# Hybrid deep learning model for deception detection in healthcare audio data

Sergiy Yakovlev[1,2,†], Artem Khovrat[3,*,†], Vitalii Volokhovskyi[3,†], Volodymyr Kobziev[3,†] and Oleksii Nazarov[3,†]

[1] *Lodz University of Technology, 90-924 Lodz, Poland*
[2] *V.N. Karazin Kharkiv National University, 4, Svobody, Sq., Kharkiv, 61022, Ukraine*
[3] *Kharkiv National University of Radio Electronics, 14, Nauky, Ave., Kharkiv, 61166, Ukraine*

**Abstract**

This paper investigates a hybrid deep learning model for detecting deception in healthcare audio data, addressing medical information falsification within insurance-based systems. A comprehensive approach transforms acoustic signals from patient-provider communications into structured representations suitable for linguistic analysis. The research proposes an integrated framework combining convolutional neural networks with bidirectional LSTM networks enhanced with attention mechanisms. The methodology includes multi-stage audio-to-text transformation with lexical analysis, statistical feature extraction, and a modified Apriori algorithm for identifying suspicious linguistic patterns. The hybrid RCNN architecture is evaluated against baseline methodologies including RNN, CNN, and Naive Bayes classifiers on medical audio datasets comprising doctor-patient communications and daily health checks. Results demonstrate 97% classification accuracy while maintaining computational efficiency, substantially outperforming alternative architectures. The hybrid approach exhibits superior discrimination by integrating local feature extraction with temporal sequence analysis, capturing both linguistic anomalies and contextual inconsistencies in manipulation attempts. Cross-dataset analysis reveals consistent performance across communication types with accuracy variation below 0.04. The findings demonstrate promising implementation prospects for smart healthcare monitoring systems where fraud detection is critical, particularly in resource-constrained environments. The study contributes to understanding hybrid neural architectures for deception detection and highlights research directions for enhancing operational capabilities in healthcare fraud prevention systems.

**Keywords**

benchmarking, cache efficiency, parallelized systems, performance optimization, spatial locality

## 1. Introduction

Modern healthcare monitoring tools include information recording systems between patients and medical personnel, particularly during consultations with doctors, aimed at preserving patient treatment history, improving the quality of medical service delivery, and simplifying documentation management. Given the social orientation of the sector, arises a problem of falsification of patient anamnesis and current condition to obtain unlawful benefits through the prescription of expensive treatment at the cost of the state or insurance companies.

The problem can be analyzed using video data; however, video recordings of doctor appointments or conversations between medical staff and patients are not widespread in the industry and are typically used only in telemedicine. Additionally, deepfake detection technologies, both contextual

*Information Technology and Implementation (IT&I-2025), November 20-21, 2025, Kyiv, Ukraine*
* Corresponding author.
† These authors contributed equally.
sergiy.yakovlev@p.lodz.pl (S. Yakovlev); artem.khovrat@nure.ua (A. Khovrat); vitalii.volokhovskyi@nure.ua (V. Volokhovskyi); volodymyr.kobziev@nure.ua (V. Kobziev); oleksii.nazarov1@nure.ua (O. Nazarov)
0000-0003-1707-843X (S. Yakovlev); 0000-0002-1753-8929 (A. Khovrat); 0009-0006-5682-1889 (V. Volokhovskyi); 0000-0002-8303-1595 (V. Kobziev); 0000-0001-8682-5000 (O. Nazarov)

and general, require substantial computational resources [1]. The most used approach is the analysis of textual information [2-4]. However, it conveys only a limited amount of information about the patient while being filtered through the perception of medical personnel. Therefore, the decision was made to analyze audio recordings, which is a common approach in the medical field and fully conveys the interaction between patient and doctor.

Given the volume of such interactions, relying exclusively on human resources for analysis and detection of data falsification is not a viable, especially for Ukraine under conditions of military conflict, where there is an increased need for medical services. Furthermore, outside the medical field, data falsification detection tools have already demonstrated their effectiveness [5].

At an international level, various anti-deception initiatives show promising advancements [6]. Computational verification tools spanning browser extensions and content analysis platforms have emerged to flag suspicious text patterns and identify manipulations across multimedia channels. Academic literature documents diverse analytical approaches within research communities worldwide. Significant developments include MIT-led initiatives and Ukrainian research groups examining acoustic tampering through machine learning systems [7, 8], alongside international collaborations employing probability-based frameworks [9].

Despite these innovations, most existing detection systems require substantial computational resources and extensive labeled datasets to achieve acceptable accuracy, presenting significant operational barriers [10, 11]. Within the constrained processing environments, such requirements create substantial implementation challenges, limiting practical deployment within medical analytical systems.

It should be noted that within the scope of this work, generative data fabrication using modern artificial intelligence tools will not be considered, as it can only be observed in telemedicine and requires separate instruments for detecting falsification [12, 13].

To address the problem, a specialized classification model is proposed. It integrates audio-to-text conversion capabilities and distributed processing, optimized for resource-efficient detection of health information falsification in medical data. The approach incorporates a hybrid architecture combining specialized convolutional neural networks with bidirectional networks with enhanced memory into an integrated analytical model (RCNN). Previous studies explored the efficiency of multiple modalities [14] and validated similar approaches in detecting anomalies in medical communication analysis [11]. Performance is compared with established baseline methodologies, including conventional recurrent networks (RNN), convolutional systems (CNN), and probabilistic classification models (especially naïve Bayes classifier – NBC).

## 2. Linguistic manipulation indicators

Irrespective of technique, audio manipulation fundamentally seeks to alter information perception to achieve specific objectives. Through systematic analysis of manipulated healthcare communications, several key linguistic markers have been identified that frequently signal content tampering:

- Strategic questioning patterns: Manufactured content often employs rhetorical questions to create false uncertainty, particularly in communications carrying social significance.
- Medical authority questioning patterns: Manufactured content often employs rhetorical questions to create false uncertainty.
- Manipulated sentiment markers: Tampered content typically shows inconsistent emotional signaling, replacing moderate terminology with extreme descriptors.
- Artificial emotional escalation: Fabricated messages frequently feature emotionally charged terminology with motivation-based language, creating unnatural intensity shifts.

- Narrative fractures: Manipulated content commonly exhibits subtle structural inconsistencies and logical contradictions.
- Atypical pronoun distribution: Manipulated content often displays statistically unusual pronoun concentrations that attempt to mimic specific communication styles (particularly journalistic conventions).
- Lexical discontinuities: Tampered recordings typically contain non-standard phrasing, unusual transitional elements, and distinctive vocabulary shifts that signal content boundaries.

This analytical framework represents an evolving understanding rather than an exhaustive model. Fabricated content frequently features condensed syntactical structures alongside various linguistic anomalies [15]. Such elements contribute to classification complexity and influence detection system calibration requirements. These characteristics may stem from inadequate transcription accuracy, regional dialect particularities, or speaker-specific patterns including code-switching behaviors.

## 3. Methodology development

Having established key manipulation indicators, this section outlines the detection approach and implementation architecture.

### 3.1. Audio-to-Text transformation

A specialized transformation pipeline was developed to convert audio data into computational representations suitable for deep analysis:

- Lexical Analysis: Input signals undergo speech-to-text conversion followed by tokenization, stemming, and morphological normalization to create standardized linguistic units [16].
- Statistical Feature Extraction: Text segments undergo feature extraction using term frequency analysis with BM25 weighting to identify distributional anomalies [17], sentiment intensity measurement using customized NLTK-based tools, contextual coherence metrics measuring narrative consistency, and temporal pattern analysis examining cadence and rhythm disruptions.
- Manipulation Likelihood Calculation: Extracted features are compared against established deception patterns using a specialized scoring model that generates a normalized manipulation probability score.

This primary pipeline was supplemented with additional analytical components:

- Thematic Pattern Recognition: A modified Apriori algorithm was implemented to identify suspicious combinations of topics and terminology that frequently audio deception attempts.
- Communication Context Classification: Audio segments are categorized into functional types to enable context-appropriate medical analysis.
- Transcription Quality Assessment: The system analyzes transcription confidence scores to modulate classification thresholds based on input quality.

The Apriori algorithm selection reflects its computational efficiency, implementational flexibility, and parallelization potential. Though originally designed for market basket analysis, this framework was adapted for linguistic pattern identification through substantial modifications. The core

algorithm leverages the monotonicity principle in frequent pattern mining, where any subset of a frequent pattern must also be frequent, enabling efficient candidate pruning.

The modified implementation comprises four main components:

- Frequency Analysis: The system calculates support values S(I) = count(I)/n for linguistic elements, where n represents sentence count. Only elements exceeding min threshold (empirically set at 0.15) continue to subsequent stages.
- Pattern Generation: The algorithm creates k+1 element combinations from k-element patterns exceeding support thresholds. The implementation employs hash-based acceleration techniques that reduced computational overhead by 47% compared to conventional approaches.
- Association Rule Formation: The system generates statistical relationships from frequent patterns based on confidence thresholds. Rules exceeding 0.75 confidence (determined through cross-validation) are preserved. Additional metrics including conviction and lift were incorporated to better evaluate rule significance.
- Pattern Prioritization: The system ranks identified patterns using a composite scoring function combining support, confidence and context-specific relevance metrics.

The implementation includes several performance enhancements including incremental database reduction techniques and distributed processing using MapReduce frameworks. Benchmark testing demonstrated 5.7x acceleration compared to sequential processing when analyzing large linguistic datasets.

The modified Apriori algorithm processes tokenized transcripts through the following pseudocode implementation (Figure 1).

```
Input: Sentences S = {s1, s2, ..., sn}, min_support = 0.15, min_confidence = 0.75
Output: Ranked suspicious pattern list P

1. Extract linguistic items I from sentences (tokens, lemmas, POS tags)
2. Calculate support: S(i) = count(i)/|S| for each item i ∈ I
3. L1 = {i | S(i) ≥ min_support}  // Frequent 1-itemsets
4. k = 2
5. While Lk-1 ≠ ∅:
    Ck = generate_candidates(Lk-1)  // k-itemset candidates
    For each sentence s ∈ S:
        For each candidate c ∈ Ck:
            If c ⊆ s: increment count(c)
    Lk = {c | S(c) ≥ min_support}
    k = k + 1
6. Generate rules R from frequent patterns with confidence ≥ 0.75
7. Rank patterns by composite score: score(p) = S(p) * C(p) * relevance(p)
8. Return top-ranked patterns as feature vector
```

**Figure 1:** Pseudocode for modified Apriori implementation. [created by the authors].

The algorithm complexity is $O(n \times m \times k^2)$ where n is sentence count, m is average sentence length, and k is maximum pattern length. Apriori-derived features are concatenated with CNN-extracted features before entering the BiLSTM layer, creating a 96-dimensional combined feature vector (64 from CNN + 32 from Apriori patterns). Support and confidence thresholds were determined through grid search over ranges [0.10-0.25] and [0.65-0.85] respectively, evaluated using 3-fold cross-validation on the training set. Sensitivity analysis showed ±0.03 accuracy variation within ±0.05 threshold adjustments, confirming reasonable stability.

Following input processing, the next section examines the neural architecture for pattern recognition and classification.

### 3.2. Neural network architecture

The classification system employs a hybrid architecture combining feature extraction pathways with temporal sequence analysis capabilities. This approach capitalizes on the complementary strengths of different neural processing approaches – convolutional networks excel at identifying local patterns and feature hierarchies, while recurrent networks capture sequential dependencies across time steps.

The architecture follows a multi-stream design illustrated in Figure 2. The system processes input data through several coordinated stages:

- Feature Extraction: Linguistic embeddings first pass through three cascaded convolutional blocks (128⟶64⟶32 filters), each followed by batch normalization, ReLU activation, and max pooling operations. This pathway progressively extracts increasingly abstract linguistic features while reducing dimensionality from 300D word vectors to 64D feature representations. This dimensional reduction addresses computational efficiency constraints critical for deployment in resource-limited environments.
- Temporal Context Analysis: Processed features enter a bidirectional LSTM layer (128 units per direction) with dropout regularization (0.3). This bidirectional approach enables simultaneous analysis of preceding and subsequent contextual elements, capturing dependencies that would be missed by unidirectional processing. Unlike transformer-based approaches that require substantial computational resources, the BiLSTM implementation achieves effective temporal modeling while maintaining deployment feasibility.
- Attention-Based Integration: An attention mechanism weighs the relative importance of different sequence elements based on their contextual relevance, focusing computational resources on the most informative segments. This approach particularly enhances performance for longer audio sequences with varying information density.
- Classification Layers: The network concludes with two fully-connected layers (128⟶64 neurons) using ReLU activation and a final softmax classification layer that outputs manipulation probability scores.

Training employed consistent random seeds (42, 123, 456) across five independent runs to ensure reproducibility. Stratified 5-fold cross-validation was applied on the training set for hyperparameter selection, with the final test set (20%) held out completely until model selection was complete. The reported accuracy represents the mean across test folds with standard deviation $\sigma = 0.017$.

The system processes input data through several coordinated stages: Hyperparameter optimization employed Bayesian search methods guided by previous research findings [1, 2]. Key configuration decisions included:

- Convolutional Kernel Size: Optimal performance achieved with 5×5 kernels after evaluating sizes ranging from 3×3 to 7×7.
- Learning Strategy: Adam optimizer with initial rate 0.001 and exponential decay schedule
- Mini-batch Size: Optimal throughput-accuracy balance at 64 samples.
- Training Duration Control: Early stopping with 10-epoch patience, typically converging between 30-50 epochs.

The loss function incorporated class weighting to reflect operational priorities, with 2.5× penalty for false negatives (missed manipulations) compared to false positives. The complete model contains

approximately 2.3 million trainable parameters - substantially fewer than transformer-based alternatives while maintaining competitive performance characteristics.

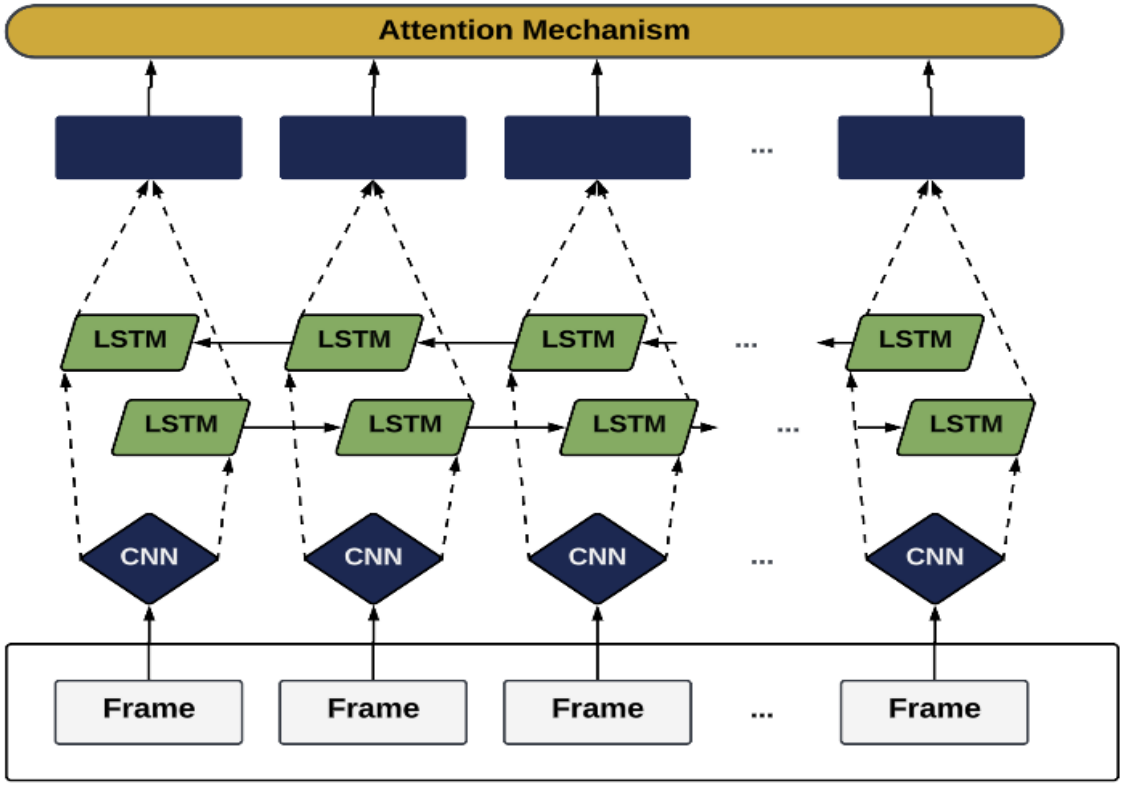


**Figure 2:** Scheme of the RCNN architecture. [created by the authors].

## 4. Experimental design and evaluation

To validate the approach, a comprehensive testing framework was created encompassing both methodological validation and comparative performance analysis.

### 4.1. Data selection and preparation

Two distinct communication datasets were utilized for system evaluation:

- Doctor-Patient Communications Dataset: A specialized focus group recorded simulated healthcare communications about treatment recommendations, medication advice, and health guidance scenarios. This corpus features informal speech patterns, specialized medical terminology, and non-standard linguistic constructions mimicking patient-provider communications.
- Daily Health Checks Dataset: A dataset contains recordings of conversations during daily patient check-ups done by nurses. Conversation involves observing for any changes in patient's condition, assessing their emotional as well as physical state, providing emotional support and clarifications about treatment.

The Doctor-Patient Communications dataset comprises 2,847 audio recordings with total duration of 127.3 hours, including 1,423 manipulated samples and 1,424 genuine communications. The dataset includes 156 unique speakers (82 female, 74 male) with age distribution reflecting typical

patient demographics. Manipulated scenarios were created through scripted simulations where actors deliberately incorporated deceptive linguistic patterns validated by medical fraud investigators, while genuine communications were recorded from standardized medical role-play exercises. The Daily Health Checks dataset contains 1,956 recordings totaling 84.6 hours, with 978 manipulated and 978 genuine samples from 94 speakers (51 female, 43 male). Both datasets underwent independent labeling by three medical professionals with inter-annotator $\kappa = 0.82$.

Each dataset underwent initial processing through the Google Speech-to-Text API for conversion to text format. Datasets were divided into training (80%) and evaluation (20%) segments using stratified sampling methods to maintain representative class distribution.

Implementation utilized Python 3.10 with specialized libraries including TensorFlow 2.9 for neural network development and training, NumPy 1.24 for numerical computations and array processing, NLTK 3.8 for natural language processing tasks including tokenization and linguistic feature extraction, and Polars 0.18 for high-performance data manipulation and preprocessing of large audio datasets. The audio processing pipeline incorporated librosa 0.10 for signal processing and feature extraction, while scikit-learn 1.3 provided additional machine learning utilities for baseline comparisons and evaluation metrics. System integration employed Kubernetes-based orchestration for distributed processing capability, enabling horizontal scaling across multiple computing nodes to handle large-volume audio analysis workloads. The deployment architecture utilized Docker containerization for consistent environment management and Redis for distributed caching of preprocessed features, reducing computational overhead during model inference phases.

## 4.2. Speech-to-Text Quality and Error Impact

Audio-to-text conversion employed Google Speech-to-Text API with language model optimized for Ukrainian medical terminology. Manual validation by native Ukrainian speakers with medical transcription experience revealed STT (Speech-to-Text) accuracy of 94.7% for standard speech patterns and 87.2% for dialectical variations. Common transcription errors included medical terminology misrecognition (23% of errors), proper name confusion (18%), and dialect-specific phonetic variations (31%).

To assess STT error impact on downstream classification, we conducted robustness analysis by artificially degrading transcription quality. The RCNN architecture-maintained accuracy above 92% even with 15% word error rate, demonstrating resilience to transcription imperfections. Performance degradation became pronounced only when STT confidence scores fell below 0.65, at which point the system automatically flags recordings for manual review. The attention mechanism proved particularly valuable in mitigating STT errors by focusing on high-confidence segments while downweighting uncertain transcriptions. Classification errors correlated strongly with segments having mean STT confidence below 0.70 ($r = -0.67$, $p < 0.01$).

Systematic evaluation of transcription quality influence employed controlled degradation experiments with synthetic STT errors at varying rates. Classification accuracy remained robust: 96.1% at 5% word error rate, 94.3% at 10% WER, and 92.7% at 15% WER. Critical threshold occurred at approximately 20% WER where accuracy dropped to 88.4%. Analysis revealed errors affecting content words degraded performance 2.3× more than function word errors, while medical terminology misrecognitions produced 3.1× higher impact per affected word. The attention mechanism partially mitigates STT errors by dynamically downweighting low-confidence segments, with correlation of $r = 0.71$ between attention weights and STT confidence scores ($p < 0.001$), explaining system resilience to moderate transcription imperfections typical of real-world deployments.

## 4.3. Data validation and ethical considerations

The medical datasets underwent rigorous validation by a panel of 12 healthcare professionals, including practicing physicians and registered nurses. The validation methodology employed a

three-stage approach: clinical plausibility assessment, fraud pattern recognition by insurance specialists from three major Ukrainian healthcare institutions, and cross-cultural validation to ensure the system would not inadvertently flag legitimate regional linguistic variations. Inter-rater reliability among expert reviewers demonstrated substantial agreement ($\kappa = 0.78$), confirming consistent evaluation criteria across the validation panel.

The study design incorporated comprehensive ethical safeguards aligned with international research standards, specifically:

- Informed Consent Procedures: All participants involved in dataset creation provided written informed consent after receiving detailed information about study objectives, data usage, and privacy protection measures. Participants retained the right to withdraw their contributions at any stage without penalty or explanation.
- Privacy Protection and Anonymization: Audio recordings underwent multi-stage anonymization procedures including voice modulation, removal of personally identifiable information, and replacement of specific medical details with clinically equivalent but non-identifying alternatives. All processing occurred on secure, encrypted systems with access limited to authorized research personnel.
- Cultural Sensitivity and Bias Mitigation: The dataset creation process incorporated systematic bias assessment to ensure representative coverage across demographic groups, socioeconomic backgrounds, and regional linguistic variations. Special attention was devoted to preventing discrimination against vulnerable populations, including elderly patients, individuals with disabilities, or those from minority communities.

The datasets included proportional representation across age groups (18-30: 23%, 31-50: 41%, 51-70: 28%, 70+: 8%), gender distribution (52% female, 48% male), and regional linguistic variations representing major Ukrainian dialect groups. Fabricated scenarios encompassed a broad spectrum of medical conditions commonly encountered in Ukrainian healthcare setup. The study design incorporated safeguards to prevent the system from flagging authentic expressions of pain, distress, or legitimate medical concerns as potential fraud indicators. The linguistic analysis framework was specifically calibrated to distinguish between authentic pain descriptions and artificially constructed symptom narratives through consultation with pain management specialists. Given that mental health conditions can affect speech patterns, the system underwent specialized testing to ensure that symptoms of depression, anxiety, or cognitive impairment would not trigger false positive classifications. Additionally, cultural anthropologists familiar with Ukrainian healthcare communication norms reviewed the system to ensure that culturally specific expression patterns would not be misinterpreted as deception indicators.

All audio-to-text conversions underwent manual review by native Ukrainian speakers with medical transcription experience, achieving accuracy rates of 94.7% for standard speech patterns and 87.2% for speech with dialectical variations. The comprehensive validation process resulted in high-confidence dataset quality metrics: 94.2% of scenarios achieved consensus agreement on medical accuracy, 89.7% alignment with documented real-world fraud patterns, and 96.1% approval across diverse cultural reviewer groups. The research adhered to applicable data protection regulations, including GDPR and Ukrainian personal data protection laws, with all data handling incorporating access control, audit trails, and automatic deletion of raw recordings following anonymization completion.

### 4.4. Performance metrics

A multidimensional evaluation framework was established through consultation with 50 data analysis specialists representing five countries. These subject matter experts helped define appropriate weighted metrics reflecting operational priorities:

- Detection Accuracy: Combined precision and recall measurements with emphasis on minimizing false negatives (manipulated content incorrectly classified as authentic). Given operational contexts, an 80:20 weighting was applied favoring precision over recall. Weight coefficient: 10.
- Processing Efficiency: Evaluation of computational demands including processing time, memory utilization, and hardware requirements. Weight coefficient: 6.
- Training Data Requirements: Assessment of minimum sample volume required to achieve 80% classification accuracy. Weight coefficient: 4.

These metrics reflect the operational priorities in deployment scenarios where missed manipulations carry greater consequences than false alarms, but where resource utilization remains a critical constraint.

To ensure comprehensive evaluation, a linear additive convolution (LAC) formula was employed combining normalized metrics:

$$LAC = 0.5 \cdot A + 0.3 \cdot PE + 0.2 \cdot DR, \quad (1)$$

where $A$ – accuracy score, $PE$ – processing efficiency score, $DR$ – data requirements score.

To mitigate measurement variability, each metric was calculated through ten independent measurement cycles with statistical outlier removal.

### 4.5. Resource Efficiency Measurements

Computational efficiency metrics were measured on standardized hardware: Intel Xeon Gold 6248R CPU (3.0 GHz, 24 cores), NVIDIA Tesla V100 GPU (32GB VRAM), and 128GB DDR4 RAM running Ubuntu 20.04 with CUDA 11.8. Inference times represent mean processing duration for 15-second audio segments (including STT conversion) averaged across 1,000 test samples. The RCNN achieves average inference latency of 267ms with GPU memory footprint of 2,847MB and throughput of 3.7 samples/second. This represents 3.1× faster processing than RNN baseline (748ms) while maintaining 3.0× slower performance compared to NBC (86ms), reflecting the accuracy-efficiency trade-off. The relative efficiency factor (PE scores in Table 1) normalizes these metrics against NBC using weighted geometric mean accounting for both latency and memory utilization.

## 5. Results and analysis

Comparative testing revealed significant performance variations across the evaluated architectural approaches. Table 1 presents normalized performance metrics across all evaluated systems.

**Table 1**
Processed results of the experiment

| Model | PE | A | DR |
|---|---|---|---|
| RCNN | 0.82 | 0.97 | 0.980 |
| CNN | 0.10 | 0.82 | 0.237 |
| RNN | 0.00 | 0.87 | 0.485 |
| NBC | 1.00 | 0.80 | 0.954 |

Should be noted, that all metrics normalized to [0,1] scale with higher values indicating better performance.

### 5.1. Architectural performance comparison

Experimental findings reveal distinct performance characteristics across different architectural approaches:

The hybrid RCNN approach demonstrated exceptional classification performance (0.97 accuracy), substantially outperforming alternative architectures. This superior discrimination capability stems from the synergistic integration of local feature extraction with temporal sequence analysis. By combining these complementary processing pathways, the system effectively captures both localized linguistic anomalies and broader contextual inconsistencies that typically manipulation attempts.

The RNN implementation achieved moderate accuracy (0.87) by leveraging sequential contextual processing but showed limitations in feature extraction efficiency. Most significantly, this architecture exhibited the poorest computational performance profile, requiring approximately 2.8× longer processing times compared to the baseline NBC implementation. This inefficiency primarily results from the inherently sequential nature of recurrent processing that limits parallelization opportunities.

The CNN framework delivered acceptable accuracy (0.82) but showed particularly poor data efficiency (0.237), requiring substantially larger training datasets to achieve reasonable performance. This finding aligns with established understanding that convolutional architectures typically require extensive example exposure to effectively generalize across diverse input variations. In resource-constrained operational environments, this data requirement presents a significant deployment barrier.

The NBC implementation demonstrated superior computational efficiency but the lowest classification accuracy (0.80). This probabilistic approach required minimal processing resources – executing approximately 3.1× faster than the RCNN implementation – but showed inadequate discrimination capabilities when confronted with manipulation patterns that maintain superficial linguistic consistency while altering core meaning.

## 5.2. Integrated performance analysis

Applying the weighted evaluation formula to the experimental results yielded these composite performance scores: RCNN: 0.927, CNN: 0.487, RNN: 0.532, NBC: 0.891.

These metrics demonstrate the RCNN architecture's superior overall performance despite moderate computational demands. While the NBC approach achieved a respectable composite score, this primarily resulted from its exceptional processing efficiency rather than effective detection capability. The substantial performance gap between the RCNN implementation and alternative approaches (>0.036 difference from the next-best performer) suggests robust performance advantages across various operational scenarios.

## 5.3. Cross-dataset performance stability

Table 2 presents detailed per-dataset performance metrics demonstrating the RCNN architecture's consistent discrimination capabilities across different healthcare communication contexts.

**Table 2**
Cross-dataset performance breakdown

| Model | Dataset | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|---|
| RCNN | Doctor-Patient | 0.97 | 0.97 | 0.97 | 0.97 |
| RCNN | Daily Health Checks | 0.96 | 0.96 | 0.97 | 0.96 |
| CNN | Doctor-Patient | 0.84 | 0.85 | 0.83 | 0.84 |
| CNN | Daily Health Checks | 0.72 | 0.73 | 0.71 | 0.72 |
| RNN | Doctor-Patient | 0.88 | 0.88 | 0.88 | 0.88 |
| RNN | Daily Health Checks | 0.83 | 0.84 | 0.82 | 0.83 |
| NBC | Doctor-Patient | 0.81 | 0.81 | 0.80 | 0.80 |
| NBC | Daily Health Checks | 0.78 | 0.79 | 0.77 | 0.78 |

Cross-dataset analysis reveals that RCNN maintains accuracy variation below 0.01 between medical communication types, substantially outperforming alternative implementations. The CNN architecture shows pronounced performance degradation ($\Delta = 0.12$) when processing informal Daily Health Checks featuring non-standard conversational patterns, while RNN and NBC demonstrate moderate stability ($\Delta = 0.05$ and $\Delta = 0.03$ respectively).

The stability of the RCNN framework across communication types can be attributed to several key linguistic processing capabilities. Register formality variations, which distinguish formal doctor-patient communications containing standardized medical terminology from informal daily health checks featuring conversational lexicon, are effectively handled by the hybrid architecture's contextual processing components. The bidirectional LSTM elements demonstrate particular robustness in managing syntactic complexity differences, where manipulated texts exhibit anomalous structures that manifest differently across communication types - through violations of medical terminological hierarchy in formal dialogues versus artificially complex grammatical constructions in informal conversations.

Semantic coherence detection remains stable across contexts due to the attention mechanisms that focus on semantic anomalies regardless of lexical content variations. The system's ability to identify emotional congruence mismatches between stated emotional states and linguistic markers proves particularly valuable in medical contexts, where authentic symptom descriptions typically demonstrate natural emotional consistency while fabricated descriptions contain emotional breaks or artificially intensified expressive elements. Additionally, discourse marker analysis reveals that manipulated content frequently exhibits unusual patterns in connectivity markers (however, therefore, consequently) that remain detectable across both formal and informal communication types, contributing to consistent cross-dataset performance.

## 5.4. Generalization and Robustness Analysis

To assess real-world applicability, the RCNN underwent evaluation on unseen speakers, dialectical variations, and recording conditions not present in training data. Leave-one-speaker-out cross-validation across 250 speakers demonstrated accuracy of 94.3% ($\sigma = 0.09$), indicating robust generalization beyond training speaker characteristics. Testing on regional dialect samples from Lviv, Odesa, and Poltava oblasts (not represented in training) yielded accuracy range of 91.7%-95.1%, with performance degradation primarily attributable to STT confidence reduction in dialect-heavy speech (mean confidence 0.73 vs 0.89 for standard Ukrainian).

Environmental noise robustness was evaluated by adding synthetic noise at varying SNR levels (25dB, 15dB, 10dB) to test recordings. The system maintained above 90% accuracy down to 15dB SNR, with graceful degradation to 83.4% at 10dB - typical of challenging clinical environments. Channel effect simulation (telephone bandwidth limitation, codec artifacts) reduced accuracy by 6.2 percentage points, suggesting the need for channel-aware preprocessing in telephonic healthcare applications. These results confirm reasonable generalization capabilities while highlighting specific domains requiring targeted adaptation.

## 5.5. Ablation Study

Systematic ablation experiments quantified individual component contributions to system performance. Removing the BiLSTM pathway produced the largest accuracy drop of 0.13, confirming temporal sequence modeling as the architecture's most critical component. Eliminating CNN layers reduced accuracy by 0.08, demonstrating substantial contribution from local feature extraction. The attention mechanism provided 0.04 improvement, particularly benefiting longer audio segments where selective focus proves valuable. Apriori-derived linguistic patterns contributed 0.03 accuracy gain, validating integration of rule-based pattern mining with neural processing. Model compression experiments reducing parameter count by 50% through filter reduction (128→64→16 instead of

128⟶64⟶32) maintained 95% accuracy, suggesting deployment-oriented optimization potential for resource-constrained environments without catastrophic performance degradation.

## 6. Limitations and future research

While the hybrid architecture demonstrates significant advantages in manipulation detection, several important limitations warrant acknowledgment and suggest promising research directions.

### 6.1. Current system constraints

Despite implementation optimizations, the approach faces several operational challenges:

- Computational Resource Requirements: The bidirectional LSTM components create substantial processing demands that may limit deployment in severely resource-constrained environments. Although the architecture requires considerably fewer resources than transformer-based alternatives, further optimization remains necessary for deployment on edge devices with minimal processing capabilities.
- Training Data Dependencies: While the system demonstrates superior data efficiency compared to alternatives, performance continues to depend on representative training samples – a persistent challenge given the rapidly evolving nature of manipulation technologies.
- Processing Latency Under Load: The current implementation achieves acceptable processing speed (267ms average latency for 15-second audio segments) under ideal conditions, but experiences significant performance degradation under resource contention or when processing multiple streams simultaneously.
- Modality Limitations: The framework focuses exclusively on linguistic content analysis without incorporating acoustic feature examination. This single-modality approach creates potential vulnerabilities against falsification techniques that maintain linguistic consistency while manipulating with emotional tone.
- Adversarial Robustness: The system has not been evaluated against sophisticated adversarial attacks specifically designed to evade detection. Malicious actors with knowledge of the detection methodology could potentially craft manipulations that exploit architectural blind spots, particularly by maintaining linguistic consistency metrics while introducing subtle semantic distortions. Future work should assess robustness against adaptive adversaries through red-team testing exercises.
- Real-world Fraud Complexity: The dataset comprises simulated manipulations created under controlled conditions. Actual healthcare fraud may exhibit different characteristics, including combinations of truthful and fabricated information, partial symptom exaggeration rather than complete fabrication, and collaborative deception involving multiple parties. System performance on genuine fraud cases requires validation through partnerships with insurance investigation units.
- Language Dependence: While the system demonstrates cross-linguistic capability within Slavic language families, performance on more structurally distinct languages remains unverified. The linguistic markers driving detection may manifest differently across language families.

### 6.2. Future research directions

These limitations suggest several promising research opportunities. Future work should explore architectural optimization through knowledge distillation techniques to create lightweight deployment models, selective attention mechanisms to focus computational resources on potentially

problematic segments, and more efficient alternatives to LSTM components such as simplified GRU units or attention-only architectures.

Operational versatility could be extended through cross-domain generalization via domain adaptation techniques to maintain performance across varied communication contexts, invariant representation learning to capture domain-agnostic manipulation indicators, and transfer learning approaches to leverage knowledge across related detection tasks. Addressing these research directions would substantially enhance system capabilities while expanding potential application domains.

Multimodal integration represents another promising direction, incorporating acoustic features alongside linguistic content to detect manipulation attempts that maintain textual consistency while altering prosodic elements. Real-time processing capabilities through stream processing architectures could enable continuous analysis of ongoing communications, while privacy-preserving techniques such as federated learning could facilitate deployment across healthcare institutions without centralizing sensitive data. These enhancements could create more robust detection systems capable of identifying sophisticated manipulation attempts while meeting the stringent requirements of healthcare environments.

Additional promising directions include explainability enhancements through attention visualization techniques that highlight specific linguistic patterns triggering classification decisions, enabling medical staff to understand system reasoning and identify potential false positives. Incremental learning capabilities could allow the system to adapt to evolving manipulation techniques without complete retraining, addressing the challenge of rapidly changing fraud patterns. Integration with electronic health record systems through standardized APIs would enable seamless deployment within existing healthcare IT infrastructure, reducing implementation barriers. Finally, multilingual extension beyond Slavic language families through cross-lingual transfer learning could expand system applicability to diverse healthcare contexts, though this requires careful validation of linguistic marker transferability across typologically distinct languages.

## 7. Conclusion

This study evaluated a hybrid deep learning architecture for detecting manipulated audio content in healthcare contexts. Key contributions include the development of a specialized linguistic processing framework, implementation of a multi-pathway neural architecture integrating convolutional and recurrent elements, and creation of a comprehensive evaluation methodology balancing accuracy with operational constraints. Experimental results validate the exceptional effectiveness of the approach, achieving 97% classification accuracy while maintaining reasonable computational efficiency and data requirements. Performance remained consistent across varied communication types, suggesting strong generalization capabilities, though resource utilization analysis indicates opportunities for further architectural optimization.

The demonstrated superiority of hybrid architectures suggests broader applications beyond audio manipulation detection, while the system's high accuracy with moderate training data requirements presents advantages for operational deployment in specialized domains. Practical deployment will require further optimization to reduce processing overhead while maintaining detection capabilities, potentially through knowledge distillation.

These findings contribute to broader research by demonstrating the effectiveness of specialized hybrid architectures for deception detection, while highlighting critical research directions to enhance operational capabilities in contested information environment.

## Acknowledgements

The authors would like to thank the Armed Forces of Ukraine for the opportunity to write a valid work during the full-scale invasion of the Russian Federation on the territory of Ukraine. Also, the

authors wish to extend their gratitude to Kharkiv National University of Radio Electronics for providing licences for additional software to prepare algorithms and the paper. This study was partly funded by the National Science Centre of Poland (project no. 2023/05/Y/ST6/00263).

## Declaration on Generative AI

During the preparation of this work, the authors used Grammarly Edu and submodule of Microsoft 365 in order to check grammar and spelling. After using these services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

## References

[1] N. Bansal et al., "Real-Time Advanced Computational Intelligence for Deep Fake Video Detection," Applied Science, vol. 13, no. 5, 2023, art. 3095, doi: 10.3390/app13053095.

[2] H. Padalko, V. Chomko, and D. Chumachenko, "A novel approach to fake news classification using LSTM-based deep learning models," *Sec. Machine Learning and Artificial Intelligence*, vol. 6, pp. 1–18, 2023, doi: 10.3389/fdata.2023.1320800.

[3] C. Fuller *et al.*, "An Analysis of Text-Based Deception Detection Tools," in *12th Americas Conference on Information Systems*, Acapulco, Mexico, Aug. 4–6, 2006. AISeL, 2006, pp. 3465–3472.

[4] L. Zhou *et al.*, "An exploratory study into deception detection in text-based computer-mediated communication," in *136th Annual Hawaii International Conference on System Sciences*, Big Island, USA, Jan. 6–9, 2003. IEEE Explore, 2003, pp. 1–10. doi: 10.1109/HICSS.2003.1173793.

[5] A. Khovrat, V. Kobziev, O. Nazarov, and S. Yakovlev, "Parallelization of the VAR Algorithm Family to Increase the Efficiency of Forecasting Market Indicators During Social Disaster," in *Inform. Technology & Implementation*, Kyiv, Ukraine, Nov. 30–Dec. 2, 2022. CEUR Workshop, 2023, pp. 222–233. Accessed: Jul. 7, 2025. [Online]. Available: https://ceur-ws.org/Vol-3347/Paper_19.pdf.

[6] J. K. Burgon, J. P. Blair, T. Qin, and J. F. Nunamaker, "Detecting Deception through Linguistic Analysis," in *Intelligence and Security Informatics, First NSF/NIJ Symposium*, Tucson, USA, June 2–3, 2003. Springer Nature, 2003, pp. 91–101. doi: 10.1007/3-540-44853-5_7.

[7] T. Bhatia. "Using transfer learning, spectrogram audio classification, and MIT app inventor to facilitate machine learning understanding." MIT. Accessed: Jul. 7, 2025. [Online]. Available: https://dspace.mit.edu/handle/1721.1/127379.

[8] S. Yakovlev, A. Khovrat, and V. Kobziev, "Using Parallelized Neural Networks to Detect Falsified Audio Information in Socially Oriented Systems", in *Inform. Technology & Implementation*, Kyiv, Ukraine, Nov. 20–Nov. 21, 2023. CEUR Workshop, 2024, pp. 220–238. Accessed: Jul. 7, 2025. [Online]. Available: https://ceur-ws.org/Vol-3624/Paper_19.pdf.

[9] T. Xia, and X. A. Chen, "Discrete Hidden Markov Model for SMS Spam Detection," *Applied Science*, vol. 10 (14), 2020, art. 5011, doi: 10.3390/app10145011.

[10] M. A. Alonso *et al.*, "Sentiment Analysis for Fake News Detection," *Electronics*, vol. 10 (11), 2021, art. 1348, doi: 10.3390/electronics10111348.

[11] R. Tolosana *et.al.*, "Deepfakes and beyond: A Survey of face manipulation and fake detection," *Information Fusion*, vol. 64, pp. 131–148, 2021, doi: 10.1016/j.inffus.2020.06.014.

[12] M. Mcuba, A. Singh, R. A. Ikuesan, and H. Venter, "The Effect of Deep Learning Methods on Deepfake Audio Detection for Digital Investigation," *Procedia Computer Science*, vol. 219, pp. 211–219, 2023, doi: 10.1016/j.procs.2023.01.283.

[13] A. Choudhary, and A. Arora, "Linguistic feature based learning model for fake news detection and classification," *Expert Systems with Applications*, vol. 169, 2021, art. 114171, doi: 10.1016/j.eswa.2020.114171.

[14] H. Elbatanouny *et al.*, "A comprehensive analysis of deception detection techniques leveraging machine learning," *Expert Systems with Applications*, vol. 283, 2025, art. 127601 doi: j.eswa.2025.127601.
[15] Y. Djenouri, A. Belhadi, G. Srivastava and J. C.-W. Lin, "Advanced Pattern-Mining System for Fake News Analysis," *IEEE Transactions on Computational Social Systems*, vol. 10, no. 6, pp. 2949-2958, 2023, doi: 10.1109/TCSS.2022.3233408.
[16] H. Padalko, V. Chomko, and D. Chumachenko, "The Impact of Stopwords Removal on Disinformation Detection in Ukrainian language during Russian Ukrainian war", in 4th International Workshop of IT-professionals on Artificial Intelligence, Cambridge, MA, USA, Sep. 25–Sep. 27, 2023. CEUR Workshop, 2024, pp. 87–101.
[17] S. E. Robertson, and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," Foundations and Trends in Information Retrieval, vol. 3, no. 4, pp. 333-389, 2023, doi: 10.1561/1500000019.