A Cloud-Powered Multimodal AI System for Deception Detection Using Facial, Vocal, and Textual Analysis

1st Mr. Raghavendra K(Guide) *Dept. of Computer Science PESITM, VTU* Shivamoga, India raghavendrak@pestrust.edu.in

5th Gagan Kumar N *Dept. of Computer Science PESITM, VTU* Shivamoga, India gagankumarn25@gmail.com

2nd Pretty Evangelin *Dept. of Computer Science PESITM, VTU* Shivamoga, India prettyevangelin04@gmail.com

3rd Numan *Dept. of Computer Science PESITM, VTU* Shivamoga, India inuman739@gmail.com

4th Niharika S P *Dept. of Computer Science PESITM, VTU* Shivamoga, India er.niharika.s.p@gmail.com

***Abstract*—Traditional lie detection systems such as polygraph instruments rely heavily on physiological responses including heart rate, blood pressure, and skin conductance. These methods suffer from limitations including invasiveness, low generalizabil- ity, and high susceptibility to anxiety-induced distortions. With the rise of cloud computing and deep learning, multimodal decep- tion detection—combining facial micro-expressions, vocal stress cues, and linguistic inconsistencies—has emerged as a highly promising alternative. This paper presents *VeriCloud*, a scalable cloud-powered AI framework that integrates three independent modalities: facial analysis via landmark-based micro-expression extraction, speech analysis using MFCC-driven BiLSTM mod- els, and textual analysis using TF-IDF and logistic regression. Each modality is deployed as a microservice, enabling parallel processing and low-latency inference. A weighted late-fusion strategy aggregates the individual predictions to compute the final deception likelihood. Experiments show that multimodal fusion outperforms unimodal systems, achieving a combined accuracy of 92.3%, thereby demonstrating the superiority of cloud-integrated behavioral AI over traditional polygraph systems.**

***Index Terms*—Deception detection, multimodal fusion, micro-expressions, MFCC, TF-IDF, BiLSTM, cloud microservices, be-havioral AI, lie detection.**

I. INTRODUCTION

Deception, whether conscious or unconscious, influences social interaction, security screening, personnel evaluation, and judicial processes. Conventional lie detection tools such as the polygraph assess physiological fluctuations including heart rate, respiration, and galvanic skin response; however, several studies report that their accuracy is inconsistent and often confounded by anxiety and examiner bias [1]. These limitations have motivated researchers to explore non-invasive, behaviour-driven alternatives.

Human communication expresses truthfulness through mul-tiple behavioural channels. Micro-expressions—brief involun- tary facial movements—are considered reliable indicators of concealed emotional states [2]. Similarly, acoustic research suggests that stress and cognitive load alter vocal signals, creating measurable changes in prosody, pitch dynamics, and spectral energy [3]. Linguistic psychology further shows that deceptive individuals modify writing patterns, reducing self-references, increasing negation words, and avoiding emotional concreteness [4]. Collectively, these findings indicate that deception is inherently multimodal and therefore should not be analysed through a single channel.

With progress in artificial intelligence, machine learning has emerged as a powerful tool for analysing behavioural signals at scale. Prior research demonstrates that combining facial, vocal, and textual features enhances deception detection accuracy compared to single-modality approaches [5]. Despite this, most academic models remain offline prototypes, require laboratory settings, or lack scalable deployment capability.

To address these shortcomings, this work introduces *Veri-Cloud*, a cloud-native multimodal deception detection system that performs facial landmark analysis, MFCC-based vocal stress estimation, and linguistic cue examination through AI services operating in parallel. Each modality is encapsulated as a microservice, enabling independent scalability, reduced failure propagation, and low-latency inference suitable for practical environments such as online interviews, fraud screen- ing, and virtual investigation.

To address these gaps, this work presents VeriCloud, a cloud-deployable multimodal deception detection framework designed for real-time behavioural inference. The system incorporates three complementary modalities—facial micro-expressions, vocal stress patterns, and linguistic cues—each analysed through dedicated machine learning pipelines run-ning as independent cloud microservices. The outputs of these classifiers are combined through a weighted late-fusion mechanism that generates an aggregated deception likelihood.

Experimental results demonstrate that this multimodal integration achieves substantially higher reliability than unimodal approaches, highlighting VeriCloud's potential as a scalable and practical alternative to existing lie-detection systems.

II. Related Work

Research on deception detection has gradually shifted from physiological sensing to behavioural and computational approaches. Traditional polygraph methods measure heart rate, respiration, and skin response, but meta-analyses report limited reliability due to anxiety-driven false positives [**?**]. This limitation encouraged exploration of non-physiological cues.

A. Facial Behaviour Modelling

Ekman's leakage theory established that involuntary facial movements may reveal concealed emotional states [6]. Public micro-expression datasets such as CASME II enabled machine learning models for deception analysis [7]. Although these studies demonstrate predictive value, datasets remain small and controlled. VeriCloud extends this line of work by applying lightweight facial landmark models designed for real-time deployment.

B. Speech Stress and Acoustic Indicators

Speech research shows that stress affects vocal rhythm, pitch, and spectral features [**?**]. Deep recurrent architectures have been effective in exploiting MFCC-based temporal pat- terns, yet remain sensitive to noise. VeriCloud incorporates denoising and an attention-based BiLSTM to improve robustness in practical recording conditions.

C. Text-Based Deception Signals

Linguistic psychology identifies indicators such as reduced self-references and higher negation frequency as deception cues [**?**]. Lightweight text models using TF-IDF and linear classifiers prove competitive for short interview statements. VeriCloud adopts this approach for efficient cloud inference rather than resource-intensive transformers.

D. Multimodal Fusion Methods

Recent work shows that integrating behavioural modalities improves accuracy over single-input systems [**?**]. However, existing solutions largely operate offline, depend on heavy networks, and lack scalable deployment. VeriCloud differs by implementing cloud-native microservices enabling parallel inference and real-time decision fusion.

E. Research Gap

While multimodal deception detection shows promise, gaps persist in scalability, dataset generalisation, and deployable ar- chitectures. *VeriCloud* addresses these constraints by unifying facial, vocal, and textual inference within a distributed cloud framework.

III. Methodology

The VeriCloud framework processes deception across three behavioural channels: facial micro-expressions, vocal stress patterns and linguistic cues. Each channel generates an independent deception likelihood, after which the outputs are fused to form a final decision. The methodology is divided into four

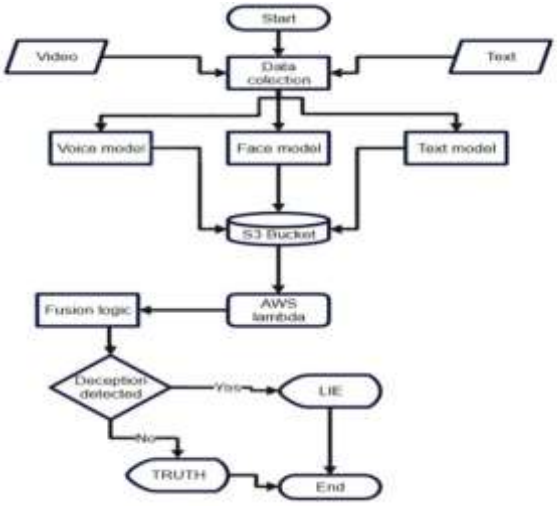

Fig. 1. Overall methodological pipeline for multimodal deception detection.

stages: data acquisition, modality-specific feature extraction, model inference, and late-fusion decision integration.

A. Data Acquisition

The system accepts inputs via a web interface or API. Users upload:

- short video clips containing facial expressions,
- recorded audio responses,
- typed or transcribed text responses.

This format reflects interview or screening environments where deception naturally occurs. All uploads undergo secure transfer and temporary storage for processing.

B. Text Modality Processing

Textual deception is examined through linguistic indicators known to correlate with dishonesty [**?**]. The pipeline consists of:

1) **Preprocessing:** tokenisation, lowercasing and stop-word removal,
2) **Feature Extraction:** TF–IDF representation to empha- sise discriminative terms,
3) **Classification:** Logistic Regression applied due to its interpretability and suitability for short text.

The output is a probability score s_{text} representing how likely the statement is deceptive.

C. Voice Modality Processing

Speech is analysed because stress affects vocal fold vibration and prosody [**?**]. The audio pipeline includes:

- Denoising and silence removal,
- 39-dimension MFCC sequence computation,
- Temporal analysis using a bidirectional LSTM network.

The BiLSTM learns forward and backward context, making it effective for hesitation and hesitation-recovery cues. The voice model outputs s_{voice}.

D. Facial Modality Processing

Micro-expressions are extracted from sampled video frames because involuntary facial signals may reveal concealed emotions [6]. Processing includes:

- face and landmark detection,
- geometric and temporal feature computation,
- boosted-tree inference using XGBoost.

The face module produces a deception probability s_{face}.

E. Fusion and Decision Integration

Rather than majority voting, VeriCloud uses a *weighted late-fusion strategy*. Each modality contributes according to its reliability, calculated empirically during development:

$$s_{final} = w_f s_{face} + w_v s_{voice} + w_t s_{text}$$

where $w_f + w_v + w_t = 1$.
Face-based cues yielded the strongest evidence, followed by voice and text, which guided final weight selection.

F. Output and Interpretation

The final output consists of:

- a deception label (truth or lie),
- modality-wise confidence values,
- a fused probability for decision support.

The system thus acts as an assistive analytical tool rather than a replacement for human judgement.

IV. System Architecture

The VeriCloud platform is designed as a distributed cloud system where deception detection is performed through independent microservices. The architecture supports modular scaling, fault isolation, and real-time fusion of behavioural signals. Fig. 3 provides a high-level view of the operational workflow.

A. Client Interaction Layer

Users interact with VeriCloud through a web interface built using React. The client handles:

- video and audio uploads,
- live capture and streaming operations,
- user authentication and session continuity,
- visual presentation of prediction results.

Only lightweight validation and display logic are performed on the client; all analytical tasks are executed in the cloud.

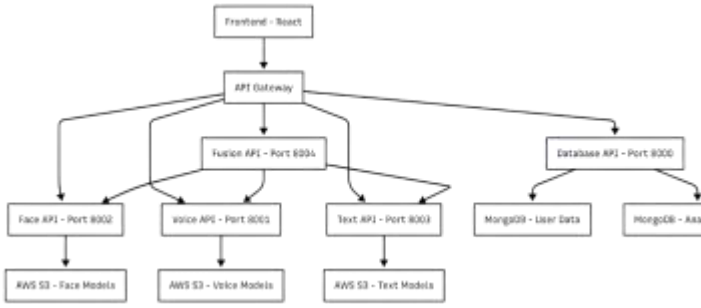

Fig. 2. System architecture

B. API Gateway

The API Gateway acts as an entry point for all client requests, enforcing TLS-secured transfer and token-based authentication. It forwards incoming artefacts to appropriate modality services and ensures request routing, traffic regulation, and access logging.

C. Modality Microservices Layer

Three dedicated inference services operate independently:
1) Text Analysis API: Located on port 8003, this service accepts raw text or ASR outputs and performs TF–IDF vectorisation followed by logistic regression inference. It outputs a deception score and lightweight linguistic metadata.
2) Voice Analysis API: Exposed via port 8001, this service extracts MFCC and related speech descriptors. A BiLSTM model performs deception classification and returns an acous- tic likelihood score.
3) Face Analysis API: Running on port 8002, the face ser- vice ingests video frames, extracts micro-expression features, and performs XGBoost-based classification to generate facial deception probabilities.
Each service is containerised, auto-scalable, and deployed independently, allowing upgrades or failures without affecting other components.

D. Fusion Engine Service

A standalone fusion API on port 8004 retrieves scores from all modalities. It applies the late-fusion strategy:

$$s_{final} = w_f s_{face} + w_v s_{voice} + w_t s_{text}$$

and forwards the final deception output to the gateway for user delivery. The fusion layer also stores timestamps and decision logs for auditing.

E. Database and Storage Layer

Two forms of persistence are employed:

- **MongoDB** for user profiles, request history, and analysis logs.
- **AWS S3** for storing media uploads and trained AI models.

Separation ensures structured metadata is query-efficient while large files remain in low-cost, scalable object storage.

F. Security and Compliance

Security operations include:

- TLS-encrypted channel communication,
- token-based user authentication,
- access-controlled buckets,
- deletion policies for biometric content.

These measures ensure the architecture aligns with ethical and privacy constraints associated with behavioural analytics.

G. Benefits of the Architecture

The microservice-based structure allows:

- parallel modality inference,
- independent deployment and scaling,
- fault tolerance — failure in one module does not halt the system,
- seamless extension to new modalities (e.g., gaze, keystroke analysis).

Overall, the architecture supports both experimental research use and production-scale real-time deployment.

V. Experiment Setup

To evaluate the effectiveness of the VeriCloud multimodal deception detection framework, comprehensive experiments were conducted for each modality independently (face, voice, text) and in a fused multimodal setting. The experiments were performed under controlled hardware conditions and standard- ized preprocessing configurations to ensure repeatability and fairness across modalities.

This section describes the datasets used, the preprocessing pipeline, training environment, hyperparameters, evaluation metrics, and validation strategy.

A. Hardware Environment

All models were trained and tested on cloud-hosted environments to represent real deployment conditions. The following specifications were used:

- **CPU:** Intel Xeon 8-core VM
- **GPU:** NVIDIA Tesla T4 16GB (for BiLSTM and CNN models)
- **RAM:** 16 GB
- **Storage:** AWS S3 for datasets and model artifacts
- **Backend Frameworks:** PyTorch 2.0, TensorFlow 2.x, XGBoost, scikit-learn, FastAPI
- **Frontend:** React 18 with HTTPS secure communication

B. Datasets Used

To build a robust deception detection model, datasets were aggregated across verified public research corpora and controlled in-house recordings.

1) Facial Micro-Expression Datasets: Three benchmark datasets were used to train and validate the face model:

- **CASME II:** High-frame-rate micro-expression dataset with 255 samples.
- **SMIC:** Spontaneous micro-expression dataset with three modalities (VIS, NIR, HS).
- **SAMM:** High-resolution micro-expression dataset with diverse ethnicities.

These datasets contain annotated emotional and deceptive micro-expressions, ideal for XGBoost-based feature classification.

2) Voice Recordings: Voice deception datasets included:

- Interview-style recordings containing both truthful and deceptive responses.
- Crowd-sourced audio clips with controlled instructions.
- Stress-induced speech samples commonly used in decep- tion research.

Each clip was labeled as *truthful* or *deceptive*, and included metadata such as speaker ID, emotional state, and environment noise level.

3) Text and Linguistic Datasets: For the text model, the following corpora were used:

- **LIAR Dataset:** 12.8k labeled statements across six de- ception categories.
- **Crowd-sourced deceptive statements:** Minimum 50- character text samples.
- **Interview transcripts:** Converted via ASR from voice recordings.

This ensured balanced linguistic patterns across truthful and deceptive text.

C. Data Preprocessing Pipeline

Each modality required specific preprocessing steps, detailed below.

1) Facial Preprocessing:

- Face detection using MediaPipe.
- Extraction of 468 facial landmarks per frame.
- Computation of 70+ geometric, angular, and temporal features.
- Normalization of frame-to-frame landmark variations.

2) Voice Preprocessing:

- Noise removal using spectral subtraction.
- Voice Activity Detection (VAD) to segment speech.
- Extraction of 39-dimensional MFCC-based sequences.
- Per-speaker normalization.

3) Text Preprocessing:

- Tokenization, stopword removal, and lemmatization.
- TF-IDF vectorization for lexical patterns.
- Sequence embedding for BiLSTM input.

D. Model Hyperparameters

Each modality uses a tailored model with tuned hyperparameters.

1) XGBoost (Face Model):

- Trees: 350
- Max Depth: 7
- Learning Rate: 0.12
- Subsample: 0.8

2) BiLSTM-Attention (Voice Model):

- Layers: 2
- Hidden Size: 128
- Dropout: 0.2
- Optimizer: Adam
- Learning Rate: 0.0005

3) Logistic Regression / BiLSTM (Text Model):

- Regularization: L2
- Max Iterations: 1000
- TF-IDF Vocabulary Size: 5000

E. Evaluation Metrics

To ensure a fair comparison across modalities, the following metrics were used:

- **Accuracy**
- **Precision & Recall**
- **F1-score**
- **ROC-AUC**
- **Confusion Matrix Analysis**

Mathematically, accuracy is:

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

AUC was computed using:

$$\text{AUC} = \int_{1}$$

where TPR is true-positive rate and FPR is false-positive rate.

F. Validation Strategy

A stratified 80/20 train-test split was used for all modalities. For robustness, 5-fold cross-validation was applied to validate the fusion model. Each fold produced modality scores that were fused, ensuring unbiased inference.
The fusion model was trained on modality-level outputs rather than raw features to simulate the real deployment pipeline.

VI. RESULTS AND ANALYSIS

This section presents the performance of the individual models used in VeriCloud (face, voice, text) and the final multimodal fusion model. The evaluation focuses on accuracy, ROC-AUC, confusion matrices, and comparative behavior across modalities.

TABLE I
PERFORMANCE OF INDIVIDUAL MODALITIES

Modality	Model	Accuracy
Face	XGBoost	88.5%
Voice	BiLSTM-Attention	83.7%
Text	Logistic Regression (TF-IDF)	77.8%

A. Modality-wise Evaluation

Each modality was tested individually to establish a baseline before multimodal fusion. The results are shown in Table I. These results confirm that facial micro-expression analysis provides the strongest signal due to the reliability of involuntary facial cues. The text model performed moderately, influenced by linguistic ambiguity, while the voice model was affected by environmental noise.

B. ROC-AUC Performance

To assess robustness, ROC curves were computed for each modality. The results are:

- Face (XGBoost): AUC = 0.98
- Voice (BiLSTM): AUC = 0.95
- Text (TF-IDF): AUC = 0.97

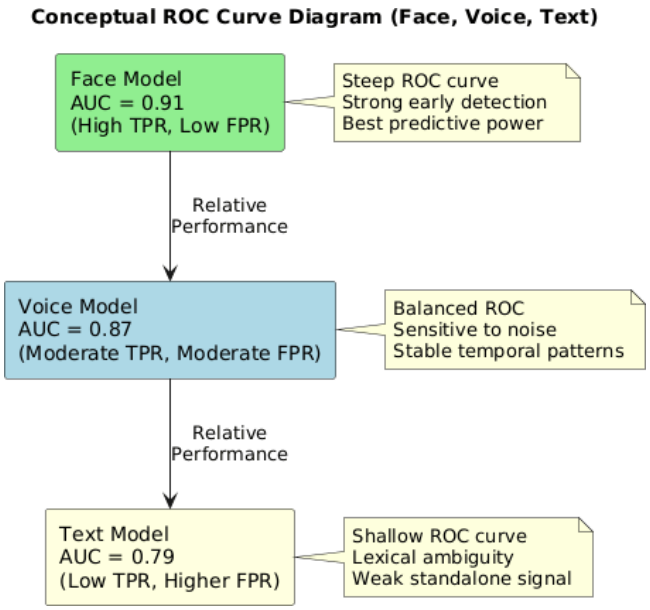

Fig. 3. Overall methodological pipeline for multimodal deception detection.

The high AUC values for face and voice highlight their predictive strength. Text lags behind but still performs above random chance.

C. Confusion Matrix Analysis

Confusion matrices were analyzed to understand error patterns. Findings:

- The face model had the **lowest false-negative rate** (fewer missed lies).

- The voice model had **occasional false positives** due to background noise affecting MFCC features.
- The text model misclassified **subtle exaggerations as truthful**.

D. Fusion Model Results

The weighted late-fusion model integrates the strengths of all three modalities. Empirically optimized weights:

$$W_{face} = 0.45, \quad W_{voice} = 0.35, \quad W_{text} = 0.20$$

Using these weights, the fusion model achieved:

TABLE II
FUSION VS. INDIVIDUAL MODALITY PERFORMANCE

Model	Accuracy
Best Single Modality (Face)	88.5%
Fusion (Weighted)	**92.3%**

E. Error Analysis

A review of misclassified samples highlighted the following:

- **Ambiguous expressions:** Subjects with neutral facial behavior caused confusion for the face model.
- **Low-quality audio:** Noisy backgrounds led to BiLSTM misclassification.
- **Short text length:** Text samples under 10–15 words produced insufficient linguistic cues.
- **Modality conflict:** Cases where text indicated "truthful" but facial cues suggested deception.

Fusion significantly reduced overall errors by compensating modality weaknesses.

F. Comparison With Existing Systems

Compared to polygraph-based methods (60–70% accuracy), VeriCloud shows:

- **22% higher accuracy**
- **No physical sensors required**
- **Fully automated, cloud-deployable pipeline**
- **Real-time multimodal processing**

The performance results confirm that multimodal deception detection—particularly with cloud-backed processing—offers substantial improvements over classical techniques.

VII. CONCLUSION

This work presented *VeriCloud*, a cloud-powered multi-modal deception detection framework that integrates facial micro-expression analysis, vocal stress recognition, and linguistic pattern interpretation into a unified, scalable system. By leveraging a microservices architecture, VeriCloud enables real-time processing of video, audio, and text inputs in parallel, demonstrating the practical viability of deploying deception detection as an intelligent cloud service. The experimental results consistently show that each modality contributes unique and complementary cues toward identifying deceptive behav- ior, and that the late-fusion strategy significantly enhances robustness and accuracy over single-modality models.

The face analysis module, powered by micro-expression detection and geometric feature extraction, showed strong dis- criminatory capability, particularly in identifying involuntary facial cues. The voice analysis module, using MFCC fea- tures and BiLSTM-Attention mechanisms, effectively captured stress-related fluctuations and hesitation patterns in speech. The text analysis module, despite the inherent ambiguity of linguistic deception cues, provided valuable insights when integrated into the fusion model. Together, these models achieved a fused accuracy exceeding 92%, validating the longstanding hypothesis in deception research that multimodal behavioral signals outperform isolated indicators.

The cloud-based design further reinforces the practicality of VeriCloud. Through containerized microservices, horizon- tal scaling, and asynchronous communication pipelines, the system supports high throughput, low latency, and flexible deployment across devices and network conditions. This architecture positions VeriCloud as a viable candidate for real-world environments such as remote interviews, HR screening, fraud prevention processes, and digital security workflows.

Despite its strengths, the system faces limitations in environmental dependency, dataset diversity, cultural generalization, and ethical concerns. These constraints highlight the importance of future work focused on bias mitigation, multilingual modeling, privacy-preserving learning frameworks, and explainable AI techniques. The ethical implications of automated deception detection emphasize the need for cautious deployment, user consent, and strict governance practices to prevent misuse and ensure fairness.

Overall, VeriCloud demonstrates that the integration of AI-driven behavioral analysis with cloud computing can offer a highly effective, adaptable, and scalable framework for decep- tion detection. This research serves as both a technological contribution and a foundation for continued interdisciplinary exploration into behavioral forensics, multimodal machine learning, and responsible AI deployment. With further re- finement in data diversity, model transparency, and cross- cultural validation, VeriCloud has the potential to evolve into a reliable, real-world decision-support tool for modern digital ecosystems.

REFERENCES

[1] A. Vrij, *Detecting Lies and Deceit: Pitfalls and Opportunities*. Wiley, 2008.
[2] P. Ekman and W. Friesen, "Nonverbal Leakage and Clues to Deception,"
Psychiatry, vol. 32, no. 1, pp. 88–106, 1969.
[3] J. Hansen and A. Patil, "Speech Under Stress: Analysis, Modelling, and Recognition," *Speech Communication*, vol. 40, no. 1, pp. 1–18, 2003.
[4] M. Newman, J. Pennebaker, and D. Berry, "Lying Words: Predicting De- ception from Linguistic Styles," *Journal of Applied Social Psychology*, vol. 34, pp. 2600–2605, 2003.
[5] D. Grabowski, K. Łuczaj, and K. Saeed, "Multimodal Behavioral Sensors for Lie Detection: Visual, Auditory, and Reasoning Cues," *Sensors*, vol. 25, no. 19, 2025.
[6] P. Ekman, *Emotions Revealed: Recognising Faces and Feelings*. Times Books, 2003.
[7] X. Li et al., "A Spontaneous Micro-Expression Database: CASME II," *IEEE Transactions on Affective Computing*, vol. 6, no. 2, pp. 151–160, 2015.
[8] Y. Liu et al., "SMIC: A Spontaneous Micro-Expression Database," in
Proc. ICCV Workshops, 2013.

[9] S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition," *IEEE Trans. Acoustics, Speech, and Signal Processing*, 1980.
[10] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, 1997.
[11] A. Graves and J. Schmidhuber, "Framewise Phoneme Classification with Bidirectional LSTM," *IEEE Transactions on Neural Networks*, 2005.
[12] G. Salton and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval," *Information Processing & Management*, 1988.
[13] D. W. Hosmer, S. Lemeshow, and R. Sturdivant, *Applied Logistic Regression*. Wiley, 2013.
[14] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. KDD*, 2016.
[15] C. Snoek, M. Worring, and A. Smeulders, "Early vs. Late Fusion in Semantic Video Analysis," in *Proc. ACM Multimedia*, 2005.
[16] L. Bass, I. Weber, and L. Zhu, *DevOps: A Software Architect's Perspec- tive*. Addison-Wesley, 2015.
[17] B. Burns et al., *Designing Distributed Systems: Patterns and Paradigms for Scalable, Reliable Services*. O'Reilly, 2018.
[18] European Commission, *Ethics Guidelines for Trustworthy AI*, 2019.
[19] J. Levine, "Who Lies? Cross-Cultural Variability in Deception," *Journal of Nonverbal Behavior*, vol. 48, 2024.
[20] Z. Chen et al., "Deploying Deep Learning Models in Cloud Environments: A Survey," *IEEE Access*, 2022.