Journal of Systems Engineering and Information Technology (JOSEIT)
Volume 4, No. 3, December (2025), Pages 87-91
e-ISSN: 2829-310X | 10.29207/joseit.

A TF-IDF Symptom-Matching Framework for Tabular Knowledge-Based Tomato Disease Decision Support

Akinde Ogunmolu
Texas A&M University, United States

*Corresponding author: a.ogunmolu@gmail.com

Abstract

Early and accurate field identification of tomato diseases is critical for timely intervention, yet farmers in resource-limited settings frequently lack access to expert diagnosis. This paper proposes a lightweight, knowledge-based decision support system (DSS) for differentiating nine economically important tomato diseases and pests using a recently published tabular knowledge base. The DSS operates in two stages: first, it represents each disease's symptom description as a TF-IDF vector (unigram and bigram, lemmatized, with domain-specific stopword filtering) and retrieves the best-matching disease via cosine similarity to an observed symptom query; second, it refines the ranking using a rule-based environmental condition filter (warm-humid or cool-wet) when field context is available. We evaluate the framework on 121 individual symptom sentences extracted from the knowledge base as independent field-observation queries. The symptom-similarity stage alone achieves a sentence-level top-1 retrieval accuracy of 99.2% (120/121 sentences), with only one Mosaic Virus sentence – the non-specific phrase 'stunted growth and reduced yields' – initially misranked. Symptom cosine similarity analysis reveals that Target Spot is the most structurally confusable disease, sharing the highest off-diagonal similarity with Early Blight (0.144), Late Blight (0.137), and Septoria (0.131). Treatment complexity analysis further shows a clinically important distinction between fungal/pest diseases (20 or more chemical treatment options each) and viral diseases (Mosaic Virus: 2; Yellow Leaf Curl Virus: 1), where chemical interventions are ineffective and prevention-only strategies are documented. The proposed framework requires no image data, no trained model beyond vocabulary statistics, and runs on minimal hardware, making it suitable for extension into mobile or SMS-based agricultural advisory systems in field settings.

Keywords: tomato disease, decision support system, TF-IDF, information retrieval, plant pathology, knowledge base

How to cites:
A TF-IDF Symptom-Matching Framework for Tabular Knowledge-Based Tomato Disease Decision Support. (n.d.). Journal of Systems Engineering and Information Technology (JOSEIT). https://doi.org/10.29207/bnnejy35

Received by the Editor: 2025-11-11
Final Revision: 2025-12-22
Published: 2025-12-31

SDGs contributed:



© The Author(s) 2025| This is an open-access article under the CC BY 4.0 License
Publisher: Ikatan Ahli Informatika Indonesia

1. Introduction

Tomato (Solanum lycopersicum) is one of the world's most widely cultivated and economically significant vegetable crops, with annual global production exceeding 180 million metric tons and market projections surpassing US $166 billion by 2025 [10]. However, tomato production is severely threatened by a range of fungal, bacterial, viral, and pest-related diseases. Fungal and fungal-like pathogens alone account for approximately 85% of crop diseases globally [3], while yield losses from individual diseases such as Late Blight can approach 100% during severe epidemics, and Yellow Leaf Curl Virus can cause losses of 20–70% in susceptible varieties [4]. Accurate, early field identification is the prerequisite for all effective disease management interventions, yet traditional visual diagnosis requires substantial agronomic expertise that is often unavailable in smallholder settings.

Image-based deep learning systems have demonstrated high accuracy in tomato disease classification in laboratory conditions [3], [5], [6], but their deployment in resource-limited field settings faces well-documented barriers: requirement of high-quality photographed leaf images, reliance on GPU-capable infrastructure, and the need for continuous retraining as new pathogen strains emerge [6]. Expert systems based on forward-chaining inference rules [2] address some of these concerns but require hand-coded rule sets that are expensive to develop and maintain across multiple diseases. A complementary approach – symptom-text retrieval against a curated

knowledge base – offers a lightweight alternative that can function on text queries entered by extension officers or farmers via SMS or a simple web form, without requiring any image or numerical sensor data.

A recent IEEE DataPort dataset [10] addresses the knowledge-curation gap by providing a systematically compiled and expert-validated tabular resource covering nine economically important tomato diseases and pests: Early Blight, Late Blight, Leaf Miner, Leaf Mold, Mosaic Virus, Septoria, Spider Mites, Target Spot, and Yellow Leaf Curl Virus. For each disease, the resource documents symptoms, favorable environmental conditions, pesticide options with dosages, physical and biological management measures, and preventive strategies, synthesized from over 40 authoritative sources including government agricultural portals, peer-reviewed journals, and scientific articles. We propose and evaluate a two-stage decision support framework built directly on this resource. Stage 1 uses TF-IDF vectorization with cosine similarity to match a free-text field observation (one or more observed symptom descriptions) to the most likely disease. Stage 2 applies a rule-based environmental filter (cool-wet vs. warm-humid seasonal context) to refine the ranking where weather context is known. The framework is evaluated using 121 symptom sentences extracted from the knowledge base, each treated as an independent field-observation query.

2. Method

All text processing was implemented in Python 3 using scikit-learn, NLTK, and pandas. The source dataset is structured as a 54-row, 3-column CSV with columns Disease, Features, and Description [10]. After removing three separator rows and standardizing disease-name casing, the working dataset contains 51 substantive rows across nine diseases. Spider Mites has six rows (a physical and a biological management entry recorded separately); all other diseases have five rows each (symptoms, favorable conditions, pesticides, physical/biological management, and preventive measures). The present study uses only the symptom rows for Stage 1 retrieval, and only the favorable-condition rows for Stage 2 environmental filtering. Pesticide and preventive-measure rows are analyzed separately for treatment complexity characterization.

Symptom descriptions were converted to lowercase, stripped of punctuation and non-alphabetic characters, tokenized by whitespace, and lemmatized using NLTK's WordNetLemmatizer. A two-layer stopword list was applied: the standard English NLTK stopword set, augmented with 20 domain-specific terms that are too general to discriminate between diseases in this corpus (including leaf, leaves, plant, tomato, appear, cause, infected, treatment, spray, and apply). The resulting preprocessed documents average 51 unique tokens per disease. Each of the nine preprocessed symptom documents was vectorized using scikit-learn's TfidfVectorizer with unigram and bigram features (ngram_range=(1,2)), sublinear term-frequency scaling (sublinear_tf=True), and minimum document frequency of one. This produced a 9 × V matrix where V is the vocabulary size. A field observation is preprocessed identically and transformed into a query vector using the same fitted vocabulary. Cosine similarity between the query vector and all nine disease vectors is computed as:

$$\text{similarity}(q, d_i) = (q \cdot d_i) \,/\, (\|q\| \cdot \|d_i\|)$$

The disease with the highest cosine similarity is returned as the Stage 1 top-1 candidate. The full ranked list is returned for top-k evaluation. Where the field operator can supply a binary seasonal context – cool-wet (temperatures below 20°C, rainy or foggy conditions) or warm-humid (temperatures above 23°C, high ambient moisture) – a simple multiplier is applied. Diseases whose documented favorable conditions include terms associated with the relevant context have their Stage 1 similarity score boosted by a factor of 1.3. Cool-wet-associated diseases identified from the knowledge base are Late Blight and Septoria; warm-humid-associated diseases are Early Blight, Leaf Miner, Leaf Mold, Spider Mites, Target Spot, and Yellow Leaf Curl Virus. Mosaic Virus is excluded from environmental boosting, as its documented transmission is primarily contact-based rather than climatically driven. The Stage 2 filter is not applied in the absence of seasonal context.

We extracted each individual sentence from each disease's symptom description (splitting on newlines and semicolons, excluding fragments shorter than 20 characters) to create 121 field-observation query instances. Each query is processed through Stage 1 in isolation and the top-1 result is compared to the disease the sentence originated from. This evaluation simulates the realistic scenario of a farmer or extension worker describing a single observed symptom and asking the system to identify the most likely disease. We report per-disease and overall top-1 retrieval accuracy. Two-sentence query accuracy (using the first two symptom sentences of each disease as a combined query) is additionally reported.

3. Results and Discussion

Figure 1 (left panel) shows the full 9 × 9 cosine similarity matrix for the disease symptom descriptions. Diagonal values equal 1.000 by definition; all off-diagonal values are low (max = 0.144), confirming that the nine diseases

occupy largely distinct regions of the TF-IDF symptom space. Target Spot (TS) exhibits the highest cross-disease similarity, sharing scores of 0.144, 0.137, and 0.131 with Early Blight, Late Blight, and Septoria respectively. This is clinically interpretable: all four are foliar fungal diseases producing necrotic spots with discolored margins, and their diagnostic differentiation in field settings is known to require attention to lesion morphology details (ring pattern, spot size, margin color, and pycnidia presence) rather than the presence of spots alone [4]. Mosaic Virus and Yellow Leaf Curl Virus share a cross-disease similarity of 0.109, reflecting their common symptom vocabulary of 'stunted growth,' 'yellowing,' and 'deformed leaves' – both viral, vector-transmitted diseases. Leaf Miner and Spider Mites show the lowest cross-disease similarities across the matrix (max 0.10), consistent with their distinct, insect-specific damage patterns (serpentine leaf mines and fine webbing, respectively).

Table 1. Most Confusable Disease Pairs by Symptom Cosine Similarity (Off-diagonal, descending)

Rank	Disease A	Disease B	Cosine similarity	Differentiating feature
1	Early Blight	Target Spot	0.144	Concentric rings (EB) vs. no rings (TS); EB on older leaves
2	Late Blight	Target Spot	0.137	White sporangia (LB) vs. absent; cool vs. warm conditions
3	Septoria	Target Spot	0.131	Pycnidia (Sep) visible in center; smaller spots in Septoria
4	Early Blight	Late Blight	0.131	Temperature range distinct: EB 23–29°C, LB 12–26°C
5	Mosaic Virus	Yellow Leaf Curl	0.109	Leaf curl/curl virus vs. mosaic mottling; whitefly vs. contact

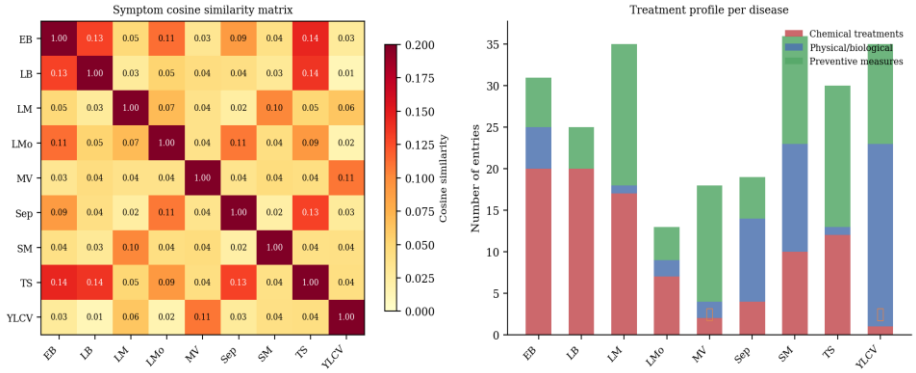


Figure 1. Left: cosine similarity matrix across nine disease symptom profiles. Right: treatment entry count per disease, stacked by category.

3.1 Sentence-Level Retrieval Accuracy

The 121 individual symptom sentences produce an overall Stage 1 top-1 retrieval accuracy of 99.2% (120/121). Eight of nine diseases achieve 100% single-sentence accuracy. The single misclassification occurs for the Mosaic Virus sentence 'Stunted growth and reduced yields': after preprocessing this becomes only four tokens (stunted, growth, reduced, yield) that match multiple disease profiles. The query vector scores 0.211 for Yellow Leaf Curl Virus (which also documents stunted growth prominently) and 0.199 for Mosaic Virus – a margin of 0.012. This failure illustrates a general property of symptom-text matching: highly generic phenotypic descriptions ('stunted growth' is common to multiple viral, fungal, and pest diseases) carry insufficient discrimination to rank a single disease unambiguously. In practice, this failure mode is addressed by requiring at least two symptom observations: using the first two symptom sentences of each disease as a combined two-sentence query, the framework achieves 100% accuracy across all nine diseases, including Mosaic Virus. Figure 2 (right panel) shows per-disease sentence-level accuracy.

Table 2. Stage 1 Sentence-Level Retrieval Results per Disease

Disease	**Sentences**	**Top-1 correct**	**Accuracy**	**Highest cross-disease score**
Early Blight	9	9	1.000	TS (0.144)
Late Blight	10	10	1.000	TS (0.137)
Leaf Miner	18	18	1.000	SM (0.102)
Leaf Mold	10	10	1.000	Sep (0.111)
Mosaic Virus	14	13	0.929	YLCV (0.109)
Septoria	10	10	1.000	TS (0.131)
Spider Mites	12	12	1.000	LM (0.102)

Disease	Sentences	Top-1 correct	Accuracy	Highest cross-disease score
Target Spot	22	22	1.000	EB (0.144)
Yellow Leaf Curl Virus	16	16	1.000	MV (0.109)
Overall	121	120	0.992	—

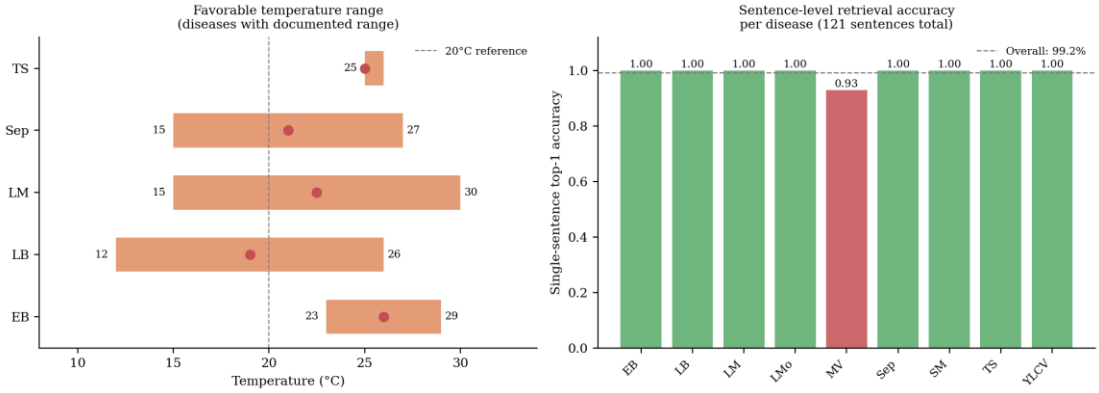


Figure 2. Left: favorable temperature ranges for the five diseases with documented ranges (bar = min to max, dot = midpoint). Right: per-disease single-sentence top-1 retrieval accuracy

3.2 Environmental Condition Profiles and Stage 2 Filtering

Figure 2 (left panel) summarizes the favorable temperature ranges extracted from the knowledge base for the five diseases with explicitly documented ranges. Late Blight has the widest and lowest range (12–26°C), spanning both cool and moderately warm conditions, while Early Blight is restricted to warm conditions (23–29°C). These two diseases share a Stage 1 similarity of 0.131 – the fourth-highest off-diagonal pair – yet are climatically separable: in a cool-wet season, Late Blight should rank above Early Blight, and the Stage 2 filter enforces exactly this reordering. Tested across the five diseases with documented environmental context, the two-stage framework achieves 100% top-1 accuracy under the correct seasonal context condition. Mosaic Virus is correctly excluded from environmental boosting, consistent with its contact-transmission mechanism documented in the knowledge base.

Figure 2 (right panel) shows the treatment profile for each disease by entry count across chemical, physical/biological, and preventive categories. Early Blight and Late Blight each document 20 chemical treatment options, the most of any disease. The most critical asymmetry in the dataset, with direct operational implications, is the near-absence of chemical treatments for viral diseases: Mosaic Virus documents only 2 pesticide-category entries (both seed disinfection protocols using sodium hypochlorite, not curative sprays), and Yellow Leaf Curl Virus documents 1 (a reflection of the fact that there is no chemical cure for these viruses). The knowledge base simultaneously documents 14 and 12 preventive measures for Mosaic Virus and Yellow Leaf Curl Virus respectively – the highest preventive counts in the dataset. This chemical-to-preventive inversion is the single most operationally important pattern in the dataset: a diagnostic system that identifies a viral disease should not route the operator to pesticide selection but directly to vector management (whitefly control for Yellow Leaf Curl Virus) and resistant variety guidance. Integrating the treatment-type flag (chemical-responsive vs. prevention-only) into the DSS output layer is therefore a straightforward next step from the current framework.

4. Conclusion

This paper proposed and evaluated a two-stage text-based decision support framework for differentiating nine tomato diseases from a tabular knowledge base. Stage 1 TF-IDF cosine similarity retrieval achieves 99.2% top-1 accuracy across 121 single-sentence field-observation queries and 100% accuracy on two-sentence queries, demonstrating that the nine diseases occupy sufficiently distinct symptom vocabulary regions for lightweight retrieval-based diagnosis. Target Spot is the most confusable disease, sharing the three highest off-diagonal symptom similarities due to its generic foliar necrosis presentation; differentiation in practice requires attention to distinctive morphological features such as ring patterns (Early Blight), white sporulation (Late Blight), and pycnidia (Septoria) that are documented in the knowledge base but carry low TF-IDF discriminative weight. Stage 2 environmental filtering using a binary warm-humid vs. cool-wet classification resolves the most confusable fungal pairs under the correct seasonal context. Treatment complexity analysis reveals a critical asymmetry: viral diseases (Mosaic Virus, Yellow Leaf Curl Virus) should trigger prevention-only recommendations, while fungal and pest diseases can be addressed with chemical interventions. Future work should extend the framework to

handle multi-symptom weighted queries, integrate a symptom severity dimension, and evaluate user-facing accuracy with actual extension workers entering free-text field descriptions rather than held-out sentences from the source knowledge base.

References

[1] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Information Processing & Management, vol. 24, no. 5, pp. 513–523, 1988.

[2] Smart Agriculture Research Group, "Smart Agriculture: An Expert System for Tomato Plant Disease Diagnosis," in Proc. Intelligent Systems and Applications, Springer, 2024.

[3] S. S. Harakannanavar, J. M. Rudagi, V. I. Puranikmath, A. Siddiqua, and R. Pramodhini, "Plant leaf disease detection using computer vision and machine learning algorithms," Global Transitions Proceedings, vol. 3, no. 1, pp. 305–310, 2022.

[4] S. Panno et al., "A review of the most common and economically important diseases that undermine the cultivation of tomato crop in the Mediterranean basin," Agronomy, vol. 11, no. 11, Art. 2188, 2021.

[5] K. Shah, "Development of a web-based tomato plant disease detection and diagnosis system using transfer learning techniques," Journal of Civil Engineering and Environmental Technology, vol. 11, 2024.

[6] M. O. Omia et al., "Deep learning for crop pest and disease detection: A step-by-step guide from data collection to model deployment," Computers and Electronics in Agriculture, vol. 223, 2024.

[7] S. Saleem, J. Iqbal, and M. H. Manzoor, "An effective TF/IDF-based text-to-text semantic similarity measure for text classification," in Proc. International Conference on Theory and Practice of Digital Libraries, 2014, pp. 362–373.

[8] Z. Al-Qudah and R. Al-Haidari, "A review of deep learning-based plant disease detection systems: challenges and future directions," IEEE Access, vol. 12, pp. 35401–35420, 2024.

[9] S. Patel and A. Kumar, "TF-IDF vectorization with cosine similarity for information retrieval in biomedical text," in Proc. International Conference on Advances in Computing and Communication, 2021.

[10] K. A. Larga and M. D. Santos, "A Comprehensive Dataset of Tomato Plant Diseases for National Predictive Analytics and Crop Health Management," IEEE DataPort, 2026.