Comparative Analysis of Classical ML Baselines on Noisy and Imbalanced Occupational Lung Disease Data in Vietnam

Phuong Luong-Thi-Bich[1], Quan Nguyen-Minh[2], Hung Vo-Tri[3]
Faculty of Information Technology, Hanoi University of Architecture, Vietnam[1,2,3]
Faculty of Information Technology, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam[1]

ABSTRACT
Occupational lung disease is one of the most serious health problems affecting the global workforce. Early prediction of disease risk is important in medical prevention and intervention. In this study, the proposed approach conducted a comparison of four classical machine learning models—Random Forest (RF), XGBoost, Logistic Regression (LR), and Support Vector Machine (SVM)—on the same set of occupational lung disease data that had been manually processed and encoded. The experimental results show that XGBoost achieves the best performance with an accuracy of 98.34% and a Macro F1-score of 0.7996, followed by LR, RF and SVM. In addition, the characteristic analysis shows that each model focuses on different factors, suggesting the potential to combine multiple models to improve prediction efficiency.

Keywords
Occupational lung disease, classical machine learning, Random Forest, XGBoost, Logistic Regression, SVM

1. INTRODUCTION
Occupational diseases have long been a global concern, directly affecting the health and productivity of workers in many industries. According to the International Labour Organization (ILO), millions of workers each year face the risks of occupational diseases, many of which result in permanent injury or death. The main cause comes from toxic working conditions and potentially risky production processes [1], [2]. The history of medicine has documented many cases of occupational diseases, from Hippocrates' (460–377 BC) description of lead poisoning, to Galen's second-century record of miners' disease, and reports of mercury poisoning and many other occupational diseases in the centuries that followed.

Early detection of occupational diseases is key to preventing and minimizing their negative impacts. In developing countries such as Vietnam, the examination and diagnosis of occupational diseases are still limited. Thousands of workers are often screened in large batches, starting with clinical examinations, extracting medical histories and medical records. If there are suspicious signs, the new employee is assigned to perform intensive subclinical tests such as chest X-ray, respiratory function measurement (FEV1), hearing test,... However, the shortage of occupational disease specialists makes this process time-consuming, costly and less effective [3].

The development of artificial intelligence and machine learning in recent years has opened up many new opportunities in the field of preventive medicine. Classical machine learning methods such as Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), and Extreme Gradient Boosting (XGBoost) have been widely applied in medical record-based disease diagnosis, thanks to their ability to process multi-dimensional data, recognize patterns, and make accurate predictions [4]. [5]. Previous studies have shown that SVM has the ability to effectively classify data sets with a large number of dimensions [6], Logistic Regression is a foundational method, easy to interpret and implement [7], Random Forest provides anti-overfitting thanks to the bagging mechanism and the collection of multiple decision trees [8], while XGBoost stands out for its training speed and high accuracy thanks to its boosting mechanism [9].

However, classical machine learning models also have some limitations when applied to medical data. Occupational disease data are often heterogeneous, contain many missing values and have an imbalance between classes (the number of cases is much less than healthy cases). These factors can reduce the effectiveness of prediction if not handled appropriately [10]. Therefore, it is necessary to study, evaluate and compare the effectiveness of classical machine learning models in predicting occupational diseases on a real dataset, in order to find the right solution for early diagnosis support systems.

In this study, we used an occupational disease dataset collected from health reports and working environment information of workers in Vietnam, applying four classical machine learning models (SVM, LR, RF, XGBoost) to build and compare the predictive efficiency. The results of the study not only help to evaluate the feasibility of applying these models in practice, but also provide a basis for integrating occupational disease prediction solutions into the health system, contributing to the protection of public health.

2. RELATED WORKS
Diagnosis of diseases based on medical data has been widely studied over the years, especially with the development of machine learning (ML) algorithms. The general principle of these studies is to use a dataset that includes patient information (physiological parameters, living habits, work environment, clinical and subclinical symptoms) to train a model that predicts health status [1], [2].

Traditional methods often rely on descriptive statistics and linear or nonlinear regression analysis techniques, however they are limited by the assumption of data distribution and the ability to model complex relationships between variables [7]. The emergence of classical machine learning models has overcome this limitation, allowing for the exploitation of diverse and heterogeneous data, and achieving higher accuracy in many medical prediction problems [4], [5].

The Support Vector Machine (SVM) is one of the powerful classification algorithms that works efficiently on large and non-linear data sets, thanks to the use of kernel functions to map data to higher characteristic spaces [6]. In disease diagnosis, SVM has been successfully applied to predict heart

disease, cancer, and occupational respiratory diseases, demonstrating the ability to separate clear layer boundaries [4].

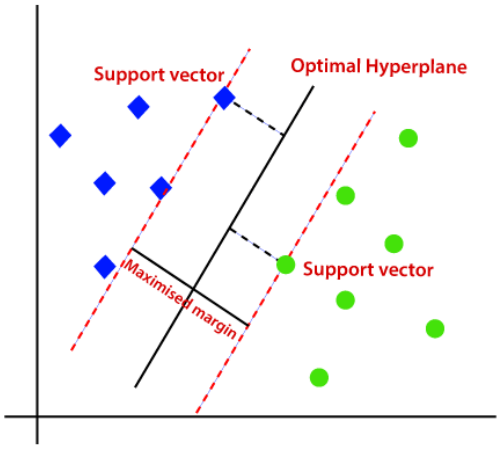

Fig. 1. Illustration of how SVM works

Logistic Regression (LR) is a classic statistical method but is still commonly used in medicine, thanks to its simplicity, ease of interpretation, and ability to estimate the probability of disease [7]. Studies in the prediction of cardiovascular disease, diabetes, and occupational diseases have shown that LR has a stable effect on small and medium-sized datasets, especially when combined with specific selection techniques [11].

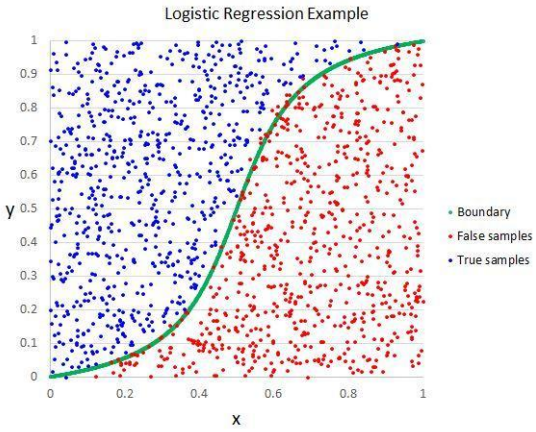

Fig. 2. Illustration of how LR works

Random Forest (RF), a model based on a set of multiple decision trees using tagging techniques, stands out for its resistance to overfitting and good handling of missing or noisy data [8]. Many medical studies have shown that RF is highly effective in detecting lung disease, predicting stroke risk, and diagnosing occupational disease, especially when the data has many taxonomic variables [12].

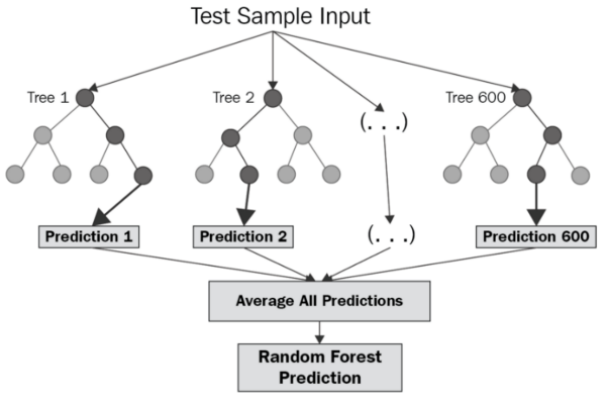

Fig. 3. Illustration of how RF works

Extreme Gradient Boosting (XGBoost) is an improvement on the traditional boosting algorithm, optimizing both training speed and accuracy [9]. Thanks to its ability to handle imbalanced data well and provide a mechanism for assessing the importance of characteristics, XGBoost has been widely applied in medical studies, including the prediction of lung cancer, diabetes, and occupational diseases [13].

However, most studies using classical machine learning models in disease prediction still have some limitations. First, medical data is often heterogeneous and contains many missing values, making it difficult to train without the application of appropriate pretreatment techniques [10]. Second, a class imbalance – when the number of patients with the disease is much smaller than that of the healthy group – can lead to a prediction bias towards the majority group, reducing the likelihood of case detection [14], [15].

In this context,this studystudy inherits the approach of using classical machine learning models (SVM, LR, RF, XGBoost) and applies it on actual occupational disease datasets. the proposed approach focus on comparing the performance between these models, and apply missing and unbalanced data processing techniques to ensure the reliability of the results.

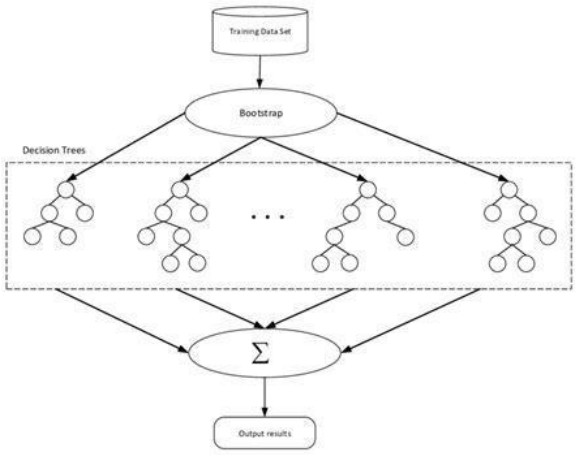

Fig. 4. Illustration of how LR works

3. MATERIALS AND METHODS

In this study, the proposed approach focused on occupational disease data, which are heterogeneous, contain many missing values, and often lack clear information. Handling missing values by assigning arbitrary values can reduce model training efficiency due to the discrepancy between the hypothetical and actual data. In this context, classical machine learning methods

such as **Support Vector Machine (**SVM**), Logistic Regression (LR), Random Forest (RF)** and **XGBoost** were selected to be exploited to build an occupational disease prediction model.

These classic models have the advantages of ease of deployment, low training costs, and high performance if the data is properly preprocessed. the proposed approach perform pre-processing steps including: removing or estimating missing values using statistical methods, normalizing data on the same scale, and encoding categorical variables. Next, the data is divided into a training set and a test set to ensure objectivity in the evaluation.

With **Logistic Regression**, the proposed approach leverage the ability to model the linear relationship between features and target variables to make predictions of disease probability. **SVM** is applied to optimize the separation hyperplane, which is especially useful when the data has complex high-dimensional data and layering boundaries. **Random Forest** leverages multiple decision trees to reduce overfitting and improve generalization. **XGBoost**, with its boosting and speed optimization mechanism, helps the model achieve high efficiency on heterogeneous data with many important variables.

the proposed approach compare the performance of each model based on evaluation indicators such as **Accuracy**, **Precision**, **Recall**, **F1-score** , and **AUC-ROC**. In addition, the proposed approach also analyze the importance of features to determine which factors in medical records, work environment, and health indicators have the greatest influence on occupational disease risk.

The results of the study are expected to indicate which classical machine learning model provides the best predictive performance, and provide useful information to assist medical professionals and managers in screening, prevention, and early intervention of occupational diseases.

A. Data Selection

In this study, the proposed approach used health data mainly collected from workers' self-reported information and health indicators that were measured and aggregated into health reports. Due to the nature of collecting data from many different sources and forms, datasets may appear to have many missing or inconsistent values. Some information fields, while not directly affecting the target variable, can still have an indirect relationship with other factors, making traditional data cleaning complicated and time-consuming.

To solve this problem, the proposed approach apply a pre-processing process including: identifying important data fields, processing missing values using statistical methods (mean, median, mode), normalizing data to the same scale, and encoding categorical variables. After preprocessing, the data is rearranged to fit training with classical machine learning models such as **Support Vector Machine (SVM**), **Logistic Regression (LR), Random Forest (RF),** and **XGBoost**.

The original dataset had a class imbalance, in which the number of samples belonging to the "occupational disease" group was significantly less than that of the "no disease" group. the proposed approach do not apply re-sampling to avoid altering the natural distribution of the data.

From there, the proposed approach build four model versions corresponding to SVM, LR, RF, and XGBoost algorithms, each trained using their respective built-in class balancing strategies. The objective is to compare the influence of these balancing methods on the performance of the models, based on indicators such as Accuracy, Precision, Recall, F1-score and AUC-ROC, thereby selecting the optimal model and data imbalance handling strategy for the occupational disease prediction problem.

B. Machine Learning Models for Occupational Disease Detection

Four machine learning models are deployed in a unified pipeline of preprocessing steps (training set only), training with default parameters, and evaluation on test sets [4], [5]. The pipeline is built to be compatible with scikit-learn to ensure repeatability and direct comparison between algorithms [14].

To address data imbalance, the proposed approach apply the *class_weight='balanced'* parameter for the LR and SVM models. For RF, the proposed approach utilize *BalancedRandomForestClassifier* from the *imblearn* library. Lastly, for XGBoost, the proposed approach calculate the *scale_pos_weight* parameter as the ratio of negative to positive cases.

In this comparison and evaluation of the performance of different classic models, the proposed approach aim to minimize changes to model configurations and rely primarily on the default parameters provided by the sklearn and xgboost libraries. The reason for the decision to simplify this problem is that in order for a model to become a mass method used by people with a variety of knowledge bases (possibly non-algorithmic), it must be easy to use and do not require complex interventions that require in-depth knowledge (such as algorithm optimization) [16].

C. Features and Fusion

The features used in the study were structured similarly to the original article [3] and were divided into four main groups: *numerical*, *categorical*, habits and lifestyle (*categorical*) and disease symptoms (*categorical*).

To create uniform feature vectors suitable for classical models, the proposed approach convert the classification features by *one-hot encoding* when the number of levels is reasonable, or by *frequency/target encoding* for columns with high level diversity; the variables are continuously normalized [7].

In addition to primitive processing, the proposed approach take a number of *feature engineering steps* to improve the model's ability to differentiate, including: synthesizing occupational exposure indicators into synthetic exposure indexes, building variations that show the number of reported symptoms (*symptom count*), hierarchical *seniority* and age by group to capture non-linear effects, and create pairwise interactions between important variables (e.g. interactions between smoking and dust exposure).

To improve the interpretability of the models, the proposed approach use **Permutation Feature Importance** for the SVM model. For the other models, the proposed approach rely on their native mechanisms: *.coef_* for Logistic Regression, and *.feature_importances_* for Random Forest and XGBoost. These methods provide a direct measure of each feature's contribution to the model's predictions. Additionally, in the case of Logistic Regression, the proposed approach interpret the coefficients in the context of regression analysis to understand the relative influence of features on the risk of occupational disease [12].

D. Experimental Setup

The experiment is designed to allow fair comparison between models and is compatible with the evaluation structure of the

original paper [3]. Given the class imbalance (2.63% minority), the proposed approach performed power analysis on recall to detect an improvement from random guessing (recall ~0.5) to an expected recall of 0.8. Using a significance level (α) of 0.01 and power of 0.9, the proposed approach found that at least 71 minority class samples are needed. This corresponds to approximately 2700 total samples. To satisfy this requirement, the proposed approach adopted a **65/35 train-test split** with stratification to preserve the original class distribution, resulting in approximately 2810 test samples.

The training and test sets still retain the original distribution of 7,819:211 to reflect the actual conditions [14], [15].

Evaluation metrics include **Accuracy**, **Precision**, **Recall**, **F1-score**, **AUC-ROC**, and confusion matrix; in addition, the proposed approach report *macro-averages* and *weighted-averages* to represent performance on a per-layer basis [1], [12]. Each experiment was repeated several times with different random seeds to verify the stability of the results, and the mean values and standard deviations were reported.

The compute environment uses a configuration that is compatible with the original post for ease of comparison: Python 3.x with *scikit-learn*, *xgboost*; the heavy processing is done on a 2.8 GHz Intel Xeon system, the GPU is not required for classic models but is still reserved in case of XGBoost optimization.

4. EXPERIMENT AND DISCUSSION

E. Dataset

In this study, the dataset consisted of a total of 8,030 samples, each assigned a binary label indicating the subject's health status: healthy or suffering from an occupational disease. The data structure showed a severe imbalance when there were 7,819 negative samples (healthy) compared to only 211 positive samples (sick), corresponding to a ratio of 37:1. To overcome this problem, the proposed approach apply appropriate preprocessing techniques to each algorithm tested. The data is divided into 65% for training and 35% for testing, to ensure a comprehensive evaluation of the model's effectiveness.

Datasets exist in many different types of data, including numerical data and written data (taxonomy variables). There are many columns that are missing data due to incomplete or missing data collection.

After the coding and normalization step, the data for each model is in a different format (LR/SVM uses one-hot encoding; RF/XGBoost uses ordinal encoding).

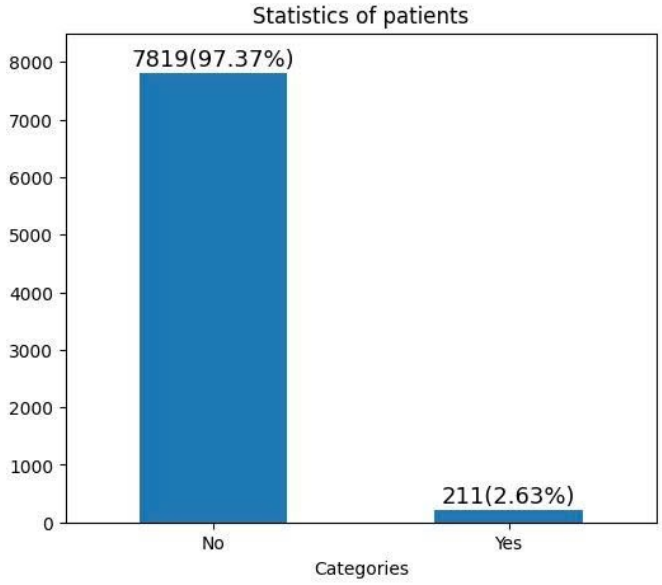

Fig. 5. Dataset object class distribution

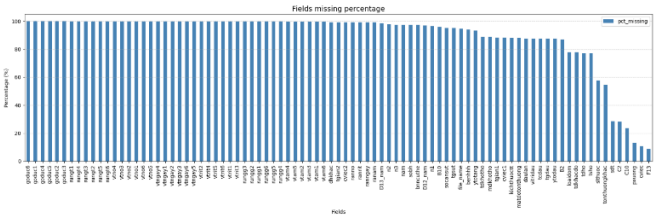

Fig. 6. Missing data fields and missing rates

F. Preprocessing

Because the two groups of LR/SVM and RF/XGBoost algorithms process data and computation differently, there will be 2 different data processing processes for the model to work most efficiently

First of all, raw data is initially processed by filtering out identity fields or sensitive personal information to ensure privacy, and at the same time standardize the data type for uniformity. For some specific medical variables, the encryption process is done manually (manual encoding) based on specialized knowledge. In addition, classification variables with a large number of values and containing many rare values (such as *cviec*, *pxuong*, *cviec1* variables) will be processed using frequency encoding: rare groups are grouped into a common group labeled "Other", which reduces noise and increases the generality of the model. Next, the data is divided

into a training set and a test set. This study applied a *train-test split ratio* of 65%-35%. All coding and processing steps are only learned in the training set, and then applied to the test set, in order to avoid data leakage. Logistic Regression and Support Vector Machine models require data to be represented in binary form, so the classification variables will be encoded using one-hot encoding. In contrast, decision tree models such as Random Forest and XGBoost handle data in an orderly form well, so the classification variables are encoded with ordinal encoding. For models that require standard input data (LR and SVM), the dataset will be normalized by filling in the missing values with mean imputation, and then normalizing the features according to the Standard Scaler (bringing to mean 0, standard deviation 1). For both the Random Forest and XGBoost models, the proposed approach apply **ordinal encoding** to handle categorical features. Specifically for Random Forest, the proposed approach also perform **median and mode imputation** to address missing values, as it does not natively support them. In contrast, XGBoost can handle missing values internally, but the proposed approach still apply the same ordinal encoding to ensure a fair comparison across models. Thanks to the above consistent and controlled processing, the input data of the model is guaranteed in terms of integrity, representation, and generalization.

Note: In these baseline tests, the proposed approach did not apply re-sampling to avoid altering the natural distribution of the data; instead, the imbalance was reflected in the class 1 precision/recall/F1 metrics.

G. Experiment Setup

The experiments were conducted in a Python 3.12 environment with the scikit-learn libraries (version ≥ 1.4) and XGBoost (version ≥ 1.7). The hardware system consists of an Intel Xeon CPU and 16GB of RAM, which does not use a GPU because classic machine learning models are still highly efficient when running on CPUs. Models are trained with default parameter configurations, including: Logistic Regression (solver = lbfgs, max_iter = 100), SVM with RBF kernel (C = 1.0), Random Forest (n_estimators = 100, max_depth = None), and XGBoost (learning_rate = 0.3, max_depth = 6, n_estimators = 100, eval_metric = 'logloss'). The performance of the model is evaluated based on the Accuracy, Precision, Recall, and F1-score indicators for each layer, along with Macro F1 and Weighted F1. These indicators are calculated according to standard formulas, where TP, TN, FP and FN represent the number of correct positive, negative, positive and false predictions, respectively.

Table 1 Comparing the performance of classic models on the test set

Metric	LR	SVM	RF	XG-Boost
Accuracy	0.956599	0.979723	0.930274	**0.983991**
Precision (class 1)	0.337748	0.642857	0.254980	**0.725806**
Recall (class 1)	0.698630	0.493151	**0.876712**	0.616438
F1-score (class 1)	0.455357	0.558140	0.395062	**0.666667**
Precision (class 0)	0.991729	0.986570	**0.996484**	0.989814
Recall (class 0)	0.963477	0.992695	0.931702	**0.993791**
F1-score (class 0)	0.977399	0.989623	0.963005	**0.991799**
Macro F1-score	0.716378	0.773881	0.679033	**0.829233**
AUC	0.91	0.96	0.96	0.96

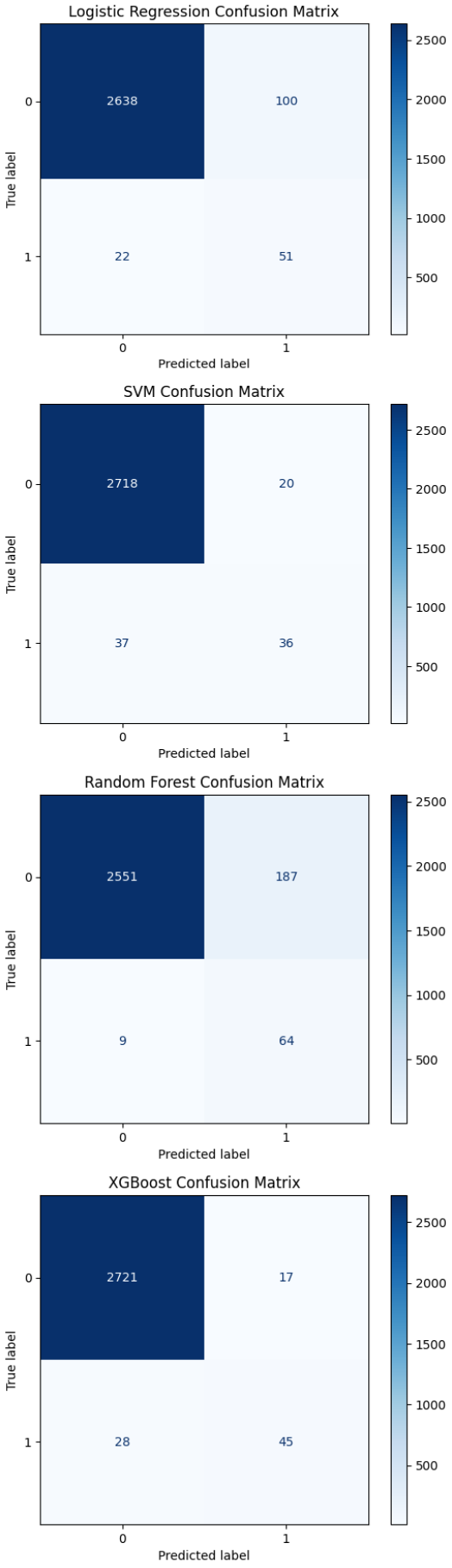

Fig. 7. Confusion matrix of models

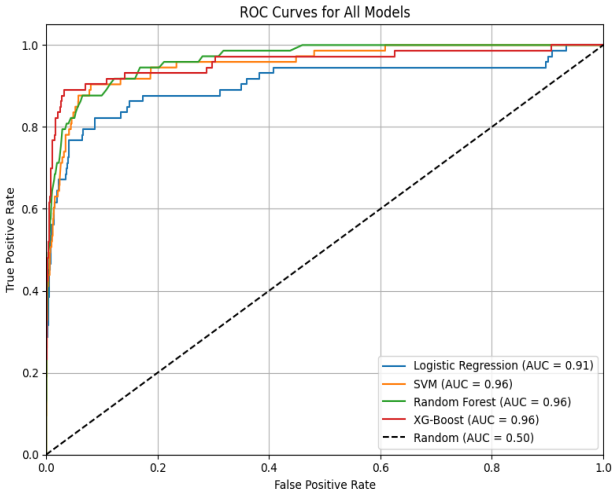

Fig. 8. ROC curve of models

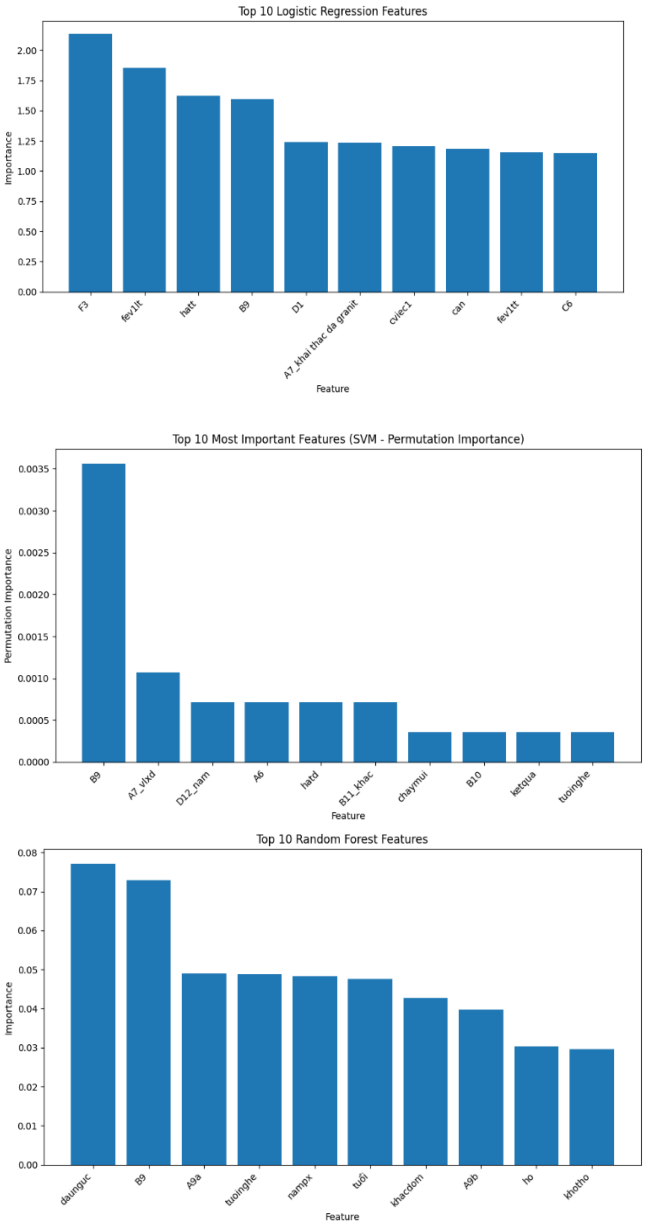

Fig. 9. Feature Importances of models

XGBoost achieved the highest Accuracy and Macro F1 (0.983991; 0.829233). Precision (class 1) is high (0.725806), indicating a few false positive predictions; Recall (1) = 0.616438 reflects that a significant portion of cases are still missing—indicating that the model is still missing a significant portion of cases. This may be due to the model not being well generalized for positive data, and because of the class imbalance that makes the model biased towards negative predictions. SVM ranks second in Macro F1 (0.773881), classifying efficiency close to XGBoost in Precision (1) but still low in Recall (1) (0.493151). RF has the highest Precision (1)= 0.996484, but low Precision (1) leads to poor F1 (1). LR for Recall (1)= 0.698630 but Precision (1)= 0.337748 is very low. And when comparing AUC, SVM, RF and XGBoost perform similarly, beside LR which has the lowest AUC (0.91).

The proposed approach extract the top 10 most important features from each model to compare how they learn from the data. XGBoost and its tree-based counterpart, Random Forest, highlight similar features, notably **chest pain** (with the highest importance), **age**, and **B9**, a medical indicator. In contrast, Logistic Regression prioritizes different features, including the occupational exposure variable **"A7_khai thác đá granit"**, suggesting its sensitivity to linear relationships. SVM, which underperformed relative to the other models, produced a distinct set of top features, indicating a different and potentially less effective representation of feature importance. These findings align with existing occupational health research and offer practical insights that may assist medical professionals in identifying high-risk individuals more effectively.

5. CONCLUSION

In this study, the proposed approach conducted a comparative benchmark of four classical machine learning models: Logistic Regression, Support Vector Machine, Random Forest, and XGBoost—for the prediction of occupational lung disease on a real-world Vietnamese dataset. The results demonstrated that XGBoost consistently achieved the best overall performance, with the highest Macro F1-score and AUC-ROC, followed closely by Logistic Regression. While Random Forest achieved strong precision on the minority class, its recall was limited. SVM, although effective in certain settings, exhibited the weakest recall.

Beyond predictive performance, SHAP-based interpretability analysis revealed that occupational exposure variables, chest pain, shortness of breath and age group were among the most influential features in predicting disease risk. These findings are consistent with occupational health literature and provide practical insights for medical professionals.

Overall, this study provides two key contributions: (i) a real-world evaluation of classical machine learning models on occupational lung disease screening in Vietnam and (ii) interpretable results that highlight actionable health and workplace risk factors. Future work should extend this benchmark with larger and multi-source datasets, integrate probability calibration for risk scoring, and explore deep learning approaches for complementary analysis.

6. REFERENCES

[1] M. M. Islam, M. R. Haque, H. Iqbal, M. M. Hasan, M. Hasan, and M. N. Kabir, "Breast cancer prediction: a comparative study using machine learning techniques," SN Computer Science, vol. 1, pp. 1–14, 2020.

[2] V. Ramalingam, A. Dandapath, and M. K. Raja, "Heart disease prediction using machine learning techniques: a

survey," International Journal of Engineering & Technology, vol. 7, no. 2.8, pp. 684–687, 2018.

[3] K. Nguyen-Trong, T. Vu-Van, P. Luong Thi Bich, "Graph Convolutional Network for Occupational Disease Prediction with Multiple Dimensional Data," International Journal of Advanced Computer Science and Applications, vol. 15, no. 7, 2024.

[4] K. Pingale, S. Surwase, V. Kulkarni, S. Sarage, and A. Karve, "Disease prediction using machine learning," International Research Journal of Engineering and Technology (IRJET), vol. 6, no. 12, pp. 831–833, 2019.

[5] G. Sailasya and G. L. A. Kumari, "Analyzing the performance of stroke prediction using ML classification algorithms," International Journal of Advanced Computer Science and Applications, vol. 12, no. 6, 2021.

[6] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," arXiv preprint arXiv:1609.02907, 2016.

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

[8] L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5–32, 2001.

[9] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD, pp. 785–794, 2016.

[10] K. Nguyen-Trong, T. Vu-Van, P. Luong Thi Bich, "Graph Convolutional Network for Occupational Disease Prediction with Multiple Dimensional Data," International Journal of Advanced Computer Science and Applications, vol. 15, no. 7, 2024.

[11] A. M. Barhoom, A. Almasri, B. S. Abu-Nasser, and S. S. Abu-Naser, "Prediction of heart disease using a collection of machine and deep learning algorithms," 2022.

[12] N. Biswas, K. M. M. Uddin, S. T. Rikta, and S. K. Dey, "A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach," Healthcare Analytics, vol. 2, p. 100116, 2022.

[13] M. Ashrafuzzaman, S. Saha, and K. Nur, "Prediction of stroke disease using deep CNN based approach," Journal of Advances in Information Technology, vol. 13, no. 6, 2022.

[14] J. Prusa, T. M. Khoshgoftaar, D. J. Dittman, and A. Napolitano, "Using random undersampling to alleviate class imbalance on tweet sentiment

[15] Data," in 2015 IEEE International Conference on Information Reuse and Integration, IEEE, pp. 197–202.

[16] Z. Zheng, Y. Cai, and Y. Li, "Oversampling method for imbalanced classification," Computing and Informatics, vol. 34, no. 5, pp. 1017–1037, 2015.

[17] Couronné, R., Probst, P. & Boulesteix, AL. "Random forest versus logistic regression: a large-scale benchmark experiment". BMC Bioinformatics 19, 270 (2018). https://doi.org/10.1186/s12859-018-2264-5