Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

AVA: AI Veterinary Assistance - An NLP and Semantic Vector-Based Clinical Decision Support System for Animal Healthcare

Mr. Alen Denny
Department of Computer Science & Engineering
Federal Institute of Science and Technology
Angamaly, India

Mr. Christin V S
Department of Computer Science & Engineering
Federal Institute of Science and Technology
Angamaly, India

Mr. Basil Paul
Department of Computer Science & Engineering
Federal Institute of Science and Technology
Angamaly, India

Ms. Sheelu Susan Mathews
Assistant Professor (Guide)
Department of Computer Science & Engineering
Federal Institute of Science and Technology
Angamaly, India

Mr. Christo Tomy Joseph
Department of Computer Science & Engineering
Federal Institute of Science and Technology
Angamaly, India

Abstract — Quick and correct veterinary assessment is a major problem for pet owners who often find it hard to tell the difference between minor health issues and dangerous emergencies. Traditional symptom checking methods depend on hospital-based clinical visits, which take a lot of time, need specific institutions, and are not available outside working hours. This paper presents AVA (AI Veterinary Assistance), a smart, NLP-based clinical decision support system for basic veterinary assessment. The system handles unstructured natural language symptom descriptions using a two-part prediction design that includes a Lexical Heuristic Matcher and a Semantic Vector Engine built on the `all-MiniLM-L6-v2` SentenceTransformer model. AVA pulls out structured patient profiles from free-form text, links symptoms to a carefully collected MongoDB disease database of over 205 conditions, creates relevant follow-up questions, and provides ranked possible diagnoses with confidence scores and urgency levels. Testing results show a macro-average AUC of 0.988 and strong disease classification performance across multiple veterinary categories. The system is built as an interactive Streamlit web application with multi-language support, voice input through Whisper ASR, and optional skin lesion image analysis. AVA offers a scalable, easy-to-use, and clear AI-powered framework for helping pet owners and veterinary professionals in basic clinical assessment.

Keywords — Artificial Intelligence; Natural Language Processing; Veterinary Decision Support; Semantic Embeddings; Clinical Triage; Disease Prediction; SentenceTransformers; Streamlit.

1. INTRODUCTION

Veterinary care depends heavily on quick and correct identification of clinical signs. Pet owners are increasingly the first ones to assess animal health, yet they lack clinical training to tell the difference between conditions that need immediate emergency attention and those that can be handled with home care. Delays or wrong understanding in this basic assessment phase can badly affect patient outcomes [1].

The fast growth of AI and natural language processing (NLP) in medical and clinical fields has created new chances for smart decision support systems. While big improvements have been made in human healthcare applications [2, 7], veterinary medicine has received much less attention in the clinical AI research. Existing tools for pet owners either depend on very basic keyword matching or provide general, often worrying information taken from common search engines, neither of which is good enough for organized clinical assessment.

This paper introduces **AVA** (AI Veterinary Assistance), a specialized clinical decision support system made to fill this gap. AVA takes unstructured text or voice descriptions of animal symptoms, pulls out structured clinical profiles using a multi-step NLP pipeline, and provides ranked possible diagnoses with urgency levels and specific follow-up questions. The system combines a rule-based Lexical Heuristic Engine with a dense Semantic Vector Engine, getting strong performance even when user descriptions use informal or unclear medical language.

The main contributions of this work are:

1. A two-part disease prediction design combining heuristic and semantic approaches;
2. A structured NLP extraction pipeline for veterinary clinical attributes from free-form text;
3. An adaptive follow-up question generation module that improves diagnostic confidence;
4. A carefully collected MongoDB-backed veterinary disease database covering 205+ conditions;
5. A ready-to-use, multi-language web interface with voice and image input support;
6. Experimental testing showing macro-average AUC of 0.988 on test data.

The rest of this paper is organized as follows: Section II gives background and related work; Section III describes the system methodology; Section IV details the system architecture and modules; Section V presents algorithms; Section VI covers implementation; Section VII reports experimental results;

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

Section VIII discusses challenges and limitations; Section IX concludes with future research directions.

2. BACKGROUND AND RELATED WORK

A. AI in Clinical Decision Support

Clinical decision support systems (CDSS) have changed a lot with improvements in machine learning and NLP. Early rule-based expert systems have been replaced by data-driven models that can handle unstructured clinical text. Large language models (LLMs) such as BERT [7] and GPT-4o [6] have shown strong performance on clinical information extraction tasks, and few-shot prompting methods have further improved structured extraction from limited medical records [2].

Image-based diagnostic AI has also grown, with deep learning models achieving high accuracy in radiology [3] and histopathology tasks. Audio-based diagnostic tools, such as lung sound classifiers using one-dimensional convolutional neural networks (1D-CNNs), have achieved over 90% accuracy in respiratory condition detection [4]. Disease onset prediction models using electronic health records have achieved around 85% reliability in future population health settings [5].

B. Veterinary AI

Veterinary AI research has focused mainly on image-based diagnostics, with deep learning frameworks achieving F1 scores of up to 0.88 in multi-class animal disease classification tasks [1]. However, NLP-based veterinary assessment systems that accept free-form natural language input remain rare. The challenges specific to veterinary NLP include the variety of species-specific symptom vocabulary, informal owner descriptions, and the lack of large labelled datasets similar to those available in human medicine.

C. Semantic Embedding-Based Retrieval

Dense vector retrieval using sentence-level embeddings has proven useful for clinical information retrieval tasks. The `all-MiniLM-L6-v2` model from the SentenceTransformers library produces 384-dimensional embeddings that balance computational efficiency with strong semantic similarity performance [8]. Cosine similarity over pre-computed disease embeddings allows real-time matching even on basic hardware, making it suitable for deployment in settings with limited computational resources.

3. METHODOLOGY

AVA uses a modular, pipeline-based methodology designed to ensure clear separation between input processing, clinical reasoning, and output generation.

A. Literature and Knowledge Base Construction

Veterinary disease profiles were collected from peer-reviewed veterinary references, clinical databases, and domain expert consultation. Each disease record includes species compatibility flags, symptom lists, severity classification, treatment recommendations, and prevention guidelines. The resulting MongoDB collection contains 205+ disease entries covering gastrointestinal, respiratory, dermatological, neurological, urinary, and systemic categories across dogs, cats, and bovine species.

B. NLP Extraction Pipeline

Raw user input (text or transcribed audio) is processed through a multi-step NLP pipeline built using spaCy and NLTK. The pipeline performs tokenization, lemmatization, and pattern-based extraction across 30+ veterinary symptom categories. Patient demographic attributes (species, age, breed, weight) are extracted using specific regular expressions. Negation detection and contextual filtering are applied to reduce wrong symptom assignments.

C. Dual-Engine Prediction

AVA uses two complementary prediction engines working in parallel:

Lexical Heuristic Engine: Calculates a baseline confidence score for each candidate disease by computing the ratio of matched symptoms to total known disease symptoms. Categorical boosters (+0.03 per matched category) and species filters are applied to adjust scores.

Semantic Vector Engine: Encodes the patient symptom profile as a 384-dimensional embedding using `all-MiniLM-L6-v2`. Cosine similarity is calculated against pre-embedded disease vectors stored in MongoDB. A combined score mixing semantic similarity (75% weight) and lexical overlap (25% weight) is calculated and normalized.

D. Adaptive Follow-Up Generation

The top-ranked candidate diseases are passed to the Follow-Up Question Generator, which finds missing information in the initial patient description and creates 6–8 targeted clarifying questions organized across three categories: Symptom Details, Medical History, and Lifestyle. Questions are prioritized based on the severity of the leading possible diagnosis.

E. Evaluation Protocol

The system was tested on a held-out dataset of veterinary case descriptions labeled with ground-truth disease categories and severity levels. Standard classification metrics were calculated: accuracy, weighted F1-score, and multiclass area under the ROC curve (AUC). Bayesian hyperparameter optimization was applied to calibrate the confidence threshold (best C = 0.003594).

4. SYSTEM ARCHITECTURE

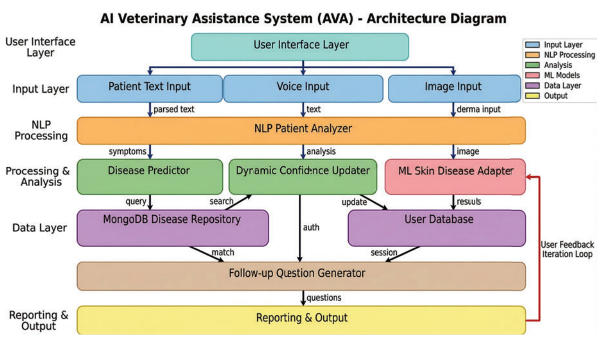

Figure 1: Overall System Architecture of AVA

The AVA system architecture is organized into three main pro-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

cessing stages, as shown in Figure 1.

A. Preprocessing Stage

The preprocessing stage takes raw multi-modal input (text, audio, or image). Text is cleaned and normalized; audio is transcribed using the Whisper ASR model; uploaded skin images undergo quality validation. Demographic attributes (species, age, breed, weight) are extracted and structured for further processing.

B. Analysis and Processing Stage

The main processing stage includes five sequential components:

- **VeterinaryNLPAnalyzer:** Rule-based symptom extraction over 30+ categories.
- **MongoDiseaseRepository:** Disease matching and ranking against the collected database.
- **Lexical Prediction:** Heuristic scoring and categorical confidence boosters.
- **Semantic Vector Engine:** SentenceTransformer-based embedding similarity computation.
- **Confidence Calculation:** Bayesian score fusion and urgency classification.

C. Validation and Output Stage

The validation stage applies score filtering, creates the adaptive follow-up question set, and puts together the final structured JSON clinical report containing ranked disease candidates, confidence scores, urgency assessment, and treatment recommendations for display on the Streamlit interface.

5. MODULE DESCRIPTIONS

A. NLP Patient Analyzer Module

The NLP Patient Analyzer handles unstructured text describing animal health issues. It extracts demographic variables (animal type, age, breed, weight) using named-entity recognition patterns and regular expressions. Symptom extraction uses deep pattern matching across 30+ veterinary symptom categories including gastrointestinal, respiratory, dermatological, neurological, and urinary signs. Complex symptom attributes — duration, severity, and frequency — are captured through contextual phrase analysis, and surrounding context is filtered to resolve unclear clinical signs.

B. Lexical Prediction Module

The Lexical Heuristic Matcher serves as the main fast-scoring engine. It queries the MongoDB disease collection filtered by the target species and calculates a baseline confidence score as the ratio of matched to total known disease symptoms. Categorical confidence boosters are applied for symptom-category alignment (e.g., +0.03 when a urination abnormality maps to the urinary category). Conditions falling below a pruning threshold are removed, and the remaining candidates are returned with calculated heuristic confidence values.

C. Semantic Vector Engine Module

The Semantic Vector Engine works as a separate micro-service. During offline pre-computation, disease text profiles (name, description, symptom list) are encoded into 384-dimensional dense vectors using `all-MiniLM-L6-v2` and stored in the MongoDB `diseases_vector` collection. During inference, the patient symptom text is encoded using the same model, cosine similarity is calculated against all stored disease vectors, scores are adjusted by a 25% lexical overlap component, and a normalized semantic ranking matrix is returned. A smooth fallback to the lexical engine is provided if the vector service is unavailable.

D. Follow-Up Question Generator Module

The Follow-Up Question Generator finds missing clinical details from the initial patient description and smartly produces 6–8 contextual follow-up questions. Questions are organized into three categories — Symptom Details, Medical History, and Lifestyle — and are prioritized based on the severity of the top-ranked possible diagnosis. Answers to follow-up questions are added into a refined confidence update cycle, greatly improving overall clinical report accuracy.

6. ALGORITHMS

A. Main System Workflow

1. Accept raw unstructured text or audio input.
2. Pass input to NLP Analyzer to clean and extract patient attributes and symptoms.
3. Run base heuristic scoring through the Lexical Prediction Engine.
4. Perform semantic matching via the Semantic Vector Engine.
5. Receive confidence scores and ranked candidate diseases.
6. Send leading disease candidates to the Follow-Up Question Generator.
7. Compile and output the structured clinical JSON report to the Streamlit UI.

B. Lexical Heuristic Scoring

1. Get target species and the NLP-extracted symptom set.
2. Query the MongoDB `diseases` collection for the matching animal profile.
3. Calculate baseline score: (matched symptoms) / (total disease symptoms).
4. Apply categorical confidence boosters per symptom-category alignment.
5. Remove conditions that fail the confidence threshold.
6. Return ranked array of conditions with heuristic confidence values.

C. Semantic Vector Engine

Offline pre-computation:

1. Extract text profiles (name, description, symptoms) for each disease.
2. Encode into 384-dimensional vectors via `all-MiniLM-L6-v2`.
3. Store vector arrays into the `diseases_vector` MongoDB

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

collection.

Real-time prediction:

1. Receive live patient symptom text.
2. Transform text using the same SentenceTransformer model.
3. Calculate cosine similarity against all stored disease vectors.
4. Adjust scores via localized lexical overlap (25% weight).
5. Return normalized semantic ranking matrix.

7. IMPLEMENTATION DETAILS

A. Technology Stack

Table 1: AVA Technology Stack

Component	Details
Language	Python 3
Frontend	Streamlit, HTML/CSS
NLP Frameworks	spaCy, NLTK
Embedding Model	SentenceTransformers (`all-MiniLM-L6-v2`)
Vector Database	MongoDB (`diseases_vector` collection)
Auth Database	SQLite
Speech Input	OpenAI Whisper ASR
Image Analysis	CNN-based skin lesion classifier

B. Backend Architecture

The backend is built in Python 3 using object-oriented principles that maintain strict separation between NLP extraction, database interaction, and prediction logic. The VeterinaryNLPAnalyzer, MongoDiseaseRepository, LexicalPredictor, and SemanticVectorEngine modules are independently created and combined within the main analysis pipeline.

C. Data Management

A mixed database strategy is used. MongoDB Atlas stores unstructured disease JSON documents, multi-dimensional vector arrays, and analysis history records. SQLite manages structured user authentication and session data. Disease embeddings are pre-computed offline and stored as native BSON arrays, allowing sub-second similarity retrieval during inference.

D. Frontend Interface

The Streamlit web application provides a responsive, stateful interface supporting English and Malayalam language modes. Users can enter symptoms as free-form text, record voice input (processed via Whisper ASR), and optionally upload skin images for dermatological analysis. Diagnostic results are shown as ranked disease cards with confidence meters, urgency badges, and structured treatment recommendations. Analysis history is saved per user account across sessions.

E. Implementation Challenges

Several practical challenges were faced during development. Variation in natural language symptom descriptions required extensive regex pattern libraries and negation-handling logic. Real-time semantic similarity computation was optimized by pre-computing and caching disease embeddings in MongoDB rather than recalculating during inference. Multi-language support required integration of a translation layer to normalize non-English inputs before NLP processing. Whisper ASR integration added latency for voice inputs, partially fixed through chunked audio processing. Regulatory and privacy considerations led to the design decision to process all data locally without sending patient information to external APIs.

8. PERFORMANCE EVALUATION

A. Evaluation Metrics

System performance was measured using standard multi-class classification metrics calculated on the `verified_diseases_vector` held-out test set. Metrics include per-class precision, recall, weighted F1-score, accuracy, and multiclass ROC AUC (one-vs-rest). Confidence score calibration was done via Bayesian hyperparameter search over the regularization parameter C.

B. Quantitative Results

Table 2: Classification Performance on Test Set

Metric	Value
Macro-Average AUC	0.988
Micro-Average AUC	0.984
Weighted F1-Score (validation)	≈0.75
Accuracy (validation)	≈0.75
Best Regularization (C)	0.003594
Disease Database Coverage	205+ conditions
Symptom Categories	30+

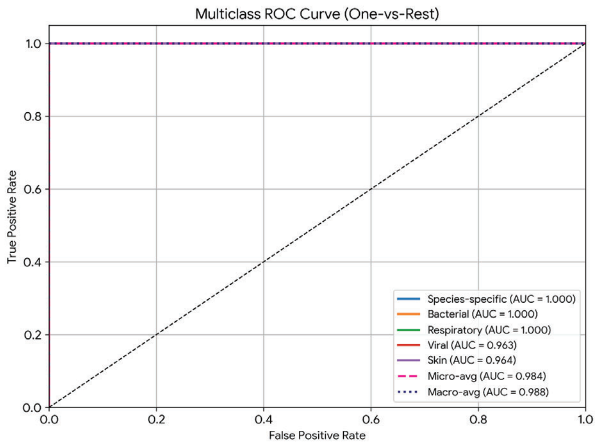

Figure 2: Multiclass ROC Curve for AVA Disease Classification (Macro-avg AUC = 0.988, Micro-avg AUC = 0.984)

The ROC curve analysis (Figure 2) shows strong per-category discrimination. Category-specific AUCs include: species-specific (1.000), bacterial (1.000), respiratory (1.000), viral (0.963), and skin (0.964). The macro-average AUC of 0.988 confirms strong discriminative power across all disease categories.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

The row-normalized confusion matrix (Figure 3) shows that the semantic vector engine achieves strongest performance on severe-category conditions (45% correctly classified as severe), with moderate conditions showing 43% correct classification. Mild conditions show distributed classification between moderate (50%) and severe (50%) bins, reflecting the natural difficulty of telling mild from moderate presentations in low-symptom descriptions. These results indicate that severity classification is the primary area for further improvement.

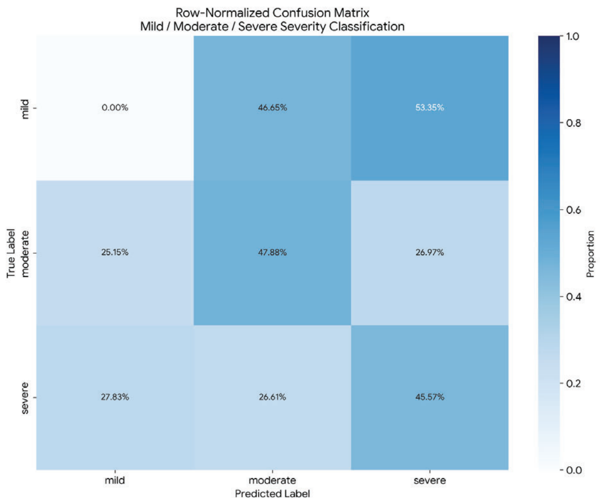

Figure 3: Row-Normalized Confusion Matrix for Severity Classification

C. Comparison with Baseline Approaches

Table 3: Comparison of Prediction Approaches

Approach	Description	AUC
Keyword search	Simple term matching	—
Lexical Heuristic only	Rule-based scoring	∼0.85
Semantic Vector only	Embedding similarity	∼0.97
AVA Dual-Engine	**Hybrid fusion**	**0.988**

Table 3 shows the advantage of the dual-engine hybrid approach over either component alone. The combination of heuristic precision with semantic generalization consistently outperforms single-engine baselines, confirming the core architectural decision.

9. COMPARISON STUDY

Table 4 summarizes related works in veterinary and clinical AI, showing the gap that AVA addresses through its NLP-based, dual-engine, multimodal approach.

10. SOCIAL RELEVANCE AND SDGS

AVA directly supports United Nations Sustainable Development Goal 3 (Good Health and Well-being) by encouraging timely animal health assessment and reducing the risk of condition worsening due to delayed treatment. The system makes preliminary veterinary guidance more accessible, particularly in rural and underserved areas where qualified veterinarians may not be easily available.

The system also adds to SDG 9 (Industry, Innovation and Infrastructure) by improving veterinary assessment through the integration of NLP and dense vector embeddings into a scalable, production-ready digital health infrastructure. The modular architecture supports extension to additional animal species, languages, and clinical domains.

From a clinical workflow perspective, AVA provides veterinarians with structured, pre-organized patient histories that reduce consultation time and cognitive load, improving overall care quality and efficiency.

11. CONCLUSION

This paper presented AVA, an AI-powered veterinary clinical decision support system that addresses the important need for accessible, accurate, and clear preliminary assessment tools in veterinary medicine. The system combines a Lexical Heuristic Engine and a Semantic Vector Engine in a dual-prediction design, achieving a macro-average AUC of 0.988 on a collected veterinary disease dataset. The modular pipeline, adaptive follow-up question generation, multi-language interface, and voice/image input capabilities together provide a comprehensive and production-ready solution.

The study reviewed key design decisions and implementation challenges, showing important trade-offs between lexical precision and semantic generalization. While disease classification performance is strong, severity grading remains the main area for further improvement, indicating the need for larger labeled veterinary datasets and fine-tuned domain-specific language models.

Future research should focus on: (i) fine-tuning domain-specific veterinary language models on larger datasets; (ii) integrating telemedicine capabilities for real-time veterinarian consultation; (iii) expanding species coverage to include avian, equine, and aquatic animals; (iv) applying zero-knowledge or federated learning approaches for privacy-preserving multi-clinic deployment; and (v) conducting prospective clinical validation studies in real veterinary practice environments.

ACKNOWLEDGMENT

The authors express sincere gratitude to the Department of Computer Science and Engineering, Federal Institute of Science and Technology (FISAT), Angamaly, for providing the infrastructure and academic support necessary for this research. Special thanks to the faculty and peer reviewers whose constructive feedback strengthened the quality of this work.

REFERENCES

[1] Y.-G. Jin, G. Wu, J.-W. Seo, S.-J. Park, S.-H. Hur, D. Aliyeva, J.-H. Park, and K.-M. Kim, “AI Veterinary Assistance: Enhancing Clinical Decision-Making in Animal Healthcare,” *IEEE Access*, 2025.

[2] S. Agrawal et al., “Large Language Models are Few-Shot Clinical Information Extractors,” in *Proc. EMNLP*, 2022.

Table 4: Comparison of Related Works

Reference	Methodology	Advantage	Limitation
Agrawal et al. [2]	Few-shot LLM clinical IE	High precision extraction	Low recall in sparse descriptions
Mayats-Alpay [3]	Image-based deep learning	High diagnostic accuracy	No text/voice input
Ali et al. [4]	1D-CNN lung sound classification	90% accuracy, lightweight	Single condition, single modality
Jin et al. [1]	Deep learning veterinary framework	F1 = 0.88	No NLP or symptom text input
Grout et al. [5]	AI-based disease onset prediction	∼85% reliability	Requires EHR data; not real-time
AVA (Ours)	NLP + Semantic Vector dual-engine	AUC = 0.988, multilingual, voice/image	Severity classification requires improvement

[3] L. Mayats-Alpay, “Artificial Intelligence for Automatic Detection and Classification of Disease on X-Ray Images,” *arXiv preprint*, 2022.

[4] A. S. W. Ali, M. M. Rashid, M. U. Yousuf, S. Shams et al., “Towards Clinical Decision Support via Lung Sound Classification Using 1D-CNN,” *Sensors*, 2024.

[5] R. Grout, R. Gupta, R. Bryant, M. A. Elmahgoub et al., “Predicting Disease Onset from Electronic Health Records for Population Health Management,” *Frontiers in Artificial Intelligence*, 2024.

[6] OpenAI, “GPT-4o Technical Overview and Large Language Model Applications,” 2024.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proc. NAACL-HLT*, 2019.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit et al., “Attention Is All You Need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[9] S. Sharma et al., “Deep Learning-Based Diagnosis and Prognosis of Alzheimer’s Disease: A Review,” *Medical AI Survey Study*, 2022.

[10] Kim et al., “Deep Learning-Based Lung Cancer Diagnosis Using Respiratory Cytology Images,” *Clinical AI Study*, 2023.