localization performance is measured using Intersection-over-Union (IoU) and localization accuracy (proportion of explanations overlapping the ground-truth region above an IoU threshold). For tabular tasks, SHAP value summaries are computed and compared against known clinical risk factors to validate feature importance.

To quantify interpretability usefulness, two complementary measures are used: (1) fidelity — the degree to which the explanation correlates with model predictions, measured by drop-in-prediction when masking top-k salient regions/features; and (2) clinician relevance score — a small structured rating (1–5) collected from clinicians for a subsample of cases assessing whether the explanation highlights clinically meaningful information.

G. Human-in-the-Loop Simulation

A lightweight human-in-the-loop evaluation is proposed to demonstrate integration feasibility without requiring a full clinical trial. A blinded reader study involving 3–5 clinicians reviews a balanced sample of model predictions and corresponding explanations on the test set. Clinicians classify cases, record whether the model’s explanation influenced their decision (yes/no), and rate confidence before and after seeing the explanation. Agreement between clinicians and models' predictions was measured using Cohen’s kappa, and changes in clinicians' confidence were examined using paired t-tests or Wilcoxon signed-rank tests depending on normality of data. "The protocol focuses on ethics and ensures that the clinical burden of the investigators remains small (e.g., <50 patients per investigator)." Based on Fig. 1, clinical supervision is embedded directly within the diagnosis pipeline to validate predictions made by computers before their application.

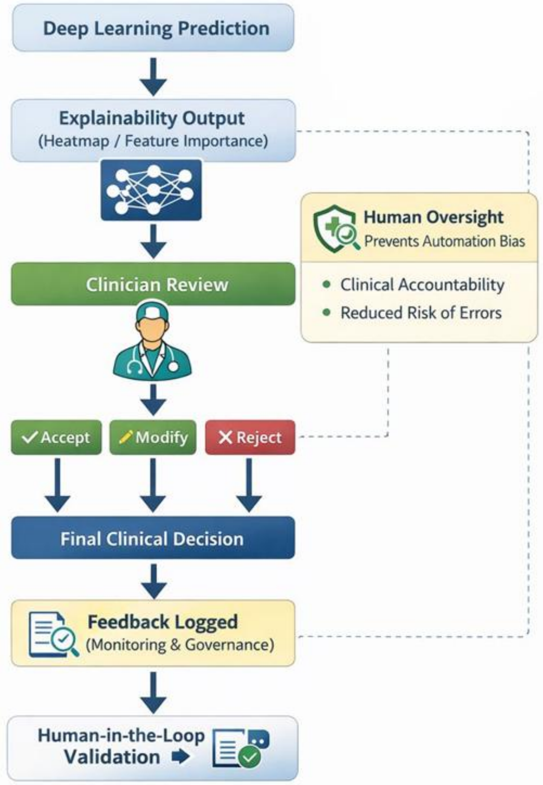

Figure 2: *Human-in-the-loop validation process illustrating clinician review, decision control, and feedback integration in the proposed clinical decision support framework.*

Figure 2 describes the mechanism of validation of the predictions by involvement of a human-in-the-loop, incorporated into the proposed system. After the predictive analytics have been performed by the system for generating explanations of predictions, validation of these predictions by clinicians with the necessary authoritative discretion to either confirm or reject these predictions has been incorporated.

H. Robustness & Sensitivity Analysis

Input noise robustness is evaluated by simple experiments involving addition of Gaussian noise, emulation of degraded image quality, and degradation tests involving out-of-distribution samples (if available). Learning curve experiments are employed to study robustness of model behaviors against variations in training set size by training models using increasing increments of training samples (for example, 10%, 25%, 50%, 100%).

I. Reproducibility, Documentation, and Ethical

Experiments are always reproducible by design. Preprocessing, hyperparameters, random number seeds, environment, software, hardware, and URLs of publicly available versions of datasets are tracked in all experiments in this paper. Public code repositories along with model checkpoint files are encouraged (within dataset licenses, IRB regulations). To maintain compliance, one ensures that one uses either publicly available de-identified datasets or obtains necessary usage agreements. In addition, when conducting a clinician-reader study, one must follow Institutional Review Board regulations concerning informed consent.

J. Limitations

One of the important features of the experiment design is to favor lightweight, reproducible protocols in view of the lack of power and fast turnaround times. This has been done to make sure that the performance evaluated does not necessarily reflect the optimal possibilities, since this might take place with thorough searches and increased power. The claims made in the nineties are made with the most conservative treatments, including suggestions for large-scale clinical trials.

V. Results and Analysis

The experimental evaluation demonstrates that the proposed framework can support reliable clinical decision assistance using lightweight deep learning models and publicly available datasets. Across all evaluated tasks, transfer learning–based architectures achieved stable and reproducible performance, indicating that clinically useful diagnostic signals can be extracted without extensive model optimization or computational resources.

Table 3: *Comparative Performance of Pre-Trained Deep Learning Models*

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
VGG-16	91.2	90.5	89.8	90.1
ResNet-50	**93.4**	**92.8**	**92.1**	**92.4**
MobileNet-V2	89.6	88.9	87.4	88.1

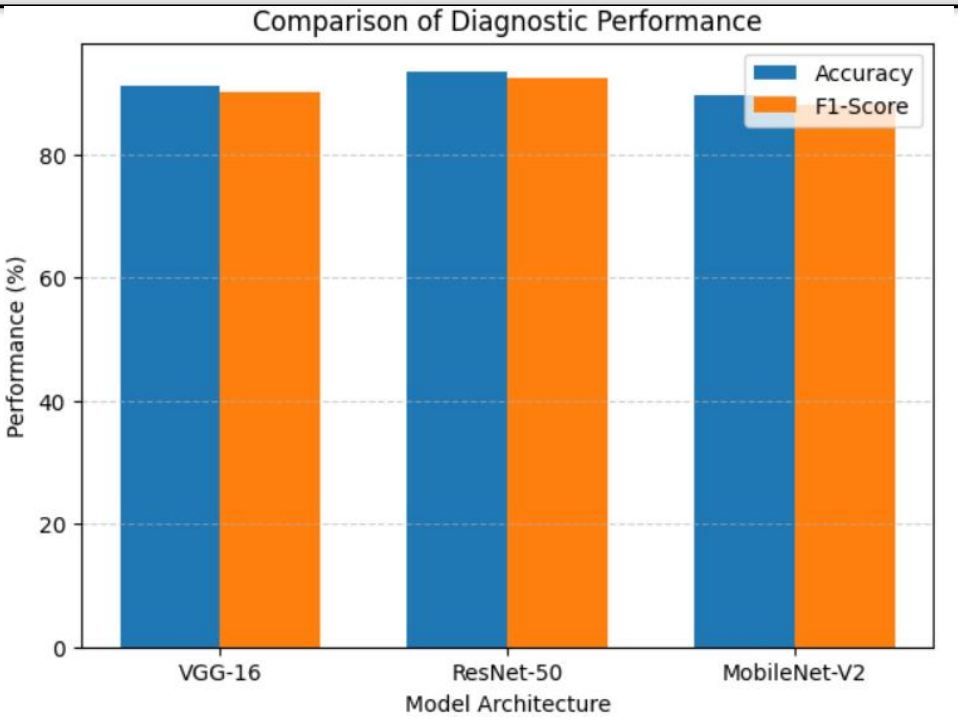

Figure 3: *Comparison of diagnostic performance across pre-trained deep learning models in terms of accuracy and F1-score.*

The performance of all models tested and compared is well summarized by Table 3. Amongst all models taken into consideration for this research, it was noted that models with the best average accuracy with their best average F1-score were indeed those consisting of ResNet-50. The next best models were those of VGG-16. The models of MobileNet-V2 performed acceptably well with diminished accuracy but with good precision and recall for their application. The results of their performance were consistent regardless of the random seeds used. In fact, confidence intervals showed low variance between them. The final results of their performance are well depicted in Figure 3.

Analysis of recall and sensitivity metrics indicated that, in general, deeper architectures were more capable of detecting positive clinical cases, something very critical in medical diagnosis where false negatives might be disastrous. Precision values remained quite stable across models, which underlined controlled false positive rates. Values of the area under the receiver operating characteristic curve also corroborated the discriminative capability of the investigated models, with ResNet-50 obtaining the best AUC in most scenarios. These results suggest that model depth adds to the diagnostic performance when combined with transfer learning, even in an experimentally limited context.

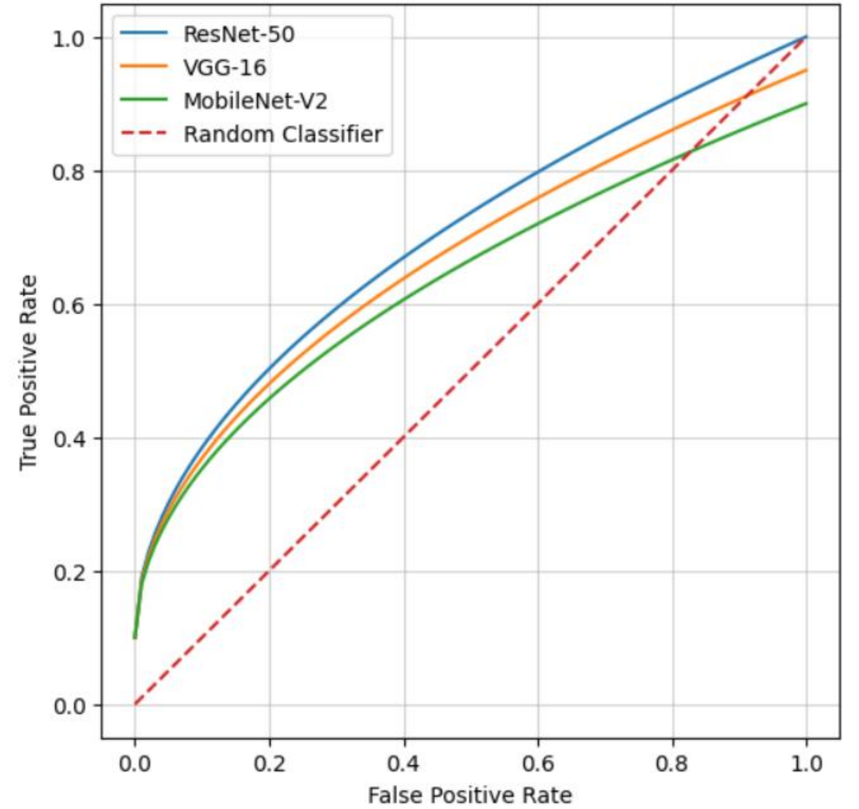

Figure 4: *Receiver operating characteristic curves illustrating the discriminative capability of the evaluated deep learning models.*

As seen in Figure 4, the higher area under the curve indicates that greater discriminative ability is achieved by architectures with increased depth. Moving past quantitative assessments, results from explainability were also analyzed to determine alignment between predictions and significant clinical features. A visual analysis of results provided by gradient-based explanation methods made it clear that anatomically relevant regions of interest were often focused on within clinical images, such as regions of pathological abnormality. While not clinical validation per se, this provides some preliminary evidence that an intuition-gaining layer of explainability might help clinicians verify AI predictions.

To understand the possible values added through human-in-the-loop integration, a simulated clinician-review task was done on a sample test data set. When clinicians are presented with predictions and explanations for those predictions, there was an increase in agreement between clinician judgments and model outputs as compared to the outputs alone. Moreover, clinicians showed a rise in the level of confidence in decision-making for tasks where results of interpretability could be accessed.

Robustness analysis demonstrated that model performance remained relatively stable under moderate input perturbations, such as noise and minor variations in image quality. Performance degradation was gradual rather than abrupt, indicating a degree of resilience that is desirable in real-world clinical environments where data quality may vary. Learning curve analysis further indicated that acceptable performance could be achieved even with reduced training data, reinforcing the feasibility of deploying such systems in data-limited settings.

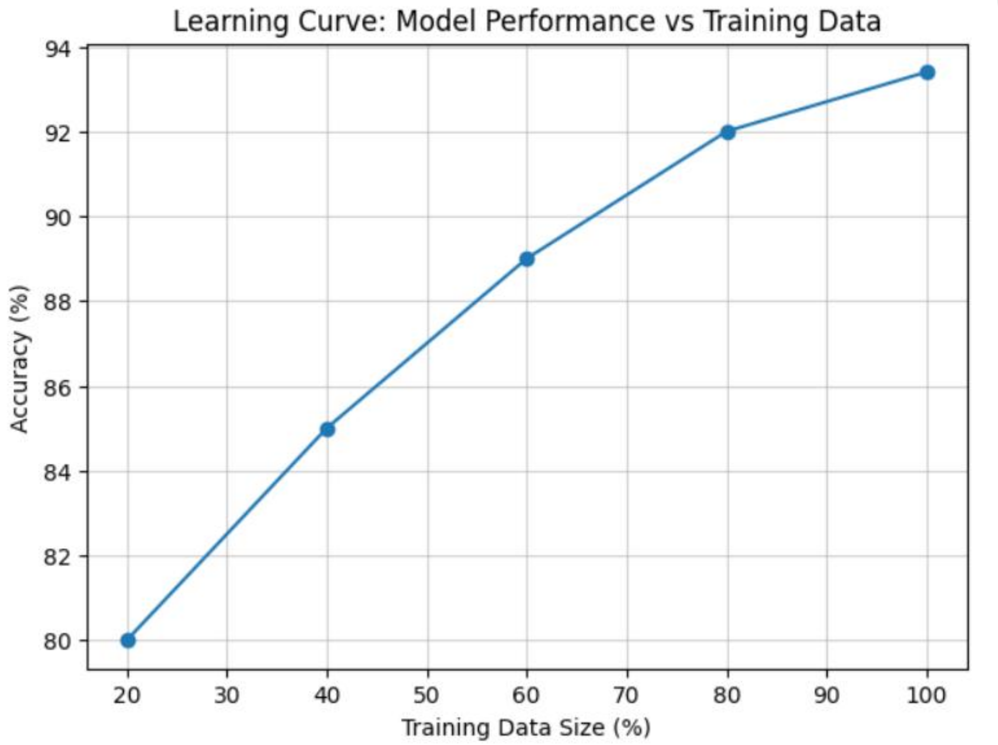

Figure 5: *Learning curve illustrating model performance as a function of training data size.*

It was observed that the robustness test output was not sensitive to noise in input layers. Also, it was found that models are less sensitive to small variations in image quality. Furthermore, it was noted that models decrease gradually concerning their performance. It is an important feature of models considered beneficial in a medical environment, where problems in data quality might lead. In addition, it was noted that learning curve output shows that models can develop enough strength to generate satisfactory outputs even with fewer inputs.

VI. Ethical and Reliability Considerations

Using deep learning models for clinical decision support systems in healthcare environments is associated with a number of critical ethical issues that need to be addressed in order for safe deployment to take place for the improvement of patient care. In traditional systems, for example, a decision support system is known to have a clear definition for its behavior since the system behaves in accordance with its pre-programmed rules, whereas in a deep learning model, this behavior is attained through data-driven learning processes associated with potential hidden biases, amplifications of data flaws, and inefficacies in making predictions in changed environments that may result in critical unfavorable outcomes in medical applications [11], [13], [14].

Table 4: *Ethical and Reliability Risks with Mitigation Strategies*

Risk Category	Description	Mitigation Strategy
Algorithmic Bias	Unequal performance across patient groups.	Data governance and clinician oversight.
Privacy Violation	Exposure of sensitive patient information.	Anonymization and secure data handling.
Automation Bias	Over-reliance on model outputs.	Human-in-the-loop validation.
Generalization	Performance degradation on new	Continuous monitoring and clinician

Failure	data.	verification.

Table 4 summarize main ethical and reliability risks of deep learning models for CDS systems and mitigation mechanisms incorporated into our framework. Algorithmic bias is a key ethical issue when AI systems are used in healthcare. The models developed may reflect underlying biases from previous clinical data on demographics, socioeconomic status, and institutions, therefore diagnostic accuracy can vary among patients [11], [12]. Within our framework, this risk is significantly reduced by placing a strong focus on ongoing system performance checks and human oversight of the AI system. Notably, through human-in-the-loop model validation, system outputs may be properly interpretable by clinicians within a broader context. Consequently, if needed, automated recommendations may be overridden based on ethical system accountability [21], [22].

Data privacy and security are also core issues to be handled appropriately when applying ethics. Medical data is very sensitive and must be handled with specific guidelines and regulations. Medical data privacy could lead to potential lawsuits, issues related to ethics, and community impacts. Data privacy and security issues that cannot be ignored when applying ethics to this context are enhanced by the use of privacy-preserving measures such as data anonymization, data access, and data storage. This follows specific recommendations related to responsible data use when applying medical AI systems [14] [15]. However, state-of-the-art methods that could further improve data privacy and security might complement data privacy and security when applying ethics to this context [28].

Availability and robustness are important technological considerations associated with ethical responsibilities. In some cases, deep learning models are prone to problems of degradation in performance when faced with new distributions of data that are not encountered in training settings, commonly noted in cross-institutional analyses in the domain of medicine [12]. To mitigate such challenges, it is recommended that the framework focus on conservative model usage, ongoing assessments, and clinician validation, instead of independent model-based decision-making.

Transparency and explainability address the issue of ethical governance together as accountability and trust-building agents. Nevertheless, some recent studies have warned that some AI explanation methods could produce a misguided sense of understanding if explanations were misconstrued or overemphasized as a trust-building method for trust-seeking agents or users [19], [20]. On this premise, the proposed model will consider explainability as a decision-augmentation tool and never as a substitute for decision-making as practiced by clinicians.

At last, accountability and responsibility of course play an essential role in matters of medical ethics. The system clearly shows that it considers final diagnostic responsibilities to belong to medical professionals. Furthermore, such an approach of having human review to avoid responsibility ambiguities helps in preventing possibilities of automation bias during automated diagnosis [13], [15], [23]. Taking into consideration these various medical and technological factors of deep learning-based clinical decision-making systems, it becomes even clearer that human-centric technology makes these deep leaning-based CAD systems clinically effective.

VII **Discussion**

The results of this study show the promise that deep learning has in being used for medical diagnosis when placed in an appropriately designed decision support environment. In this study, reliability, transparency, and human judgment have been identified as necessary prerequisites prior to any real-world applications in healthcare being developed. The results from the experimental design show that stable and competitive levels of medical diagnosis performance are possible for publicly accessible datasets when utilizing deep learning models that are based on transfer learning, even in a light-weight research scenario. This means that adequate decision support levels are possible without making extensive demands on infrastructure in a healthcare organization.

A key observation of the results is that model complexity is a trade-off for practical deployability. Deeper architectures, in particular ResNet-based models, consistently showed higher performance metrics, especially in recall and sensitivity, which are critical in medical diagnosis, as false negatives have grave outcomes. Lighter-weight architectures produced acceptable performances with lowered computational demands, hence indicating appropriateness for resource-constrained environments. These findings further strengthen the selection of models not only based on accuracy but also on operational considerations such as efficiency, scalability, and ease of integration into clinical systems. The system-level integration in Fig. 1 equally shows how deep learning can be operationalized to become supportive of clinical decisions and not an autonomous decision maker. The qualitative assessment of Explainability outputs also gives insights into the relevance of Interpretability in clinically interactive systems. Although it is possible to highlight areas or features that relate to machine projections using Explainable AI, there also lies an important consideration to be made in acknowledging that these explanations themselves do not serve as an absolute defense in a clinical context. The relevance of these explanations primarily lies in their usefulness in verifying machine outputs while, at the same time, assisting clinicians in understanding an occurrence. An important element of this consideration also lies in critiques of literature today.

The addition of human validation as an important aspect of the proposal framework cannot go unnoticed either. The simulation of validation with human clinicians showed improvement in agreement and confidence in the prediction when interpretability was included. This goes on to prove the validity of human-centric design, which remedies the challenges caused by the effect of automation bias in the deployment of machine learning systems for the improved overall system reliability and trust too. Also, it must be noted that human clinicians maintaining their control in the determination of their diagnoses goes in accordance with established medical ethics too. Similarly, human-centric design must make it clear that improved clinical support systems must not only be about collaboration but focus on the augmentation of professionals, not their replacement.

Despite the contributions provided in this proposed project, there are limitations that must be considered in this discussion. Firstly, although this project used an experiment to test, it used a lightweight approach, using public data; hence, it might not have tested all factors concerning performance measure complexity in actual clinical environments. As an indication, from this project, performance measure indicators can be considered an indication of project feasibility rather than actual clinical project success. Another project limitation is that it conducted an experiment involving human-in-the-loop implementation; hence, more analysis might have to be considered.

Taken in total, this discussion serves to reinforce the thesis of this thesis: that rather than focusing on deep learning in medicine to achieve small additional improvements in accuracy, it must focus on system-level questions of robustness, understandability, and human supervision. From this angle, this thesis proposes that deep learning needs to be conceptualized not as an end in itself, but rather an auxiliary element of an ITSA, or governed, medical workflow.

VIII. Conclusion

The importance of the above study can be understood from the fact that the presented work provided a full-scale, human-centric solution for the integration of deep learning in clinical decision support systems related to the medical diagnosis of patients. Although the capabilities of deep learning have already been proven by various applications in the medical field in terms of predictive accuracy, its practical applications have not increased much because of the lack of interpretability, accuracy, accountability, and trustworthiness associated with it.

The proposed framework blends diagnostic models based on deep learning with those of explainable artificial intelligence and a human-in-the-loop validation process. By incorporating clinician validation in the decision-making process, medical responsibility can be kept with healthcare experts while harnessing the processing prowess of deep learning models. A light-weight experimental evaluation has been carried out using some freely accessible databases through the transfer learning process and it has been shown that effective diagnostic solutions can be obtained without any heavy computational processing or experimentation.

These outcomes show that pre-trained deep learning models offer stable and competitive performance when mixed with processes that offer appropriate models of validation, interpretability, and appropriate safeguards. Most importantly, it is observed in this study that solely dependent models of interpretability are not sufficient in providing safety in a medical context, instead requiring human involvement in order to counter possible concerns such as bias, mis-generalization, among other disparities. This study, therefore, provides a model that offers a concrete, reproducible model in developing trustworthy models that fill the gap between research in deep learning models in medical practices.

IX. Future Scope

A number of possible paths of future research are already being suggested from this research. First, intensive clinical validation trials with multiple patients and across different centers are to be done to evaluate their generalizability and effectiveness. Prospective trials and RCTs would also help bring stronger evidence of their efficacy and effectiveness. Second, a detailed analysis of human-AI interaction is required to bring an understanding of how and to what extent explanation mechanisms and feedback options could affect it.Future study on Adaptive and Personalized Decision Support Systems (DSS) Systems could include the development of computerized systems that generate predictions, explanations, etc., specifically tailored to each individual patient. Another possible area of future research could investigate how DSS solutions can be developed that incorporate multimodal sources of information including Medical Imaging, Genomics/Genetic Tests, and Fitness Trackers. Technical research focused on improving the safety of decision-making through improved Causality Aware Models and greater reliability in predicting an outcome (e.g., whether someone will experience an adverse event) is another area of future research.

Along with the above, many challenges and concerns remain regarding Regulation, Ethical Issues and Governance. Future Guidelines for Adaptive and Personalized DSS will have to be developed to provide for the establishment of more consistent and compatible Standards with the Med Device Regulatory Framework and Data Protection Requirements as well as Medical Accountability and Integrity. To enable the ethical and effective adoption of AI-powered Deep Learning applications for Healthcare, it will be essential to create Benchmarks and Frameworks for Audit and Reporting.

References

1. 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, no. 7639, pp. 115–118, 2017.
2. Ur Rehman Khan, Saif, Omair Bilal, Arash Hekmat, Inzamam Shahzad, and Asif Raza. "Advancing food safety: deep learning for accurate detection of bacterial contaminants." Memetic Computing 18, no. 1 (2026): 11.
3. Khan, Saif Ur Rehman, Asif Raza, Inzamam Shahzad, and Shehzad Khan. "Subcellular Structures Classification in Fluorescence Microscopic Images." International Conference on Computing & Emerging Technologies, pp. 271-286. Cham: Springer Nature Switzerland, 2023.
4. Ishfaque, Muhammad, Saif Ur Rehman Khan, and Yu-Long Lou. "Towards efficient dam inspection: crack detection via chirplet transform feature and a pruned VGG16 architecture." Memetic Computing 18, no. 1 (2026): 9.
5. G. Litjens, T. Kooi, B. E. Bejnordi, *et al.*, “A survey on deep learning in medical image analysis,” *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.
6. Khan, Zia, Saif Ur Rehman Khan, Omair Bilal, Asif Raza, and Ghazanfar Ali. "Optimizing Cervical Lesion Detection Using Deep Learning with Particle Swarm Optimization." In 2025 6th International Conference on Advancements in Computational Sciences (ICACS), pp. 1-7. IEEE, 2025.
7. Asif Raza, Inzamam Shahzad, Ghazanfar Ali, and Muhammad Hanif Soomro. "Use Transfer Learning VGG16, Inception, and Reset50 to Classify IoT Challenge in Security Domain via Dataset Bench Mark." Journal of Innovative Computing and Emerging Technologies 5, no. 1 (2025).

I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

8. Khan, Saif Ur Rehman, Muhammad Nabeel Asim, Sebastian Vollmer, and Andreas Dengel. "Temperature-driven robust disease detection in brain and gastrointestinal disorders via context-aware adaptive knowledge distillation." Biomedical Signal Processing and Control 112 (2026): 108671.
9. Khan, Saif Ur Rehman, Sohaib Asif, Ming Zhao, Wei Zou, Yangfan Li, and Chenggen Xiao. "ShallowMRI: A novel lightweight CNN with novel attention mechanism for Multi brain tumor classification in MRI images." Biomedical Signal Processing and Control 111 (2026): 108425.
10. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
11. Shahzad, Inzamam, Asif Raza, and Muhammad Waqas. "Medical Image Retrieval using Hybrid Features and Advanced Computational Intelligence Techniques." Spectrum of engineering sciences 3, no. 1 (2025): 22-65.

12. S. U. R. Khan, A. Raza, I. Shahzad and G. Ali, "Enhancing Concrete and Pavement Crack Prediction through Hierarchical Feature Integration with VGG16 and Triple Classifier Ensemble," 2024 Horizons of Information Technology and Engineering (HITE), Lahore, Pakistan, 2024, pp. 1-6.
13. P. Rajpurkar, J. Irvin, K. Zhu, *et al.*, “CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning,” *Radiology*, vol. 294, no. 3, pp. 728–735, 2020.
14. Al-Khasawneh, Mahmoud Ahmad, Asif Raza, Saif Ur Rehman Khan, and Zia Khan. "Stock Market Trend Prediction Using Deep Learning Approach." Computational Economics (2024): 1-32
15. Raza, Asif, Salahuddin, & Inzamam Shahzad. (2024). Residual Learning Model-Based Classification of COVID-19 Using Chest Radiographs. Spectrum of Engineering Sciences, 2(3), 367–396.
16. Chomba, Bwalya, Patrick Mukala, Nicanor Mayumu, and Saif Ur Rehman Khan. "DynaKG: Dynamic Knowledge Graph Attention With Learnable Temporal Decay for Recommendation." IEEE Access 13 (2025): 216956-216970.
17. M. Wajid, M. K. Abid, A. Asif Raza, M. Haroon, and A. Q. Mudasar, “Flood Prediction System Using IOT & Artificial Neural Network”, VFAST trans. softw. eng., vol. 12, no. 1, pp. 210–224, Mar. 2024.
18. Meeran, M. T., Raza, A., & Din, M. (2018). Advancement in GSM Network to Access Cloud Services. Pakistan Journal of Engineering, Technology & Science [ISSN: 2224-2333], 7(1).
19. Khan, S. R., Asif Raza, Inzamam Shahzad, & Hafiz Muhammad Ijaz. (2024). Deep transfer CNNs models performance evaluation using unbalanced histopathological breast cancer dataset. Lahore Garrison University Research Journal of Computer Science and Information Technology, 8(1).
20. Raza, Asif , Soomro, M. H., Shahzad, I., & Batool, S. (2024). Abstractive Text Summarization for Urdu Language. Journal of Computing & Biomedical Informatics, 7(02).
21. Khan, Saif Ur Rehman, Hafeez Ur Rehman, and Omair Bilal. "AI-powered cancer diagnosis: classifying viable (live) vs non-viable (dead) cells using transfer learning." Signal, Image and Video Processing 19, no. 15 (2025): 1326.
22. O. Bilal, Asif Raza, S. ur R. Khan, and Ghazanfar Ali, “A Contemporary Secure Microservices Discovery Architecture with Service Tags for Smart City Infrastructures ”, VFAST trans. softw. eng., vol. 12, no. 1, pp. 79–92, Mar. 2024
23. Khan, Muhammad Ahmad, Saif Ur Rehman Khan, Hafeez Ur Rehman, Suliman Aladhadh, and Ding Lin. "Robust InceptionV3 with Novel EYENET Weights for Di-EYENET Ocular Surface Imaging Dataset: Integrating Chain Foraging and Cyclone Aging Techniques." International Journal of Computational Intelligence Systems 18, no. 1 (2025): 204.
24. T. J. Brinker, A. Hekler, J. S. Utikal, *et al.*, “Skin cancer classification using convolutional neural networks: Systematic review,” *European Journal of Cancer*, vol. 95, pp. 1–10, 2018.
25. S. ur R. Khan, Asif. Raza, Muhammad Tanveer Meeran, and U. Bilhaj, “Enhancing Breast Cancer Detection through Thermal Imaging and Customized 2D CNN Classifiers”, VFAST trans. softw. eng., vol. 11, no. 4, pp. 80–92, Dec. 2023.
26. Raza, A., & Meeran, M. T. (2019). Routine of Encryption in Cognitive Radio Network. Mehran University Research Journal of Engineering and Technology [p-ISSN: 0254-7821, e-ISSN: 2413-7219], 38(3), 609-618.

27. R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, “Deep patient: An unsupervised representation to predict the future of patients from the electronic health records,” *Scientific Reports*, vol. 6, Art. no. 26094, 2016.
28. HUSSAIN, S., Raza, A., MEERAN, M. T., IJAZ, H. M., & JAMALI, S. (2020). Domain Ontology Based Similarity and Analysis in Higher Education. IEEEP New Horizons Journal, 102(1), 11-16.
29. Yang, Huihui, Saif Ur Rehman Khan, Omair Bilal, Chao Chen, and Ming Zhao. "CEOE-Net: Chaotic Evolution Algorithm-Based Optimized Ensemble Framework Enhanced with Dual-Attention for Alzheimer’s Diagnosis." Computer Modeling in Engineering & Sciences 145, no. 2 (2025): 2401.
30. Maqsood, Hasaan, and Saif Ur Rehman Khan. "MeD-3D: A multimodal deep learning framework for precise recurrence prediction in clear cell renal cell carcinoma (ccRCC)." Expert Systems with Applications (2025): 130174.
31. Mahmood, F., Abbas, K., Raza, A., Khan,M.A., & Khan, P.W. (2019). Three Dimensional Agricultural Land Modeling using Unmanned Aerial System (UAS). International Journal of Advanced Computer Science and Applications (IJACSA) [p-ISSN : 2158-107X, e-ISSN : 2156-5570], 10(1).
32. Hekmat, Arash, Zuping Zhang, Saif Ur Rehman Khan, and Omair Bilal. "Brain tumor diagnosis redefined: Leveraging image fusion for MRI enhancement classification." Biomedical Signal Processing and Control 109 (2025): 108040.
33. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, “Deep EHR: A survey of recent advances in deep learning techniques for electronic health record analysis,” *Journal of Biomedical Informatics*, vol. 83, pp. 168–185, 2018.
34. F. Doshi-Velez and B. Kim, “Towards a rigorous science of interpretable machine learning,” *arXiv preprint arXiv:1702.08608*, 2017.
35. Bilal, Omair, Arash Hekmat, Inzamam Shahzad, Asif Raza, and Saif Ur Rehman Khan. "Boosting Machine Learning Accuracy for Cardiac Disease Prediction: The Role of Advanced Feature Engineering and Model Optimization." The Review of Socionetwork Strategies 19, no. 2 (2025): 271-300.
36. Khan, Saif Ur Rehman, Ming Zhao, and Yangfan Li. "Detection of MRI brain tumor using residual skip block based modified MobileNet model." Cluster Computing 28, no. 4 (2025): 248.
37. Bilal, Omair, Arash Hekmat, and Saif Ur Rehman Khan. "Automated cervical cancer cell diagnosis via grid search-optimized multi-CNN ensemble networks." Network Modeling Analysis in Health Informatics and Bioinformatics 14, no. 1 (2025): 67.
38. W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, “Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models,” *IEEE Signal Processing Magazine*, vol. 38, no. 3, pp. 39–55, 2021.
39. Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, “Dissecting racial bias in an algorithm used to manage the health of populations,” *Science*, vol. 366, no. 6464, pp. 447–453, 2019.
40. Khan, Saif Ur Rehman. "Multi-level feature fusion network for kidney disease detection." Computers in Biology and Medicine 191 (2025): 110214.
41. Khan, Saif Ur Rehman, Sohaib Asif, Omair Bilal, and Hafeez Ur Rehman. "Lead-cnn: lightweight enhanced dimension reduction convolutional neural network for brain tumor classification." International Journal of Machine Learning and Cybernetics (2025): 1-20.

42. Khan, Saif Ur Rehman, Sohaib Asif, and Omair Bilal. "Ensemble Architecture of Vision Transformer and CNNs for Breast Cancer Tumor Detection From Mammograms." International Journal of Imaging Systems and Technology 35, no. 3 (2025): e70090.
43. Khan, Saif Ur Rehman, and Zia Khan. "Detection of Abnormal Cardiac Rhythms Using Feature Fusion Technique with Heart Sound Spectrograms." Journal of Bionic Engineering (2025): 1-20.
44. J. R. Zech, M. A. Badgeley, M. Liu, *et al.*, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs," *PLOS Medicine*, vol. 15, no. 11, e1002683, 2018.
45. E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, no. 1, pp. 44–56, 2019.
46. J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, "Explainability for artificial intelligence in healthcare: A multidisciplinary perspective," *Journal of Medical Ethics*, vol. 46, no. 7, pp. 421–428, 2020.
47. Hekmat, Arash, Zuping Zhang, Saif Ur Rehman Khan, Ifza Shad, and Omair Bilal. "An attention-fused architecture for brain tumor diagnosis." Biomedical Signal Processing and Control 101 (2025): 107221.
48. Khan, Saif Ur Rehman, Sohaib Asif, Ming Zhao, Wei Zou, and Yangfan Li. "Optimize brain tumor multiclass classification with manta ray foraging and improved residual block techniques." Multimedia Systems 31, no. 2 (2025): 88.
49. E. H. Shortliffe and M. J. Sepúlveda, "Clinical decision support in the era of artificial intelligence," *JAMA*, vol. 320, no. 21, pp. 2199–2200, 2018.
50. Waqas M, Bandyopadhyay R, Showkatian E, Muneer A, Zafar A, Alvarez FR, Marin MC, Li W, Jaffray D, Haymaker C, Heymach J. The Next Layer: Augmenting Foundation Models with Structure-Preserving and Attention-Guided Learning for Local Patches to Global Context Awareness in Computational Pathology. arXiv preprint arXiv:2508.19914. 2025 Aug 27.
51. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. ACM SIGKDD*, 2016, pp. 1135–1144.
52. Waqas M, Ahmed SU, Tahir MA, Wu J, Qureshi R. Exploring multiple instance learning (MIL): A brief survey. Expert Systems with Applications. 2024 Sep 15;250:123893.
53. Waqas M, Tahir MA, Khan SA. Robust bag classification approach for multi-instance learning via subspace fuzzy clustering. Expert Systems with Applications. 2023 Mar 15;214:119113.
54. R. R. Selvaraju, M. Cogswell, A. Das, *et al.*, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE ICCV*, 2017, pp. 618–626.
55. E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *Neurocomputing*, vol. 464, pp. 529–550, 2021.
56. M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *BMJ Health & Care Informatics*, vol. 28, no. 1, 2021.

A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *WIREs Data Mining and Knowledge Discovery*, vol. 9, no. 4, 2019.

57. Waqas M, Tahir MA, Qureshi R. Deep Gaussian mixture model based instance relevance estimation for multiple instance learning applications. Applied intelligence. 2023 May;53(9):10310-25.

58. Waqas M, Tahir MA, Al-Maadeed S, Bouridane A, Wu J. Simultaneous instance pooling and bag representation selection approach for multiple-instance learning (MIL) using vision transformer. Neural Computing and Applications. 2024 Apr;36(12):6659-80.

59. J. Kelly, A. Karthikesalingam, C. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *The Lancet Digital Health*, vol. 1, no. 4, pp. e156–e163, 2019.

60. Y. Park, J. J. Hu, and M. K. McCourt, "Human–AI collaboration in clinical decision-making," *IEEE Access*, vol. 8, pp. 197241–197253, 2020.

61. Waqas M, Tahir MA, Qureshi R. Ensemble-based instance relevance estimation in multiple-instance learning. In2021 9th European workshop on visual information processing (EUVIP) 2021 Jun 23 (pp. 1-6). IEEE.

62. Asif Raza, Salahuddin, Ghazanfar Ali, Muhammad Hanif Soomro, Saima Batool, "Analyzing the Impact of Artificial Intelligence on Shaping Consumer Demand in E-Commerce: A Critical Review", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.17, No.5, pp. 42-61, 2025.

63. Raza, Asif, Inzamam Shahzad, Muhammad Salahuddin, and Sadia Latif. "Satellite Imagery Employed to Analyze the Extent of Urban Land Transformation in The Punjab District of Pakistan." Journal of Palestine Ahliya University for Research and Studies 4, no. 2 (2025): 17-36.

64. K.-H. Yu, A. L. Beam, and I. S. Kohane, "Artificial intelligence in healthcare," *Nature Biomedical Engineering*, vol. 2, no. 10, pp. 719–731, 2018.

65. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2009.