Intelligent Classification and Identification of Similar Respiratory System Diseases

Zhuoyang Liu

Information Systems, University of New South Wales, Sydney, Australia
Z5471767@ad.unsw.edu.au

Abstract. This study focuses on symptom classification models for the four most common respiratory diseases (COVID, FLU, COLD, and ALLERGY). The aim is to address the challenge of distinguishing between similar and troublesome respiratory illnesses while maintaining accuracy and minimizing unnecessary time wastage. Based on the project's publicly available dataset, the models underwent standardization, cleaning, type unification, severity encoding, hierarchical segmentation, and standardized preprocessing. Three models were compared: Bernoulli Naive Bayes (an interpretable baseline), RBF-SVM (a traditional strong baseline), and MLP. In cases of imbalanced datasets, probability calibration and class threshold scanning were employed to find a balance between precision and recall. A macro-averaging metric was used to avoid the majority class "drowning" effect. In the overall output, RBF-SVM showed stable performance in macro F1 and macro AUC; after calibration and threshold adjustment, MLP improved the precision of easily confused classes while maintaining overall accuracy. The main confusion and low precision in the models stemmed from the significant overlap of features between COVID and COLD. This article provides a method for solving related problems using neural networks. The accuracy of these models cannot be fully trusted and requires subsequent manual verification to ensure accuracy.

Keywords: Respiratory Symptoms, Class Non-uniformity, Probability Calibration, Feature Overlap.

1 Introduction

Respiratory diseases are among the most common illnesses in daily life [1, 2]. Their numerous types and highly overlapping symptoms make differentiation difficult without relevant knowledge [3, 4, 5]. This high degree of overlap in symptoms and features poses a challenge to early screening [6, 7]. The accuracy of minority classes and classes with similar symptoms is often masked by the overall picture. To address these issues, this study used different models and employed a core workflow of unified preprocessing → transparent baseline → strong baseline → small neural network. Probability calibration and classification thresholds were subsequently used to achieve a balance between high precision and recall.

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_68

The project aims to build a classification model that can distinguish common respiratory diseases over four weeks, making it easier for people to identify their disease type and decide whether to seek medical attention. Eliminating the need for appointments saves time for those in cities with limited medical facilities, as most respiratory diseases are not severe and can be managed with medication.

The dataset selection and processing were consistent across all models. The project used (dataset name), performed column cleaning, standardized data types, and encoded feature severity. A 70/15/15 split and class weights were used for stratification to reduce data imbalance.

The project initially chose Bernoulli NB to verify separability, RBF-SVM as a strong nonlinear baseline, and small MLPs (ReLU, Dropout, early stopping) to provide end-to-end capabilities and visualization; the key was calibration + threshold scanning to adapt to different business costs. The entire project visualized every key node and output, making it easier to observe and the obvious data fluctuations made subsequent changes easier.

2 Dataset Description and Processing

The sample data used in this study came from the publicly released respiratory symptoms and disease types dataset (COVID/FLU/COLD/ALLERGY symptoms dataset, publicly anonymized data, used only for research and teaching reproduction) on the Kaggle platform. The data consisted entirely of patient-reported symptoms and signs, which were statistically analyzed and denoised before being uploaded to Kaggle. The dataset contained 44,453 entries, with 21 fields. One field was designated as the type, categorizing the data into four classes (COVID/FLU/COLD/ALLERGY). The other 20 fields represented signs and conditions. The data collection period was not concentrated across different seasons, and the uneven distribution of disease data increased the difficulty of subsequent identification.

The dataset mainly includes the following feature fields: COUGH, MUSCLE_ACHES, TIREDNESS, SNEEZING, PINK_EYE, SORE_THROAT, RUNNY_NOSE, HEADACHE, CHEST_PAIN, BREATHING_DIFFICULTY, LOSS_OF_TASTE/LOSS_OF_SMELL (abnormal taste/smell), NAUSEA/VOMITING/DIARRHEA (nausea/vomiting/diarrhea), etc.

To avoid overlapping studies of some features, some fields are graded. This study uses 0/1/2/3 to represent different degrees. In specific implementation, Yes/No is uniformly mapped to 1/0; if Mild/Moderate/Severe levels exist, they are mapped to 1/2/3. All field names are uniformly padded with leading and trailing whitespace. The dataset classification employed stratified sampling to maintain a relatively consistent label ratio across subsets, typically 70%/15%/15%, with 15% for the training set and 15% for the test set. In the 15% test set (6668 data points), ALLERGY=2,457, COLD=154, COVID=307, FLU=3,750. This highly imbalanced data distribution severely impacted accuracy and recall, a problem that needs to be addressed in subsequent model development.

In addition, this paper performed standard processing on the dataset, including column cleaning, type standardization, missing value imputation, and numerical scaling, to ensure dataset consistency and usability.

3 Model

Model and The overall models used in this study were completed and compared under the same preprocessing pipeline. Naive Bayes (NB/BNB) was used as a transparent baseline, kernel support vector machine (RBF-SVM) as a traditional strong baseline, and a multilayer perceptron (MLP) with dropout and early stopping mechanisms was used as the main neural network model.

The output of the neural network model and SVM was calibrated using CalibratedClassifierCV to make subsequent model adjustments, such as threshold changes, easier. The MLP structure uses a two-layer ReLU fully connected configuration with Dropout, with Dropout values of 0.3/0.2 for the 256-128 hidden layers, and a 4-dimensional softmax output layer. The model optimizer used was Adam, with a batch size of 256 and a maximum training iteration of 40. Overfitting and stable convergence were addressed using EarlyStopping(monitor="val_loss", patience=4, restore_best_weights=True) and ReduceLROnPlateau (monitor="val_loss", factor=0.5, patience=2). RBF-SVM Hyperparameter Selection: 5-fold hierarchical cross-validation with grid search targeting macro-F1; candidate set $C \in \{0.1, 0.5, 1, 2\}$, $\gamma \in \{\text{"scale"}, 0.01, 0.05, 0.1\}$. $C = 0.5$ and γ = "scale" were ultimately selected, and the system was retrained on the full training set. BernoulliNB (BNB): Laplace smoothing coefficient $\alpha \in \{0.1, 0.5, 1\}$, retrained after selecting the optimal αon the validation set. MLP: Two layers of ReLU fully connected (256–128) + Dropout (0.3/0.2), output layer 4-way Softmax; Adam, batch size 256, maximum 40 rounds; EarlyStopping(monitor="val_loss", patience=4, restore_best_weights=True) and ReduceLROnPlateau(monitor="val_loss", factor=0.5, patience=2) control overfitting and convergence stability. Probability calibration and threshold post-processing: CalibratedClassifierCV is used to perform posterior probability calibration on SVM and MLP, and class thresholds are scanned on the validation set to adjust Precision/Recall under different business costs. The benefits of cost-sensitive/order class loss or mixed loss in imbalanced tasks can also be considered [7, 8]. The output of the neural network model and SVM was calibrated using CalibratedClassifierCV to make subsequent model adjustments, such as threshold changes, easier. The MLP structure uses a two-layer ReLU fully connected configuration with Dropout, with Dropout values of 0.3/0.2 for the 256-128 hidden layers, and a 4-dimensional softmax output layer. The model optimizer used was Adam, with a batch size of 256 and a maximum training iteration of 40. Overfitting and stable convergence were addressed using EarlyStopping (monitor="val_loss", patience=4, restore_best_weights=True) and ReduceLROnPlateau (monitor="val_loss", factor=0.5, patience=2). RBF-SVM Hyperparameter Selection: 5-fold hierarchical cross-validation with grid search targeting macro-F1; candidate set $C \in \{0.1, 0.5, 1, 2\}$, $\gamma \in \{\text{"scale"}, 0.01, 0.05, 0.1\}$. $C = 0.5$ and γ = "scale" were ultimately selected, and the system was retrained on the

full training set. BernoulliNB (BNB): Laplace smoothing coefficient $\alpha \in \{0.1, 0.5, 1\}$, retrained after selecting the optimal α on the validation set. MLP: Two layers of ReLU fully connected (256–128) + Dropout (0.3/0.2), output layer 4-way Softmax; Adam, batch size 256, maximum 40 rounds; EarlyStopping(monitor="val_loss", patience=4, restore_best_weights=True) and ReduceLROnPlateau(monitor="val_loss", factor=0.5, patience=2) control overfitting and convergence stability. Probability calibration and threshold post-processing: CalibratedClassifierCV is used to perform posterior probability calibration on SVM and MLP, and class thresholds are scanned on the validation set to adjust Precision/Recall under different business costs. The benefits of cost-sensitive/order class loss or mixed loss in imbalanced tasks can also be considered [9, 10].

4 Experimental

4.1 Experimental Setup

The three models used throughout the experiment were compared in the same environment to ensure feasibility: Python 3.11, scikit-learn 1.5+, TensorFlow/Keras 2.12+, and Matplotlib/Seaborn. All random sources were seeded at 42.

To mitigate class imbalance, class weighted classifiers (such as RBF-SVM) were uniformly set to class_weight="balanced". Macro-average metrics (macro-Precision/Recall/F1, macro-AUC, macro ROC) were used in reporting and threshold selection to avoid systematic bias caused by majority class dominance. Regarding the potential impact of metric weighting and different metric choices on the conclusions, this paper further points out their potential biases and trade-offs in the discussion.

4.2 Experimental Results and Analysis

Accuracy, macro-Precision/Recall/F1, and macro-AUC were measured during the runtime phase, and the results were visualized. The visualization produced confusion matrices in different colors and one-to-many ROC curves (4 class curves and a single macro curve), as well as the learning curve of the MLP. After probability calibration and class threshold adjustment, the MLP improved the precision of easily confused classes without changing the overall accuracy, demonstrating a feasible trade-off path for calibration awareness.

Table 1. Overall performance on the test set

Model	Accuracy	Macro-Precision	Macro-Recall	Macro-F1	Macro-AUC
Naive Bayes (BNB, Transparent Baseline)	0.88–0.90	0.62–0.68	0.72–0.78	0.66–0.72	0.95–0.96
RBF-SVM	0.9331	0.7521	0.9621	0.8171	0.9863
MLP（Baseline）	0.9273	0.7325	0.8870	0.7835	≈0.99
MLP(Threshold post-processing)	0.9273	0.7424	0.9575	0.8066	≈0.99

Among all the initial results, SVM-RBF, as shown in Table 1 and Figure 1 , performed best in terms of macroscopic capability and stability (Accuracy≈0.933, macro-Precision≈0.752, macro-Recall≈0.962, macro-F1≈0.817, macro-AUC≈0.986). The MLP baseline was generally close to SVM (Accuracy≈0.927, macro-F1≈0.784, macro-AUC≈0.99). The study found that adjusting the classification threshold in the neural network significantly improved accuracy. However, it was discovered that excessively high thresholds could cause a small number of data points to become null values, and some data with a precision of 1 after a higher threshold might lose their reference value. As shown in Figure 2, after all selections are completed, the overall accuracy remains basically unchanged. (Accuracy≈0.927, macro-Precision≈0.742, macro-Recall≈0.958, macro-F1≈0.807, macro-AUC≈0.99), resulting in a balancing decision of "trading higher accuracy for some recall." This balancing strategy is consistent throughout the model. When adjusting the COLD thresholds, Precision≈0.48 and Recall≈0.97, false positives are detected. After adjusting the thresholds, Precision≈0.63 and Recall≈0.25 (more cautious, fewer false positives). This strategy is also more in line with the response strategies for epidemics, influenza, or common diseases, prioritizing accuracy over recall.

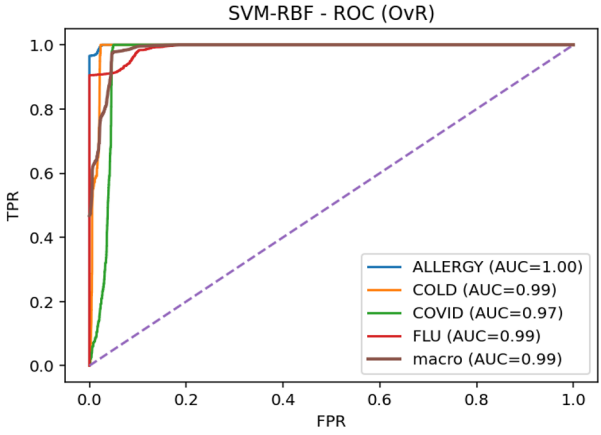

Fig. 1. SVM-RBF – ROC on the test set (Picture credit: Original)

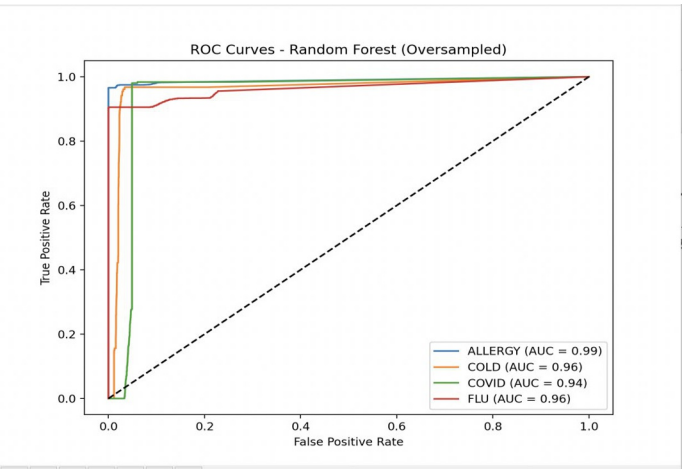

Fig. 2. MLP – ROC on the test set after probability calibration (Picture credit: Original)

In the confusion matrices shown in Figures 3, 4, and 5, the diagonal cells for ALLERGY and FLU are almost full, while misclassifications for COVID and COLD are mainly concentrated between each other because of high feature overlap, such as cough, runny nose, and fatigue, which are essentially the same symptoms of respiratory diseases. Therefore, more refined features may be introduced later to differentiate them, such as the duration of fever and the presence of medical history. Excluding these two categories, the macro AUC of SVM-RBF is approximately 0.986, and the macro AUC of MLP is approximately 0.99. The ROC curves for all four categories are generally close to the upper left corner, with ALLERGY/FLU approaching absolute accuracy and COLD/COVID around 0.97. The study found that even with feature overlap, the ranking ability did not decrease.

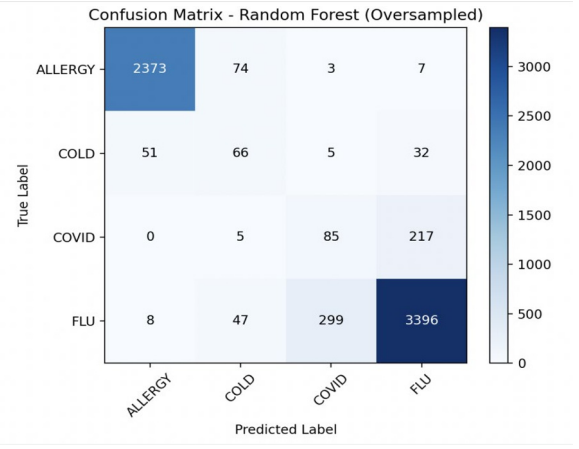

Fig. 3. Normalized confusion matrix on the test set (Picture credit: Original)

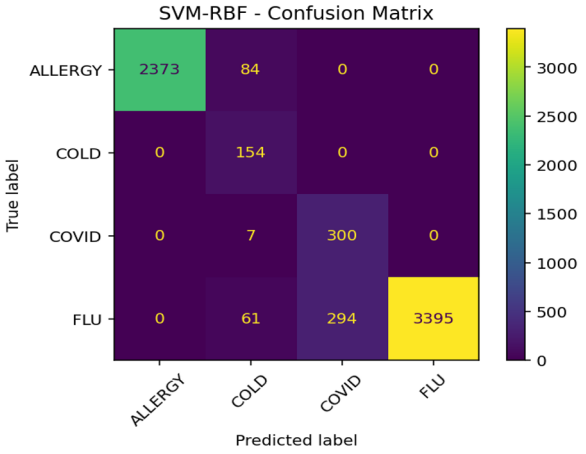

Fig. 4. SVM-RBF — Confusion Matrix on the test set (Picture credit: Original)

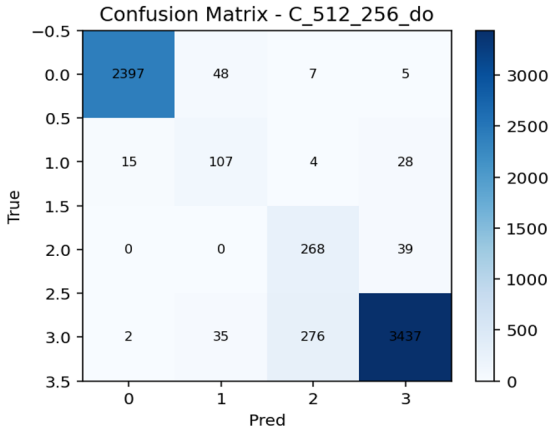

Fig. 5. MLP — Confusion Matrix on the test set (Picture credit: Original)

Overall, the three models in the study achieved the best accuracy after adjusting the thresholds of their neural networks. The different models formed a mutually beneficial cycle in the experiments. In the study, BNB provided a transparent and low-cost baseline, indicating good data separability; RBF-SVM was optimal in terms of macroscopic metrics and robustness, suitable for scenarios requiring a strong baseline and stable performance; MLP was more engineering-friendly in terms of probability calibration, adjustable thresholds, and training visualization, achieving higher accuracy in a few classes without sacrificing overall accuracy through post-processing with classification thresholds. For COLD and COVLD, the high accuracy achieved after multiple data captures and manual differentiation of the other two diseases, supported by a large amount of data, made it a viable option.

At the deployment level (remote triage/public health screening), it is recommended to output calibrated probabilities and set thresholds according to the scenario, while monitoring the majority class dominant risk with macro indicators; these practices are consistent with the credible deployment recommendations for medical AI. For scenarios with extreme imbalance or significantly asymmetric costs, the benefits of cost-sensitive/order class loss or mixed loss can be further evaluated.

5 Conclusion

This study compared the ability of BernoulliNB, RBF-SVM and lightweight MLP to distinguish COVID, FLU, COLD and ALLERGY in cases of class imbalance and high symptom overlap. Under a unified process (hierarchical segmentation, class weight training, macro-average indicators), RBF-SVM performed the most robustly (Accuracy≈0.933, macro-F1≈0.817, macro-AUC≈0.986); the overall accuracy of MLP was similar (≈0.927) and the macro AUC was higher (≈0.99). After probability calibration and class-based threshold adjustment, MLP improves the accuracy of easily confused categories without sacrificing overall accuracy; BNB provides a transparent and low-cost baseline for separability. Errors are mainly concentrated between COVID and COLD; ALLERGY and FLU are more clearly distinguished. For deployment, the workflow provides reproducible preprocessing, calibrated risk-sensitive thresholds, and interpretable diagnostics based on confusion matrix/ROC. In the future, time and context features will be introduced to explore cost-sensitive/ordinal loss and integration, quantify uncertainties beyond calibration, and validate on prospective external data.

References

1. Wang, M., Yang, B., Liu, Y., Yang, Y., Ji, H., Yang, C.: Emerging Infectious Disease Surveillance Using a Hierarchical Diagnosis Model and the Knox Algorithm. Scientific Reports 13, 19836 (2023)
2. Wang, X., Zhao, W., Chen, Z., Liu, J., Kang, B., Guo, W.: AI Pathogen Type Discrimination Model for Pneumonia Based on Multidimensional Clinical Data. Journal of Clinical Emergency 25(7), 336–342 (2024)
3. van den Goorbergh, R., van Smeden, M., Timmerman, D., Van Calster, B.: The Harm of Class Imbalance Corrections for Risk Prediction Models: Illustration and Simulation Using Logistic Regression. Journal of the American Medical Informatics Association (2022)
4. Gao, R., Li, T., Tang, Y., Xu, Z., Kammer, M., Antic, S.L., Sandler, K., Moldonado, F., Lasko, T.A., Landman, B.: A Comparative Study of Confidence Calibration in Deep Learning: From Computer Vision to Medical Imaging. arXiv:2206.08833 (2022)
5. Ouyang, C., Wang, S., Chen, C., Li, Z., Bai, W., Kainz, B., Rueckert, D.: Improved Post-Hoc Probability Calibration for Out-of-Domain MRI Segmentation. arXiv:2208.02870 (2022)
6. Harbecke, D., Chen, Y., Hennig, L., Alt, C.: Why Only Micro-F1? Class Weighting of Measures for Relation Classification. arXiv:2205.09460 (2022)

7. Carriero, A., Luijken, K., de Hond, A., Moons, K.G.M., Van Calster, B., van Smeden, M.: The Harms of Class Imbalance Corrections for Machine-Learning-Based Prediction Models: A Simulation Study. arXiv:2404.19494 (2024)
8. Dimitriadis, T., Dümbgen, L., Henzi, A., Puke, M., Ziegel, J.: Honest Calibration Assessment for Binary Outcome Predictions. arXiv:2203.04065 (2022)
9. Lekadir, K., Feragen, A., Frangi, A.F., et al.: FUTURE-AI: International Consensus Guideline for Trustworthy and Deployable AI in Healthcare. arXiv:2309.12325 (2023)
10. Sadi, A.A., Chowdhury, L., Jahan, N., Rafi, M.N.S., Chowdhury, R., Khan, F.A., Mohammed, N.: LMFLOSS: A Hybrid Loss for Imbalanced Medical Image Classification. arXiv:2212.12741 (2022)