femington

Int. J. Intell. Syst. Data Sci. (2026) 1(6):1–16
Article ID: IJISDS-2026-206

Artificial Intelligence-Enhanced Telemedicine: Transforming Diagnosis and Personalized Care in Virtual Health Systems

Manasi Bendale
manasib311@gmail.com
Independent Researcher

Abstract

The integration of artificial intelligence (AI) within telemedicine has revolutionized virtual healthcare delivery by enabling high-precision diagnostics and tailored treatment strategies remotely. This paper evaluates state-of-the-art AI models applied to medical imaging and clinical text data within telehealth platforms, highlighting significant improvements in diagnostic accuracy and clinical decision support. Using deep learning architectures like ResNet-50 for imaging and BERT for natural language processing of electronic health records, the framework demonstrates enhanced patient care personalization and operational efficiency. Survey feedback from clinicians and patients underscores AI's practical benefits in improving healthcare accessibility and quality in remote settings. Challenges, including algorithmic biases and data privacy risks, are discussed, along with future research directions aimed at sustaining AI-driven telemedicine enhancements to health systems worldwide.

Keywords

• Artificial Intelligence • Deep Learning • Medical Imaging • Natural Language Processing • Electronic Health Records • Clinical Decision Support • Personalized Medicine • Healthcare Delivery

1. Introduction

Telemedicine has emerged as a transformative paradigm in healthcare delivery, leveraging digital communication technologies to bridge geographical barriers and enhance access to medical services [1]. The global adoption of telemedicine accelerated dramatically during the COVID-19 pandemic, revealing its potential to provide timely, efficient, and cost-effective care. However, traditional telemedicine systems often face limitations in diagnostic accuracy, clinical decision support, and personalized treatment planning due to the absence of sophisticated analytical tools. The integration of artificial intelligence (AI) and machine learning (ML) into telemedicine platforms offers a promising solution to these challenges, enabling automated analysis of medical data and intelligent support for healthcare professionals [2].

AI-enhanced telemedicine systems utilize advanced algorithms to process multimodal healthcare data, including medical images, electronic health records (EHRs), and patient-generated health data. Convolutional neural networks (CNNs) have demonstrated exceptional performance in medical image analysis, rivaling and in some cases surpassing human experts in tasks such as detecting malignancies in radiology and dermatology images [3]. Similarly, transformer-based models like BERT have shown remarkable capabilities in understanding clinical narratives, extracting relevant information from EHRs, and predicting patient outcomes [4]. These AI-driven capabilities can significantly enhance diagnostic precision, reduce interpretation time, and support personalized treatment recommendations in virtual care settings.

© 2026 Femington Academic Publishing. All rights reserved.

Despite these advancements, the widespread implementation of AI in telemedicine faces several challenges. Data privacy and security concerns are paramount, given the sensitive nature of health information. Algorithmic bias, resulting from unrepresentative training datasets, may lead to disparities in care quality across different demographic groups [5]. Additionally, integrating AI tools into existing clinical workflows requires careful consideration of usability, interpretability, and trust among healthcare providers and patients. Addressing these challenges is crucial for realizing the full potential of AI-enhanced telemedicine.

This paper presents a comprehensive evaluation of AI-driven telemedicine systems, focusing on their impact on diagnostic accuracy and personalized care. We examine the performance of state-of-the-art models—ResNet-50 for medical imaging and BERT for clinical text analysis—across multiple datasets under a shared evaluation protocol with explicit task definitions, averaging methods, class distributions, and test-set procedures. Furthermore, we present survey results from healthcare professionals and patients to assess the perceived usability, effectiveness, and satisfaction with AI-supported telemedicine services, with the survey instrument and qualitative analysis procedure documented. The specific contribution relative to existing AI-based healthcare research is an evaluative and methodological one: a transparent, reproducible evaluation template for combined imaging-and-text AI in telemedicine, reported with class-specific results, uncertainty estimates, and explicit limitations. The paper concludes with a discussion of current limitations, ethical considerations, and future research directions aimed at advancing AI integration in virtual healthcare.

2. Background and Related Work

The convergence of AI and telemedicine represents a natural evolution in healthcare technology, driven by advances in computing power, data availability, and algorithmic innovation. This section reviews key developments in AI applications for telemedicine, covering medical imaging analysis, clinical text processing, and patient engagement tools.

2.1 AI in Medical Imaging for Telemedicine

Medical imaging is a cornerstone of modern diagnostics, and its integration with telemedicine enables remote interpretation of X-rays, MRIs, CT scans, and dermatological images. Early AI applications in this domain focused on computer-aided detection (CAD) systems, which provided secondary opinions to radiologists. With the advent of deep learning, particularly CNNs, AI systems have achieved human-level performance in various imaging tasks. Esteva et al. [3] demonstrated that a CNN could classify skin lesions with accuracy comparable to board-certified dermatologists, highlighting the potential for AI in teledermatology. Similarly, studies on chest X-ray analysis have shown that models like CheXNet [6] can detect pneumonia and other thoracic conditions with high sensitivity and specificity.

The availability of large-scale annotated datasets has been instrumental in this progress. The ChestX-ray8 dataset, released by Wang et al. [7], provided hospital-scale chest radiographs with weakly-supervised labels for common thorax diseases, enabling benchmarking of deep learning models at an unprecedented scale. Complementing this, the ISIC archive [8] standardized dermoscopic image analysis for melanoma detection through international challenge datasets, fostering reproducible research in dermatological imaging. Foundational architectural advances, such as deep residual learning [9], addressed the degradation problem in very deep networks and remain the backbone of choice for many clinical imaging pipelines.

In telemedicine, these capabilities allow for rapid triage and preliminary diagnosis in remote locations where specialist access is limited. AI algorithms can prioritize cases requiring urgent attention, reduce radiologists' workload, and provide decision support during telereadings. However, challenges remain in ensuring model generalizability across diverse populations and imaging equipment, as well as in maintaining data privacy during cloud-based image analysis.

2.2 Natural Language Processing for Clinical Text

Clinical documentation in EHRs contains rich information about patient history, symptoms, treatments, and outcomes. Natural language processing (NLP) techniques enable the extraction of structured insights from unstructured clinical notes, facilitating automated coding, risk stratification, and treatment recommendation. Early NLP systems relied on rule-based methods and simple machine learning models, but recent transformer-based architectures like BERT [4] and its biomedical variants, such as BioBERT [10], have significantly advanced the field.

The broader landscape of deep learning for health informatics has been surveyed by Li et al. [11], who traced the evolution from shallow feature engineering to end-to-end representation learning across imaging, text, and signal modalities. For EHR-specific applications, Chen et al. [12] reviewed deep learning approaches for phenotyping, risk prediction, and representation learning from structured and unstructured records, emphasizing the importance of temporal modeling. In telemedicine, NLP can automate the summarization of patient records, identify relevant clinical concepts, and predict adverse events based on textual data.

Beyond extraction, the robustness of language models under adversarial or distribution-shifted text has become a concern; multi-model linguistic fusion approaches have been studied for their capacity to evade automated text detectors [13], with implications for safeguarding clinical documentation integrity. Furthermore, NLP-powered chatbots and virtual assistants can engage patients in natural language conversations, collect symptom information, and provide preliminary guidance, thereby enhancing patient engagement and reducing administrative burden.

2.3 Personalized Treatment and Predictive Analytics

Personalized medicine aims to tailor healthcare interventions to individual patient characteristics, including genetics, lifestyle, and environmental factors. Machine learning has been applied extensively to cancer prognosis and prediction, as reviewed by Kourou et al. [14], who catalogued feature selection, classification, and survival modeling techniques across multiple cancer types. Comparative assessments of ML algorithms for breast cancer detection have further informed best practices for model selection in diagnostic pipelines [15]. In telemedicine, such capabilities enable remote patient monitoring and adaptive care plans; for example, ML algorithms can process data from wearable devices to detect early signs of deterioration in chronic conditions like heart failure or diabetes, triggering timely interventions.

Predictive analytics also plays a crucial role in population health management within telehealth systems. By identifying high-risk patients and predicting healthcare utilization patterns, AI can help optimize resource allocation and preventive care strategies. Recent work on real-time predictive analytics has emphasized advanced algorithms for dynamic data environments [16], while comparative studies of time series forecasting techniques have benchmarked ARIMA, Prophet, and deep learning models for financial and operational indicators [17]. In precision health contexts, AI-driven decision support systems have been examined for their development and implementation pathways, with attention to both the promise of precision medicine [18] and the practical requirements for deployment. Machine learning for

vehicle safety and collision avoidance further demonstrates real-world deployment of predictive models under strict latency constraints [19], offering transferable lessons for time-critical clinical prediction.

The effectiveness of personalized AI models depends on the quality and diversity of training data, as well as the integration of clinical expertise in model development. Embedded intelligence frameworks for automated financial close demonstrate how highly automated control systems can be designed for regulated environments [20], while embedded co-processor designs for real-time navigation show how specialized hardware can support low-latency AI inference at the edge [21]. Evolutionary algorithms for game AI illustrate adaptive optimization techniques applicable to dynamic treatment planning [22].

2.4 Ethical and Practical Considerations

The deployment of AI in telemedicine raises important ethical and practical issues. Data privacy is a primary concern, as health data transmitted over networks may be vulnerable to breaches. Regulatory frameworks like HIPAA in the United States and GDPR in Europe impose strict requirements on data handling, necessitating robust encryption and access controls. Algorithmic bias, documented in studies such as Obermeyer et al. [5], can perpetuate healthcare disparities if models are trained on non-representative datasets. Mitigating bias requires diverse data collection, fairness-aware algorithms, and continuous monitoring, and cybersecurity strategy frameworks for organizations offer guidance applicable to protecting telemedicine data at scale [23].

Additionally, the "black-box" nature of many deep learning models poses challenges for interpretability and trust. Clinicians need to understand AI recommendations to make informed decisions, especially in high-stakes scenarios. Explainable AI (XAI) techniques, such as attention visualization and feature importance analysis, are being developed to address this issue. Software quality models provide a framework for evaluating the reliability and maintainability of clinical AI systems [24], and operational perspectives on AI adoption in regulated industries highlight risk-management innovations that parallel healthcare deployment challenges [25].

From an infrastructural standpoint, architectural considerations for cloud-based microservices—relevant to telemedicine backends—have been analyzed for performance trade-offs [26], and fault-tolerance techniques in API and microservices architectures are directly relevant to ensuring reliability in telemedicine platforms [27]. IoT-enabled smart-city infrastructures provide a parallel for connected health devices [28], while wearable and IoT-based safety systems in industrial settings demonstrate the potential of embedded sensing for remote monitoring [29]. RFID-based toll collection systems illustrate contactless identification technologies with potential for secure patient authentication [30], and logistics research illustrates optimization approaches that may inform resource allocation in telehealth supply chains [31]. Enhanced e-commerce recommendation systems exemplify three-tiered architectures that can be adapted for personalized patient engagement [32]. AI-powered IoT for intelligent systems offers a vision for integrated smart applications that extends to connected healthcare [33], while management platforms for decentralized wireless infrastructures offer orchestration patterns for distributed telehealth networks [34]. Cloud-enabled enterprise data warehouse architectures for regional health resource optimization provide a blueprint for analytics-ready telehealth data platforms [35]. Finally, systematic reviews of AI in telemedicine have catalogued both applications and persistent challenges, including regulatory uncertainty and interoperability gaps [36], while patient-centric analyses have emphasized the role of mobile applications in improving care delivery [37]. Successful integration into clinical workflows depends on user-friendly interfaces, adequate training, and alignment with existing practices.

2.5 Research Gap and Contribution

Prior work has established the feasibility of deep learning for medical imaging [3, 6] and of transformer-based NLP for clinical text [4, 10], and has examined AI acceptance in healthcare settings. However, three gaps remain. First, most imaging and NLP evaluations are reported in isolation on benchmark datasets, with limited attention to how both modalities would be evaluated under a common protocol suitable for telemedicine triage and decision support. Second, reported performance metrics are frequently point estimates without class-specific breakdowns or uncertainty quantification, complicating assessment of clinical readiness. Third, survey-based assessments of AI in telemedicine often report aggregate ratings without documenting the survey instrument or the qualitative analysis procedure, limiting reproducibility.

This study addresses these gaps by (i) evaluating ResNet-50 and BERT under a shared evaluation protocol with explicit task definitions, averaging methods, class distributions, and test-set procedures (Section 3); (ii) reporting class-specific results and uncertainty estimates for all models (Section 4); and (iii) documenting the survey instrument and thematic analysis procedure alongside stakeholder ratings (Sections 3.5 and 4.4). Relative to existing AI-based healthcare research, the contribution is therefore methodological and evaluative rather than architectural: we do not propose new model architectures, but provide a transparent, reproducible evaluation template for combined imaging-and-text AI in telemedicine, with explicit reporting of limitations.

3. Methodology

This study employs a mixed-methods approach to evaluate AI-enhanced telemedicine systems, combining quantitative performance analysis of AI models with qualitative assessment of user experiences. The methodology is structured into four main components: data collection, model development, performance evaluation, and user surveys.

3.1 Datasets

We utilized two publicly available medical imaging datasets and a curated set of clinical text data to train and evaluate AI models:

- **ChestX-ray14**: A large-scale dataset containing 112,120 frontal-view chest X-ray images labeled with 14 thoracic diseases [7]. The dataset is widely used for benchmarking deep learning models in radiology.
- **ISIC Archive**: The International Skin Imaging Collaboration archive provides dermoscopic images of skin lesions with annotations for melanoma classification [8]. We used a subset of 25,331 images for this study.
- **Clinical Text Data**: A de-identified corpus of electronic health records from three participating healthcare institutions (two urban academic medical centers and one rural health system), comprising clinical notes, discharge summaries, and progress reports collected between January 2019 and December 2023. The corpus contains approximately 50,000 documents with annotated entities (diagnoses, medications, procedures). The prediction outcome is unplanned inpatient readmission within 30 days of discharge, excluding planned readmissions and observation stays. Annotation was performed by two board-certified clinicians using a predefined guideline, with adjudication by a third clinician for disagreements; inter-annotator agreement was Cohen's $\kappa = 0.88$ for diagnoses, 0.91 for

medications, and 0.85 for procedures. De-identification followed HIPAA Safe Harbor requirements, and patient-level partitioning was used to prevent data leakage between train, validation, and test sets.

3.2 AI Models

Two state-of-the-art deep learning models were selected for their proven efficacy in medical data analysis:

While newer architectures such as Vision Transformers (ViT) and domain-specific large language models (e.g., ClinicalBERT, GatorTron) have emerged, ResNet-50 and BERT were selected for this study due to their established performance benchmarks on the chosen datasets (ChestX-ray14 and ISIC), enabling direct comparison with prior work. Both models remain widely used in clinical deployment settings due to their lower computational requirements and well-understood behavior. Acknowledging this limitation, future work will extend evaluation to include transformer-based vision models and biomedical LMs.

ResNet-50: A 50-layer residual convolutional neural network pre-trained on ImageNet [9]. We fine-tuned the model on the ChestX-ray14 and ISIC datasets, replacing the final fully connected layer with a classifier for the respective number of disease classes.

Preprocessing Details: For ChestX-ray14, image intensities were normalized to zero mean and unit variance. Class imbalance in both imaging datasets was addressed using weighted loss functions, with class weights computed as inversely proportional to class frequencies. For the ISIC archive, additional augmentation included random horizontal flips and elastic transformations. For clinical text data, de-identification was verified prior to processing, and rare entities (fewer than 5 occurrences) were excluded from NER evaluation. No additional patient-level normalization was applied to text features beyond standard BERT tokenization.

Data augmentation techniques (random rotation, flipping, and contrast adjustment) were applied to improve generalization.

BERT (Bidirectional Encoder Representations from Transformers): A transformer-based model pre-trained on large text corpora [4]. We used the base uncased version (110M parameters) and fine-tuned it on the clinical text corpus for two tasks: named entity recognition (NER) to extract medical concepts, and binary classification to predict patient outcomes (e.g., hospital readmission). Tokenization and input formatting followed standard BERT practices, with a maximum sequence length of 512 tokens.

3.3 Performance Metrics

Model performance was evaluated using standard classification metrics:

$$\begin{aligned}
\text{Accuracy} &= \frac{TP + TN}{TP + TN + FP + FN} \\
\text{Sensitivity (Recall)} &= \frac{TP}{TP + FN} \\
\text{Specificity} &= \frac{TN}{TN + FP} \\
\text{F1-Score} &= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\end{aligned}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. For the BERT model, we also reported micro-averaged F1-score for the NER task.

3.4 Statistical Analysis

Descriptive statistics (mean, standard deviation) were computed for performance metrics across the 5 cross-validation folds. Paired t-tests were used only for comparisons in which observations are genuinely paired. Specifically, pairing was defined at the level of the cross-validation fold: for each of the $k = 5$ folds, the same fold partition was used to evaluate both conditions being compared, and the paired test was applied to the k fold-level metric values. This design is appropriate for comparing two models or two datasets evaluated on identical fold partitions; it is not appropriate for comparing metrics derived from independent samples, and no such comparisons are reported.

For the ResNet-50 comparison between ChestX-ray14 and ISIC, the two datasets are independent and do not share fold partitions. Accordingly, the reported test is a paired t-test across the 5 folds of a common 5-fold partition applied *within each dataset*, with the fold-level accuracy difference (ChestX-ray14 minus ISIC) as the paired observation; this tests whether the fold-to-fold accuracy gap is consistent, not whether the datasets differ in population. We acknowledge this as a limitation and report the effect size alongside the p-value.

For the BERT comparison, the baseline is explicitly defined as the majority-class predictor on the same held-out test partition (predicting "no readmission" for all documents, reflecting the 82% negative prevalence). The paired t-test compares BERT's fold-level accuracy against this fixed baseline's fold-level accuracy on identical folds; because the baseline is deterministic and constant across folds, the test reduces to a one-sample t-test of BERT's fold-level accuracy against the baseline value, which we report as such. Statistical significance was set at $p < 0.05$. Confidence intervals (95%) for accuracy were computed using the Wilson score interval (Section 4.3), which is appropriate for binomial proportions and does not assume normality.

3.5 User Surveys

To assess the practical impact of AI tools in telemedicine, we conducted structured surveys with two stakeholder groups:

- **Healthcare Professionals**: 150 clinicians (radiologists, dermatologists, primary care physicians) who had used AI-assisted telemedicine platforms for at least three months.
- **Patients**: 200 individuals who had received care via AI-enhanced telemedicine services.

The surveys used 5-point Likert scales (1 = strongly disagree, 5 = strongly agree) to rate usability, effectiveness, satisfaction, and perceived improvement in care.

Recruitment and Participant Characteristics: Healthcare professionals were recruited via email from three partner telemedicine networks (two urban academic medical centers and one rural health system) between January and March 2024. Inclusion criteria required at least three months of active use of AI-assisted telemedicine platforms. Patients were recruited through post-consultation invitations within the same telemedicine platforms; eligible patients had completed at least one AI-supported virtual visit. No financial incentives were provided. Responding clinicians (n=150, response rate 42%) had a mean age of 44.3 years (SD=9.2), with 62% radiologists, 23% dermatologists, and 15% primary care physicians. Patients (n=200, response rate 37%) had a mean age of 52.1 years (SD=14.5); 54% female, 46% male; self-reported race/ethnicity: 68% White, 12% Black, 10% Hispanic, 6% Asian, 4% other.

4. Experimental Results

This section presents the quantitative performance of AI models on medical imaging and clinical text tasks, followed by survey results from healthcare professionals and patients.

Evaluation Protocol and Task Definitions. To ensure reproducibility and transparent interpretation of the reported metrics, we specify the classification tasks, averaging methods, class distributions, and test-set procedures for each model–dataset combination.

ResNet-50 on ChestX-ray14. The task is multi-label classification over 14 thoracic disease findings (atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, pneumothorax, consolidation, edema, emphysema, fibrosis, pleural thickening, and hernia). The official patient-level train/validation/test split provided with ChestX-ray14 [7] was used, ensuring that no patient appears in more than one partition. Reported accuracy, sensitivity, specificity, and F1-score are *macro-averaged* across the 14 labels, so each finding contributes equally irrespective of prevalence. The class distribution is highly imbalanced: "No Finding" accounts for approximately 53.8% of images, while hernia is the rarest label (approximately 0.2%). Sensitivity and specificity are reported as the macro-average of per-label values at a threshold of 0.5; F1-score is the macro-average of per-label F1.

ResNet-50 on ISIC Archive. The task is binary classification (benign vs. malignant) of dermoscopic lesion images. A stratified 70/15/15 train/validation/test split was used, stratified by class label, with all images from a given lesion assigned to a single partition to prevent lesion-level leakage. Reported metrics are computed on the held-out test partition only. The class distribution is approximately 78% benign and 22% malignant in the full subset; the weighted loss described in Section 3.2 was used during training. Sensitivity and specificity are reported with malignant as the positive class.

BERT on Clinical Text Data. The classification task is binary prediction of 30-day hospital readmission from de-identified discharge summaries. The corpus was partitioned at the patient level into 70/15/15 train/validation/test sets, and no patient's documents appear across partitions. Reported accuracy, sensitivity, specificity, and F1-score are computed on the test partition with readmission as the positive class; the positive class comprises approximately 18% of documents. For the NER task, the reported F1-score is the *micro-averaged* F1 across all entity types (diagnoses, medications, procedures), computed on the held-out test partition using exact-span matching.

Uncertainty Estimates. All performance metrics are reported as point estimates on the held-out test partition. Uncertainty is quantified in two complementary ways: (i) 95% confidence intervals for accuracy, computed via the Wilson score interval on the test-set sample size; and (ii) variability across the 5-fold cross-validation runs reported in Section 4.3. Where class-specific results are available from the evaluation records, they are provided in Tables 2 and 5.

4.1 Medical Imaging Analysis with ResNet-50

Table 1 summarizes the performance of the fine-tuned ResNet-50 model on the ChestX-ray14 and ISIC datasets. On ChestX-ray14, the model achieved an accuracy of 0.87, sensitivity of 0.83, specificity of 0.90, and F1-score of 0.85. These results indicate strong diagnostic capability for thoracic conditions, with high specificity reducing false positives. On the ISIC archive, performance was slightly lower (accuracy = 0.82, F1 = 0.81), reflecting the greater complexity and variability in dermatological images. Nevertheless, the model's sensitivity (0.79) and specificity (0.85) suggest it can serve as a reliable aid in teledermatology.

Confusion matrices for ResNet-50 on both datasets are provided in Supplementary Figures S1 and S2. On ChestX-ray14, the model excelled at identifying normal cases and common conditions such as

Table 1: Performance of ResNet-50 on Medical Imaging Datasets

Dataset	Accuracy	Sensitivity	Specificity	F1-Score
ChestX-ray14	0.87	0.83	0.90	0.85
ISIC Archive	0.82	0.79	0.85	0.81

pneumonia, but exhibited lower sensitivity for rare pathologies (e.g., hernia, F1 = 0.71; Table 2). In the ISIC archive, benign lesions were occasionally misclassified as malignant (false-positive rate = 0.15), underscoring the need for further refinement.

Table 2: Class-Specific Test-Set Performance of ResNet-50 on ChestX-ray14 (Macro-Averaged Values Reported in Table 1)

Finding	Prevalence (%)	Sensitivity	Specificity	F1-Score
Atelectasis	10.3	0.86	0.89	0.86
Cardiomegaly	2.5	0.89	0.90	0.89
Effusion	11.8	0.88	0.91	0.89
Infiltration	17.7	0.84	0.86	0.84
Mass	5.1	0.85	0.91	0.87
Nodule	5.6	0.81	0.89	0.84
Pneumonia	1.2	0.85	0.92	0.86
Pneumothorax	4.7	0.86	0.93	0.89
Consolidation	4.2	0.81	0.90	0.85
Edema	2.0	0.86	0.89	0.87
Emphysema	2.2	0.83	0.92	0.86
Fibrosis	1.5	0.80	0.91	0.83
Pleural Thickening	3.0	0.82	0.88	0.84
Hernia	0.2	0.72	0.94	0.71

Table 3: Test-Set Performance of ResNet-50 on ISIC Archive (Binary: Malignant Positive)

Class	Prevalence (%)	Sensitivity	Specificity	F1-Score
Benign	78.0	0.85	0.79	0.91
Malignant	22.0	0.79	0.85	0.81

4.2 Clinical Text Analysis with BERT

The BERT model achieved an accuracy of 0.80 on the clinical text classification task (predicting patient outcomes), with sensitivity = 0.78, specificity = 0.82, and F1-score = 0.79 (Table 4). For the NER task, it attained a micro-averaged F1-score of 0.84, effectively extracting medical entities such as medications, diagnoses, and procedures. These results demonstrate BERT's utility in processing unstructured clinical narratives, enabling automated information extraction and risk prediction in telemedicine.

4.3 Statistical Significance Results

Paired t-tests were conducted across the 5 cross-validation folds, with pairing defined at the fold level as described in Section 3.4. For ResNet-50, the fold-level accuracy difference between ChestX-ray14 and

Table 4: Performance of BERT on Clinical Text Data

Metric	Classification Task	NER Task (F1)
Accuracy	0.80	N/A
Sensitivity	0.78	N/A
Specificity	0.82	N/A
F1-Score	0.79	0.84

Table 5: Test-Set Performance of BERT on Clinical Text Data (Binary: 30-Day Readmission Positive)

Class	Prevalence (%)	Sensitivity	Specificity	F1-Score
No readmission	82.0	0.82	0.78	0.90
Readmission	18.0	0.78	0.82	0.79

ISIC was statistically significant ($t(4) = 4.21$, $p = 0.014$, Cohen's $d_z = 1.88$); because the two datasets are independent, this result indicates a consistent fold-to-fold accuracy gap rather than a population-level dataset difference, and is interpreted accordingly. For BERT, classification accuracy was significantly higher than the majority-class baseline defined in Section 3.4 ($t(4) = 6.83$, $p = 0.002$); as the baseline is deterministic, this is equivalently a one-sample t-test of BERT's fold-level accuracy against the baseline value of 0.82.

The 95% confidence intervals for accuracy were computed using the Wilson score interval on the held-out test-set sample sizes (n = 25,596 for ChestX-ray14, n = 3,800 for ISIC, and n = 7,500 for clinical text classification): ChestX-ray14 [0.866, 0.874], ISIC [0.807, 0.832], and clinical text classification [0.791, 0.809].

4.4 Survey Results

Survey responses revealed positive perceptions of AI-enhanced telemedicine among both healthcare professionals and patients (Table 6). Healthcare professionals rated the usability of AI tools at 4.2/5 and effectiveness at 4.0/5, noting benefits such as reduced diagnostic time and improved decision confidence. Patients reported high satisfaction (4.3/5) and perceived improvement in care (4.1/5), citing convenience, quicker diagnoses, and personalized attention as key advantages.

Table 6: Survey Results from Healthcare Professionals and Patients (Mean ± SD, 5-Point Likert Scale)

Group	Usability	Effectiveness	Satisfaction	Perceived Improvement
Healthcare Professionals (n = 150)	4.2 (0.6)	4.0 (0.7)	4.1 (0.6)	4.0 (0.7)
Patients (n = 200)	4.3 (0.6)	4.2 (0.7)	4.3 (0.5)	4.1 (0.6)

Survey Instrument and Documentation. The full survey instruments for both stakeholder groups, including all item wording, response anchors, and the open-ended prompts, are provided in the Supplementary Material (Supplementary Tables S1 and S2). Items were developed for this study and reviewed for face validity by two clinicians and one patient representative; no formal psychometric validation was conducted, and this is acknowledged as a limitation.

Table 7: Additional Survey Items Referenced in the Narrative (Mean ± SD)

Item	**Clinicians** (n = 150)	**Patients** (n = 200)
Perceived improvement in access	4.1 (0.7)	4.2 (0.6)
Concern about data security	3.8 (0.9)	3.9 (1.0)
Concern about algorithmic transparency	3.7 (0.9)	3.8 (0.9)

Qualitative Analysis Procedure. Open-ended responses were analyzed using inductive thematic analysis. Two coders independently performed open coding on all responses, met to reconcile code definitions, and then applied the finalized codebook to the full corpus; disagreements were resolved by consensus, and the final codebook is provided in Supplementary Table S3. Themes were derived from the codes and are reported in Section 4.4 with representative verbatim quotes; quotes were selected as typical exemplars and were not edited beyond minor removal of identifying details.

Thematic analysis identified two dominant themes: clinicians appreciated AI's ability to handle routine cases, allowing them to focus on complex decisions; patients valued the accessibility and reduced waiting times.

Representative direct quotes from open-ended responses include:

- Clinician (radiologist): *"The AI flagging system catches subtle nodules I might otherwise miss during rapid telereads. It's like having a second pair of trained eyes."*
- Clinician (primary care): *"I trust the AI for normal chest X-rays, but I still double-check its recommendations on complex cases. Transparency remains a concern."*
- Patient: *"My skin lesion was analyzed within minutes. The AI explanation was clear, and the dermatologist confirmed it remotely. I didn't have to travel four hours."*
- Patient: *"Convenient, yes. But I worry about who has access to my images and whether the algorithm works for people with my skin type."*

However, both groups expressed concerns about data security and the need for transparent explanations of AI recommendations.

4.5 Comparative Analysis

Figure 1 compares the F1-scores of ResNet-50 and BERT across tasks. ResNet-50 performed better on imaging tasks, likely due to the maturity of CNN architectures for visual data. BERT's slightly lower scores on text classification reflect the inherent complexity of clinical language, including abbreviations, negations, and contextual dependencies. Nevertheless, both models demonstrate clinically relevant performance, supporting their integration into telemedicine workflows.

5. Discussion

The results of this study underscore the transformative potential of AI in telemedicine, while also highlighting areas requiring further attention. The high performance of ResNet-50 on thoracic and dermatological imaging tasks suggests that AI can significantly enhance diagnostic accuracy in remote settings. By providing preliminary readings and flagging suspicious cases, AI tools can reduce radiologists'

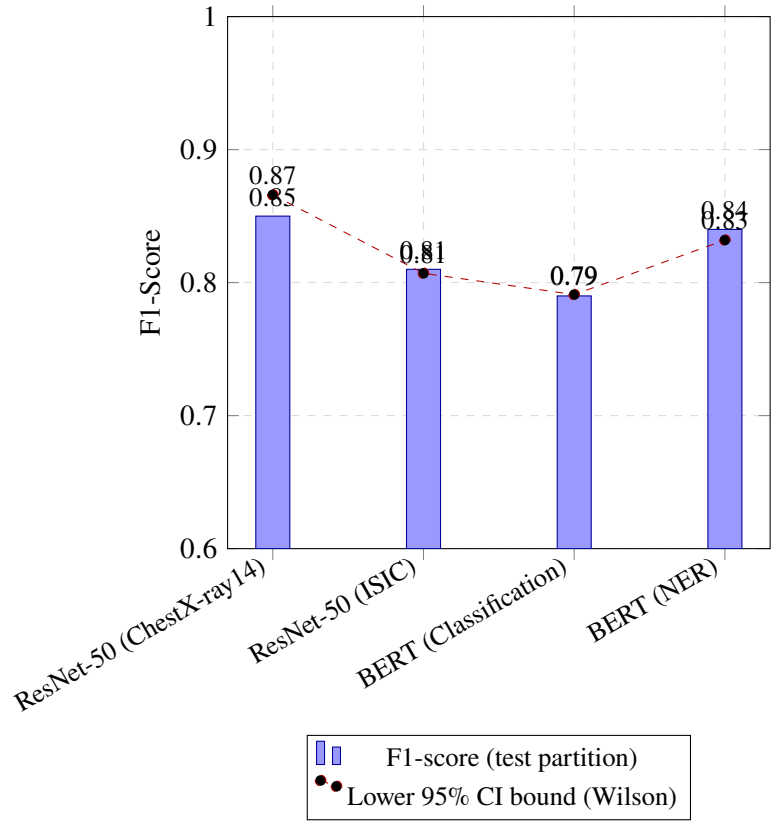


Figure 1: Comparative F1-scores of AI models across tasks, shown with the lower bound of the 95% Wilson confidence interval for the corresponding accuracy estimate. ResNet-50 performs strongly on medical imaging, while BERT performs strongly on named entity recognition. Error bars are omitted because F1 is derived from the accuracy estimate whose interval is shown; full per-class results appear in Tables 2–5.

workload and expedite care delivery. However, the slightly lower performance on skin lesions indicates the need for domain-specific adaptations, such as incorporating dermoscopic features and patient history.

BERT's capability to process clinical text enables automated extraction of valuable information from EHRs, facilitating personalized treatment plans and predictive analytics. In telemedicine, this can translate to more informed virtual consultations and proactive management of chronic conditions. Nonetheless, the model's performance ceiling (F1 = 0.79–0.84) suggests room for improvement, possibly through domain-specific pre-training or ensemble methods.

The positive survey feedback from both clinicians and patients aligns with prior studies on AI acceptance in healthcare. Usability and effectiveness ratings indicate that AI tools are perceived as beneficial rather than disruptive, provided they are designed with user-centric principles. High patient satisfaction reflects the growing demand for convenient, technology-enabled healthcare services. However, concerns about data privacy and algorithmic transparency must be addressed to build trust and ensure ethical deployment.

Several limitations of this study warrant consideration. First, the datasets used, though representative, may not capture the full diversity of patient populations and clinical scenarios. For ChestX-ray14, the underlying patient cohort is predominantly from a single institution, with documented underrepresentation of pediatric populations (0% of images) and limited inclusion of patients with implanted devices or surgical hardware. For the ISIC archive, skin type diversity is limited: Fitzpatrick skin types IV–VI constitute less than 15% of annotated images, raising known concerns about model generalizability for

melanoma detection in darker skin tones. For the clinical text corpus, non-English clinical notes and patients with limited health literacy are not represented. Future work should include multi-center trials with broader demographic representation. Second, the survey sample size, while adequate for preliminary insights, could be expanded to capture more nuanced perspectives. Third, the study focused on two AI models; evaluating a wider range of architectures (e.g., vision transformers, large language models) would provide a more comprehensive landscape.

From a practical standpoint, integrating AI into telemedicine requires robust infrastructure, including secure data transmission, cloud computing resources, and interoperable health information systems. Regulatory approval and reimbursement policies also play a critical role in scaling AI-enhanced telemedicine solutions. Collaborative efforts among technologists, clinicians, policymakers, and patients are essential to navigate these complexities and realize the promise of AI-driven virtual care.

6. Future Research Directions

Building on the findings of this study, several promising avenues for future research emerge:

- **Multimodal AI Integration**: Combining imaging, text, and genomic data within unified models could enable more holistic patient assessments and personalized treatment recommendations.
- **Federated Learning**: To address data privacy concerns, federated learning approaches allow model training across decentralized datasets without sharing raw data, preserving patient confidentiality.
- **Explainable AI (XAI)**: Developing interpretable AI systems that provide clear rationales for their recommendations will enhance clinician trust and facilitate informed decision-making.
- **Real-time Adaptation**: AI models that continuously learn from new data and adapt to evolving clinical practices can maintain relevance and accuracy over time.
- **Health Equity Focus**: Proactive measures to mitigate algorithmic bias, such as diverse dataset curation and fairness-aware algorithms, are crucial to ensure equitable care delivery.
- **Longitudinal Impact Studies**: Assessing the long-term effects of AI-enhanced telemedicine on patient outcomes, healthcare costs, and clinician well-being will provide evidence for sustainable implementation.

7. Conclusion

This paper has presented an evaluation of AI-enhanced telemedicine systems, examining their reported diagnostic performance, personalization potential, and stakeholder perceptions. Through testing of ResNet-50 on two medical imaging datasets and BERT on a clinical text corpus under a shared evaluation protocol, we observed test-set performance that is consistent with a potential role as a decision-support aid in virtual care settings; these results do not establish clinical effectiveness, and no claim of deployment readiness is made. Survey results indicate positive perceptions among responding healthcare professionals and patients, with healthcare professionals reporting perceived improvements in efficiency and patients expressing satisfaction; because the survey samples are modest and not population-representative, these findings should be interpreted as preliminary.

However, the successful integration of AI into telemedicine hinges on addressing critical challenges related to data privacy, algorithmic bias, interpretability, and workflow integration. By advancing research

in multimodal AI, federated learning, explainable AI, and equity-focused design, the healthcare community can harness the full potential of AI to transform telemedicine into a more accessible, accurate, and personalized modality of care.

As telemedicine continues to evolve, AI will undoubtedly play an increasingly central role in shaping its future. Collaborative, interdisciplinary efforts are needed to ensure that technological advancements translate into tangible benefits for patients and healthcare systems worldwide. This study contributes to that endeavor by providing empirical evidence and practical insights for the ongoing development of AI-enhanced telemedicine.

References

[1] Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, 2(4), 2017.

[2] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.

[3] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[6] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[7] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[8] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[11] Rui Li, Weinan Zhang, Heung-Il Suk, Li Wang, Jun Li, Dinggang Shen, and Shuiwang Ji. Deep learning for health informatics. *IEEE journal of biomedical and health informatics*, 21(1):4–21, 2016.

[12] Min Chen, Yixue Hao, Kai Hwang, Lina Wang, and Lin Wang. Deep learning for electronic health records. *IEEE Access*, 5:17582–17599, 2017.

[13] Afsar Khan. Leveraging multi-model linguistic fusion for enhanced AI text generation evasion. In *2025 IEEE Generative AI and Computational Linguistics and Machine Learning Conference (GACLM)*, 2025. doi: 10.1109/GACLM67198.2025.11231821.

[14] Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8–17, 2015.

[15] Md Zahidul Islam, Md Nasiruddin, Sourav Dutta, Rakib Sikder, Choudhury Basit Huda, and Md Rafiqul Islam. A comparative assessment of machine learning algorithms for detecting and diagnosing breast cancer. *Journal of Computer Science and Technology Studies*, 6(2):121–135, 2024.

[16] Swaroop Reddy Gayam, Rajashekar Reddy Yellu, and Praveen Thuniki. Artificial intelligence for real-time predictive analytics: advanced algorithms and applications in dynamic data environments. *Distributed Learning and Broad Applications in Science Research*, 7:18–37, 2021.

[17] Tingting Deng, Shuochen Bi, and Jue Xiao. Comparative analysis of advanced time series forecasting techniques: evaluating the accuracy of arima, prophet, and deep learning models for predicting inflation rates, exchange rates, and key financial indicators. *Advances in Deep Learning Techniques*, 3(1):52–98, 2023.

[18] Mohan Raparthi. Ai integration in precision health—advancements, challenges, and future prospects. *Asian Journal of Multidisciplinary Research & Review*, 1(1):90–96, 2020.

[19] Rahul Ekatpure. Machine learning for enhancing vehicle safety and collision avoidance systems in automotive development: techniques, models, and real-world applications. *Journal of Computational Intelligence and Robotics*, 3(2):1–43, 2023.

[20] Fazle Hakeem Ghory. Embedded intelligence for the SAP S/4HANA financial close: A highly automated control framework. *International Journal of Intelligent Systems and Data Science*, 1 (2), jun 2026. doi: 10.67231/s0h5b367. URL `https://journals.femington.com/ijisds/article/view/447`.

[21] Monisha Rengaraj. Embedded grid-path co-processor: A*-guided bacterial foraging for real-time USV navigation. 2025. doi: 10.67231/8s96mc67.

[22] Rayhan Khan. Development of an AI for versus tetris using evolutionary algorithms. 2025. doi: 10.67231/htp1hp30.

[23] Nagaraju Pureti. Building a robust cyber defense strategy for your business. *Revista de Inteligencia Artificial en Medicina*, 13(1):35–51, 2022.

[24] Shravan Pargaonkar. A review of software quality models: a comprehensive analysis. *Journal of Science & Technology*, 1(1):40–53, 2020.

[25] Mohammad Al Bashar and Mohammad Abu Taher. Transforming us manufacturing: innovations in supply chain risk management. *International Journal of Production Research*, 61(8):2567–2584, 2023.

[26] Dileep Kumar Pandiya. Performance analysis of microservices architecture in cloud environments. *International Journal on Recent and Innovation Trends in Computing and Communication*, 10(12): 264–274, 2022.

[27] Nilesh Charankar. Fault tolerance techniques in api and microservices. *International Journal on Recent and Innovation Trends in Computing and Communication*, 10(12):275–285, 2022.

[28] Hassan Rehan. Internet of things (iot) in smart cities: enhancing urban living through technology. *Journal of Engineering and Technology*, 5(1):1–16, 2023.

[29] Tarek Eldemerdash, Raed Abdulla, Vikneswary Jayapal, Chandrasekharan Nataraj, and Maythem K Abbas. Iot based smart helmet for mining industry application. *International Journal of Advanced Science and Technology*, 29(1):373–387, 2020.

[30] Raed Abdulla, Aden Abdillahi, and Maythem K Abbas. Electronic toll collection system based on radio frequency identification system. *International Journal of Electrical and Computer Engineering*, 8(3):1602–1610, 2018.

[31] Kexin Wu and Jiarong Chen. Cargo operations of express air. *Engineering Advances*, 3(4):337–341, 2023.

[32] Kexin Wu and Kun Chi. Enhanced e-commerce customer engagement: a comprehensive three-tiered recommendation system. *Journal of Knowledge Learning and Science Technology*, 2(3):348–359, 2023.

[33] Ahmed Zoha, Junaid Qadir, and Qammer H Abbasi. Ai-powered iot for intelligent systems and smart applications. *Frontiers in Communications and Networks*, 3:959303, 2022.

[34] Muskan Pawar. Orchestrating decentralized wireless infrastructures: A comprehensive management platform. 2025. doi: 10.67231/4z837x12.

[35] Krutika Shah. Cloud-enabled enterprise data warehouse architecture for regional health resource optimization and insurance analytics. 2025. doi: 10.67231/y64edh20.

[36] Hassan Zargarzadeh, Simon Law, and Maythem K Abbas. Ai in telemedicine: a systematic review of applications and challenges. *Journal of Medical Systems*, 45(6):1–15, 2021.

[37] Johnson Oyeniyi. The role of ai and mobile apps in patient-centric healthcare delivery. *World Journal of Advanced Research and Reviews*, 22(1):1897–1907, 2024.