*engineering proceedings*

MDPI

*Proceeding Paper*

# Classification of Infectious and Parasitic Diseases by Smart Healthcare System †

**Junwei Yang [1], Teerawat Simmachan [2], Subij Shakya [3] and Pichit Boonkrong [4,*]**

[1] Department of Clinical Epidemiology and Biostatistics, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok 10400, Thailand; junwei.yan@mahidol.ac.th
[2] Thammasat University Research Unit in Statistical Theory and Applications, Thammasat University, Pathum Thani 12120, Thailand; teerawat@mathstat.sci.tu.ac.th
[3] Department of Food and Nutrition, University of Helsinki, 00790 Helsinki, Finland; subij.shakya@helsinki.fi
[4] College of Biomedical Engineering, Rangsit University, Pathum Thani 12000, Thailand
* Correspondence: pichit.bk@rsu.ac.th
† Presented at the 2025 IEEE 5th International Conference on Electronic Communications, Internet of Things and Big Data, New Taipei, Taiwan, 25–27 April 2025.

## Abstract

We developed a machine-learning model for the International Classification of Diseases, 10th Revision (ICD-10) classification using data from 5108 patients. Nine features, including age, gender, BMI, and vital signs, were extracted to classify the top three ICD-10 categories: intestinal infections, tuberculosis, and other bacterial diseases. Decision trees, random forest, and XGBoost models were tested using the synthetic minority over-sampling technique (SMOTE) and class weights to minimize class imbalance. Five-fold cross-validation was used using the training and testing datasets in a data ratio of 80:20. The random forest model with class weights showed the best performance. Shapley additive explanations (SHAP) analysis highlighted body-mass index (BMI), gender, and pulse as key features. The developed model showed potential for enhancing ICD-10 classification through real-time and personalized medical applications.

**Keywords:** disease classification; feature importance; ICD-10 codes; machine learning; smart healthcare

check for updates

Academic Editors: Teen-Hang Meen, Shu-Han Liao and Cheng-Fu Yang

Published: 1 September 2025

**Citation:** Yang, J.; Simmachan, T.; Shakya, S.; Boonkrong, P. Classification of Infectious and Parasitic Diseases by Smart Healthcare System. *Eng. Proc.* **2025**, *108*, 14. https://doi.org/10.3390/engproc2025108014

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

Infectious and parasitic diseases are significant global health challenges, particularly in regions with limited healthcare resources. Implementing the International Classification of Diseases, 10th Revision (ICD-10) classification with a data-driven approach enables precise identification of disease patterns, aiding in early detection, resource allocation, and policy-making. Integrating the method into a smart healthcare system ensures improved decision-making, better patient outcomes, and proactive disease management.

The integration of machine learning (ML) with ICD-10 disease classification has shown significant potential in automating the coding process, which is traditionally manual and time-consuming. ML models support clinical decision-making by providing accurate and timely disease classifications, which is crucial for patient care and treatment planning [1–5]. ML models significantly reduce the time and effort required for manual ICD-10 coding, improving both efficiency and accuracy in clinical settings [1,3,6]. Nevertheless, the class imbalance plays a massive role in model performance. To handle the class imbalance and enhance model performance, the synthetic minority over-sampling technique (SMOTE) and

class weights are widely used [7–12]. Implementing ML in disease classification based on ICD-10, decision trees (DTs), random forest (RF), and extreme gradient boosting (XGBoost) are often utilized. In healthcare, DT models are particularly valued for their intuitive representation, ease of understanding, and nonparametric nature, which are appropriate for exploratory knowledge discovery [2,3]. RF classifiers have been widely used in various disease classification tasks, demonstrating high accuracy and robustness [13,14]. The use of the RF algorithm in ICD-10 coding has been explored in several studies, highlighting its effectiveness and challenges [15,16]. XGBoost models consistently show higher accuracy and better performance metrics compared with traditional statistical models and other ML algorithms [3,17,18]. XGBoost has been used for the automatic classification of diseases based on ICD-10 codes [17,18]. Shapley additive explanations (SHAP) are used to interpret model predictions, making the results more understandable for clinical practitioners [6,19]. Data-driven models in infectious and parasitic disease classification are vital for modern healthcare systems because they enable efficient resource allocation, early disease detection, and precise patient care.

Based on the previous studies, we combined class imbalance and feature extraction for the classification performances of DT, RF, and XGBoost models. To interpret the results from the best classifier, SHAP values were utilized. The experimental results were analyzed for the possible implementation of the developed model.

## 2. Machine-Learning Model

To develop a data-driven ML model in the ICD-10 classification scheme, data were collected from hospital records to ensure ethical compliance and comprehensive coverage. Pre-processing techniques including class-imbalance strategies and feature importance were used using DT, RF, and XGBoost to classify disease categories.

### *2.1. Data Collection*

Secondary data comprising ICD-10 records was retrieved from Nong Han Hospital, Udon Thani, Thailand, from 2020 to 2022. This study received ethical approval under an exemption category from the Ethics Committee at Rangsit University, Thailand, in compliance with the principles of the Helsinki Declaration. This ensured ethical adherence while utilizing pre-existing clinical data for research purposes.

ICD-10 is a medical classification list by the World Health Organization (WHO) for systematically categorizing diseases, health conditions, and related medical procedures [20,21]. The ICD-10 system uses hierarchical alphanumeric codes for diseases, aiding clinical documentation and analysis. This study focuses on codes A00–A99, covering infectious and parasitic diseases of major public health concern.

As shown in Table 1, classes 1, 2, and 4 (intestinal infectious diseases, tuberculosis, and other bacterial diseases) were selected for classification due to their high prevalence, collectively accounting for the majority (98.88%) of the patients. This ensures that the model focuses on the most impactful categories in the dataset. Additionally, these classes represent significant public health concerns, allowing the classification model to aid in resource allocation, early diagnosis, and disease management. The remaining classes are less represented, making them less suitable for robust modeling.

The data collected were pre-processed, involving data cleansing, transformation, and imputation. *K*-fold cross-validation (CV) was conducted to ensure reliable evaluation by rotating training and validation sets. The class imbalance was minimized using class weights and SMOTE. The mean decrease in accuracy (MDA) from RF was used for feature extraction and ranking by their impact on accuracy.

**Table 1.** Frequency distribution of patients across 11 different ICD-10 classes ($n$ = 5108).

| Code Range | Class Label | Infectious and Parasitic Diseases | Count | Percentage |
|---|---|---|---|---|
| A00–A09 | 1 | Intestinal infectious diseases | 1851 | 36.24 |
| A15–A19 | 2 | Tuberculosis | 2148 | 42.05 |
| A20–A28 | 3 | Certain zoonotic bacterial diseases | 33 | 0.65 |
| A30–A49 | 4 | Other bacterial diseases | 1052 | 20.60 |
| A50–A64 | 5 | Infections with a predominantly sexual mode of transmission | 14 | 0.27 |
| A65–A69 | 6 | Other spirochetal diseases | 2 | 0.04 |
| A70–A74 | 7 | Other diseases caused by chlamydia | 0 | 0.00 |
| A75–A79 | 8 | Rickettsioses | 4 | 0.08 |
| A80–A89 | 9 | Viral infections of the central nervous system | 3 | 0.06 |
| A90–A91 | 10 | Dengue fever | 1 | 0.02 |
| A92–A99 | 11 | Arthropod-borne viral fevers and viral hemorrhagic fevers | 0 | 0.00 |

As the three classes were focused, the missing values were screened and removed. The number of patients was 2888. The descriptive statistics of five continuous features including age, BMI, temperature, respiratory rate (RR), and pulse are presented in Table 2. Four features include gender, status, blood, and occupation. Females accounted for 56.44% (1630), while males represented 43.56% (1258). Regarding marital status, 72.82% (2103) were married, 26.14% (755) were single, and 1.04% (30) belonged to other categories. Dominant blood types were B (69.56%, 2009), followed by O (19.36%, 559), A (8.52%, 246), and AB (2.56%, 74).

**Table 2.** Descriptive statistics for continuous features ($n$ = 2888).

| Statistics | Age | BMI | Temp | RR | Pulse |
|---|---|---|---|---|---|
| Mean | 50.2230 | 22.3070 | 36.9539 | 20.4278 | 93.5760 |
| Standard deviation (SD) | 21.1889 | 4.7420 | 0.8576 | 2.4679 | 18.6243 |
| Minimum | 0.0000 | 11.7922 | 34.5000 | 16.0000 | 24.0000 |
| P25 | 35.0000 | 19.1111 | 36.5000 | 20.0000 | 80.0000 |
| P50 | 54.0000 | 21.8750 | 36.7000 | 20.0000 | 90.0000 |
| P75 | 67.0000 | 24.9770 | 37.0000 | 20.0000 | 104.0000 |
| Maximum | 122.0000 | 50.5351 | 42.0000 | 48.0000 | 182.0000 |

With 2888 patients for a 3-class disease classification task, a standard train-test dataset split in 80:20 resulted in 2310 training samples and 578 testing samples. However, relying solely on training and testing datasets led to biased results, especially for imbalanced datasets. To improve reliability and generalizability, a five-fold CV was used by training on four folds and validating on the fifth, rotating across all folds.

To deal with class imbalance, the concepts of class weight and SMOTE were introduced to investigate whether one can enhance the classification performance of DTs, RF, and XGBoost models.

- Class weight: Class imbalance is challenging in classification tasks, leading to biased predictions favoring majority classes. To address this, adjusting class weights in the loss function are assigned to minority classes depending on importance, encouraging models to better capture patterns from underrepresented groups [9,10,12]. We used Scikit-learn to balance class weights using Equation (1).

$$w_c = \frac{N}{C \times N_c} \quad (1)$$

where $w_c$ is the weight for class $c$, $N$ is the total samples, $C$ is the number of classes, and $N_c$ is the class count. Comparisons of models with and without class weights are conducted to assess performance improvements.

- SMOTE: The training set exhibited class imbalance, with class 1 being overrepresented (1088 samples), while classes 2 and 4 were minority classes (572 samples each). After applying SMOTE, all classes were balanced to contain 1088 samples each. Consequently, the total size of the training set increased from 2232 to 3264 samples. SMOTE effectively mitigated the class-imbalance issue by oversampling the minority classes, ensuring an equal representation of all classes in the training set, which is critical for improving classification model performance on imbalanced datasets [7,8,11,12].

### *2.2. Feature Selection Techniques*

MDA is widely used for assessing feature importance in ML models [22]. In this study, the MDA from original and SMOTE training sets were evaluated and then fed into DT, RF, and XGBoost models. Nine features were extracted from the training datasets. The number of features was iteratively varied from 1 to 9 to determine the optimal features. To compare the different sets of features, the features selected by the least absolute shrinkage and selection operator (LASSO), ridge, and MDA techniques were considered for different data balancing strategies.

### *2.3. Models*

Regarding the ICD-10 disease classification based on the dataset, we considered three classes of the majority, i.e., A00–A09 (intestinal infections), A15–A19 (tuberculosis), and A30–(other bacterial diseases). The classifiers and their hyperparameter are described as follows.

By handling both numeric and categorical data, DT models provide interpretable predictions, aiding healthcare professionals in identifying disease categories accurately while adapting to complex, hierarchical structures of the ICD-10 classification system [2–4]. The tree splits the features at each node based on thresholds or categorical values to create decision paths.

RF is an ensemble learning method used to classify infectious and parasitic diseases based on ICD-10 codes [15,16]. It builds multiple DTs during training and combines their outputs to improve predictive accuracy and reduce overfitting. Each tree is constructed using a random subset of features and samples, ensuring robustness to noise and class imbalance.

XGBoost is a powerful ML algorithm designed for classification and regression tasks. It builds an ensemble of DTs sequentially to optimize performance by minimizing errors through gradient descent. In ICD-10 classification, XGBoost effectively handles the challenge of imbalanced datasets in infectious and parasitic disease predictions by assigning higher importance to minority classes [17,18].

To obtain the best parameter combination for each model, a grid search hyperparameter tuning strategy was applied in the training process. Each parameter set was evaluated using a five-fold CV, with the area under the curve (AUC) as the selection metric. The selected best parameters for each model are as follows.

- DTs:= {max_depth = 10, min_samples_leaf = 3, min_samples_split = 5};
- RF:= {max_depth = 15, min_samples_split = 5, n_estimators = 60};
- XGBoost:= {learning_rate = 0.2, max_depth = 5, n_estimators = 70}.

### *2.4. Model Performance*

Feature importance based on MDA provides information on the key predictors contributing to the model, while class weighting is used to address data imbalance by ensuring

equitable representation of all classes. To evaluate the impact of the techniques on model performance, a 3 × 3 confusion matrix for ICD-10 disease classification was used to derive essential metrics, including accuracy, precision, recall (sensitivity), specificity, F1-score, balanced accuracy, and AUC [23,24]. In this study, $AUC_{macro}$ was used. A higher AUC shows better discrimination between disease groups, particularly in handling imbalanced datasets. The average AUC for the C class was defined as follows.

$$AUC_{macro} = \frac{1}{C}\sum_{i=1}^{C} AUC_i \quad (2)$$

Finally, the Friedman test and Nemenyi post hoc test were performed for the comparison of model performance at the significance level $\alpha = 0.05$.

### *2.5. Feature Importance*

SHAP is a framework for interpreting ML models by quantifying feature importance using SHAP values from cooperative game theory [6,19]. It explains how each feature contributes to the model's prediction. In classifying infectious and parasitic diseases using ICD-10 codes, SHAP assigns importance to each feature by evaluating its marginal contribution across feature combinations. In this study, SHAP was evaluated only from the best model.

## 3. Results

Python programming (version 3.10.9, packaged by conda-forge, MSC v.1916 64-bit AMD64) was utilized to conduct the analysis, incorporating three imbalance handling strategies (original, class weighting, and SMOTE) and three ML models (DT, RF, and XGBoost). The implementation relied on several key libraries, including pandas (1.5.2), scikit-learn (1.6.1), scipy (1.9.3), statsmodels (0.13.2), polars (0.20.26), and PyTorch (2.1.2+cu118). Cumulative feature selection was applied, ranging from one to nine features. In total, 81 distinct model combinations were generated and systematically compared. Feature selections affected the classification performances, and the feature importance was evaluated by SHAP.

### *3.1. Selected Features*

Implementing the feature-selection method, the ordered sets of features were listed as follows.

- LASSO = {Temp, Age, Gender, BMI, RR, Blood, Pulse}
- Ridge = {Temp, Age, Gender, BMI, RR, Blood, Pulse}
- MDA_Original = {Temp, Age, Gender, BMI, RR, Blood, Pulse, Occ, Status}
- MDA_SMOTE = {Temp, Age, Gender, BMI, Blood, Pulse, RR, Occ, Status}

The LASSO provided significant features whilst the ridge, MDA_Original, and MDA_SMOTE kept all features. Setting the threshold value of 0.001, all four feature selections contained the same first seven features. The MDA for features in both the original and SMOTE-enhanced were compared in Figure 1. The top four features (temperature, age, gender, and BMI) showed the same rank in both datasets, indicating consistent importance across sampling methods. Among the nine features, occupation and marital status were the least important. To assess the effect of feature inclusion, features were sequentially added to the RF model based on importance rankings. Performance was evaluated at each step to examine the incremental contribution of each feature.

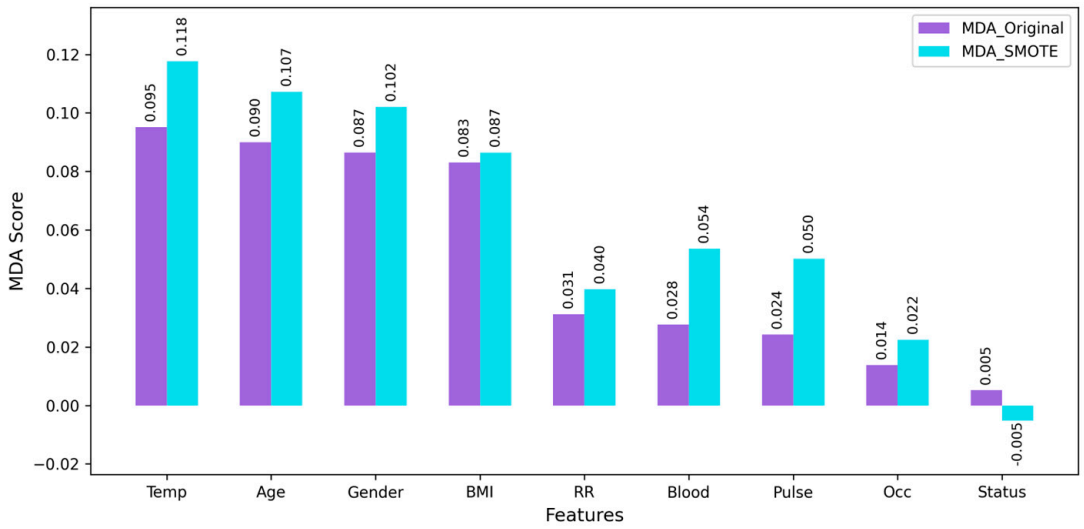



**Figure 1.** Comparison of MDA scores from original and SMOTE train sets across nine features.

### *3.2. Model Comparison*

Using the combination of feature-selection techniques and class-imbalance strategies, model performances, especially accuracy, F1-score, and AUC were compared with the best combination of the classifier. The classification performances of the DT, RF, and XGBoost models across original, class weights and SMOTE training sets with different sets of features were compared. As shown in Figure 2, the outstanding scenarios are compared across different classifiers and class-imbalance strategies. By Friedman test at $\alpha = 0.05$, there is no significant improvement of each classifier across different class-imbalance strategies. By the Wilcoxon pairwise test across different classifiers, RF and XGBoost models showed approximately the same performance while DT models showed the lowest values for all evaluation metrics.

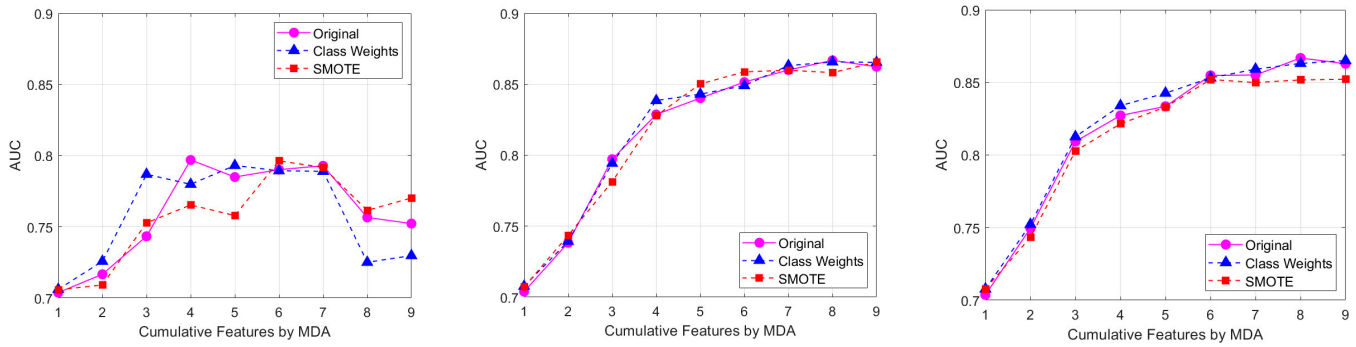



**Figure 2.** AUC comparison in the three ML models across different imbalance strategies against cumulative features: (**a**) DT, (**b**) RF, and (**c**) XGBoost models.

#### 3.2.1. Effect of Feature Selection

Figure 2 shows the AUCs of three ML models under different class-imbalance strategies, evaluated against cumulative features ranked by MDA. Sequentially, the inclusion of cumulative features based on MDA showed improvement in AUC, particularly with the initial four features. The performance metrics showed a plateau or a decline, suggesting diminishing returns or potential overfitting. The DT model achieved optimal performance with 4–6 features, the RF model performed best with 6–8 features, while the XGBoost model maintained a high AUC with 7–9 features.

3.2.2. SHAP Analysis for Feature Importance

To interpret the experimental results and identify the best classifier, such as RF with class weights, SHAP was used to identify features influencing predictions and more explainable predictions for certain classes. The SHAP summary plots highlight the most influential features in classifying three disease groups (Figure 3). The left plot shows gender, BMI, temperature, and age as the important features with the highest mean SHAP values across all classes, indicating their strong impact on predictions. The right plot demonstrates the magnitude of SHAP values of each feature. High feature values generally increase the prediction probability, particularly for gender, BMI, and pulse which show significant class-separation influence in the classifiers.

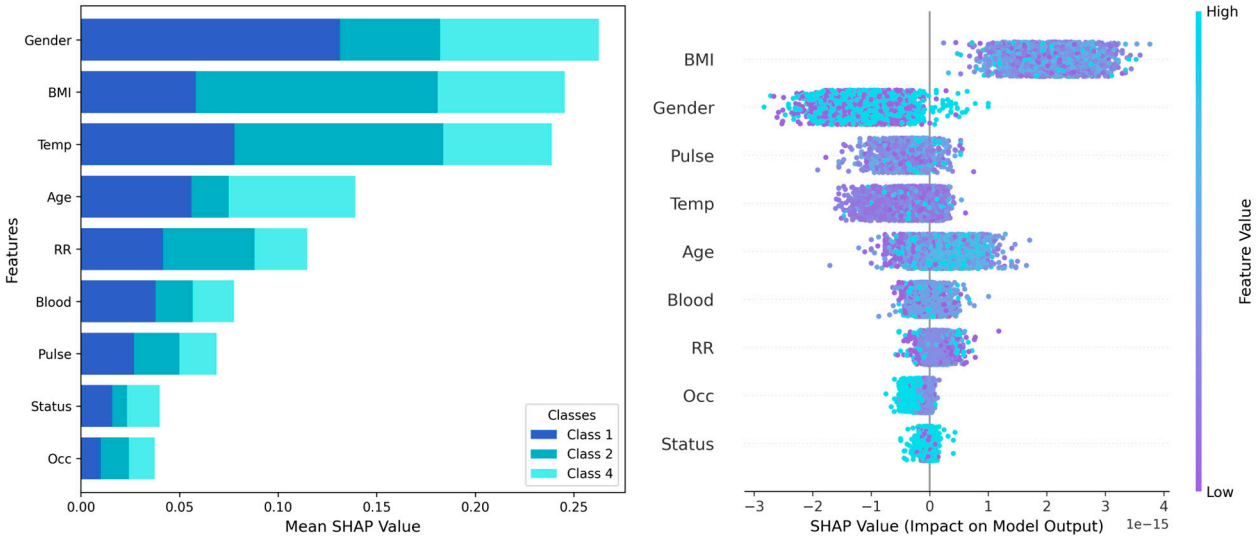



**Figure 3.** Mean SHAP values (**left**) and features' impact on model output (**right**).

## 4. Discussion

The capability in classification among various scenarios was investigated in this study. The impact of class-imbalance strategies, feature selection, and possible application was examined as follows.

### *4.1. Impact of Class Imbalance*

The use of class weights slightly improved the performance of the RF and XGBoost models, while SMOTE had little effect, and the DT model performed best on the original dataset without class weights. The small number of features limited the model's effectiveness. Adding more clinical features, such as blood pressure, comorbidity, symptom duration, past medical history, medication usage, lifestyle factors (e.g., smoking, alcohol drinking, diet), genetic markers, and environmental exposures improve the models.

### *4.2. Relationship Between Features and Diseases*

The relationship between the important features played a crucial role in classifying intestinal infections, tuberculosis, and bacterial diseases. Elevated temperature and abnormal BMI were significant indicators of infections, while age was critical for identifying risk groups, especially for tuberculosis. Gender differences influenced disease prevalence, with several infections showing gender-specific patterns. The effectiveness of healthcare ML models significantly impacted the number of features used. While a small number of features limited the model's ability to capture complex patterns, feature-selection techniques effectively identified the most relevant features, improving model performance and

interpretability. Balancing the quantity and quality of features was crucial for developing robust and accurate healthcare ML models [1–6,9].

### *4.3. Application in Smart Healthcare*

Class-imbalance strategies and feature-selection techniques enhanced the classification of infectious and parasitic diseases based on ICD-10 codes for smart healthcare systems [7–10,12]. These methods improved model performance by addressing uneven class distributions and identifying key predictors including temperature, age, gender, and BMI. Balanced data and optimized features were used to accurately detect disease categories such as intestinal infections, tuberculosis, and bacterial diseases. This approach supported early diagnosis, resource allocation, and better public health management in healthcare systems.

## 5. Conclusions

We introduced class weights and SMOTE to identify three classifiers in DT, RF, and XGBoost models with imbalanced data. The ratio of training and testing dataset was 80:20 and a five-fold CV was implemented. Different sets of features from LASSO, ridge, and iterative MDA were fed into the ML models. The RF model, adjusted for class weights, showed the best performance in disease classification. Feature-selection methods with MDA were used to identify predictors such as temperature, age, gender, and BMI. These factors were essential for differentiating between intestinal infections, tuberculosis, and bacterial diseases. Although class weights slightly boosted the model performance, the SMOTE showed limited benefits. The results of this research highlighted the significance of feature extraction and tackling class imbalance in the classification of medical data. Future research is necessary to integrate additional clinical features, including blood pressure, comorbidities, and lifestyle factors, to improve model accuracy and generalizability. ML techniques enable healthcare systems to create more accurate and efficient disease classifiers, ultimately aiding in early diagnosis, resource management, and personalized patient care.

**Author Contributions:** J.Y. performed the programming, investigation, and visualization. T.S. and P.B. contributed to the conceptualization, methodology, and formal analysis. T.S. and S.S. were responsible for validation. P.B. and S.S. prepared the original draft, while T.S. and P.B. reviewed and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** This study received ethical approval under an exemption category from the Ethics Committee at Rangsit University, Thailand (DPE. No. RSUERB2024-032), in compliance with the principles outlined in the Helsinki Declaration.

**Informed Consent Statement:** As the research involved de-identified secondary data, informed consent was waived by the Ethics Committee at Rangsit University and Nong Han Hospital.

**Data Availability Statement:** Data available from the corresponding author on request.

**Acknowledgments:** The authors would like to acknowledge Nong Han Hospital, Udon Thani for providing the medical dataset used in this study. The constructive comments from the anonymous reviewers are deeply appreciated for improving the overall quality and clarity of this manuscript.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References


1. Sammani, A.; Bagheri, A.; van der Heijden, P.G.M.; Te Riele, A.S.J.M.; Baas, A.F.; Oosters, C.A.J.; Oberski, D.; Asselbergs, F.W. Automatic multilabel detection of ICD10 codes in Dutch cardiology discharge letters using neural networks. *NPJ Digit. Med.* **2021**, *4*, 37. [CrossRef] [PubMed]
2. Kansal, A.; Gao, M.; Balu, S.; Nichols, M.; Corey, K.; Kashyap, S.; Sendak, M. Impact of diagnosis code grouping method on clinical prediction model performance: A multi-site retrospective observational study. *Int. J. Med. Inform.* **2021**, *151*, 104466. [CrossRef] [PubMed]
3. Tran, Z.; Verma, A.; Wurdeman, T.; Burruss, S.; Mukherjee, K.; Benharash, P. ICD-10 based machine learning models outperform the Trauma and Injury Severity Score (TRISS) in survival prediction. *PLoS ONE* **2022**, *17*, e0276624. [CrossRef] [PubMed]
4. Tran, Z.; Zhang, W.; Verma, A.; Cook, A.; Kim, D.; Burruss, S.; Ramezani, R.; Benharash, P. The derivation of an International Classification of Diseases, Tenth Revision--based trauma-related mortality model using machine learning. *J. Trauma Acute Care Surg.* **2022**, *92*, 561–566. [CrossRef] [PubMed]
5. Mayya, V.; King, C.; Vu, G.T.; Gurupur, V. Empirical Study of Feature Selection Methods in Regression for Large-Scale Healthcare Data: A Case Study on Estimating Dental Expenditures. *IEEE Access* **2024**, *12*, 153564–153579. [CrossRef]
6. Diao, X.; Huo, Y.; Zhao, S.; Yuan, J.; Cui, M.; Wang, Y.; Lian, X.; Zhao, W. Automated ICD coding for primary diagnosis via clinically interpretable machine learning. *Int. J. Med. Inform.* **2021**, *153*, 104543. [CrossRef] [PubMed]
7. Kosolwattana, T.; Liu, C.; Hu, R.; Han, S.; Chen, H.; Lin, Y. A self-inspected adaptive SMOTE algorithm (SASMOTE) for highly imbalanced data classification in healthcare. *BioData Min.* **2023**, *16*, 15. [CrossRef] [PubMed]
8. Ilham, A.; Kindarto, A.; Kareem Oleiwi, A.; Khikmah, L.; Fathurohman, A.; Dias Ramadhani, R.; Abdunnasir Jawad, S.; April Liana, D.; Mutiar, A. CFCM-SMOTE: A Robust Fetal Health Classification to Improve Precision Modelling in Multi-Class Scenarios. *Int. J. Comput. Digit. Syst.* **2024**, *16*, 471–486. [CrossRef]
9. Almagro, M.; Unanue, R.M.; Fresno, V.; Montalvo, S. ICD-10 coding of Spanish electronic discharge summaries: An extreme classification problem. *IEEE Access* **2020**, *8*, 100073–100083. [CrossRef]
10. Mahmoodi, N.; Shirazi, H.; Fakhredanesh, M.; DadashtabarAhmadi, K. Automatically weighted focal loss for imbalance learning. *Neural Comput. Appl.* **2024**, *37*, 4035–4052. [CrossRef]
11. Simmachan, T.; Boonkrong, P. Effect of Resampling Techniques on Machine Learning Models for Classifying Road Accident Severity in Thailand. *J. Curr. Sci. Technol.* **2025**, *15*, 99. [CrossRef]
12. Fahmi, A.; Muqtadiroh, F.A.; Purwitasari, D.; Sumpeno, S.; Purnomo, M.H. A Multi-Class Classification of Dengue Infection Cases with Feature Selection in Imbalanced Clinical Diagnosis Data. *Int. J. Intell. Eng. Syst.* **2023**, *15*, 176–192.
13. Lutimath, N.; Sharma, N.; Byregowda, K. Prediction of heart disease using biomedical data through machine learning techniques. *EAI Endorsed Trans. Pervasive Health Technol.* **2021**, *7*, e3. [CrossRef]
14. Mahajan, R.; Arora, R. Smart Healthcare Solutions: IoT Integration for Sustainable Management of Kidney Diseases Leveraging Machine Learning. In Proceedings of the International Conference on Sustainable Development through Machine Learning, AI and IoT, Cham, Switzerland, 27–28 April 2024; pp. 244–253.
15. Le Lay, J.; Alfonso-Lizarazo, E.; Augusto, V.; Bongue, B.; Masmoudi, M.; Xie, X.; Gramont, B.; Clarier, T. Prediction of hospital readmission of multimorbid patients using machine learning models. *PLoS ONE* **2022**, *17*, e0279433. [CrossRef] [PubMed]
16. Zikos, D.; DeLellis, N. Comparison of the Predictive Performance of Medical Coding Diagnosis Classification Systems. *Technologies* **2022**, *10*, 122. [CrossRef]
17. Hasan, M.N.; Hamdan, S.; Poudel, S.; Vargas, J.; Poudel, K. Prediction of length-of-stay at intensive care unit (icu) using machine learning based on mimic-iii database. In Proceedings of the IEEE Conference on Artificial Intelligence (CAI), Santa Clara, CA, USA, 5–6 June 2023; pp. 321–323.
18. Falter, M.; Godderis, D.; Scherrenberg, M.; Kizilkilic, S.E.; Xu, L.; Mertens, M.; Jansen, J.; Legroux, P.; Kindermans, H.; Sinnaeve, P.; et al. Using natural language processing for automated classification of disease and to identify misclassified ICD codes in cardiac disease. *Eur. Heart J.-Digit. Health.* **2024**, *5*, 229–234. [CrossRef] [PubMed]
19. Liu, J.; Li, R.; Yao, T.; Liu, G.; Guo, L.; He, J.; Guan, Z.; Du, S.; Ma, J.; Li, Z. Interpretable Machine Learning Approach for Predicting 30-Day Mortality of Critical Ill Patients with Pulmonary Embolism and Heart Failure: A Retrospective Study. *Clin. Appl. Thromb./Hemost.* **2024**, *30*, 1–13. [CrossRef] [PubMed]
20. WHO. ICD-10 Version: 2019. Available online: https://icd.who.int/browse10/2019/en (accessed on 11 January 2025).
21. CDC. Classification of Diseases, Functioning, and Disability. Available online: https://www.cdc.gov/nchs/icd/icd-10-cm/index.html (accessed on 11 January 2025).
22. Simmachan, T.; Wongsai, S.; Lerdsuwansri, R.; Boonkrong, P. Impact of COVID-19 Pandemic on Road Traffic Accident Severity in Thailand: An Application of K-Nearest Neighbor Algorithm with Feature Selection Techniques. *Thail. Stat.* **2025**, *23*, 129–143.

23. Hanmanop, S.; Jongsiri, T.; Khongtan, M.; Tengjongdee, S.; Chaitan, C.; Pechprasarn, S.; Boonkrong, P. Integrating Magnetic Resonance Imaging and Deep Learning Networks for Brain Tumor Classification. In Proceedings of the Biomedical Engineering International Conference (BMEiCON), Chon Buri, Thailand, 21–24 November 2024; pp. 1–5.
24. Boonkrong, P.; Simmachan, T.; Sittimongkol, R.; Lerdsuwansri, R. Data-Driven Approach in Provincial Clustering for Sustainable Tourism Management in Thailand. *Thail. Stat.* **2025**, *23*, 481–500.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.