**Task 4: Critical Analysis**

**1. TF-IDF vs. One-Hot Encoding: Performance Discussion**

**Why TF-IDF Might Outperform One-Hot Encoding:**

* **Semantic Weighting**:  
  TF-IDF gives more importance to words that are frequent in a particular document but rare across others. This weighting can help differentiate subtle differences between diseases better than a simple presence/absence in one-hot.
* **Dimensional Density**:  
  TF-IDF vectors are usually denser and better capture textual nuance, while one-hot vectors are highly sparse and binary — which can be limiting when distinguishing between diseases that share symptoms.
* **Redundancy Reduction**:  
  One-hot encoding may treat similar words as completely unrelated (e.g., “fever” vs. “fevers”), while TF-IDF (especially when combined with preprocessing like stemming) helps mitigate this.

**When One-Hot Might Perform Better:**

* If the vocabulary is small and very specific (e.g., highly standardized symptom checklists), one-hot might avoid overfitting compared to TF-IDF's more sensitive weights.
* One-hot can be more interpretable in clinical rules or rule-based systems.

**2. Clinical Relevance of the Results**

* The **best model** (KNN with TF-IDF and cosine similarity, k=7) achieved **64% accuracy and 49.4% F1-score**, significantly outperforming other configurations.
* This suggests that **TF-IDF is capturing clinically meaningful symptom groupings** better than raw one-hot encoding.
* While we categorized each disease into 4 broader classes manually, the higher F1-scores from TF-IDF indicate that its representation **more closely aligns with these real-world clinical groupings**.
* **Cosine similarity**—which measures vector orientation rather than distance—worked best, likely because diseases can share sets of symptoms in varying degrees (angle of symptom distribution matters more than absolute presence/absence or magnitude).

**3. Limitations of Both Encoding Methods**

**One-Hot Encoding:**

* **Lacks Frequency Information**:  
  It only records presence/absence, ignoring how often a symptom appears in a description.
* **High Dimensionality**:  
  Can lead to very sparse vectors, especially in symptom-rich datasets, which weakens performance in distance-based models like KNN.
* **No Context Awareness**:  
  Treats all symptoms as equally important, regardless of rarity or specificity.

**TF-IDF Encoding:**

* **Still Bag-of-Words Based**:  
  It doesn’t capture word order, negations, or syntactic context (e.g., “no fever” is treated the same as “fever”).
* **Susceptible to Noise**:  
  If the text includes irrelevant or common non-diagnostic words, TF-IDF can be misled unless stop-word removal is applied.
* **Assumes Text Availability**:  
  TF-IDF works only when textual descriptions are available; doesn’t help with purely structured symptom data.

**Conclusion**

TF-IDF, particularly with cosine similarity, proves to be a better representation of symptom profiles for disease categorization in this dataset. Its performance edge shows that term relevance and frequency are crucial in distinguishing clinically meaningful patterns. However, both methods have limitations—future work could explore **context-aware encodings** (e.g., BERT embeddings) or hybrid models combining **structured data + text** for improved clinical decision support.