Predicting Human Diseases from Symptom Profiles Using Multi-Class Classification Models

Aabu Yousuf Raj, Md Rakibul Hasan
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
aabu.yousuf.raj@g.bracu.ac.bd, md.rakibul.hasan5@g.bracu.ac.bd

***Abstract*—This study evaluates automated multi-class disease prediction for humans using a public-domain dataset of 4,000 symptom-profile records across three disease classes (Typhoid, Diabetes, Gonorrhea) with six symptom-based feature groups. Eight supervised classifiers were assessed under two encoding strategies: label encoding for tree-based models and one hot encoding for distance-based models. Feature relevance was determined using two chi-square methods: an interpretable contingency-table approach and scikit-learn's SelectKBest(chi2). We compared performance across three scenarios: (a) full-feature training, (b) top-3 features from SelectKBest, and (c) top-3 features augmented with the most discriminative feature from the contingency test (Discharge). Full-feature models achieved high accuracy (LR: 0.96, GBM: 0.95), while top-3 feature models showed reduced performance (accuracy: 0.71–0.76). Including Discharge in the top-4 set restored accuracy to near full-feature levels (0.94–0.96). Logistic Regression and GBM are recommended for field deployment due to their robust performance.**

***Index Terms*—human disease prediction, symptom profiles, machine learning, multi-class classification, feature selection, clinical decision support**

I. Introduction

Early and accurate detection of human diseases is crucial for improving patient outcomes and reducing complications, especially in resource-constrained settings. Traditional diagnosis often depends on clinical examinations and laboratory tests, which can be time-consuming, costly, and not always readily available in rural or low-resource areas. Data-driven approaches using supervised machine learning offer a practical support tool by predicting likely disease labels from patient-reported symptom profiles, helping clinicians and health workers prioritize cases, guide timely referrals, and initiate early interventions. This study develops a reproducible machine learning pipeline for predicting human diseases from categorical symptom profiles, focusing on: (1) systematic preprocessing and encoding, treating ”nil” entries as informative; (2) chi-square based feature selection to identify compact symptom subsets; (3) evaluation of eight classifiers (tree-based and distance based) on full and reduced feature sets; and (4) evidence-based recommendations for model and feature strategies tailored for low-resource veterinary diagnostics. Our contributions include: a tailored preprocessing strategy for categorical symptom data, dual chi-square feature-ranking methods for complementary insights, an empirical comparison showing Logistic Regression and Gradient Boosting Machines excel with full features, and methodological recommendations to avoid feature-selection leakage and ensure reproducibility. This pipeline supports efficient, interpretable diagnostic assistants for clinical use.

II. Related Work (Literature Review)

Machine learning has been increasingly adopted for human clinical decision support, including symptom-based disease prediction and digital triage (online symptom checkers). In these systems, patients enter reported symptoms (often categorical and incomplete), and the model outputs a likely diagnosis (or shortlist) and/or a recommended care level. Recent systematic reviews consistently report that symptom checkers show high variability in both diagnostic and triage accuracy across tools and conditions, and they emphasize the need for rigorous, transparent evaluation before real-world use [1], [3]. Large-scale evaluations using standardized case vignettes further demonstrate that triage performance differs widely between apps and has not consistently improved over time, reinforcing that model design, symptom representation, and evaluation methodology strongly influence safety and utility [2].

Beyond symptom checkers, many studies formulate symptom-to-disease mapping as a supervised classification problem on structured/tabular data, where classical models such as Logistic Regression, SVM, kNN, Decision Trees, Random Forests, and gradient-boosting ensembles remain competitive and practical for clinical datasets [4], [5]. Reviews of machine learning in disease diagnosis highlight that tree ensembles and boosting methods often perform strongly on tabular clinical inputs because they capture nonlinear symptom interactions, while simpler linear baselines (e.g., Logistic Regression) can provide robustness and interpretability when features are well-encoded [4], [6].

A recurring challenge in symptom-based prediction is feature selection and questionnaire length. Since real users may not complete long symptom forms, many works aim to identify compact symptom subsets that preserve predictive power, commonly using univariate statistical tests (e.g., chi-square for categorical features) or model-based importance ranking [4]. However, methodological guidance stresses that feature selection is a form of preprocessing and must be performed using training data only (or within each cross-validation fold) to avoid data leakage and overly optimistic performance estimates [7], [8]. In addition, healthcare modeling guidance recommends robust validation practices (e.g., careful cross-validation design and reporting) because evaluation choices can significantly affect measured performance and clinical reliability [8].

Motivated by these findings, our work focuses on a reproducible symptom-based pipeline for multi-class human disease prediction, emphasizing careful handling of categorical symptom profiles (including explicit "Nil/None" values), chi-square-based feature analysis, and systematic comparison of multiple classifiers under both full-feature and reduced-feature and settings.

III. Dataset and Methodology

A. Dataset Overview

We utilized a CC0 Public Domain symptom–disease dataset containing 4,000 records with seven columns:

- **Discharge**: categorical symptoms related to bodily discharge
- **Feelings and Urge**: symptoms such as fatigue, hunger, and fever
- **Pain and Infection**: symptoms such as blurred vision and infections
- **Physical Conditions**: physical symptoms such as bloody diarrhea and rashes
- **Critical Feelings**: critical symptoms including seizures and confusion
- **Critical**: binary indicator (*Critical* / *Not Critical*)
- **Disease**: target variable {Typhoid, Diabetes, Gonorrhea}

The dataset contained no null values. Explicit "Nil" values were retained because they can carry informative clinical signals. No significant class imbalance was observed; a minor imbalance for Gonorrhea was handled via per-class evaluation metrics.

B. Preprocessing and Encoding

Two encoding strategies were employed based on model requirements:

- **Label Encoding**: for tree-based models (Decision Tree, Random Forest, GBM, XGBoost)
- **One-Hot Encoding**: for distance/linear models (KNN, SVM, Logistic Regression, MLPClassifier)

Feature correlation analysis revealed no high linear correlation between columns, eliminating the need for redundancy removal.

C. Chi-Square Feature Selection

Two complementary chi-square methods were implemented.

Method 1: Contingency Table Analysis. For each categorical feature, contingency tables against *Disease* were constructed using `scipy.stats.chi2_contingency`. The results were:

- **Discharge**: $\chi^2 = 2268.59$, $p < 0.001$
- **Feelings and Urge**: $\chi^2 = 1498.49$, $p < 0.001$
- **Physical Conditions**: $\chi^2 = 1493.13$, $p < 0.001$
- **Critical Feelings**: $\chi^2 = 806.85$, $p < 0.001$
- **Critical**: $\chi^2 = 21.53$, $p < 0.001$
- **Pain and Infection**: $\chi^2 = 0.40$, $p = 1.000$ (not significant)

TABLE I
Full-Feature Model Performance

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.96	0.96	0.95	0.96
GBM	0.95	0.95	0.94	0.95
SVM	0.95	0.95	0.94	0.95
Decision Tree	0.94	0.94	0.94	0.94
Random Forest	0.94	0.94	0.94	0.94
XGBoost	0.94	0.94	0.94	0.94
MLPClassifier	0.94	0.94	0.94	0.94
KNN	0.92	0.92	0.91	0.92

Method 2: SelectKBest Analysis. We applied `SelectKBest`(χ^2) to the encoded features, yielding the following ranking:

1) **Feelings and Urge**: Score = 746.76
2) **Physical Conditions**: Score = 263.89
3) **Critical Feelings**: Score = 170.47
4) **Discharge**: Score = 124.92
5) **Critical**: Score = 20.01
6) **Pain and Infection**: Score = 0.0004

Both methods agreed that *Pain and Infection* lacks discriminative power. Differences in the ranking of other features are expected due to encoding effects and feature cardinality.

D. Experimental Design

Models evaluated included Decision Tree, Random Forest, GBM, XGBoost, KNN, SVM, Logistic Regression, and MLPClassifier. Performance was assessed using accuracy, precision, recall, and F1-score across three feature scenarios:

- **Full-feature training**: all six symptom features
- **Top-3 features**: selected by `SelectKBest`(χ^2)
- **Top-4 features**: Top-3 plus *Discharge*

IV. Results

A. Full-Feature Results

Table I presents the performance of all models using the complete feature set.

Logistic Regression achieved the highest performance (96% accuracy), followed by GBM and SVM at 95%. This indicates that both linear and ensemble-based models effectively capture diagnostic patterns from the symptom-profile features.

B. Reduced-Feature Results

Table II summarizes model performance under reduced feature settings (Top-3 vs. Top-4 features).

TABLE II
REDUCED-FEATURE MODEL PERFORMANCE (TOP-3 VS. TOP-4 FEATURES)

Model	Top-3 Features		Top-4 Features	
	Accuracy	F1-score	Accuracy	F1-score
Logistic Regression	0.76	0.74	0.94	0.93
GBM	0.76	0.74	0.95	0.94
Random Forest	0.75	0.74	0.94	0.93
KNN	0.71	0.71	0.88	0.87

The Top-3 feature set resulted in substantial performance degradation (71–76% accuracy), particularly affecting class-wise recall. However, adding *Discharge* to form the Top-4 set restored performance to near full-feature levels (88–95% accuracy), validating the contingency analysis findings reported in Section III-C.

V. DISCUSSION

A. Feature Selection Insights

The dramatic performance recovery when adding Discharge to the Top-3 features highlights the importance of combining multiple feature selection methods. The contingency table approach captured Discharge's high discriminative power through its direct relationship with disease classes, while SelectKBest's ranking was influenced by encoding effects and category cardinalities.

B. Methodological Considerations

Several important methodological concerns emerged: Feature Selection Leakage: Chi-square selection must be computed only on training data to avoid information leakage. This study recommends training-fold-restricted feature selection for deployment scenarios.

Evaluation Consistency: Different test set sizes across experiments (1,429 vs. 953 samples) suggest inconsistent split ting, which complicates direct comparisons. Stratified cross validation is recommended for robust evaluation.

Model Calibration: For clinical decision-making, probability calibration should be assessed, particularly for Logistic Regression and GBM models.

C. Practical Implications (Human Context)

The Top-4 feature set (Feelings and Urge, Physical Conditions, Critical Feelings, Discharge) offers a practical balance between diagnostic accuracy and data collection burden. This minimal viable feature set could facilitate deployment in resource-constrained settings while maintaining high predictive performance. This should be positioned as a decision-support tool for preliminary screening or triage, not as a substitute for clinical diagnosis.

VI. LIMITATIONS AND FUTURE WORK

The study's limitations include the moderate dataset size restricting generalization across diverse herds, use of default hyperparameters potentially underestimating model capacity, reliance on univariate feature selection missing multivariate interactions, and inconsistent test partitions complicating comparisons. Future work should focus on implementing nested cross validation for leakage-free evaluation, hyperparameter optimization through systematic search, SHAP analysis for clinical interpretability, and validation on external datasets through field trials. Additionally, investigating cost-sensitive learning for asymmetric misclassification costs and ensemble methods could further improve system effectiveness.

VII. CONCLUSION

This study developed a robust machine learning pipeline for predicting human diseases from categorical symptom profiles. Logistic Regression achieved the highest full-feature accuracy (96%), with GBM following at 95%. The dual chi-square approach revealed that while Pain and Infection lacks predictive value, Discharge provides crucial discriminative information often overlooked by standard feature selection methods.

The Top-4 feature subset (including Discharge) restored near-optimal performance while reducing data collection requirements, making it suitable for low-resource veterinary applications. The pipeline, emphasizing Logistic Regression or GBM with proper feature selection methodology, provides an efficient and interpretable diagnostic tool that balances accuracy with practical deployment considerations.

For successful field deployment, the system requires proper validation protocols, probability calibration, and integration with existing veterinary workflows to support improved animal welfare and agricultural productivity.

REFERENCES

[1] W. Wallace, E. Chan, K. Chidambaram, M. Hanna, and A. Iqbal, "The diagnostic and triage accuracy of digital and online symptom checker tools: a systematic review," *npj Digital Medicine*, vol. 5, 2022.

[2] M. L. Schmieding, S. M. M. Kopka, M. Schmidt, and T. Rieger, "Triage accuracy of symptom checker apps: Systematic review and benchmarking," *JMIR mHealth and uHealth*, 2022.

[3] E. Riboli-Sasco, A. Iftikhar, and colleagues, "Triage and diagnostic accuracy of online symptom checkers: a systematic review," *Journal of Medical Internet Research (JMIR)*, 2023.

[4] M. M. Ahsan, K. D. Gupta, and colleagues, "Machine-learning-based disease diagnosis: A comprehensive review," 2022.

[5] N. H. Alhumaidi and colleagues, "The use of machine learning for analyzing real-world data in disease prediction and management: systematic review," 2025.

[6] A. Rațiu and colleagues, "Machine learning in clinical decision making," *Applied Sciences*, 2026.

[7] scikit-learn Developers, "Common pitfalls and recommended practices (Data leakage; feature selection on training data only)," scikit-learn documentation.

[8] D. Wilimitis and colleagues, "Practical considerations and applied examples of cross-validation in health care predictive modeling," 2023.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. VanderPlas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[10] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[11] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.

[12] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD '16)*, 2016, doi: 10.1145/2939672.2939785.

[13] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995, doi: 10.1007/BF00994018.

[14] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967, doi: 10.1109/TIT.1967.1053964.

[15] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[16] M. B. Kursa and W. R. Rudnicki, "Feature selection with the Boruta package," *Journal of Statistical Software*, vol. 36, no. 11, pp. 1–13, 2010, doi: 10.18637/jss.v036.i11.

[17] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Wiley, 2013.

[18] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986, doi: 10.1038/323533a0.

[19] scikit-learn Developers, "MLPClassifier — scikit-learn documentation," scikit-learn API reference, accessed 2026.