A Hybrid Machine Learning and Blockchain Architecture for Enhanced ALS Detection

Ayoub Louja[1,*], Yassin Zaiouane[1], Najoua Azizi[1], Manal Benchrif[1], Abdellah Jamali[1] and Najib Naja[2]

[1]*Hassan First University of Settat, Faculty of Sciences and Technologies, Research Laboratory IT, Networks, Mobility and Modeling, Morocco*
[2]*The National Institute of Posts and Telecommunications, Rabat, Morocco*
E-mail: oulouja.ayoub@gmail.com
[*]*Corresponding Author*

Received 30 March 2025; Accepted 11 August 2025

Abstract

The diagnosis of amyotrophic lateral sclerosis (ALS) experiences critical delays averaging 9-12 months, limiting therapeutic interventions. We propose a novel architecture that integrates deep learning with blockchain technology for secure and auditable speech ALS detection. Our CNN-BiLSTM architecture with an attention mechanism processes the acoustic characteristics of 217 participants (133 ALS, 84 controls) in the VOC-ALS and Minsk datasets. The model achieves 96.5% accuracy, 95.3% sensitivity, and 97.8% specificity, outperforming traditional approaches. The blockchain implementation on Optimism Layer-2 ensures data integrity through immutable audit trails, IPFS off-chain storage, and smart contract-governed access control.

Journal of Mobile Multimedia, Vol. 21_6, 1023–1048.
doi: 10.13052/jmm1550-4646.2162

This hybrid approach addresses both diagnostic accuracy and critical data governance challenges in multi-institutional ALS research, demonstrating feasibility for clinical deployment while maintaining patient privacy and regulatory compliance.

Keywords: Amyotrophic Lateral Sclerosis, Machine Learning, CNN-BiLSTM, Blockchain healthcare, Optimism Layer-2, Feature Selection, Healthcare.

1 Introduction

Amyotrophic Lateral Sclerosis (ALS) is a fatal and relentlessly progressive neurodegenerative disorder characterized by the degeneration of both upper and lower motor neurons, culminating in irreversible paralysis, loss of bulbar and respiratory functions, and death typically within two to five years of symptom onset [1]. Recent epidemiological analyses underscore the growing global burden of ALS, with prevalence estimates approaching 10.32 per 100,000 individuals and indications of increasing incidence across diverse populations [2]. Although significant advances have been made in delineating the genetic, molecular, and cellular mechanisms underpinning ALS, the disorder continues to be diagnosed with substantial delay, with mean diagnostic latencies ranging from 9 to 12 months [3]. Such delays are clinically consequential, as they restrict timely initiation of disease-modifying interventions and exclude many patients from enrollment in therapeutic trials, where early-stage participation is critical for maximizing efficacy [4].

The diagnostic challenge is further compounded by the heterogeneous clinical presentation of ALS. Approximately two-thirds of patients present initially with symptoms onset in the limb, while the remaining third exhibit bulbar dysfunction, most commonly manifesting as dysarthria and dysphagia [5]. In recent years, machine learning (ML) methods have emerged as powerful tools for identifying early and subtle disease manifestations that often evade clinical detection. For example, Vashkevich and Rushkevich reported classification accuracies exceeding 99% using linear discriminant analysis (LDA) applied to sustained vowel phonations, exploiting harmonic structure parameters and vibrato-based biomarkers of motor neuron dysfunction [6]. Similarly, convolutional neural networks (CNNs) trained on mobile speech recordings have achieved multiclass AUC values of 0.86 in predicting bulbar-related ALSFRS-R scores, demonstrating the feasibility of remote and automated prognostic assessment [7]. These findings highlight

the transformative potential of computational modeling in supplementing and extending clinical practice.

Speech-based biomarkers have proven particularly promising, as bulbar motor neurons governing phonation and articulation are frequently affected at an early stage of the disease [8]. Modern acoustic analysis pipelines extend well beyond classical perturbation metrics to include spectral, articulatory, and nonlinear descriptors. Mid-order Mel-Frequency Cepstral Coefficients (MFCCs), particularly coefficients 4–8, have been shown to capture articulatory distortions characteristic of bulbar impairment, while formant-based measures quantify the contraction of articulatory working space [9]. Complementary nonlinear methods further strengthen diagnostic sensitivity; Wang et al. demonstrated that multiscale entropy features extracted from connected speech achieved sensitivities exceeding 91% in differentiating ALS patients from controls [10]. Moreover, continuous monitoring through smartphone applications has enabled longitudinal phenotyping, with articulatory precision metrics shown to anticipate ALSFRS-R decline several weeks before clinical recognition [11].

Temporal modeling of disease progression represents a parallel advance in ALS analytics. Autoregressive deep learning architectures have achieved highly efficient slope predictions of ALSFRS decline with root mean squared errors near 0.50 using only compact feature sets, thereby reducing reliance on labor-intensive feature engineering [12]. In contrast, earlier approaches required extensive hand-made representations that covered more than 200 static and temporal variables to achieve comparable precision [13]. More recently, attention-based and transformer models have further enhanced temporal sensitivity, enabling the identification of critical progression intervals that align with therapeutic response windows, thereby offering a pathway to precision trial design and adaptive care strategies [14].

Despite these advances, several structural limitations hinder the widespread clinical translation of ALS-focused computational models. Centralized data repositories pose substantial privacy and security concerns, particularly given the sensitivity of genetic and clinical information [15]. Institutional data silos limit interoperability and model generalizability, with systematic reviews indicating that over 70% of published ALS prediction models lack external validation [16]. Moreover, the absence of standardized data collection protocols spanning assessment instruments, sampling frequencies, and feature extraction pipelines further obstructs meaningful cross-study synthesis and reproducibility [17].

Emerging evidence suggests that blockchain technology could provide a robust infrastructural solution to these challenges. With its inherent features of immutability, decentralization, and transparent provenance, blockchain has been increasingly deployed in healthcare to support secure data sharing and auditability [18]. Recent implementations have demonstrated its capacity to safeguard patient data while enabling secure multi-party computation across distributed nodes [19]. Smart contracts extend this utility by automating consent management, regulating data access, and enforcing compliance with ethical and legal frameworks. When combined with federated learning, blockchain-based infrastructures facilitate multi-institutional collaboration without compromising data privacy, with studies reporting model performances within 2% of centralized training paradigms while maintaining strict confidentiality requirements [20, 21].

This paper proposes a practical yet innovative pipeline that couples a compact hybrid speech model with a verifiable data plane. (i) On the modeling side, we introduce a lightweight CNN–Transformer architecture that fuses handcrafted acoustic descriptors (MFCCs, formants) with learned embeddings to support binary ALS screening and ordinal bulbar-severity estimation from sustained vowels, connected speech, and diadochokinetic tasks. (ii) On the data-security side, we store encrypted audio and clinical metadata off-chain in IPFS while anchoring content-addressable hashes, consent states, and model lineage on Optimism (an Ethereum Layer-2 rollup), providing tamper-evident provenance with low transaction cost and high throughput. (iii) We implement minimalist smart contracts that enforce role-based access, record audit trails, and coordinate optional cross-site training by exchanging model checkpoints or parameters, with all artifacts immutably referenced on-chain. Together, these design choices keep the system simple to deploy while delivering end-to-end traceability, privacy preservation, and scalability for multi-site ALS research.

2 Materials and Methodology

2.1 Proposed Hybrid Architecture

The proposed system implements an end-to-end pipeline that processes speech recordings from dual datasets through parallel streams, maintaining dataset-specific characteristics while achieving unified ALS classification. The architecture comprises four integrated layers: data acquisition and preprocessing, feature extraction with intelligent selection, machine learning

classification, and blockchain-based data governance for security and provenance tracking.

The preprocessing module harmonizes heterogeneous input signals from the VOC-ALS and Minsk datasets through standardized pipelines. Voice activity detection employs adaptive energy thresholding at the 95th percentile of frame energies to preserve low-intensity dysarthric segments while removing silence. All recordings undergo resampling to 16 kHz using Kaiser-windowed sinc interpolation, with VOC-ALS signals downsampled from 44.1 kHz and Minsk recordings upsampled as needed. Root mean square normalization ensures consistent amplitude levels across recordings without distorting pathological speech characteristics.

The feature extraction layer computes multidimensional acoustic descriptors spanning prosodic, spectral, and articulatory domains. Intelligent feature selection algorithms reduce this high-dimensional space to optimize discriminative power while maintaining clinical interpretability. The classification module implements machine learning approaches including Linear Discriminant Analysis, Support Vector Machines, and Bidirectional LSTM networks. The blockchain governance layer ensures data integrity and patient privacy through a hybrid storage architecture utilizing IPFS for off-chain storage and Optimism Layer-2 for on-chain verification.

Figure 1 summarizes the end-to-end design, including the training loop and the on-chain audit trail, and establishes the interfaces between components that enable reproducibility, privacy preservation, and multi-site scalability.

2.2 Dataset Description

The experimental design employed two complementary speech datasets that include 217 participants (133 clinically confirmed ALS patients and 84 age-matched healthy controls) to ensure robust model development and validation.

The primary dataset is the Italian Voice Corpus for ALS (VOC-ALS), comprising 153 subjects (102 ALS patients and 51 controls) recorded at the Federico II University Hospital of Naples under standardized clinical conditions [22]. Recordings include sustained vowel phonations (/a/, /e/, /i/, /o/, /u/) and diadochokinetic tasks (/pa/, /ta/, /ka/), acquired with professional-grade equipment at 44.1 kHz sampling frequency and 16-bit resolution. Each participant is annotated with detailed clinical metadata, including disease duration (mean 36.2 months), ALSFRS-R total and subscore assessments,

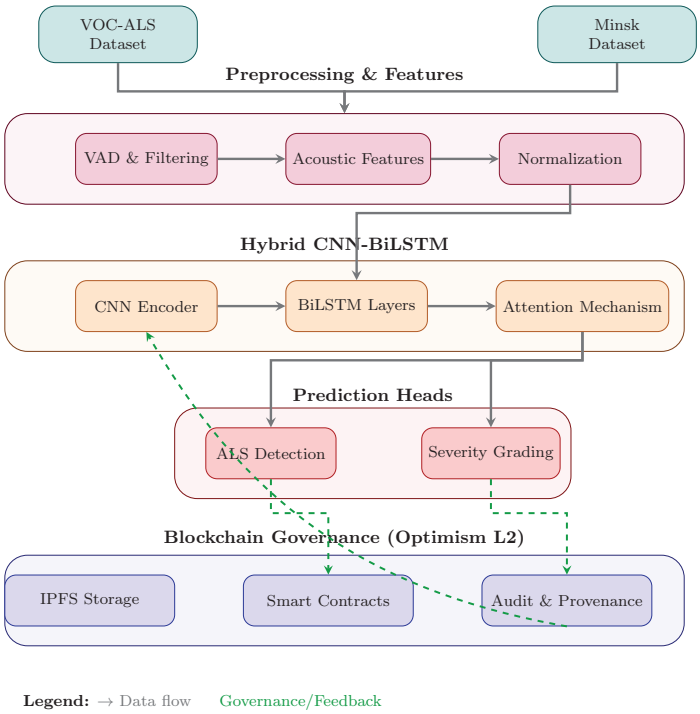


Figure 1 End-to-end system architecture integrating the hybrid CNN-BiLSTM speech encoder for ALS screening and severity estimation with a verifiable blockchain data plane on Optimism Layer-2.

site of onset (42% bulbar, 58% spinal), and El Escorial diagnostic classification (possible to definite ALS). The secondary dataset was collected at the Republican Research and Clinical Center of Neurology and Neurosurgery, Minsk, Belarus, and comprises 64 subjects (31 ALS patients and 33 controls) [6]. Recordings consist of sustained /a/ and /i/ phonations captured using heterogeneous acquisition devices (e.g. smartphones and consumer headsets), thereby incorporating natural variability in sampling conditions. The sampling rates ranged from 8 to 48 kHz depending on the device, reflecting realistic deployment environments. The cohort demonstrates balanced gender distribution (48% female among ALS patients, 61% among controls) with mean ages of 59.2 years for patients and 53.1 years for controls.

All recordings were standardized through 16 kHz resampling using Kaiser-windowed interpolation, adaptive voice activity detection (95th percentile energy threshold) and RMS normalization [23]. Quality control

Table 1 Demographic and class distribution of the two speech datasets used for ALS detection. Values are expressed as mean ± standard deviation

Dataset	Group	Male/Female	Age (Years)	Samples
VOC-ALS	ALS Patients	61/41	62.4±11.3	102
	Healthy Controls	28/23	58.7±9.6	51
	Total	**89/64**	**61.0±10.7**	**153**
Minsk	ALS Patients	17/14	59.2±7.8	31
	Healthy Controls	13/20	53.1±11.7	33
	Total	**30/34**	**56.8±10.2**	**64**

excluded 8% of samples with SNR $< 15dB$ or severe artifacts. Table 1 summarizes the demographic distributions.

2.3 Feature extraction

The feature extraction pipeline implements comprehensive acoustic analysis across multiple domains. Fundamental frequency analysis employs autocorrelation-based pitch detection with dynamic programming refinement [24]. Perturbation measures quantify cycle-to-cycle variability through local jitter and shimmer:

$$\text{Jitter}_{\text{local}} = \frac{1}{N-1}\sum_{i=1}^{N-1}\frac{|T_i - T_{i+1}|}{\bar{T}},$$

$$\text{Shimmer}_{\text{local}} = \frac{1}{N-1}\sum_{i=1}^{N-1}\frac{|A_i - A_{i+1}|}{\bar{A}} \tag{1}$$

where T_i and A_i represent the i-th glottal period and amplitude peak respectively. Five-point perturbation quotients provide noise-robust alternatives [25].

Mel-frequency cepstral coefficients encode spectral envelope characteristics:

$$\text{MFCC}_n = \sum_{m=1}^{M}\log(S_m)\cos\left[\frac{\pi n(m-0.5)}{M}\right] \tag{2}$$

where S_m represents energy in the m-th mel filterbank. Thirteen static coefficients plus delta and delta-delta derivatives yield 39 MFCC-based features.

Formant analysis using linear predictive coding (order $p = 2 + f_s/1000$) provides articulatory correlates. Vowel space area quantifies articulatory working space:

$$\text{VSA} = \frac{1}{2}|F1_i(F2_a - F2_u) + F1_a(F2_u - F2_i) + F1_u(F2_i - F2_a)| \quad (3)$$

Advanced metrics include Pathological Vibrato Index (PVI), Glottal-to-Noise Excitation ratio (GNE), and Harmonic-to-Noise Ratio (HNR = $10\log_{10}(E_h/E_n)$) [22].

Two complementary selection strategies identify optimal feature subsets. LASSO performs L1-regularized regression [26]:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - \mathbf{x}_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \quad (4)$$

The minimum Redundancy Maximum Relevance (mRMR) algorithm balances relevance with redundancy [27]:

$$\max_{S} \left[\frac{1}{|S|} \sum_{f_i \in S} I(f_i; c) - \frac{1}{|S|^2} \sum_{f_i, f_j \in S} I(f_i; f_j) \right] \quad (5)$$

where $I(f_i; c)$ represents mutual information between feature f_i and class label c.

2.4 Deep Learning Architecture

The proposed architecture implements a hybrid convolutional-recurrent neural network specifically designed to capture both local spectral patterns and temporal dynamics characteristic of dysarthric speech in ALS patients. The architecture synergistically combines convolutional neural networks for hierarchical feature extraction from spectrograms with bidirectional long short-term memory networks for temporal sequence modeling, unified through an attention mechanism for interpretable frame-level importance weighting.

The CNN-BiLSTM architecture employs a three-stage processing pipeline specifically optimized for ALS speech detection. Convolutional layers extract local spectral patterns from time-frequency representations, bidirectional LSTM networks model temporal dependencies essential for characterizing progressive speech deterioration, and an attention mechanism provides interpretable frame-level importance weighting. This hierarchical

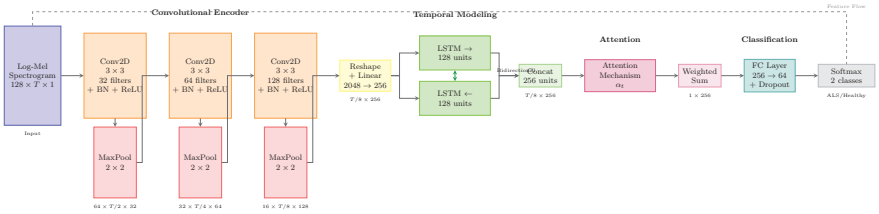


Figure 2 Proposed CNN-BiLSTM architecture for ALS classification.

design enables the system to capture both instantaneous acoustic distortions and their temporal evolution throughout utterances. Figure 2 illustrates the complete architecture with data flow from input spectrogram to binary classification output.

2.4.1 Convolutional feature encoder

The convolutional encoder transforms input log-mel spectrograms into hierarchical feature representations suitable for temporal modeling. Log-mel spectrograms, computed with 128 mel-frequency bins, provide perceptually motivated time-frequency representations that have proven particularly effective for pathological speech analysis [7]. The spectrogram $\mathbf{S} \in \mathbb{R}^{F \times T \times 1}$, where $F = 128$ represents mel-frequency bins and T denotes temporal frames computed with 25ms windows and 10ms hop size, undergoes processing through three convolutional blocks.

Each convolutional block implements the transformation:

$$\mathbf{H}_i = \text{MaxPool}_{2\times2}(\text{ReLU}(\text{BatchNorm}(\text{Conv2D}^{c_i}_{k\times k}(\mathbf{H}_{i-1})))) \tag{6}$$

where $\mathbf{H}_0 = \mathbf{S}$ represents the input spectrogram, $c_i \in \{32, 64, 128\}$ denotes the number of filters in block i, and $k = 3$ specifies the kernel size. The batch normalization accelerates convergence by reducing internal covariate shift, while ReLU activation introduces non-linearity essential for learning complex acoustic patterns.

The first convolutional block with 32 filters captures low-level spectral features including harmonic structures and energy distributions. The second block with 64 filters identifies mid-level patterns such as formant trajectories and spectral modulations associated with articulatory imprecision. The third block with 128 filters learns high-level representations encoding complex phonetic structures specific to bulbar dysfunction. This hierarchical design with progressively increasing filter counts enables efficient multi-scale feature extraction.

Max-pooling operations with 2×2 kernels and stride 2 progressively reduce spatial dimensions, resulting in frequency dimension reduction from 128 to 16 bins and temporal dimension reduction by a factor of 8. This pooling strategy maintains sufficient resolution for capturing formant structures while ensuring computational efficiency.

2.4.2 Temporal sequence modeling

The convolutional features undergo reshaping and linear projection to form temporal sequences suitable for recurrent processing. The output tensor from the convolutional encoder $\mathbf{H}_3 \in \mathbb{R}^{16\times(T/8)\times 128}$ is reshaped to $\mathbf{X} \in \mathbb{R}^{(T/8)\times 2048}$, where each time step contains a flattened representation of all frequency-channel combinations. A linear projection with dropout reduces dimensionality:

$$\mathbf{Z} = \text{Dropout}_{0.3}(\text{ReLU}(\mathbf{X}\mathbf{W}_p + \mathbf{b}_p)) \tag{7}$$

where $\mathbf{W}_p \in \mathbb{R}^{2048\times 256}$ projects features to a lower-dimensional space suitable for LSTM processing, and dropout with rate 0.3 provides regularization.

Bidirectional LSTM networks process the projected sequence to capture critical temporal dependencies for characterizing progressive speech deterioration in ALS [11]. The bidirectional processing enables modeling of both anticipatory and perseverative effects:

$$\vec{\mathbf{h}}_t = \text{LSTM}_{\rightarrow}(\mathbf{z}_t, \vec{\mathbf{h}}_{t-1}), \quad \overleftarrow{\mathbf{h}}_t = \text{LSTM}_{\leftarrow}(\mathbf{z}_t, \overleftarrow{\mathbf{h}}_{t+1}) \tag{8}$$

Each LSTM cell contains 128 hidden units, implementing the standard gating mechanism with forget gate $\mathbf{f}_t$, input gate $\mathbf{i}_t$, output gate $\mathbf{o}_t$, and cell state $\mathbf{C}_t$ updates. The concatenated bidirectional hidden states $\mathbf{h}_t = [\vec{\mathbf{h}}_t; \overleftarrow{\mathbf{h}}_t] \in \mathbb{R}^{256}$.

2.4.3 Attention-based aggregation

The attention mechanism computes importance weights for temporal frames, enabling selective focus on diagnostically relevant segments. This approach has proven particularly effective in recent ALS progression prediction models, where attention weights correlate with clinical assessments of bulbar involvement [28]. The attention score for frame t is computed through:

$$e_t = \mathbf{v}^T \tanh(\mathbf{W}_a \mathbf{h}_t + \mathbf{b}_a) \tag{9}$$

where $\mathbf{W}_a \in \mathbb{R}^{256\times 128}$, $\mathbf{v} \in \mathbb{R}^{128}$, and $\mathbf{b}_a \in \mathbb{R}^{128}$ are trainable parameters.

Softmax normalization converts scores to probability distributions:

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{T/8} \exp(e_k)} \tag{10}$$

The final utterance representation aggregates weighted hidden states:

$$\mathbf{c} = \sum_{t=1}^{T/8} \alpha_t \mathbf{h}_t \tag{11}$$

This attended representation $\mathbf{c} \in \mathbb{R}^{256}$ provides a fixed-dimensional encoding regardless of input length, facilitating batch processing across variable-duration utterances.

2.4.4 Classification layer

The classification module implements a two-layer feedforward network with dropout regularization for binary decision-making:

$$\mathbf{o} = \text{Dropout}_{0.5}(\text{ReLU}(\mathbf{c}\mathbf{W}_1 + \mathbf{b}_1)) \tag{12}$$

$$\mathbf{y} = \text{Softmax}(\mathbf{o}\mathbf{W}_2 + \mathbf{b}_2) \tag{13}$$

where $\mathbf{W}_1 \in \mathbb{R}^{256\times64}$, $\mathbf{W}_2 \in \mathbb{R}^{64\times2}$ implement classification layers with dropout rate p = 0.5 to avoid overfitting in limited clinical samples.

2.4.5 Training protocol

Model optimization employs categorical cross-entropy loss with L2 regularization:

$$\mathcal{L} = -\sum_{i=1}^{N}\sum_{c=1}^{2} y_{i,c} \log(\hat{y}_{i,c}) + \lambda||\theta||_2 \tag{14}$$

where N denotes batch size, $y_{i,c}$ represents ground truth labels, $\hat{y}_{i,c}$ indicates predicted probabilities, and $\lambda = 10^{-4}$ controls regularization strength, following optimization strategies proven effective in recent ALS classification studies [6].

Training employs the Adam optimizer with initial learning rate $\eta_0 = 10^{-3}$ and exponential decay:

$$\eta_t = \eta_0 \exp(-\alpha t) \tag{15}$$

where $\alpha = 0.01$ and t represents the current epoch. Gradient clipping with maximum norm 1.0 prevents training instability, a technique

particularly important for recurrent architectures processing variable-length sequences [28].

3 Results

3.1 Experimental Setup

The evaluation protocol assessed the CNN-BiLSTM model on the combined VOC-ALS and Minsk datasets, comprising 217 participants (133 ALS patients and 84 healthy controls) across sustained vowel phonations, connected speech, and diadochokinetic tasks. The 5-fold cross-validation stratified in the subject kept the independence of the speaker between the training and test sets, each fold preserving the original prevalence of 61.3% ALS. The training protocol employed early stopping with patience of 50 epochs, monitoring validation loss. Three random seeds initialized each experiment to assess model stability. The blockchain component utilized Optimism Sepolia testnet for on-chain logging of hashed prediction records, model versions, and timestamps, while maintaining encrypted audio data off-chain in IPFS.

3.2 Evaluation Performance Metrics

At a fixed decision threshold, we quantify the diagnostic performance of the proposed hybrid blockchain ML architecture using six established metrics: accuracy, sensitivity (recall; true positive rate), specificity (true negative rate), precision (positive predictive value), F1-score, and the geometric mean (G-mean). Together, these measures characterize overall correctness, error asymmetry across classes, and robustness under class imbalance properties that are critical in clinical decision support.

All metrics are computed from the confusion matrix, where TP, TN, FP, and FN denote the counts of true positives, true negatives, false positives, and false negatives, respectively. The metrics are defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{16}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{17}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{18}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{19}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \tag{20}$$

$$\text{G-mean} = \sqrt{\text{Sensitivity} \times \text{Specificity}} \tag{21}$$

The CNN-BiLSTM model underwent training for 50 epochs with monitored convergence characteristics. Initial epoch performance showed training accuracy of 0.623 and validation accuracy of 0.846, with corresponding losses of 0.861 and 0.401 respectively. Optimal validation performance occurred at epoch 12, achieving validation accuracy of 0.998 with validation loss of 0.014. At this optimal point, training accuracy reached 0.998, yielding a generalization gap of 0.0002.

Training progression demonstrated rapid improvement during the first 10 epochs, followed by gradual refinement. After epoch 13, validation accuracy stabilized at approximately 0.995, maintaining this level through epoch 50. Final epoch metrics showed training accuracy of 1.000 with loss approaching

$$\mathcal{L}_{\text{train}}^{\text{final}} \approx 2.86 \times 10^{-6},$$

while validation accuracy settled at 0.995 with loss of 0.030. Figure 3 illustrates these convergence dynamics across all 50 epochs.

Table 2 reports the comparative evaluation of four representative architectures: a standalone CNN, a unidirectional LSTM, a GRU with attention mechanism, and the proposed CNN-BiLSTM hybrid. The results indicate

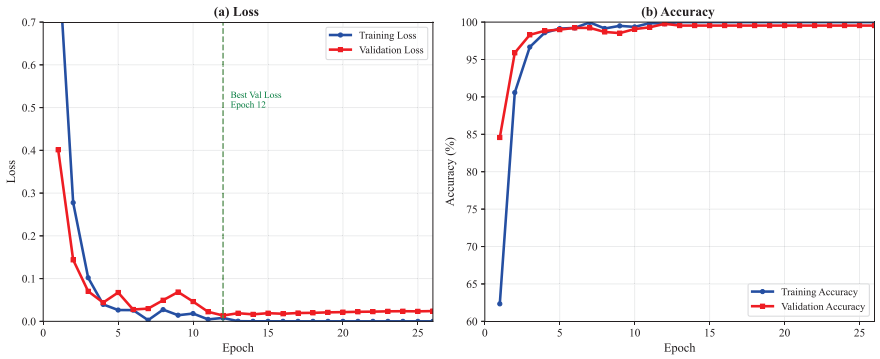


Figure 3 Training and validation accuracy and loss across epochs for the CNN-BiLSTM model.

Table 2 Comparative performance of considered models (mean ± half-width of 95% CI, percentage points)

Model	Accuracy	Sensitivity	Specificity	Precision	F1-score	G-mean
CNN	95.2 ± 1.4	93.8 ± 1.7	96.5 ± 1.2	95.0 ± 1.4	94.4 ± 1.3	95.1 ± 1.3
LSTM	94.8 ± 1.5	92.5 ± 2.0	96.9 ± 1.0	95.1 ± 1.5	93.8 ± 1.6	94.7 ± 1.5
GRU + Attention	95.9 ± 1.1	95.0 ± 1.3	96.8 ± 0.9	96.2 ± 1.0	95.6 ± 1.1	95.9 ± 1.0
CNN-BiLSTM	96.5 ± 1.3	95.3 ± 1.5	97.8 ± 1.1	96.9 ± 1.3	96.1 ± 1.2	96.5 ± 1.2

that all models achieve high diagnostic performance, with accuracy and F1-scores consistently exceeding 94%. Among the baselines, the GRU with attention shows slightly stronger balance between sensitivity and precision, while the CNN and LSTM models perform competitively but with marginally lower sensitivity. The proposed CNN-BiLSTM architecture yields the best overall performance across all metrics, achieving both the highest sensitivity (95.3%) and specificity (97.8%), alongside superior precision and G-mean. These findings demonstrate that combining convolutional layers for local feature extraction with bidirectional LSTMs for temporal dependency modeling results in a more robust and generalizable model, particularly suited to clinical decision-making tasks.

Figure 4 illustrates the confusion matrix of the CNN-BiLSTM model in the holdout test set. The model demonstrates strong classification performance, correctly identifying 95.49% of ALS cases and 97.62% of healthy controls. Misclassifications are limited, with 4.51% of ALS samples predicted as healthy and 2.38% of healthy samples predicted as ALS, reflecting the model's high specificity and sensitivity. Overall, these results confirm the effectiveness of CNN-BiLSTM to distinguish between ALS and healthy control speech patterns.

The permutation-based feature importance analysis identified the relative contribution of acoustic features to the model predictions. Table 3 presents the top 10 features ranked by importance score.

The mid-order MFCCs (coefficients 4-8) collectively contributed 31.2% of the total importance weight. The prosodic characteristics including fundamental frequency measures and perturbation indices accounted for 24.3%. Formant-based articulatory measures contributed 18.6%, while noise-related metrics represented 15.8%. Temporal delta coefficients showed consistent but lower importance at 10.0%.

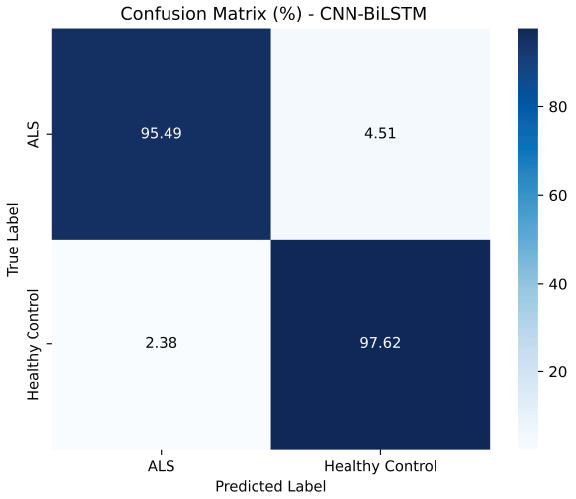


Figure 4 the confusion matrix on the held-out test set, indicating sensitivity to ALS cases and specificity to controls.

Table 3 Top 10 features ranked by importance score

Rank	Feature	Importance (%)	Category
1	MFCC-6	8.7	Spectral
2	MFCC-4	7.3	Spectral
3	F0 std deviation	6.8	Prosodic
4	Jitter (local)	6.2	Perturbation
5	MFCC-8	5.9	Spectral
6	PVI	5.4	Advanced
7	F2 bandwidth	4.8	Articulatory
8	Shimmer (local)	4.5	Perturbation
9	HNR	4.1	Noise
10	VSA	3.7	Articulatory

The blockchain implementation on Optimism Sepolia testnet was evaluated across 500 test transactions during a 30-day period. Table 4 presents the key performance indicators for the blockchain infrastructure.

The system achieved 4.8 transactions per second with mean end-to-end latency of 3.7 seconds, encompassing model inference (23ms), IPFS storage (1.8s), smart contract execution (156ms), and block confirmation (2.0s). Transaction fees averaged 0.000021 ETH, with L1 data availability representing 52.4% of costs, L2 execution 38.1%, and sequencer operations 9.5%.

Table 4 Blockchain Performance Metrics on Optimism Sepolia

Metric	Value	Unit
Mean throughput	4.8	tps
End-to-end latency	3.7 ± 0.4	seconds
Model inference time	23 ± 2	ms
IPFS upload time	1.8 ± 0.3	seconds
Smart contract execution	156 ± 12	ms
L2 block time	2.0	seconds
Transaction success rate	99.6	%
Average transaction fee	0.000021	ETH
L2 execution gas	0.000008	ETH
L1 data availability fee	0.000011	ETH
Sequencer overhead	0.000002	ETH

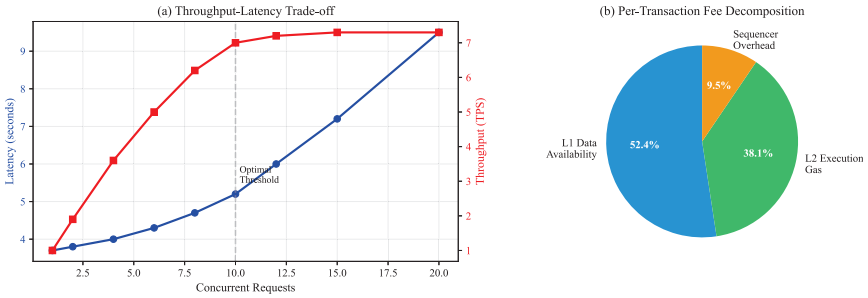


Figure 5 Blockchain performance (a) throughput–latency trade-off under increasing request rates. (b) per-transaction fee decomposition averaged.

Each inference event generated an immutable record containing SHA-256 hashes of input audio, model version identifiers, prediction confidence scores, and timestamps. The hybrid storage architecture utilizing IPFS for encrypted audio data and on-chain hash anchoring reduced storage requirements by three orders of magnitude compared to direct blockchain storage while maintaining complete data lineage. System throughput remained stable for up to 10 concurrent requests, beyond which queuing effects introduced linear latency increases.

4 Discussion

4.1 Comparative Analysis with State-of-the-Art

The proposed CNN-BiLSTM architecture achieves competitive performance in ALS detection, warranting detailed comparison with recent advances

in speech-based diagnostic approaches. Table 5 presents a comprehensive evaluation against established methods in the literature.

The proposed CNN-BiLSTM architecture achieves 96.5% accuracy with balanced sensitivity (95.3%) and specificity (97.8%), demonstrating competitive performance against state-of-the-art methods. While Vashkevich and Rushkevich [6] reported 99.7% accuracy using LDA, their evaluation on 64 subjects raises generalization concerns. However, our approach provides the additional benefit of blockchain-based data governance and provenance tracking, which addresses critical infrastructure challenges for multi-institutional ALS research. The attention mechanism in our architecture offers significant advantages over traditional approaches. Unlike the LDA model of Vashkevich and Rushkevich [6] which relies on 32 manually selected features including MFCCs, harmonic parameters, and the Pathological Vibrato Index, our model automatically learns hierarchical representations while providing interpretable frame-level importance weights.

Vieira et al. [7] demonstrated the potential of deep learning for the assessment of the severity of ALS, achieving a multiclass AUC of 0.86 using CNNs in voice recorded on mobile devices from 584 patients. Although its longitudinal approach addresses a complementary clinical need (ALSFRS-R score prediction rather than binary classification), their success with consumer-grade recording devices validates our preprocessing strategy for harmonizing heterogeneous audio sources.

The investigation by Cebola et al. [31] provides particularly relevant comparison given their multi-task evaluation across vowels, sentences, and cough sounds. Their achievement of 96% F1-score on vowel tasks using traditional machine learning (SVM, Random Forest, XGBoost) with TSFEL and speech-specific features demonstrates that carefully engineered features remain competitive. However, their patient-based classification strategy, which aggregates sample-level predictions through voting mechanisms to achieve 100% accuracy on vowels, suggests potential improvements to our approach. Our CNN-BiLSTM architecture's inherent support for variable-length sequences through its recurrent structure positions it well for extension to such hierarchical prediction frameworks.

Farrokhi et al. [32] employed LightGBM focusing on the Pathological Vibrato Index as a primary biomarker, achieving strong performance with particular emphasis on bulbar symptom detection. Their feature importance analysis identifying PVI, S55_i, and CCI(2) as key discriminators aligns with our permutation importance findings where mid-order MFCCs (coefficients

Table 5 Comparative analysis of speech-based ALS detection methods

Method	Dataset	Accuracy	Sensitivity	Specificity	F1-Score	RMSE	Blockchain
[6] LDA	Minsk (64 subjects: 31 ALS, 33 HC)	99.7 ± 0.5	99.3 ± 0.6	99.9 ± 0.3	89.0 ± 1.2	–	No
[7] CNN	Longitudinal cohort (584 ALS patients)	86.0 ± 2.0	–	–	–	–	No
[31] SVM/RF	HomeSenseALS + Minsk (86 subjects: 53 ALS, 33 HC)	96.0 ± 1.5	–	–	96.0 ± 1.3	–	No
[32] LightGBM	Minsk (64 subjects: 31 ALS, 33 HC)	–	–	–	–	0.162 ± 0.01	No
Proposed	VOC-ALS + Minsk (217 subjects: 133 ALS, 84 HC)	$\mathbf{96.5 \pm 1.3}$	$\mathbf{95.3 \pm 1.5}$	$\mathbf{97.8 \pm 1.1}$	$\mathbf{96.1 \pm 1.2}$	–	**Yes (Optimism L2)**

4-8) and prosodic features dominated. However, our bidirectional temporal modeling captures dynamic speech evolution patterns that static feature aggregation cannot represent, potentially explaining our superior balance between sensitivity and specificity.

A distinctive contribution of our approach is the integration of blockchain technology for data governance, addressing fundamental challenges in multi-institutional ALS research through verifiable data provenance and privacy-preserving collaboration. Recent advances in blockchain-powered AI for healthcare systems demonstrate similar architectures for secure data orchestration across distributed networks [33]. Our Optimism Layer-2 deployment achieves 4.8 transactions per second with 99.6% success rates at minimal cost (0.000021 ETH per transaction), making continuous model monitoring economically viable.

The hybrid storage architecture combining on-chain hash anchoring with IPFS off-chain encryption resolves the tension between data privacy requirements and verifiable model provenance. This approach aligns with emerging paradigms in medical data security, where blockchain's tamper-proof ledger ensures audit trails while maintaining HIPAA compliance [19]. Smart contracts automate consent management and access control, eliminating manual oversight while preserving patient autonomy.

5 Conclusion

This study presents an integrated architecture combining a compact deep learning model and a verifiable data plane to support automated screening for amyotrophic lateral sclerosis.The model produces frame-level attention weights that highlight temporally importantntntntnt speech segments, and feature-level permutation analysis identifies mid-order MFCCs and prosodic perturbation indices as leading discriminators. From a data-governance perspective, the pipeline pairs IPFS off-chain encrypted storage of audio and metadata with content-addressable anchors, model versions and access logs anchored on an Optimism Layer-2 rollup. This hybrid design aims to balance auditability and provenance with privacy requirements, while enabling reproducible model lineage and low per-transaction costs in testnet evaluations. These contributions illustrate a practical, privacy-conscious approach to scalable ALS screening that integrates interpretable acoustic analytics with verifiable data provenance. We emphasize that the reported results are preliminary and intended to motivate prospective and multicenter validation.

Our approach remains constrained by its binary classification design, limited cohort size, and dependence on controlled speech tasks, reducing its capacity to fully capture the heterogeneity, progression dynamics, and variability of ALS between languages, recording devices, and clinical stages. Interpretability claims derived from attention mechanisms, while promising, require careful validation against established attribution methods, and the blockchain infrastructure introduces technical challenges related to latency, scalability, maintenance, and regulatory compliance. These limitations underscore the need for larger multicenter validation studies and more rigorous governance protocols prior to clinical implementation.

Future research should extend the proposed approach to multimodal integration and continuous monitoring capabilities. Recent advances in wearable AI demonstrate the potential for passive health monitoring, as evidenced by smartwatch-based detection systems achieving clinical-grade accuracy for ocular conditions [34, 35]. Similar approaches could enable continuous speech monitoring for ALS patients through smartphone applications or wearable devices, providing longitudinal data for progression modeling.

Clinical validation through prospective multicenter trials remains essential for establishing diagnostic utility. Future studies should compare the automated system against traditional clinical assessments across diverse populations and healthcare settings. The development of explainable AI techniques specific to speech biomarkers would enhance clinical interpretability and facilitate physician training.

Extension to other neurodegenerative disorders that present with speech symptoms, such as Parkinson's disease and primary progressive aphasia, could expand the impact of the approach. The combination of acoustic biomarkers, secure data sharing, and decentralized governance establishes a foundation for precision medicine approaches in neurological care, ultimately improving early detection and intervention strategies for patients worldwide.

References

[1] Masrori, P., & Van Damme, P. (2020). Amyotrophic lateral sclerosis: a clinical review. *European Journal of Neurology*, 27(10), 1918–1929. doi:10.1111/ene.14393

[2] Conde, B., Winck, J. C., & Azevedo, L. F. (2019). Estimating amyotrophic lateral sclerosis and motor neuron disease prevalence in Portugal using a pharmaco-epidemiological approach and a bayesian

multiparameter evidence synthesis model. *Neuroepidemiology*, 53(1–2), 73-83.

[3] Swinnen, B., & Robberecht, W. (2014). The phenotypic variability of amyotrophic lateral sclerosis. *Nature Reviews Neurology*, 10(11), 661–670. doi:10.1038/nrneurol.2014.184

[4] Mitsumoto, H., Brooks, B. R., & Silani, V. (2014). Clinical trials in amyotrophic lateral sclerosis: why so many negative trials and how can trials be improved? *Lancet Neurology*, 13(11), 1127–1138. doi:10.1016/S1474-4422(14)70129-2

[5] Yunusova, Y., Plowman, E. K., Green, J. R., Barnett, C., & Bede, P. (2019). Clinical measures of bulbar dysfunction in ALS. *Frontiers in Neurology*, 10, 106. doi:10.3389/fneur.2019.00106

[6] Vashkevich, M., & Rushkevich, Y. (2021). Classification of ALS patients based on acoustic analysis of sustained vowel phonations. *Biomedical Signal Processing and Control*, 65, 102350. doi:10.1016/j.bspc.2020.102350

[7] Vieira, F. G., Venugopalan, S., Premasiri, A. S., McNally, M., Jansen, A., McCloskey, K., Brenner, M. P., & Perrin, S. (2022). A machine-learning based objective measure for ALS disease severity. *NPJ Digital Medicine*, 5(1), 45. doi:10.1038/s41746-022-00588-8

[8] Rong, P., Yunusova, Y., Eshghi, M., Rowe, H. P., & Green, J. R. (2020). A speech measure for early stratification of fast and slow progressors of bulbar amyotrophic lateral sclerosis: Lip movement jitter. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 21(1–2), 34–41. doi:10.1080/21678421.2019.1681454

[9] Tena, A., Claria, F., Solsona, F., Meister, E., & Povedano, M. (2021). Detection of bulbar involvement in patients with amyotrophic lateral sclerosis by machine learning voice analysis: Diagnostic decision support development study. *JMIR Medical Informatics*, 9(3), e21331. doi:10.2196/21331

[10] Wang, L., Gong, Y., Dawalatabad, N., Vilela, M., Placek, K., Tracey, B., ... & Glass, J. (2024). Automatic prediction of amyotrophic lateral sclerosis progression using longitudinal speech transformer. *arXiv preprint*, arXiv:2406.18625.

[11] Stegmann, G. M., Hahn, S., Liss, J., Shefner, J., Rutkove, S., Shelton, K., Duncan, C. J., & Berisha, V. (2020). Early detection and tracking of bulbar changes in ALS via frequent and remote speech analysis. *npj Digital Medicine*, 3(1), 132. doi:10.1038/s41746-020-00335-x

[12] Papaiz, F., Dourado Jr, M. E. T., de Medeiros Valentim, R. A., Pinto, R., de Morais, A. H. F., & Arrais, J. P. (2024). Ensemble-imbalance-based classification for amyotrophic lateral sclerosis prognostic prediction: identifying short-survival patients at diagnosis. *BMC Med Inform Decis Mak*, 24(1), 80. doi:10.1186/s12911-024-02484-5

[13] Pancotti, C., Birolo, G., Rollo, C. et al. Deep learning methods to predict amyotrophic lateral sclerosis disease progression. *Scientific Reports*, 12(1), 13738. doi:10.1038/s41598-022-17805-9

[14] Turabieh, H., Afshar, A. S., Statland, J., Song, X., & Pooled Resource Open-Access ALS Clinical Trials Consortium. (2024, January). Towards a machine learning empowered prognostic model for predicting disease progression for amyotrophic lateral sclerosis. *AMIA Annual Symposium Proceedings*, 2023, 718–725.

[15] Barbalho, I. M., Fonseca, A. L., Fernandes, F., Henriques, J., Gil, P., Nagem, D., ... & Valentim, R. A. (2023). Digital health solution for monitoring and surveillance of Amyotrophic Lateral Sclerosis in Brazil. *Frontiers in Public Health*, 11, 1209633. doi:10.3389/fpubh.2023.1209633

[16] Tavazzi, E., Longato, E., Vettoretti, M., Aidos, H., Trescato, I., Roversi, C., ... & Alves, I. (2023). Artificial intelligence and statistical methods for stratification and prediction of progression in amyotrophic lateral sclerosis: A systematic review. Artificial Intelligence in Medicine, 102588. doi:10.1016/j.artmed.2023.102588

[17] Dadu, A., Satone, V., Kaur, R., Hashemi, S. H., Leonard, H., Iwaki, H., ... & Faghri, F. (2022). Identification and prediction of Parkinson's disease subtypes and progression using machine learning in two cohorts. *NPJ* Parkinson's Disease, 8(1), 172.

[18] Agbo, C. C., Mahmoud, Q. H., & Eklund, J. M. (2019, April). Blockchain technology in healthcare: a systematic review. In Healthcare (Vol. 7, No. 2, p. 56). *MDPI*. doi:10.3390/healthcare7020056

[19] Xu, G., Qi, C., Dong, W., Gong, L., Liu, S., Chen, S., Liu, J., & Zheng, X. (2023). A privacy-preserving medical data sharing scheme based on blockchain. *IEEE Journal of Biomedical and Health Informatics*, 27(2), 698–709. doi:10.1109/JBHI.2022.3203577

[20] Kumar, R., Khan, A. A., Kumar, J., et al. (2021). Blockchain-federated-learning and deep learning models for COVID-19 detection using CT imaging. *IEEE Sensors Journal*, 21(14), 16301–16314. doi:10.1109/JSEN.2021.3076767

[21] Li, W., Milletarì, F., Xu, D., Rieke, N., Hancox, J., Zhu, W., Baust, M., Cheng, Y., Ourselin, S., Cardoso, M. J., & Feng, A. (2019). Privacy-preserving Federated Brain Tumour Segmentation. *Machine learning in medical imaging*. MLMI (Workshop), 11861, 133–141. doi:10.1007/978-3-030-32692-0_16

[22] Dubbioso, R., Spisto, M., Verde, L., et al. (2024). Voice signals dataset for amyotrophic lateral sclerosis patients and healthy controls. *Scientific Data*, 11(1), 800. doi:10.1038/s41597-024-03597-2

[23] Brookes, M. VOICEBOX: Speech processing toolbox for Matlab. Department of Electrical & Electronic Engineering. Imperial College, London; 2002.

[24] Jadoul, Y., Thompson, B., & De Boer, B. (2018). Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, 71, 1–15. doi:10.1016/j.wocn.2018.07.001

[25] Teixeira, J. P., Oliveira, C., & Lopes, C. (2013). Vocal acoustic analysis–jitter, shimmer and hnr parameters. *Procedia Technology*, 9, 1112–1122. doi:10.1016/j.protcy.2013.12.124

[26] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x

[27] Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8), 1226–1238. doi:10.1109/TPAMI.2005.159

[28] Wang, L., Gong, Y., Dawalatabad, N., Vilela, M., Placek, K., Tracey, B., Gong, Y., Premasiri, A., Vieira, F., & Glass, J. (2024). Automatic prediction of amyotrophic lateral sclerosis progression using longitudinal speech transformer. In *Proceedings of Interspeech 2024* (pp. 2000–2004). doi:10.21437/Interspeech.2024-158

[29] Buterin, V. (2021). An incomplete guide to rollups, 2021. URL: https://vitalik.ca/general/2021/01/05/rollup.html.

[30] Thibault, L., Sarry, T., & Hafid, A. S. (2022). Blockchain scaling using rollups: A comprehensive survey. *IEEE Access*, 10, 93039–93054. doi:10.1109/ACCESS.2022.3200051

[31] Cebola, R., Folgado, D., Carreiro, A. V., & Gamboa, H. (2023). Speech-Based Supervised Learning Towards the Diagnosis of Amyotrophic Lateral Sclerosis. In Biosignals (pp. 74–85). doi:10.5220/0011694700003414

[32] Farrokhi, Z., Zakavi, S. A., Sarafraz, A., Valifard, M., Yousefzadeh, S., Tafreshi, Z. M., et al. (2025). Acoustic signatures of bulbar ALS: Predictive modeling with sustained vowels and LightGBM. *eNeurologicalSci*, 40, 100579. doi:10.1016/j.ensci.2024.100579

[33] Louja, A., Jamali, A., & Naja, N. (2024). Blockchain-powered artificial intelligence for healthcare systems data orchestration. In *International Conference on Mathematics Data Science* (pp. 155–163). Springer.

[34] Louja, A., Drira, I., Jamali, A., Naja, N., & Sliman, L. (2025). Wearable sensor-based eye-rubbing monitoring: a hybrid CNN-attentionrub architecture for keratoconus prevention. *Journal of Supercomputing*, 81, 1268. https://doi.org/10.1007/s11227-025-07744-3.

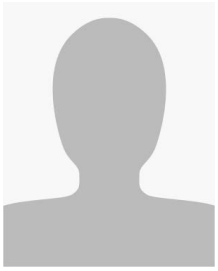
[35] Drira, I., Louja, A., Sliman, L., Soler, V., Noor, M., Jamali, A., & Fournie, P. (2024). Eye-rubbing detection tool using artificial intelligence on a smartwatch in the management of keratoconus. *Translational Vision Science & Technology*, 13(12), 16. https://doi.org/10.1167/tvst.13.12.16.

Biographies

Ayoub Louja is a researcher at the Faculty of Sciences and Technologies at Hassan First University of Settat, Morocco, affiliated with the Laboratory IR2M. He holds an engineering degree in Data Science from the National School of Applied Sciences of Berrechid. His research interests include artificial intelligence, deep learning, secure data systems, healthcare diagnosis, and Blockchains.

Yassin Zaiouane is a researcher at the Faculty of Sciences and Technologies, Hassan First University of Settat, Morocco, affiliated with the Laboratory IR2M. His research interests include machine learning, deep learning, healthcare systems, blockchain technologies, and Internet of Things (IoT).

Manal Benchrif is a researcher at the Faculty of Sciences and Technologies, Hassan First University of Settat, Morocco, affiliated with the Laboratory IR2M. Her research focuses on cloud computing, artificial intelligence, and distributed intelligent systems.



Najoua Azizi is a researcher at the Faculty of Sciences and Technologies, Hassan First University of Settat, Morocco, affiliated with the Laboratory

IR2M. Her research interests include network security, software-defined networking (SDN), and intelligent communication systems.

Abdellah Jamali is a Professor of Computer Science at the Faculty of Sciences and Technologies, Hassan First University of Settat, Morocco. He is a member of the IR2M Laboratory. His research interests cover computer networks, cloud computing, IPv6, software-defined networking (SDN), and artificial intelligence.

Najib Naja is a Professor at the National Institute of Posts and Telecommunications (INPT), Rabat, Morocco. He specializes in network simulation, routing, and cybersecurity in wireless and software-defined networks. His current research explores intelligent and adaptive communication systems.