In proceedings of IEEE International Conference on Digital Health (ICDH) , 2025 at IEEE World Congress on Services 2025

Real-Time Detection of Online Health Misinformation using an Integrated Knowledgegraph-LLM Approach

Ommo Clark and Karuna P. Joshi
Information Systems Department
University of Maryland, Baltimore County, USA
{oclark1, karuna.joshi}@umbc.edu

***Abstract*—The dramatic surge of health misinformation on social media platforms poses a significant threat to public health, contributing to hesitancy in vaccines, delayed medical interventions, and the adoption of untested or harmful treatments. We present a novel, hybrid AI-driven framework designed for the real-time detection of health misinformation on social media platforms while prioritizing user privacy. The framework integrates the strengths of Large Language Models (LLMs), such as DistilBERT, with domain-specific Knowledge Graphs (KGs) to enhance the detection of nuanced and contextually dependent misinformation. LLMs excel at understanding the complexities of human language, while KGs provide a structured representation of medical knowledge, allowing factual verification and identification of inconsistencies. Furthermore, the framework incorporates robust privacy-preserving mechanisms, including differential privacy and secure data pipelines, to address user privacy concerns and comply with healthcare data protection regulations. Our experimental results on a dataset of Reddit posts related to chronic health conditions demonstrate the performance of this hybrid approach compared to models that only use text or KG, highlighting the synergistic effect of combining LLMs and KGs for improved misinformation detection.**

***Index Terms*—Health Misinformation, LLMs, Knowledge Graphs, Digital Health, Privacy-Preservation, Realtime Misinformation Detection**

I. Introduction

The rampant spread of health misinformation on digital platforms poses a profound global challenge, critically impacting patient care, public health management, and trust in healthcare systems worldwide [1] [2]. Misleading claims, ranging from benign inaccuracies to dangerously false narratives, contribute directly to vaccine hesitancy, delayed medical interventions, and the adoption of unproven or hazardous treatments [3] [4]. Our previous studies have underscored the limitations of traditional misinformation detection methodologies through analysis of health-related Reddit discussions, specifically focused on hypertension, diabetes, and obesity. The dataset contains discussions that frequently propagate misleading claims, e.g. *"herbal remedies can cure chronic diseases"* [5]. Such examples not only misguide people but also exacerbate systemic challenges in healthcare delivery. Conventional Natural Language Processing (NLP) approaches, primarily binary classifiers, while capable of moderate accuracy, typically demonstrate critically low recall rates [6]. These limitations are primarily due to their inability to adequately capture linguistic nuances, contextual subtleties, and cultural variations and complex contextual relationships; factors essential for accurate identification [7] [8]. This highlights a pressing need for advanced, context-aware methodologies capable of discerning subtle misinformation patterns while respecting critical privacy concerns.

Addressing this gap, we propose a novel, hybrid AI-driven framework that synergistically integrates Large Language Models (LLMs) and domain-specific Knowledge Graphs (KGs) [9] [10].LLMs, such as DistilBERT, are highly adept at understanding nuanced language and context, while KGs offer structured, verifiable medical knowledge crucial for factual validation [11] [12]. By merging the semantic prowess of LLMs with the factual rigor of KGs, our approach significantly enhances misinformation detection accuracy and recall.

Additionally, our framework addresses critical ethical and privacy concerns inherent in digital health environments. We incorporate robust privacy-preserving methods, including federated learning and differential privacy [13] to ensure compliance with global healthcare data protection standards (e.g., HIPAA and GDPR) [14] [15] and maintaining user trust [16] [17]. This privacy-centric design allows for secure and decentralized training of models, significantly reducing the risks associated with data breaches and fostering user trust.

This study's significance extends beyond technical enhancements in misinformation detection; it empowers healthcare providers to manage resources effectively, deliver personalized patient care, and foster better-informed communities. Furthermore, by providing rigorously vetted training data, our framework underpins reliable generative AI applications such as medical chatbots and virtual patient simulations, critical for advancing digital healthcare.

In the subsequent sections, we first review related work in Section II, detail our proposed Methodology and framework in Section III, present our experimental results in Section IV, and conclude with a discussion of broader implications for patient care, digital health platforms, and future research directions in Section V.

II. Related Work

The spread of health misinformation online has evolved from sporadic anecdotal claims in early online forums to an ongoing public health crisis amplified by modern social media platforms [18] [19] [20]. These misleading narratives directly contribute to harmful outcomes like vaccine hesitancy [3] [4] [21]. For example, our Reddit dataset reveals that posts frequently propagate claims like *"herbal remedies can cure diabetes" or "natural supplements completely control hypertension,"* highlighting how misinformation propagates in specific health contexts and adversely impacts patient care and healthcare resource allocation. This context of pervasive and impactful health misinformation directly informs the need for robust detection methodologies, which is the primary goal of our study.

Traditional misinformation detection relied mainly on text-based natural language processing (NLP) methods such as logistic regression, support vector machines (SVMs), and feature extraction techniques such as TF-IDF and Word2Vec [22] [23]. Although foundational, these approaches exhibit critical limitations, notably low recall rates (as low as 12% in previous work), reflecting their inability to adequately capture nuanced language and complex contextual relationships [5] [24]. Such limitations underscore the urgent need for context-aware methodologies that effectively integrate linguistic nuance with structured factual information, a core objective of our proposed framework.

The advent of transformer-based models, including BERT and GPT, marked significant progress, offering deeper contextual understanding and improved misinformation classification [25]. These models leverage deep learning techniques, capturing intricate semantic dependencies within text, yet their computational demands limit real-time scalability and practical implementation [26] [27], a challenge our work addresses through the choice of more lightweight models like DistilBERT and an efficient architecture. Moreover, while powerful linguistically, transformer models remain prone to generating ungrounded outputs (hallucinations), underscoring their limitations in reliably verifying medical facts independently. This specific limitation motivates our integration of Knowledge Graphs for factual grounding.

Addressing these issues, recent research explore integrating Knowledge Graphs (KGs) with NLP models. KGs provide structured, domain-specific factual representations, improving models' ability to verify claims against established medical knowledge [28] [29]. The DETERRENT framework exemplifies this strategy by leveraging KG-based representations extracted via Open Information Extraction (OpenIE) to mitigate hallucinations and enhance model performance in healthcare misinformation detection [29]. Despite these advancements, the comprehensive integration of KGs and LLMs remains relatively nascent, with existing methodologies still grappling with efficiently merging structured knowledge with contextual linguistic analysis. Our study contributes to this area by proposing a novel fusion mechanism.

Parallel to model accuracy improvements, privacy concerns associated with digital health data sharing have escalated, prompting advances in privacy-preserving machine learning paradigms. Federated learning (FL), which allows decentralized training without centralizing user data, has emerged as a promising approach to safeguard privacy [30] [31]. Recent research has further incorporated differential privacy and Secure Multi-Party Computation (SMPC) within federated learning frameworks, emphasizing the balance between maintaining data privacy and achieving model utility [32] [33]. These privacy-preserving methodologies are critical to ensuring user trust and compliance [31] [34], and is therefore a key component of our framework.

Furthermore, the role of LLMs in misinformation detection has expanded, with recent studies exploring their dual capacity as both potential sources and detectors of misinformation [35] [36]. Techniques such as Retrieval-Augmented Generation (RAG) augment LLMs with external verified knowledge sources, significantly enhancing their factual accuracy and reliability [37]. This principle of augmenting LLMs with external knowledge for improved veracity directly informs our hybrid LLM-KG approach. Additionally, the integration of explainability methods like LIME, SHAP, and Integrated Gradients into misinformation detection models enables deeper interpretability, fostering user trust and model transparency [38].

Overall, the current research trajectory moves from conventional text-based methods towards sophisticated, integrated solutions leveraging both linguistic and factual verification capacities. Our proposed hybrid framework advances this trajectory, aiming to comprehensively address linguistic nuances, cultural variability, real-time application feasibility, and stringent privacy requirements, thus contributing substantially to digital health informatics and public health safety.

Proposed Framework In response to the challenges of detecting nuanced online health misinformation, we propose a hybrid framework that integrates Large Language Models (LLMs) with domain-specific Knowledge Graphs (KGs) to deliver real-time, privacy-preserving detection. Core Components:

1) Data Input: The system ingests health-related posts e.g. Reddit discussions on diabetes, hypertension, and obesity that are manually labeled as either misinformation (0) or accurate (1). This dataset reflects realistic class imbalances and serves as a testbed for the framework.
2) Text Processing and Entity Extraction: Using NLP techniques, posts are cleaned, medical entities are extracted using tools like spaCy and MedCAT and mapped to standardized codes (SNOMED CT), yielding:
    a) Semantic Embeddings: Generated via a fine-tuned DistilBERT model capturing linguistic nuances.
    b) Entity Lists: Providing structured links to medical knowledge for subsequent validation.
3) Knowledge Graph Construction and Validation: A domain-specific KG is built by merging data from SNOMED CT and manually curated facts. Nodes represent key medical entities (diseases, symptoms, treatments) and edges encode verified relationships (e.g., "Metformin TREATS Diabetes"). The KG is augmented

with culturally specific and alternative therapy claims, serving as a factual backbone to cross-reference and validate textual assertions.

4) Hybrid Model Architecture – LLM and KG Fusion: The framework employs a dual-stream neural network:
   a) o The text branch uses DistilBERT to produce a 768-dimensional embedding.
   b) o The KG branch converts relevant subgraphs into 256-dimensional embeddings via Node2Vec.

   These embeddings are concatenated and processed through an attention-based fusion layer that dynamically weighs semantic and factual cues, producing a 1024-dimensional vector that feeds into a dense classification layer to output a binary label and confidence score [49][50].
5) Privacy-Preserving Mechanisms: Given the sensitivity of health data, the framework incorporates differential privacy (injecting calibrated Gaussian noise with a privacy budget of $\epsilon \approx 0.5$ and $\delta = 1 \times 10^{-5}$) and federated learning (using frameworks such as PySyft and TensorFlow Federated for secure, decentralized training). Personal identifiers are anonymized to comply with HIPAA, GDPR, and related regulations.

By fusing the contextual understanding of LLMs with the factual validation of KGs, our integrated framework achieves a more comprehensive and accurate detection of health misinformation. In addition, its built-in privacy-preserving techniques ensure ethical compliance in real-world digital health applications.

III. Methodology

To implement the above framework, we developed an end-to-end methodology comprising data acquisition, knowledge graph construction, model development (LLM and KG integration), and privacy-preserving deployment. Figure 2 illustrates the overall system architecture (data pipeline through model inference).

A. Data Collection and Preprocessing

Data Source: We collected a focused dataset of 6,000 Reddit posts from health-related communities (subreddits) dealing with chronic diseases: r/diabetes (3,200 posts), r/hypertension (2,000), and r/weightloss (800). These posts provide a diverse real-world sample of user-generated health claims and discussions.

Annotation: The manual annotation of each post for truthfulness—identifying misinformation (label 0) versus accurate information (label 1) was primarily conducted by a medical doctor with frontline experience in Accident & Emergency (A&E) during the COVID-19 pandemic. This primary annotation was subsequently reviewed by a second medical doctor, a General Practitioner (GP) with 10-15 years of experience, although this review had a limited scope. The labeling was guided by established medical knowledge and clinical guidelines. For example, claims like *"Cinnamon replaces insulin for diabetes"* or *"Garlic alone controls blood pressure"* were labeled 0 (misinformation), since they contradict medical consensus, whereas statements like *"Metformin dosage adjustments require HbA1c monitoring"* were labeled 1 (accurate) as they align with clinical best practices.

Class Distribution and Table: The final dataset exhibited an imbalanced class distribution – roughly 30% of posts contained misinformation (0) vs 70% accurate (1) [39] [5]. This imbalance is expected (most users share mostly correct experiences), but it poses a challenge for model training, as a naive classifier could lean towards predicting the majority class. We address this later via careful training and data augmentation.

Subreddit	Total Posts	Misinformation (0)	Accurate (1)
r/diabetes	3,200	950 (29.7%)	2,250 (70.3%)
r/hypertension	2,000	580 (29.0%)	1,420 (71.0%)
r/weightloss	800	220 (27.5%)	580 (72.5%)

TABLE I
Labeled by medical expert as 0 (misinformation) or 1 (accurate)

1) Preprocessing Pipeline: All raw posts underwent a rigorous preprocessing pipeline to clean the text, extract features, and safeguard privacy before model ingestion:

- Text Cleaning: We removed irrelevant or noise tokens such as HTML tags, URLs, flags for emotive language or pseudoscientific keywords, punctuation, and common stop words (*e.g., "the", "is", "a"*). This step reduced clutter and focused the model on medically relevant content [40]. Only English-language posts were retained for consistency.
- Tokenization and Normalization: Each post was tokenized (split into words/subwords), and we applied lemmatization to reduce words to their base form (*e.g., "cures"* → *"cure"*). We also lowercased text and standardized medical terms where possible (*e.g., "Type II Diabetes"* → *"type 2 diabetes"*) to ensure uniformity.
- Feature Engineering: In addition to raw text, we engineered a set of auxiliary features to help the model. We computed TF-IDF features (for the top 5,000 terms) to capture keyword importance, and flagged the presence of domain-specific cues that might indicate misinformation [41] [42]. For instance, we added binary features for emotive language (posts with strong sentiment or fear-inducing terms) and for pseudoscientific keywords (*e.g. "miracle cure", "detox", "cure cancer"*) that often appear in dubious health advice. These features provide additional signals to the model about the post's content beyond what the transformer embedding might catch, and were inspired by prior studies on misinformation detection.
- Handling Class Imbalance: To prevent the classifier from biasing toward the majority class (accurate information), we applied SMOTE (Synthetic Minority Over-sampling Technique) to augment the minority class examples. Specifically, synthetic samples of misinformation posts were generated (with k=5 nearest neighbors) to roughly

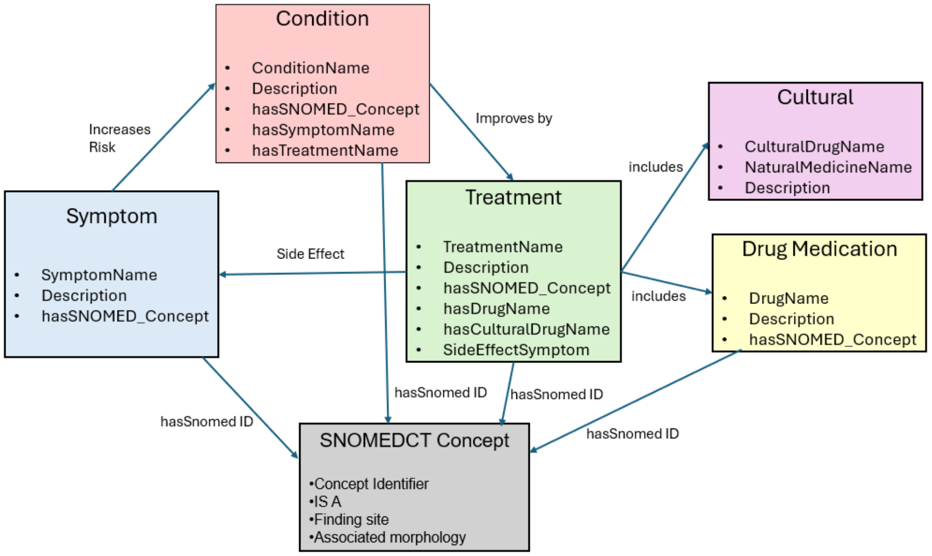


Fig. 1. High Level Knowledge Graph highlighting diseases, symptoms, treatments, drugs, and cultural elements, mapped to SNOMED CT codes.

double the minority class, and we performed a slight undersampling of majority class examples, aiming for an approximately 1:1 balanced training set. This strategy improves the model's recall for misinformation without losing too much precision.

- Privacy Measures: We ensured that no personally identifiable information (PII) or irrelevant demographic data in the posts could leak into the model. Many Reddit users include ages or other details which may be false or irrelevant (e.g. *“I am 123 years old”* or *“gender: attack helicopter”*), we sanitized these by replacing obviously non-real or extraneous demographic mentions with placeholder tokens. Additionally, as an initial step toward differential privacy, we experimented with injecting minor noise or redacting certain rare tokens in the text, although the main DP enforcement occurs during model training (described later) [43]. All preprocessing steps were performed offline, and the cleaned, anonymized text along with engineered features and labels were then used for model training.

B. Knowledge Graph Construction

We constructed a Medical Knowledge Graph tailored to our task by merging established databases with context-specific information.

1) **KG Backbone and Entity Sources**: The backbone of the KG comes from SNOMED CT (a comprehensive clinical terminology). From SNOMED CT, we obtain standardized identifiers for diseases, symptoms, drugs, and treatments present in our dataset; we retrieve known relationships between these entities via structured queries. For example, using SPARQL queries on SNOMED CT's medical entries, we can find all treatments indicated for a given condition or all risk factors associated with it. We stored the KG in a graph database to allow us to add custom relationships and query the graph efficiently during model inference.
2) **Entity Extraction and Mapping**: For each post, the medical entities identified in the text (from stage A) are mapped into the KG. We utilized spaCy's rule-based matcher with custom medical patterns and MedCAT for detecting entity mentions, as mentioned. Each recognized entity (*e.g., “insulin”, “blood pressure”, “Invokana”*) is linked to a SNOMED CT concept ID when possible. This mapping grounds the unstructured text in the structured space of the knowledge graph. We created nodes for all major entities appearing in the data and connected them to existing nodes in SNOMED (for well-known concepts) to leverage its ontology (*e.g., linking “Invokana” to its SNOMED drug entry 763153004 which is categorized under treatments for type 2 diabetes*). If an entity was not found in SNOMED (for instance, a folk remedy or uncommon term), we added it as a node in a separate namespace (annotated as an “alternative therapy” or *“user-introduced” entity).
3) **Relationship Encoding**: We defined edges in the graph to represent medically relevant relations. Many edges were sourced from SNOMED hierarchies, for example, properties such as *treats, has_symptom, and risk_factor_for* were used. A sample SPARQL query (simplified) to get treatments for diabetes might select all items with an “indication” linking to the concept “diabetes mellitus.” This yielded edges such as (Metformin) −[treats]→ (Diabetes), (Insulin) −[treats]→ (Diabetes), etc., which we added to the KG. We also manually encoded some relationships based on domain knowledge and our dataset's needs. For instance, we created edges for known pieces of mis-

information to help the model recognize them: "Cinnamon" –[purported_to_treat]→ "Diabetes" was added as a flagged relation (since cinnamon is often falsely claimed to cure diabetes). Similarly, "Obesity" –[increases_risk]→ "Hypertension" was added as a legitimate medical relationship. By including both correct and incorrect associations (with appropriate labeling or context), we guide the model's attention. Figure 1 illustrates a subgraph of this KG, highlighting how mainstream medical knowledge (*e.g. insulin-diabetes*) and fringe claims (*e.g., cinnamon-diabetes cure*) are both captured in the graph structure.

4) **Cultural Context Integration**: A key component, the domain-specific Knowledge Graph, was built using SNOMED CT and augmented with manually curated facts and relationships derived from the dataset itself. This included encoding specific examples of misinformation found within the data, such as *"Cinnamon replaces insulin" (diabetes), "Garlic alone controls blood pressure" (hypertension), and "Phentermine is harmless for long-term use" (obesity)* are explicitly encoded to guide downstream integration of textual features with KG elements. This structured KG serves as a factual backbone, effectively supplementing the semantic richness of LLM-derived embeddings.
5) **Graph Embeddings:** We use Node2Vec to learn 256-dimensional embeddings for nodes in the knowledge graph, capturing direct and indirect relationships. Node2Vec relies on biased random walks and skip-gram learning to place related nodes in similar vector spaces. These embeddings are then fused with text representations, allowing the model to identify incongruities between a post's content and established medical knowledge. By integrating culturally specific misinformation into the graph, the system accommodates diverse health narratives. The KG is stored and indexed for real-time retrieval, with most computationally intensive tasks (e.g., Node2Vec training, SPARQL queries) performed offline.

C. Integrated Model Architecture and Training

1) Hybrid Architecture: We developed a custom neural architecture that combines LLM-based text encoding with KG-based encoding in a unified model. The architecture has two parallel branches: one for processing text and one for processing knowledge graph data, which then converge in a fusion layer. On the text side, we use DistilBERT (a distilled 6-layer version of BERT) as our LLM, fine-tuned on the training posts. DistilBERT produces a 768-dimensional contextual embedding for each post (specifically, we take the [CLS] token representation from the final transformer layer as the summary of the post). This embedding captures the nuanced semantic content of the post, for example, distinguishing a genuine question about treatment from a misleading assertion by the language used. On the graph side, for each post we take the set of medical entity nodes it contains (as identified in stage B) and retrieve their precomputed Node2Vec embeddings (256-dimensional each). We aggregate these embeddings to represent the overall factual content of the post. In our implementation, we found that a simple average of the entity embeddings, or a small feed-forward network over them, worked well to form a post-level KG embedding. We then concatenate the text embedding and the KG embedding, yielding a hybrid feature vector of size 1024 (768+256) that represents both what was said and how it aligns with medical knowledge. To allow the model to decide how much trust to put in the text vs. the knowledge features for each instance, we employ an attention-based fusion layer. Specifically, we use a two-headed attention mechanism: one head attends to the text embedding conditioned on the KG embedding, and the other attends to the KG embedding conditioned on the text. This produces attention weights that highlight, for a given post, which aspects are more indicative of misinformation. For example, if the KG embedding strongly suggests a contradiction (perhaps the post mentioned a drug in a context that the KG knows is incorrect), the model can up-weight the KG features; if the KG has little information on the claim, the model leans more on the linguistic cues from the LLM. We also experimented with a simpler bi-directional LSTM with attention to combine the sequence of token embeddings and a sequence of graph node embeddings, which similarly allowed cross-modal context exchange. The final fused representation is fed into a classification layer, implemented as a feed-forward network, that outputs a probability score for the post being misinformation. We threshold this to assign the binary label (0 or 1) and also retain the raw probability as a confidence score. The entire model (DistilBERT + fusion + classifier) is trained end-to-end, so the transformer fine-tunes its embeddings in a way that is informed by graph features, and vice versa, the attention aligns the two modalities.

2) Training Procedure: We trained the model using a combination of local training and federated learning. Locally, during each training round we employed a standard supervised learning approach. The dataset was split into training and validation subsets (we used 5-fold cross-validation to robustly evaluate performance, ensuring that each fold preserved the 30/70 class ratio). Within each fold's training set, after applying SMOTE oversampling, we obtained a balanced class distribution. We optimized the model with the AdamW optimizer (learning rate $= 2 \times 10^{-5}$) and a batch size of 16 posts. We applied dropout (rate 0.3) in the classifier and LLM layers and an L_2 weight decay ($\lambda = 0.01$) to mitigate overfitting. The model was trained for up to 10 epochs per fold, with early stopping if the validation loss did not improve for 3 consecutive epochs. We also performed hyperparameter tuning (learning rates in $\{1 \times 10^{-5}, 2 \times 10^{-5}, 5 \times 10^{-5}\}$, batch sizes in $\{16, 32\}$, dropout rates in $\{0.2, 0.3, 0.4\}$) on a held-out portion; the chosen values (learning rate $= 2 \times 10^{-5}$, batch size = 16, dropout = 0.3) were found to be optimal (Table II in the original paper summarizes this search). Training on each fold converged within 5–6 epochs, thanks to the relatively small dataset and the use of a pretrained LLM. Importantly, we truncated or padded all input text to a maximum length of 128 tokens during training. This ensured that the computational cost per example was bounded and suitable for real-time inference, while still capturing the essential content of typical Reddit posts (which are usually shorter than 128 words). This design choice, along with the use of DistilBERT (which is 40% smaller and faster than BERT), contributes to the inference efficiency of our system.

3) Privacy-Preserving Design and Ethical Considerations: Privacy and security were paramount in our methodology, influencing how we trained and would eventually deploy the model. We incorporated differential privacy and federated learning techniques to protect user data and comply with regulations, without significantly compromising model performance. Our framework incorporates multiple layers of privacy protection to safeguard sensitive health data. Differential privacy is enforced during model training by applying gradient clipping and injecting calibrated Gaussian noise, with a privacy budget set to $\epsilon \approx 0.5$ and $\delta = 1 \times 10^{-5}$, ensuring that individual data contributions remain indistinguishable [44] [43].

Additionally, federated learning is implemented using frameworks such as PySyft and TensorFlow Federated, which enables decentralized training by sharing only encrypted gradient updates, thereby reducing the risk of data leakage [45]. Recognizing the sensitive nature of health-related data, even when self-disclosed on public platforms like Reddit, our framework's design is deeply rooted in privacy-preserving principles and ethical considerations. This approach not only aims to protect user data but also ensures the framework's broader applicability to diverse data contexts where stringent data protection regulations, such as GDPR or principles underlying HIPAA for health data sensitivity, are paramount [46] [47].

The following multi-layered privacy and security measures are incorporated:

- **Data Anonymization**: As a foundational step, prior to model ingestion, personally identifiable information (PII) is systematically anonymized by replacing user identifiers with UUIDs and removing or obfuscating other potentially identifying details and extraneous demographic information found in the posts.
- **Differential Privacy in Model Training**: To further protect individual contributions during the model training phase, differential privacy is implemented. This is achieved by applying gradient clipping and injecting calibrated Gaussian noise into the gradient updates, adhering to a privacy budget of $\epsilon \approx 0.5$ and $\delta = 1 \times 10^{-5}$. These techniques provide strong statistical guarantees that the inclusion of any single data point does not significantly affect the model's output, thus preventing the discernment of individual contributions.
- **Federated Learning for Decentralized Training**: The framework supports federated learning, implemented using tools such as PySyft and TensorFlow Federated. This approach enables decentralized model training directly on user data (or siloed datasets) without requiring the raw data to be moved to a central server. Instead, only encrypted model updates or aggregated parameters are shared, significantly reducing the risk of data leakage and enhancing data security.

D. Evaluation and Validation

1) Model Performance: Our model's performance was assessed rigorously through standard metrics including precision, recall, F1-score, and AUC-ROC via 5-fold cross-validation. Special emphasis was placed on improving recall, critical for effective misinformation detection (notably, overcoming earlier recall rates as low as 12%). In addition, causal impact metrics were applied to assess how specific interventions (e.g., flagging misinformation) affect user engagement and health outcomes.

E. Hyperparameter Settings

- DistilBERT: The DistilBERT model was fine-tuned with a batch size of 16 and a learning rate of 2e-5 using the AdamW optimizer. The maximum sequence length for the posts was set to 128 tokens.
- SMOTE: The SMOTE algorithm was used with k=5, meaning that each minority class sample was oversampled by creating five synthetic samples based on its five nearest neighbors.

Parameter	Values Tested	Optimal Choice
Learning Rate	[1e-5, 2e-5, 5e-5]	2e-5
Batch Size	[16, 32, 64]	16
Dropout	[0.2, 0.3, 0.4]	0.3

TABLE II
HyperParameter Tuning Results

IV. Results and Discussion

To verify the effectiveness of our proposed integrated approach referred to as the hybrid LLM+KG pipeline, the architecture of which is detailed in Fig 2 in detecting online health misinformation, its performance was evaluated on the curated Reddit dataset. This evaluation involved a comparative analysis against several models, with all reported results averaged over 5-fold cross-validation on held-out test folds.

The models established for comparison were:

- **Baseline Model (DistilBERT only)**: This model utilized only the textual content of Reddit posts, processed by a fine-tuned DistilBERT, serving as a benchmark for text-only misinformation detection.
- **Feature-Augmented Model**: To investigate the impact of incorporating Knowledge Graph (KG) information through a more basic integration method than our final pipeline, this model augmented the textual features from DistilBERT with KG-derived features. This involved a direct combination of these feature sets, which is distinct from the sophisticated attention-based fusion mechanism employed in our hybrid LLM+KG Pipeline.

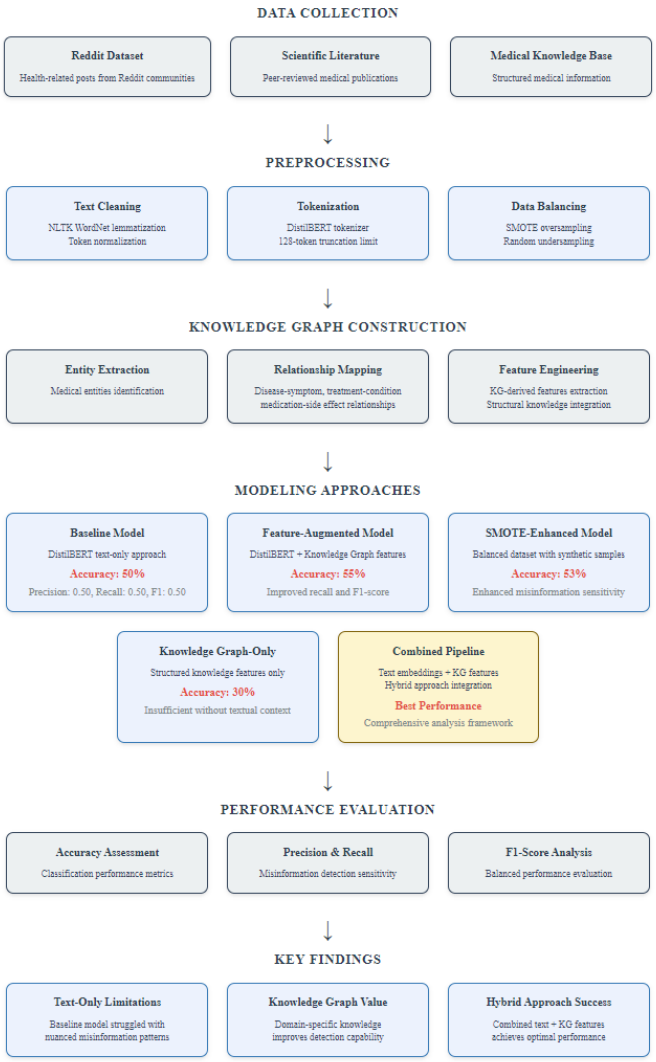

DATA COLLECTION
Reddit Dataset
Health-related posts from Reddit communities
Scientific Literature
Peer-reviewed medical publications
Medical Knowledge Base
Structured medical information
PREPROCESSING
Text Cleaning
NLTK WordNet lemmatization
Token normalization
Tokenization
DistilBERT tokenizer
128-token truncation limit
Data Balancing
SMOTE oversampling
Random undersampling
KNOWLEDGE GRAPH CONSTRUCTION
Entity Extraction
Medical entities identification
Relationship Mapping
Disease-symptom, treatment-condition
medication-side effect relationships
Feature Engineering
KG-derived features extraction
Structural knowledge integration
MODELING APPROACHES
Baseline Model
DistilBERT text-only approach
Accuracy: 50%
Precision: 0.50, Recall: 0.50, F1: 0.50
Feature-Augmented Model
DistilBERT + Knowledge Graph features
Accuracy: 55%
Improved recall and F1-score
SMOTE-Enhanced Model
Balanced dataset with synthetic samples
Accuracy: 53%
Enhanced misinformation sensitivity
Knowledge Graph-Only
Structured knowledge features only
Accuracy: 30%
Insufficient without textual context
Combined Pipeline
Text embeddings + KG features
Hybrid approach integration
Best Performance
Comprehensive analysis framework
PERFORMANCE EVALUATION
Accuracy Assessment
Classification performance metrics
Precision & Recall
Misinformation detection sensitivity
F1-Score Analysis
Balanced performance evaluation
KEY FINDINGS
Text-Only Limitations
Baseline model struggled with
nuanced misinformation patterns
Knowledge Graph Value
Domain-specific knowledge
improves detection capability
Hybrid Approach Success
Combined text + KG features
achieves optimal performance

Fig. 2. End-to-end data pipeline for the framework

- **Data-Augmented Model (SMOTE)**: The effect of addressing class imbalance in the training data was examined by applying the Synthetic Minority Oversampling Technique (SMOTE) before model training.
- **Knowledge Graph-Only Model**: To assess the efficacy of structured medical knowledge in isolation, this model was trained using only features derived from the Knowledge Graph.

The performance of our hybrid LLM+KG pipeline notably surpasses other simpler approaches achieving high accuracy and balanced precision/recall, as detailed in Table III. In particular, our integrated model obtained an accuracy of 76%, with Precision = 78%, Recall = 72%, and an F1-score = 75%. This is a notable improvement over a text-only baseline (DistilBERT without KG) which only reached about 50% accuracy with precision, recall, and F1 all around 0.50, scarcely better than random guessing. The boost in Recall for our hybrid model from 50% (baseline) to 72% is especially significant, as it indicates the model catches a far greater portion of misinformation posts, critical in a domain where missed false claims can be harmful. Meanwhile, maintaining a Precision of 78% ensures that legitimate health information is seldom misclassified, preserving user trust. The balance is reflected in the high F1-score. The Feature-Augmented Model showed only marginal improvement over the baseline (55% accuracy), and the Knowledge Graph-Only Model performed significantly lower (30% accuracy), underscoring the limitations of relying solely on structured knowledge without textual context. Training with SMOTE to address class imbalance (Data-Augmented Model) also resulted in a modest accuracy of 53%. These comparative results emphasize the performance gains achieved by the sophisticated integration in our hybrid LLM+KG Pipeline.

To put these results in context, traditional machine learning methods for similar tasks often struggled with recall (as low as 12% in some prior cases). Our approach, by incorporating

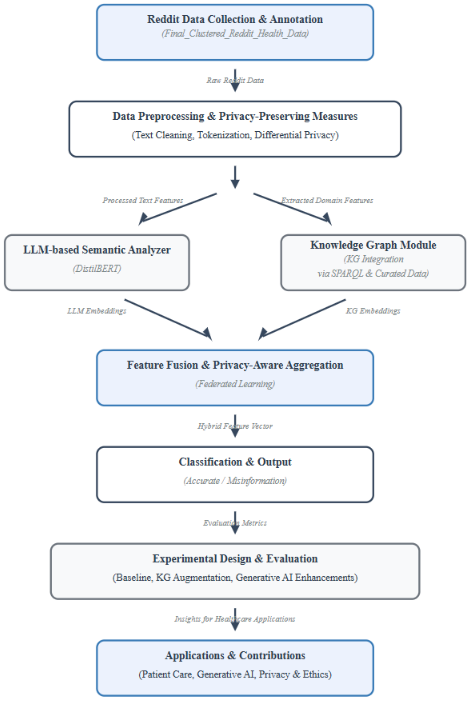

Reddit Data Collection & Annotation
(Final_Clustered_Reddit_Health_Data)
Raw Reddit Data
Data Preprocessing & Privacy-Preserving Measures
(Text Cleaning, Tokenization, Differential Privacy)
Processed Text Features
Extracted Domain Features
LLM-based Semantic Analyzer
(DistilBERT)
Knowledge Graph Module
(KG Integration
via SPARQL & Curated Data)
LLM Embeddings
KG Embeddings
Feature Fusion & Privacy-Aware Aggregation
(Federated Learning)
Hybrid Feature Vector
Classification & Output
(Accurate / Misinformation)
Evaluation Metrics
Experimental Design & Evaluation
(Baseline, KG Augmentation, Generative AI Enhancements)
Insights for Healthcare Applications
Applications & Contributions
(Patient Care, Generative AI, Privacy & Ethics)

Fig. 3. System Architecture

deep contextual and factual knowledge, elevates recall to approximately 72%, representing a significant improvement. The experimental findings clearly illustrate the effectiveness of integrating LLMs with KGs for health misinformation detection. The hybrid model not only improves raw performance metrics but does so in a balanced manner, successfully identifying most false claims without overwhelming moderators with false alerts—a crucial balance for real-world deployment.

From the results, we also observe the synergistic effect of combining modalities. An ablation analysis showed that neither component alone could achieve comparable results. The KG-only model's performance was poor (around 30-33% accuracy), and the text-only model, as previously noted (approx. 50% accuracy), was also substantially outperformed by the hybrid approach. Similarly, intermediate variants, such as enriching the text-only model with additional engineered features, yielded only slight improvements (accuracies of 53-55%). Only the combined LLM+KG model surpassed a 75% F1-score. These comparisons validate that the LLM and KG components contribute complementary strengths: the LLM captures how things are said, while the KG verifies what is said against established facts. When both signals agree (e.g., suspicious language and factual inconsistency), the model confidently flags misinformation; even when one component signals an issue and the other is neutral, the model can still detect cases that a single approach might miss.

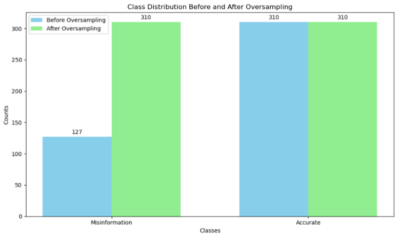


Fig. 4. Class Distribution

Statistical significance tests (two-tailed t-test over cross-validation folds) confirmed that the hybrid model's gains in F1-score over the baseline are significant (p¡0.01), suggesting the improvement is robust. Furthermore, the model's confidence scores were well-calibrated, with high-confidence predictions corresponding to a precision over 90%, useful for potential automation on the most certain detections.

To better illustrate the performance difference, consider the following table comparing the baseline model with the combined pipeline across key metrics:

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Baseline Model	50	50	50	50
Feature Augmented Model	55			
Data Augmented Model	53			
Knowledge Graph Only Model	33			
Combined Pipeline	76	78	72	75

TABLE III
HYBRID MODEL RESULT

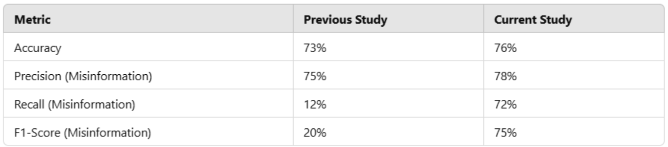
While the 5-fold cross-validation provides an estimate of the model's performance on unseen portions of the specific Reddit dataset, the study does not present results from testing on entirely separate, independently collected datasets that would demonstrate broader, out-of-distribution generalizability. The reliance on manually encoding specific misinformation examples from the dataset into the Knowledge Graph and subsequent validation on held-out portions of that same dataset means that the reported performance primarily reflects in-distribution generalization.

A. *Comparative Analysis*

The hybrid model, fusing LLM-derived embeddings with Knowledge Graph (KG) features, significantly outperforms approaches relying on only text or only KG information, as evidenced by the metrics in Table III. The substantial improvements in recall (72%) and overall F1-score (75%) for the combined pipeline, compared to the much lower scores for text-only (approx. 50% F1) and KG-only (approx. 30-33% accuracy) models, highlight that integrating domain-specific KG features via a sophisticated fusion mechanism effectively addresses the limitations of conventional NLP methods. While simpler feature augmentation provided only marginal gains over the baseline, the proposed hybrid approach delivers a more robust and balanced performance, proving more effective at capturing nuanced misinformation patterns.

Real-Time Feasibility

A primary objective of our framework is to achieve real-time



Metric	Previous Study	Current Study
Accuracy	73%	76%
Precision (Misinformation)	75%	78%
Recall (Misinformation)	12%	72%
F1-Score (Misinformation)	20%	75%

Fig. 5. Comparative Analysis

detection of health misinformation with minimal latency. Our design emphasizes computational efficiency through several key strategies: Utilizing DistilBERT—a lightweight transformer approximately 60% faster than BERT—the system processes a single post (up to 128 tokens) in tens of milliseconds on modern GPUs. Entity recognition is performed using simple rule-based methods with a compact NER model, while Node2Vec embeddings are precomputed and retrieved via constant-time database queries. Limiting the number of entities per post further controls computational load. Simulated streaming tests confirm that the entire pipeline sustains high throughput. The design supports distributed deployment, with entity extraction and KG querying running as sidecar services and LLM inference on dedicated GPU servers. Federated learning further facilitates horizontal scaling across multiple nodes.

B. Implications

The proposed framework's strong performance and design have significant implications for public health, clinical practice, and the broader digital ecosystem:

- **Public Health Surveillance**: The ability to detect health misinformation in real time facilitates early interventions, enabling public health authorities to mitigate the spread of harmful claims and support timely corrective actions. This capacity is crucial for safeguarding vulnerable populations and maintaining public trust in healthcare systems.
- **Enhanced Digital Health Platforms**: The scalable, privacy-preserving architecture offers a robust solution for content moderation on social media and health forums. By integrating LLMs with structured knowledge, the system ensures accurate identification of misleading information while adhering to data protection standards.
- **Future AI and Health Applications**: The demonstrated synergy between semantic analysis and factual verification underscores the potential of hybrid models. This approach encourages further investigation into adaptive fusion techniques and more efficient integration of structured domain knowledge, paving the way for advanced generative AI applications in digital health.

The above reinforce the framework's capacity to address both technical challenges and real-world needs in combating online health misinformation.

V. Conclusion

This study introduces a hybrid framework that integrates the semantic capabilities of Large Language Models (LLMs) with the factual rigor of domain-specific Knowledge Graphs (KGs) to detect online health misinformation. By combining a fine-tuned DistilBERT model with KG-derived features—and bolstering the approach with privacy-preserving techniques such as federated learning and differential privacy—the framework overcomes the limitations of traditional NLP methods.

Our experimental results demonstrate that the combined pipeline not only achieves balanced performance (e.g., 76% accuracy with precision of 78% and recall of 72%) but also effectively supports healthcare providers in resource management, personalized care, and timely intervention. Moreover, the integration of structured medical knowledge ensures greater interpretability and robustness, setting a new standard for scalable and ethically sound digital health solutions.

A. Limitations and Future Directions

We rightly acknowledge that despite the promising results of our integrated Knowledge Graph-LLM approach, several areas warrant further investigation to enhance its utility and robustness. Central to the framework's global applicability and generalizability, is its validation on a single, curated Reddit dataset. This includes a culturally tailored Knowledge Graph. Validation on diverse, independent external datasets is essential to affirm broader applicability beyond this specific context. Dynamic knowledge graph updates will also ensure resilience against evolving misinformation, addressing computational demands for practical real-time deployment, and maintaining a continued focus on ethical considerations such as bias and fairness. Successfully tackling these challenges, with a particular emphasis on demonstrating robust, out-of-distribution generalizability, will be vital for the framework's effective real-world application in combating health misinformation.

Lastly, to enhance the robustness and representativeness of our annotations, future work will engage a more diverse group of medical professionals, including specialists in diabetes, hypertension, and weight management from varied geographical locations such as Nigeria, the UK, and the US. This international and multi-specialty approach will help ensure a more comprehensive, culturally nuanced, and globally representative annotation of health claims. We also intend to implement formal inter-rater reliability (IRR) testing among these experts.

Acknowledgment This work was partially funded by NSF award 2310844, IUCRC Phase II UMBC: Center for Accelerated Real-Time Analytics (CARTA) and by UMBC's Cybersecurity Graduate Fellows program. The authors also acknowledge the assistance of medical experts for annotating the dataset, which was crucial to the success of this project.

References

[1] V. Suarez-Lledo and J. Alvarez-Galvez, "Prevalence of health misinformation on social media: systematic review," *Journal of medical Internet research*, vol. 23, no. 1, p. e17187, 2021.

[2] L. M. Malki, D. Patel, and A. Singh, """ the headline was so wild that i had to check": An exploration of women's encounters with health misinformation on social media," *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, no. CSCW1, pp. 1–26, 2024.

[3] S. Kisa and A. Kisa, "A comprehensive analysis of covid-19 misinformation, public health impacts, and communication strategies: Scoping review," *Journal of Medical Internet Research*, vol. 26, p. e56931, 2024.

[4] I. J. B. Do Nascimento, A. B. Pizarro, J. M. Almeida, N. Azzopardi-Muscat, M. A. Gonçalves, M. Björklund, and D. Novillo-Ortiz, "Infodemics and health misinformation: a systematic review of reviews," *Bulletin of the World Health Organization*, vol. 100, no. 9, p. 544, 2022.

[5] O. Clark, K. P. Joshi *et al.*, "Evaluating causal ai techniques for health misinformation detection," in *Causal AI for Robust Decision Making (CARD 2025) Workshop, held in conjunction with 23rd International Conference on Pervasive Computing and Communications (PerCom 2025)*, 2025.

[6] F. Alkhawaldeh, "False textual information detection, a deep learning approach," Ph.D. dissertation, University of York, 2022.

[7] H. Allam, L. Makubvure, B. Gyamfi, K. N. Graham, and K. Akinwolere, "Text classification: How machine learning is revolutionizing text categorization," *Information*, vol. 16, no. 2, p. 130, 2025.

[8] A. Mohasseb, E. Amer, F. Chiroma, and A. Tranchese, "Leveraging advanced nlp techniques and data augmentation to enhance online misogyny detection," *Applied Sciences*, vol. 15, no. 2, p. 856, 2025.

[9] F. Petroni, T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, and S. Riedel, "Language models as knowledge bases?" *arXiv preprint arXiv:1909.01066*, 2019.

[10] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, "Unifying large language models and knowledge graphs: A roadmap," *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[11] S. Hossain, A. Altarawneh, and J. Roberts, "Leveraging large language models and machine learning for smart contract vulnerability detection," *arXiv preprint arXiv:2501.02229*, 2025.

[12] D. M. Biji and Y.-W. Kim, "Evaluating the performance of large language models in classifying numerical data," in *2024 15th International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 2024, pp. 840–844.

[13] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for federated learning on user-held data," *arXiv preprint arXiv:1611.04482*, 2016.

[14] G. Alter and R. Gonzalez, "Responsible practices for data sharing." *American Psychologist*, vol. 73, no. 2, p. 146, 2018.

[15] P. Shojaei, E. Vlahu-Gjorgievska, and Y.-W. Chow, "Security and privacy of technologies in health information systems: A systematic literature review," *Computers*, vol. 13, no. 2, p. 41, 2024.

[16] P. Esmaeilzadeh *et al.*, "Privacy concerns about sharing general and specific health information on twitter: quantitative study," *JMIR Formative Research*, vol. 8, no. 1, p. e45573, 2024.

[17] S. S. Mahadik, P. M. Pawar, R. Muthalagu, N. R. Prasad, S.-K. Hawkins, D. Stripelis, S. Rao, P. Ejim, and B. Hecht, "Digital privacy in healthcare: State-of-the-art and future vision," *IEEE Access*, 2024.

[18] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi, "The spreading of misinformation online," *Proceedings of the national academy of Sciences*, vol. 113, no. 3, pp. 554–559, 2016.

[19] J. Cole, C. Watkins, and D. Kleine, "Health advice from internet discussion forums: how bad is dangerous?" *Journal of medical Internet research*, vol. 18, no. 1, p. e4, 2016.

[20] B. Swire-Thompson, D. Lazer *et al.*, "Public health and online misinformation: challenges and recommendations," *Annu Rev Public Health*, vol. 41, no. 1, pp. 433–451, 2020.

[21] J. Walker, M. Remski, and D. Beres, *Conspirituality: How new age conspiracy theories became a public health threat*. Random House Canada, 2023.

[22] Q. Su, M. Wan, X. Liu, C.-R. Huang *et al.*, "Motivations, methods and metrics of misinformation detection: an nlp perspective," *Natural Language Processing Research*, vol. 1, no. 1-2, pp. 1–13, 2020.

[23] A. Mallik and S. Kumar, "Word2vec and lstm based deep learning technique for context-free fake news detection," *Multimedia Tools and Applications*, vol. 83, no. 1, pp. 919–940, 2024.

[24] M. A. Al-Tarawneh, O. Al-irr, K. S. Al-Maaitah, H. Kanj, and W. H. F. Aly, "Enhancing fake news detection with word embedding: A machine learning and deep learning approach," *Computers*, vol. 13, no. 9, p. 239, 2024.

[25] J. Wang, X. Wang, and A. Yu, "Tackling misinformation in mobile social networks a bert-lstm approach for enhancing digital literacy," *Scientific Reports*, vol. 15, no. 1, p. 1118, 2025.

[26] S. Kukreja, T. Kumar, A. Purohit, A. Dasgupta, and D. Guha, "A literature survey on open source large language models," in *Proceedings of the 2024 7th International Conference on Computers in Management and Business*, 2024, pp. 133–143.

[27] G. Bai, Z. Chai, C. Ling, S. Wang, J. Lu, N. Zhang, T. Shi, Z. Yu, M. Zhu, Y. Zhang *et al.*, "Beyond efficiency: A systematic survey of resource-efficient large language models," *arXiv preprint arXiv:2401.00625*, 2024.

[28] B. Abu-Salih, "Domain-specific knowledge graphs: A survey," *Journal of Network and Computer Applications*, vol. 185, p. 103076, 2021.

[29] L. Cui, H. Seo, M. Tabar, F. Ma, S. Wang, and D. Lee, "Deterrent: Knowledge guided graph attention network for detecting healthcare misinformation," in *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 2020, pp. 492–502.

[30] J. Liu, J. Zhang, M. A. Jan, R. Sun, L. Liu, S. Verma, and P. Chatterjee, "A comprehensive privacy-preserving federated learning scheme with secure authentication and aggregation for internet of medical things," *IEEE journal of biomedical and health informatics*, vol. 28, no. 6, pp. 3282–3292, 2023.

[31] K. Meduri, G. S. Nadella, A. R. Yadulla, V. K. Kasula, M. H. Maturi, S. Brown, S. Satish, and H. Gonaygunta, "Leveraging federated learning for privacy-preserving analysis of multi-institutional electronic health records in rare disease research," *Journal of Economy and Technology*, 2024.

[32] M. Abaoud, M. A. Almuqrin, and M. F. Khan, "Advancing federated learning through novel mechanism for privacy preservation in healthcare applications," *IEEE Access*, vol. 11, pp. 83 562–83 579, 2023.

[33] I. Zhou, F. Tofigh, M. Piccardi, M. Abolhasan, D. Franklin, and J. Lipman, "Secure multi-party computation for machine learning: A survey," *IEEE Access*, 2024.

[34] J. Jonnagaddala and Z. S.-Y. Wong, "Privacy preserving strategies for electronic health records in the era of large language models," *npj Digital Medicine*, vol. 8, no. 1, p. 34, 2025.

[35] E. Papageorgiou, C. Chronis, I. Varlamis, and Y. Himeur, "A survey on the use of large language models (llms) in fake news," *Future Internet*, vol. 16, no. 8, p. 298, 2024.

[36] J. Lucas, A. Uchendu, M. Yamashita, J. Lee, S. Rohatgi, and D. Lee, "Fighting fire with fire: The dual role of llms in crafting and detecting elusive disinformation," *arXiv preprint arXiv:2310.15515*, 2023.

[37] C. Chen and K. Shu, "Combating misinformation in the age of llms: Opportunities and challenges," *AI Magazine*, vol. 45, no. 3, pp. 354–368, 2024.

[38] V. S. Pendyala and C. E. Hall, "Explaining misinformation detection using large language models," *Electronics*, vol. 13, no. 9, p. 1673, 2024.

[39] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[40] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[41] J. Ramos *et al.*, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, vol. 242, no. 1. Citeseer, 2003, pp. 29–48.

[42] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[43] C. Dwork, "Differential privacy," in *International colloquium on automata, languages, and programming*. Springer, 2006, pp. 1–12.

[44] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 308–318.

[45] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[46] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Machine Intelligence*, vol. 2, no. 6, pp. 305–311, 2020.

[47] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.