Development of Machine Learning Model for Diagnostic Disease Prediction Based on Laboratory Tests

Dong Jin Park
Department of Laboratory Medicine, College of Medicine, Ewha Womans University of Korea
https://orcid.org/0000-0002-2412-5292

Min Woo Park
Department of Laboratory Medicine, St.Vincent's Hospital, The Catholic University of Korea
https://orcid.org/0000-0002-3751-7519

Homin Lee
Future.lab of Korea https://orcid.org/0000-0001-9970-8073

Young-Jin Kim
Finance.Fishery.Manufacture Industrial Mathematics Center on Big Data, Pusan National University
https://orcid.org/0000-0002-7859-5521

Yeongsic Kim
Department of Laboratory Medicine, College of Medicine, The Catholic University of Korea
https://orcid.org/0000-0001-5815-5185

Young Hoon Park (✉ carrox2yh@gmail.com)
Division of Hematology, Department of Internal Medicine, College of Medicine, The Catholic University of Korea https://orcid.org/0000-0003-4516-7990

Method Article

Keywords: Machine learning, Deep neural network, Disease prediction, Laboratory test

Posted Date: April 16th, 2021

DOI: https://doi.org/10.21203/rs.3.pex-1429/v1

Abstract

Artificial intelligence is a concept that includes machine learning and deep learning. The deep learning model used in this study corresponds to DNN (deep neural network) by utilizing two or more hidden layers. In this study, MLP (multi-layer perceptron) and machine learning models (XGBoost, LGBM) were used. An MLP consists of at least three layers: an input layer, a hidden layer, and an output layer. In general, tree models or linear models using machine learning are widely used for classification. We analyzed our data by applying deep learning (MLP) to improve the performance, which showed good results. The deep learning and ML models showed differences in predictive power and disease classification patterns. We used a confusion matrix and analyzed feature importance using the SHAP value method. Here, we present a protocol to confirm that the use of deep learning can show good performance in disease classification using hospital numerical structured data (laboratory test).

Introduction

Deep learning (DL) is a subset of ML that differs from other ML processes in many ways. Most ML models perform well due to their custom-designed representation and input features. Using the input data generated through that process, ML learns algorithms, optimizes the weights of each feature, and optimizes the final prediction. DL attempts to learn multiple levels of representation using a hierarchy of multiple layers. The deep neural network (DNN) is a type of DL that uses multiple hidden layers and is renowned for analysis of high-dimensional data. In practice, the symptoms described by patients, physical examinations performed by physicians, laboratory test results, and imaging studies such as X-ray and computed tomography (CT) are generally needed to evaluate a patient's status and diagnose a specific disease. However, little research has been conducted into the predictive power and accuracy that can be achieved using laboratory data alone for the diagnosis of specific diseases. Therefore, the purpose of this study was to develop predictive models that can be used by physicians to make decisions in the hospital setting based on DL and ML using laboratory data alone, and then to validate our model through comparison of its predictions with the diagnoses of physicians. In addition, we generated an ensemble of DL and ML models to improve performance. The Shapley additive explanation (SHAP) method, which was recently developed, was used to determine the features that are important to each disease and to identify predictive relationships between diseases and features.

Reagents

No reagents are required for this protocol. We used python library. These datasets were used to construct light gradient boosting machine (LightGBM) and extreme gradient boosting (XGBoost) ML models and a DNN model using TensorFlow and Keras.

Equipment

A PC (CPU must have 8 logical processors, memory must be greater than 8 G) can be used to build a Machine Learning Model for Diagnostic Disease Prediction Based on Laboratory Tests.

Procedure

The main purpose of this study is to develop disease prediction models to quickly and

accurately turn data into diagnosis. Therefore, this study developed machine learning, deep learning,

and ensemble models for 39 diseases classification (Supplement Table S9) of patients visiting the

emergency room using 88 laboratory test parameters including blood and urine tests (Supplement Table

S1). The overall workflow of disease prediction model based on laboratory tests (DPMLT) is schematically demonstrated in Figure 1. This protocol is largely composed of 5 parts, and the third part explains the machine learning model and the deep learning model.

1.0 Data collection and preprocessing

We collected anonymized laboratory test datasets, including blood and urine test results, along with each patient's final diagnosis on discharge. We curated the datasets and selected 86 attributes (different laboratory tests) based on value counts, clinical importance-related features, and missing values. For Deep learning (DL), missing values were replaced with the median value for each disease.

2.0 Feature extraction

Feature extraction plays a major role in the creation of machine learning (ML) models.

3.0 Model selection and training

3.1 DL selection

The research in this study was conducted using a deep neural network (DNN) for structured data.

3.2 MLP (multi-layer perceptron)

All features used in this study are numeric data except for the 'sex' feature. MLP recognizes only numerical data, so we transformed the categorical feature of 'sex' into a number using LabelEncoder of the scikit-learn library. MLP does not allow for null values, so we replaced null values with the median value of each feature.

3.3 Feature normalization

Each feature had a different range. We applied a standard scale to normalize the mean and standard deviation of each feature to (0, 1) by subtracting the mean value of the feature and dividing by its standard deviation value.

3.4 Hidden layer composition

In our study, the hidden layer was comprised of two layers. We employed the Relu (rectified linear unit) activation function for each layer. We applied the dropout technique to each hidden layer, which is a simple method to prevent overfitting in neural networks.

3.5 XGBoost

XGBoost is an algorithm that overcomes the shortcomings of GBM (gradient boosting machine). The disadvantages of GBM include long learning times and overfitting problems. The most common ways to solve these problems are through parallelization and regularization. Our dataset contained null values, which MLP replaced with the corresponding median values, but XGBoost has a procedure to process null values, so utilized that procedure. The max_depth argument in XGBoost is one factor determining the depth of the decision tree. Setting max_depth to a large number increases complexity and can lead to overfitting. This study found that max_depth was optimally set to 2.

3.6 LightGBM

The difference between LightGBM and XGBoost is the method by which the tree grows. XGBoost creates a deeper level within the leaf (level-wise/depth-wise), and LightGBM generates a leaf at the same level (leaf-wise). LightGBM uses a leaf-centered tree-splitting method to split leaf nodes with the maximum loss value, creating an asymmetric tree. To avoid overfitting in LightGBM, an experiment was conducted by adjusting num_leaves and min_child_samples.

3.7 Ensemble model results (DNN, ML)

We developed a new ensemble model by combining our DL model with our two ML models to improve AI performance. We used the validation loss for model optimization.

4.0 K-Fold Cross-validation

In our study, we divided a total of 5145 datasets at a ratio of 8:2 to create the training set and test set. We set the validation data ratio to 0.2 for the training set, which was evaluated using validation loss for model optimization based on the training data. If the number of validation data is increased, the number of training data decreases, leading to a problem of high bias. We used k-fold cross validation to prevent data loss of the training set.

5.0 SHAP (Shapley Adaptive Explanations)

SHAP is an acronym for Shapley Adaptive Explanations. Relating to the Shapley value, as the name suggests. In our experiment of MLP, we can calculate SHAP value using DeepLIFT.

Troubleshooting

Regarding the overfitting issue, we separated about 1029 cases (independent data) from the total cases using python library “train test split”. Because our data set is imbalance data, we randomly and evenly selected each specific disease from the data set using stratify option.

Time Taken

The full development of disease prediction model based on laboratory tests (DPMLT) work, including design, data collection, preprocessing, and all stages, was undertaken over a period of approximately 12-15 months (2019-20).

Anticipated Results

We investigated a total of 39 specific diseases based on the International Classification of Diseases, 10th revision (ICD-10) codes. We aimed to build a new optimized ensemble model by blending a DNN (deep neural network) model with two ML models for disease prediction using laboratory test results. This study will be useful in the prediction and diagnosis of diseases.

References

1. Kwon, K., Kim, D. & Park, H. A parallel MR imaging method using multilayer perceptron. *Med Phys* 44, 6209-6224 (2017).

2. Shrikumar, A., Greenside, P. & Kundaje, A. Learning important features through propagating activation differences. in *Proceedings of the 34th International Conference on Machine Learning - Volume 70* 3145–3153 (JMLR.org, Sydney, NSW, Australia, 2017).

3. Poernomo, A. & Kang, D.K. Biased Dropout and Crossmap Dropout: Learning towards effective Dropout regularization in convolutional neural network. *Neural Netw* 104, 60-67 (2018).

4. Deng, L.*, et al.* PDRLGB: precise DNA-binding residue prediction using a light gradient boosting machine. *BMC bioinformatics* 19, 522 (2018).

5. Khan, S.H., Hayat, M. & Porikli, F. Regularization of deep neural networks with spectral dropout. *Neural Netw* 110, 82-90 (2019).

Acknowledgements

This work was supported by the Korean Society of Medical Informatics.

Figures

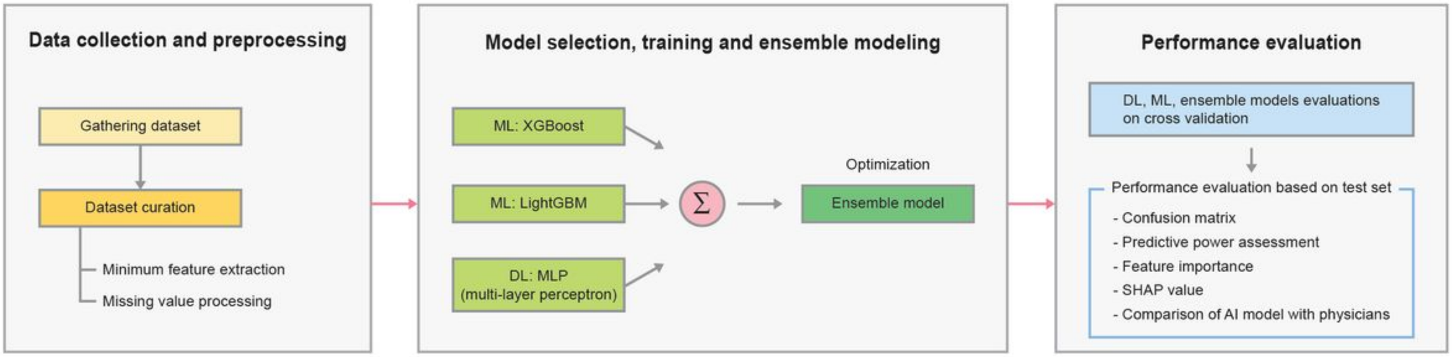
Data collection and preprocessing
Gathering dataset
Dataset curation
Minimum feature extraction
Missing value processing
Model selection, training and ensemble modeling
ML: XGBoost
ML: LightGBM
DL: MLP
(multi-layer perceptron)
Σ
Optimization
Ensemble model
Performance evaluation
DL, ML, ensemble models evaluations on cross validation
Performance evaluation based on test set
- Confusion matrix
- Predictive power assessment
- Feature importance
- SHAP value
- Comparison of AI model with physicians

Figure 1

Overall framework of DPMLT. The development of DPMLT methodology involved for three major steps. (1) Data collection and preprocessing (2) Model selection, training and ensemble modeling (3) Performance evaluation

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTable9label.docx
- SupplementaryTable1label.docx