International Journal of Internet, Broadcasting and Communication Vol.15 No.3 32-42 (2023)
http://dx.doi.org/10.7236/IJIBC.2023.15.3.32

IJIBC 23-3-5

Human Normalization Approach based on Disease Comparative Prediction Model between Covid-19 and Influenza

Janghwan Kim[1], Min-Yong Jung[2], Da-Yun Lee[3], Na-Hyeon Cho[4], Jo-A Jin[5], R. Young-Chul Kim[6*]

[1]Ph.D. candidate, Software Engineering Laboratory,
Dept. of Software and Communication Engineering, Hongik University, Republic of Korea
[2,3,4,5]B.S student, Dept. of Software and Communication Engineering,
Hongik University, Republic of Korea
[6]Professor, Software Engineering Laboratory,*
Dept. of Software and Communication Engineering, Hongik University, Republic of Korea
E-mail: [1]janghwan.kim@g.hongik.ac.kr, [2]pw050503@naver.com, [3]dana0417@naver.com,
[4]whskgus12@nate.com, [5]jinjoa1@naver.com, [6]bob@hongik.ac.kr*

Abstract

There are serious problems worldwide, such as a pandemic due to an unprecedented infection caused by COVID-19. On previous approaches, they invented medical vaccines and preemptive testing tools for medical engineering. However, it is difficult to access poor medical systems and medical institutions due to disparities between countries and regions. In advanced nations, the damage was even greater due to high medical and examination costs because they did not go to the hospital. Therefore, from a software engineering-based perspective, we propose a learning model for determining coronavirus infection through symptom data-based software prediction models and tools. After a comparative analysis of various models (decision tree, Naive Bayes, KNN, multi-perceptron neural network), we decide to choose an appropriate decision tree model. Due to a lack of data, additional survey data and overseas symptom data are applied and built into the judgment model. To protect from thiswe also adapt human normalization approach with traditional Korean medicin approach. We expect to be possible to determine coronavirus, flu, allergy, and cold without medical examination and diagnosis tools through data collection and analysis by applying decision trees.

Keywords: *COVID-19, Machine Learning, Traditional Korean Medicine,*

1. Introduction

Due to the global coronavirus epidemic, the world is spurring research to prevent infectious diseases. There are about 700 million confirmed coronavirus cases, and about 6.8 million deaths worldwide [1]. The Korean government has made various attempts to cope with the "COVID-19" crisis, including expanding the nationwide quarantine network for COVID-19. However, as of January 2023, the cumulative number of COVID-19 confirmed cases exceeded 30 million and the cumulative death toll exceeded over 30,000 [2]. According to the World Health Organization (WHO), coronaviruses are difficult to identify compared to the

Manuscript Received: May. 11, 2023 / Revised: May. 15, 2023 / Accepted: May. 18, 2023
Corresponding Author: bob@ hongik.ac.kr
Tel: +82-44-860-2477, Fax: +82-44-865-0460
Professor, Software Engineering Laboratory, Department of Software and Communication Engineering, Hongik University, Republic of Korea

flu or other bronchial-based diseases [3]. This is because the symptoms of the above diseases are very similar [4]. Therefore, physical tests such as PCR tests or self-diagnosis kits are essential to determine whether a person is infected with each disease. However, it is difficult to determine whether an infection is present through testing in cases where high medical and testing costs are borne, such as in the United States.

To solve these problems, we propose a software prediction model based on infectious disease-related symptom data learning. This method collects and analyzes data based on a decision tree model to determine the diseases of corona, flu, allergy, and cold. Through this, it is expected that the public can contribute to the prevention and early response of infectious diseases by identifying diseases without the use of testing tools.

In this paper, Chapter 2 introduces related works and stochastic machine-learning algorithms. Chapter 3 describes the comparative prediction prototype model. Chapter 4 proposes a prevention guide for pulmonary infectious diseases such as influenza and COVID-19 based on the Sasang constitution. Chapter 5 refers to conclusions and future work.

2. Related Works

2.1 A Prototype model for comparative prediction of COVID-19 and pandemic influenza

In existing studies, COVID-19 and flu viruses are conducted using Naive Bayes and K-Nearest Neighbor Classification(KNN) [5].

Training Data Use data provided by Kaggle. To process the data, the classification targets are labeled as coronavirus and influenza A, and the model is trained with seven symptoms, such as fever, cough, shortness of breath, and sore throat, as attribute values for symptom data. Figure 1 shows the mechanism for the optimization model of the existing research. Existing studies show high accuracy of over 90% using Naive Bayes and KNN algorithms.

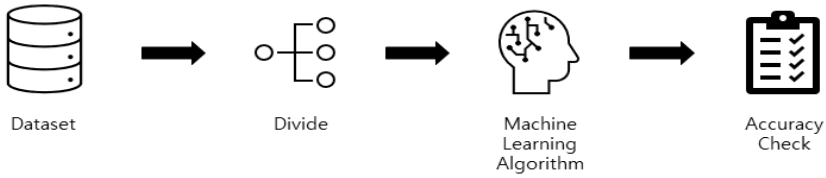

Figure 1. General Machine Learning Optimization Mechanism

2.2 A Decision Tree Classifier

Decision Tree Classifier is a machine learning supervised learning algorithm that uses a decision tree to

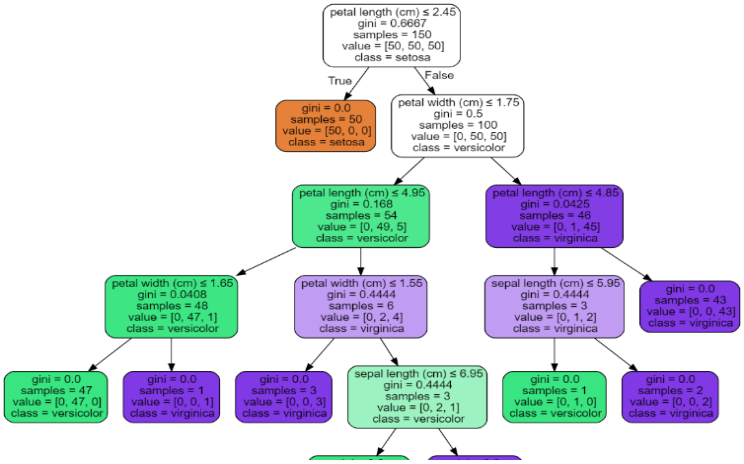

Figure 2. Visualization of Decision Tree Classifier

connect input and target variables. Both classification and regression are possible, and classification is applied when the target value is finite.

Figure 2 shows the result of the Decision Tree Classifier. It is usually made up of branches and notes like a tree. Currently, branches are represented as logical operations with true and false inputs [6]. There are typically two methods of calculating impurity in a decision tree: the Gini coefficient and the entropy. In the case of the Gini coefficient, important attributes can be identified through the corresponding values and used as a basis for classification, and only the most important attributes are determined. Reduce the training time by training the tree [7].

2.3 Naïve Bayesian

Naive Bayesian is a type of machine learning supervised learning algorithm. It is assumed that all feature values of the data used are independent of each other, and the probability of a new event is inferred and classified based on the previous and current events using Bayes' Theorem [8].

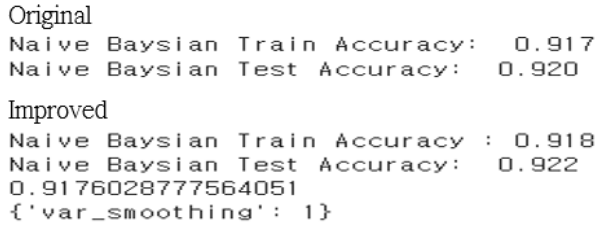

Figure 3. Improved Naive Bayesian with GridSearchCV

Figure 3 is the result of improving performance by applying GridSearchCV to the model trained with data [9]. GridSearchCV is mainly used to improve model performance in machine learning. As a result of applying the GridSearchCV algorithm by specifying a range from 0 to 10 improved accuracy from 92% to 92.2% when var_smoothing = 1.

2.4 K-Nearest Neighbors (KNN) Classifier

KNN is a type of machine learning supervised learning algorithm that uses K nearest neighbors. When new data comes in, information on the new data is predicted based on the Euclidean distance to the corresponding data through K neighbor data information of the closest existing data [10].

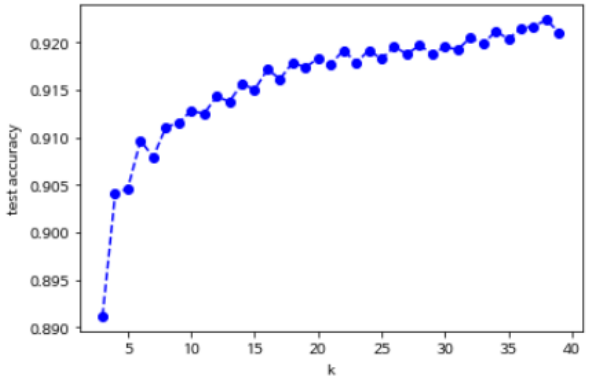

Figure 4. Performance Measurement Results on KNN

Figure 4 shows graphs that measure accuracy by applying our data to the KNN algorithm. In the case of KNN, since performance varies according to K, the number of nearest neighbors, accuracy is measured under the same conditions in the $3 \leq k \leq 40$ range. As a result, when the number k of nearest neighbors is 38, the accuracy is about 92%.

2.5 Multilayer Perceptron on Neural Network Model

The multilayer perceptron neural network has a hierarchical structure consisting of an input layer, one or more hidden layers, and finally, an output layer. Each layer of this hierarchy is connected in the input, hidden, and output directions. There is no direct connection between the connections inside each layer and between the input and output layers.

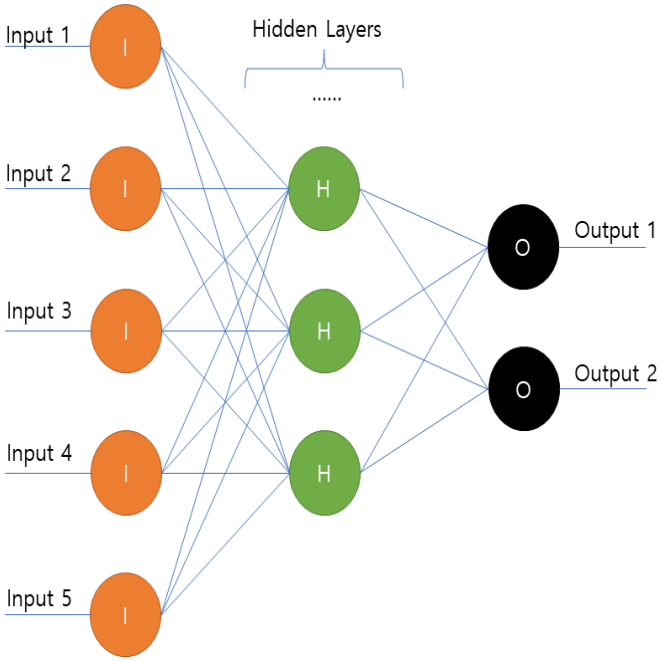

Figure 5. Multilayer Perceptron based on Neural Network

Figure 5 shows the relationship between the neural network layers and the input and output layers in the multi-perceptron neural network model. In this hierarchical structure, neurons belonging to each layer perform classification through weight calculation and activation function. The advantage of this model is that efficient learning is possible in the supervised learning environment of the classification model because even non-linear input is mapped so that linear classification is possible in the hidden layer [11].

2.6 Comparison between Diverse Models with Machine Learning Algorithms.

Table 1 shows a performance of the algorithms studied is compared to select a model suitable for the corona prediction judgment system. The comparison index is set to the classification accuracy of the test data, and the test data is 13352 data extracted 30% randomly (Random_state = 2) from the total data. Each model was created after assigning optimized hyperparameter values found through the GridSearchCV algorithm. As a result of performance comparison, all algorithms showed the same performance for prediction. Therefore, considering the possibility of future improvement, we apply the Decision Tree Classifier, which can visualize the logical structure of the algorithm, to the prototype model proposed above.

Table 1. Performance comparison of classification algorithms

ML Algorithm	Accuracy
Naive Bayesian	92.22 %
KNN	92.11 %
Decision Tree	92.02 %
Perceptron	92.07 %

2.7 Traditional Clinical Mechanism Research on Human Type Classification

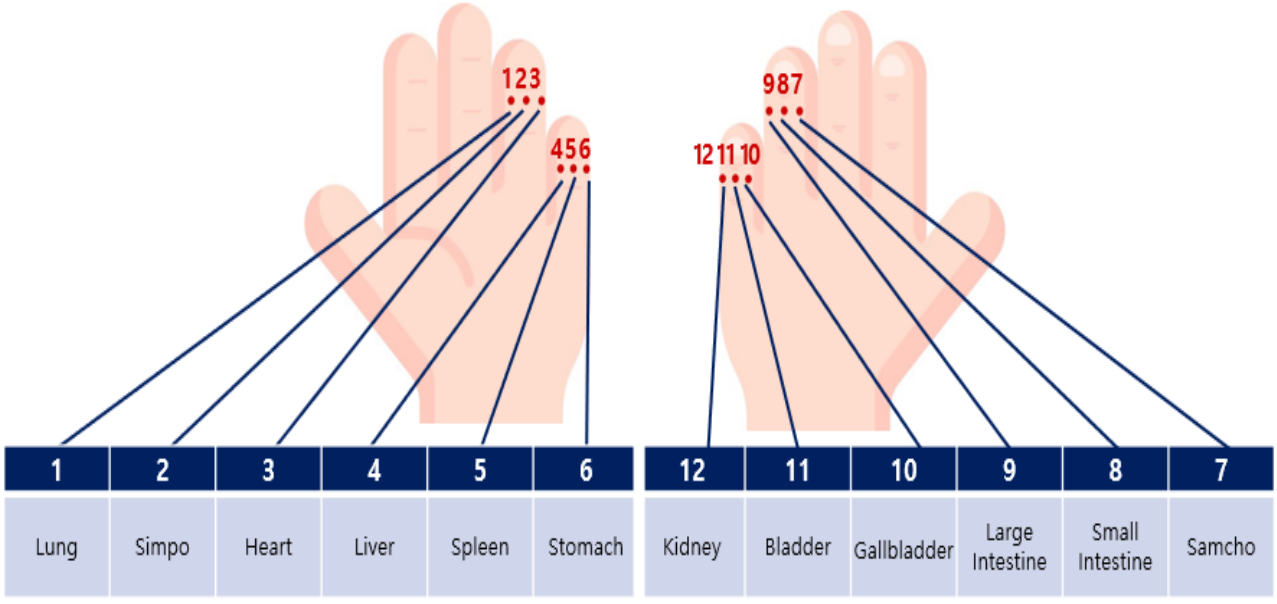

Figure 6. Korea Hand Acupuncture based on Seogeum Method

According to the eight constitutions, a person is classified according to the degree of innate activation of organs and organs such as the heart, lungs, pancreas, liver, kidneys, small intestine, large intestine, stomach, gallbladder, bladder, and sympathetic and parasympathetic nerves of the autonomic nerves. It is classified into 8 types of body constitutions [12]. A study on the meridian centered on the relationship between bioenergy and human body organs shows the result of a study that the electrical characteristics of the meridian are low and high potential [13]. Accordingly, the amount of current flowing in the living body is measured by measuring the current generated due to the potential difference, an electrical characteristic between the meridians. Figure 6 shows the parts where bioenergy can be measured for each body organ. In the order of the number indicated at each specific point, it is connected to the lungs, heart, heart, liver, spleen, stomach, trichomes, small intestine, large intestine, gall, bladder, and kidneys [14].

3. Comparative Prediction Prototype Model

3.1 Data Collection and Design

In this paper, we survey people who have had COVID-19 and the flu to use as experimental data. Through the results conducted from May 24, 2022, to May 28, 2022, and from August 21, 2022, to August 26, 2022, the type of infected mutant virus, age range (interval) of the patient, fever, cough, about 25 symptoms, including muscle pain, vomiting, sore throat, abdominal pain, diarrhea, runny nose, stuffy nose, loss of taste, loss of smell, and sneezing, are investigated. The model is trained through the data collected through this survey (163 corona cases and 25 flu cases).

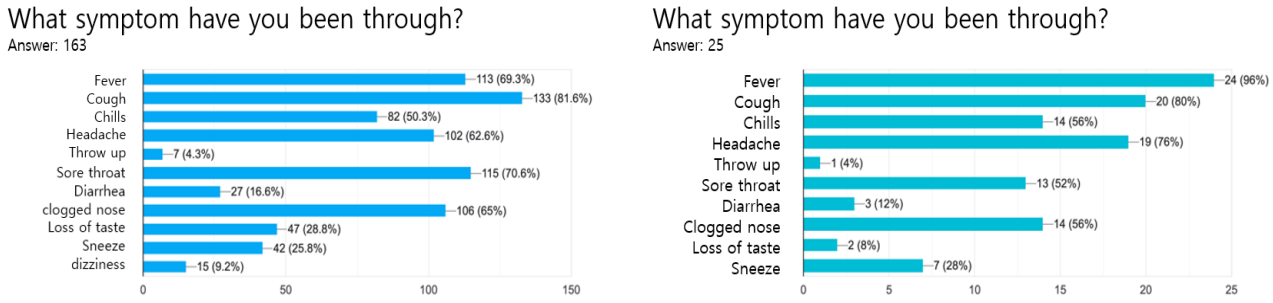

Figure 7. Survey result between Coronavirus and Flu patients

Figure 7 shows the results of a questionnaire about the symptoms of people who have had corona and flu. As a result of the survey response, fever, cough, and chills were the most common symptoms for both corona and flu, and in the case of corona, runny nose, nasal congestion, loss of taste and smell, and dizziness were higher than those of the flu.

As a result of learning with infectious disease symptom data collected through surveys, over fitting occurs in which the accuracy of the training data is high, but the error for the target data increases due to the lack of an absolute amount of data to be trained in the model. To solve this problem, add a dataset with symptom columns similar to the symptom data of the survey. The attributes of the dataset are about 20 symptoms, such as cough, muscle pain, fatigue, sore throat, runny nose, stuffy nose, fever, loss of taste, and loss of smell, and one symptom divided into severe corona, mild corona, allergy, cold, and general flu. It has a disease name column and contains a total of 44506 rows of data [15].

3.2 Data Preprocessing

Figure 8 is a heat map showing the correlation between data attribute values. Circle A(ⓐ) is a box with a thick line indicates a positive correlation, and Circle B(ⓑ) is a dotted line box which is a negative correlation. Fever, nausea, vomiting, and diarrhea show positive correlations. Itchy nose, itchy eyes, itchy throat, and ear pain show negative correlations. Since ±0.3 is a weak positive and negative linear relationship, our attribute values do not show a strong relationship. Therefore, it is judged that the problem of multicollinearity, in which some of the independent variables are expressed as a combination of other independent variables, will not occur.

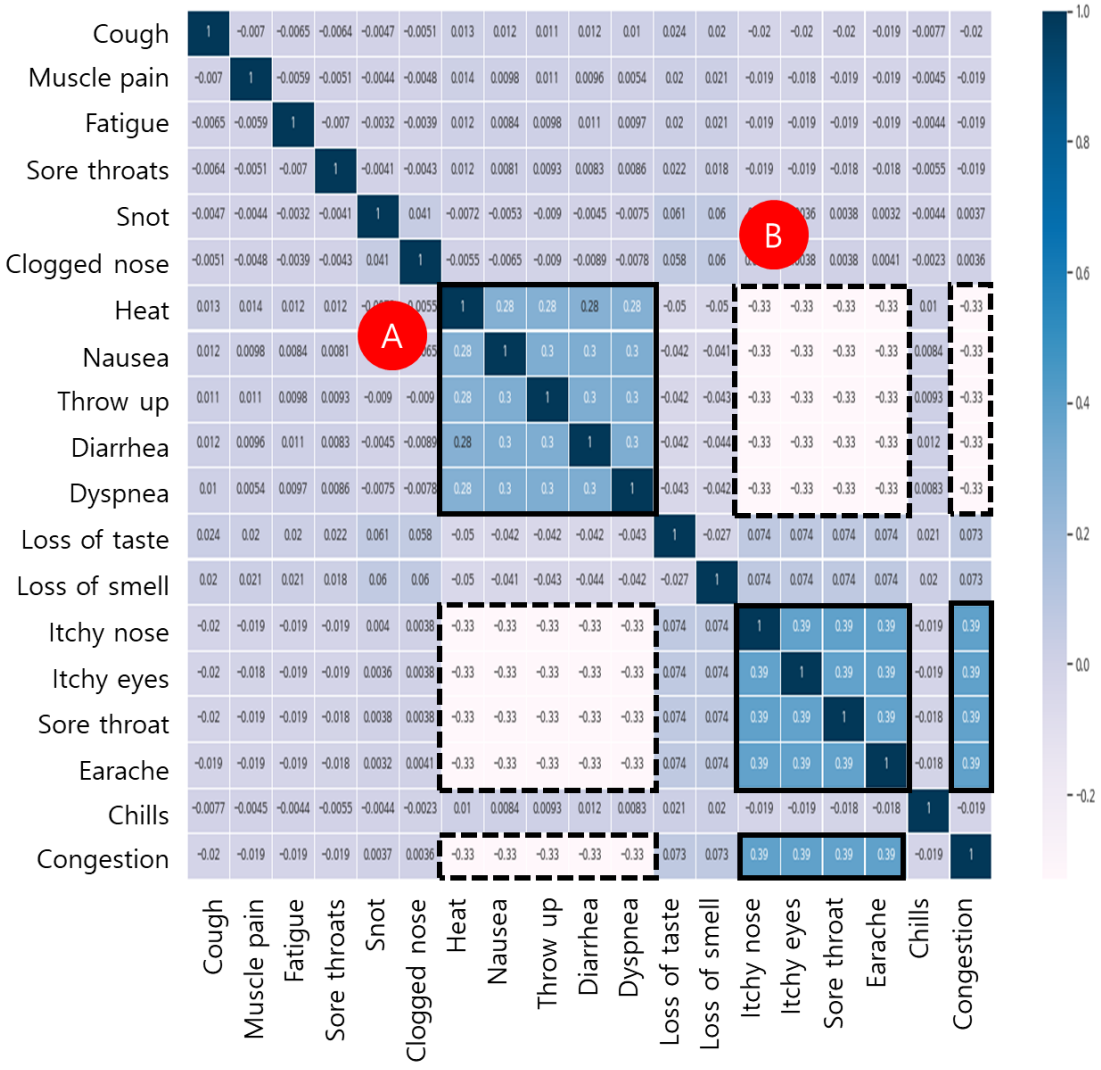

Figure 8. The correlation coefficient between variables of the x and y axis

There are a total of five disease names in the collected data: MILD COVID (mild), SEVERVE COVID (severe), COLD (cold), ALLERGY (allergy), and COMMON FLU (flu). The data were combined to construct one combined corona symptom data.

Table 2. Label encoding data conversion method

Label Encoder	Disease Type
0	ALLERGY
1	COLD
2	COMMON FLU
3	COVID

Table 2 shows the labeling of the four disease names, ALLERGY, COLD, COMMON FLU, and COVID, for data preprocessing after construction.

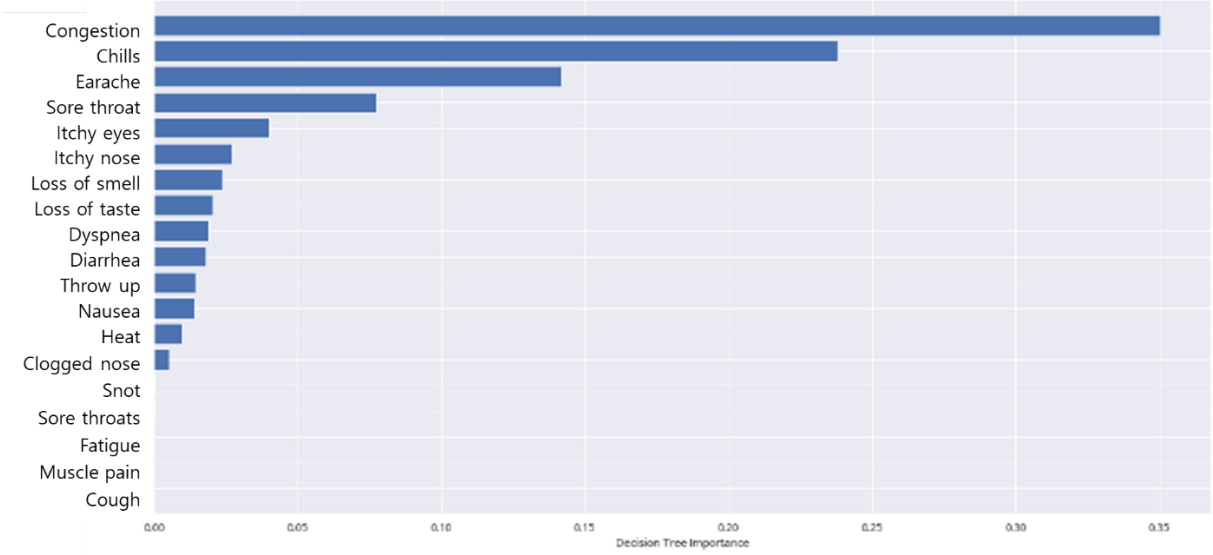

Figure 9. Variable Types of Decision Tree

Figure 9 graphs the importance of the decision tree after the decision tree. The importance of each feature is congestion, chills, earache, sore throat, itchy eyes, itchy nose, loss of smell, loss of taste, dyspnea, diarrhea, vomiting, nausea, fever, stuffy nose, runny nose, sore throat, fatigue, muscle pain, it consisted of cough, and congestion was the most important, followed by chills, ear pain, and sore throat in that order.

3.3 Structuring the Model

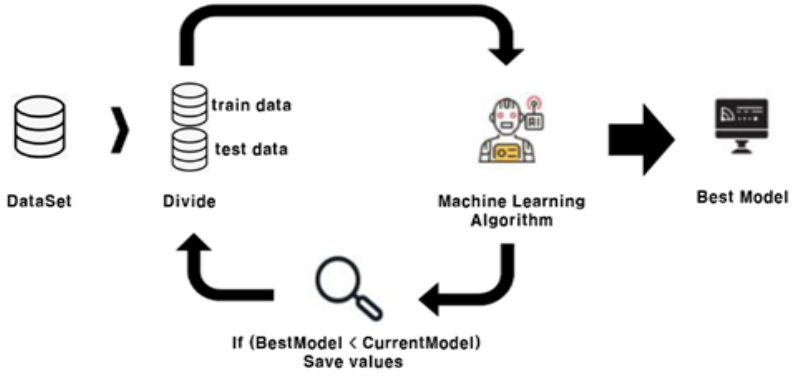

Figure 10. The Improvement Diagram of Model Decision Process

Figure 10 shows the model construction method to be applied. To maximize the model's performance, the ratio between the training dataset and the validation dataset is assigned from 1:9 to 3:7. Also, to prevent overfitting, the seed value can be generated as a random number so that the training data can be randomly selected.

In this paper, the optimal values were found and applied to the final model by comparing the model learning results based on the greedy algorithm using GreedsearchCV of the Python sklearn package for the parameters that exist for each learning algorithm and the ratio of the dataset.

```
[ hyperparameter tuning before ]
1. Decision Tree Train Accuracy:  0.942
1. Decision Tree Test Accuracy:  0.883

[ hyperparameter tuning after ]
2. Decision Tree Train Accuracy::  0.928
2. Decision Tree Test Accuracy: :  0.920
best_params : {'max_depth': 13, 'min_samples_leaf': 2}
```

Figure 11. Comparison Result of Decision Tree Classifier

We trained the model with a total of 44506 data and finally split the train data and test data at a ratio of 7:3. The first model in Figure 11 is a Decision Tree model created by default, and its accuracy is 88.3% based on the test data. The second model in Figure 4 is a model to which the GridSearchCV algorithm is applied to improve performance. The optimal parameter values found through the algorithm are max_depth=13 and min_samples_leaf=2, and the accuracy increased by about 4% from 88.3% to 92%.

3.4 Model Visualization

Figure 12 is a visualization of the judgment process of the model we created. The depth of the tree is 13 layers, and each node is set to have at least two branches. The root node is ′ear pain′, and if the symptom of each node is shown, it is connected to the right node, and in the opposite case, it is connected to the left node. We show through visualization that the prototype model is the dominant allergy and influenza classification model.

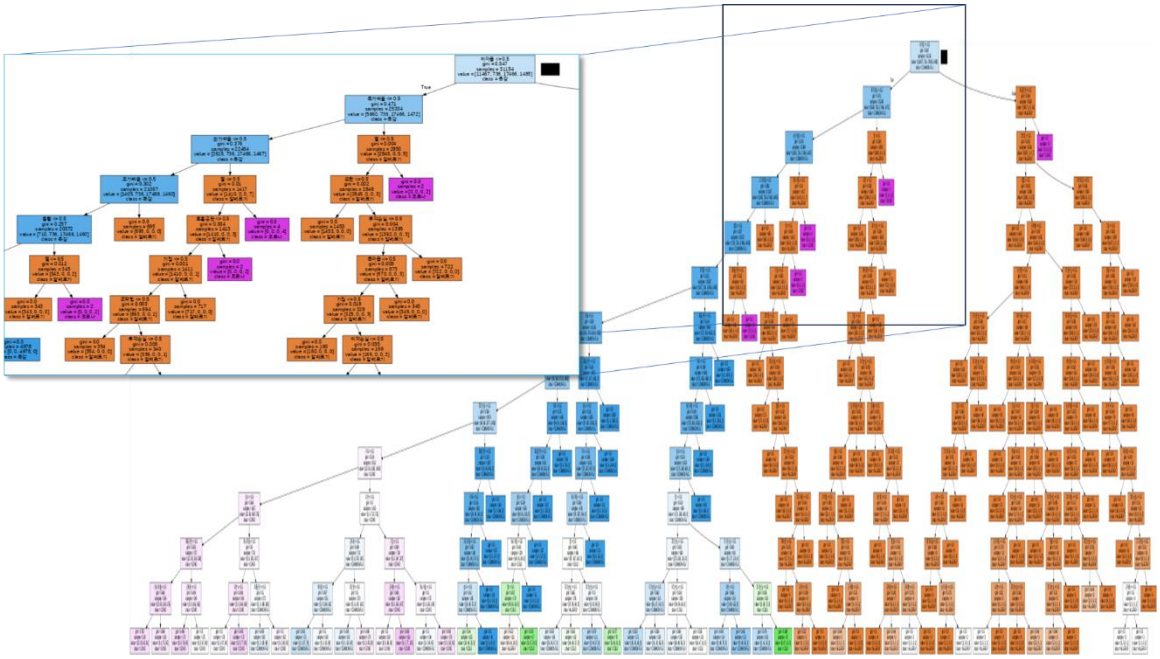

Figure 12 The Visualization Approach of Decision Tree Prototype Model

4. Bio-current pattern-based prevention guide for pulmonary infectious diseases

As shown in the above experiment results, four diseases (corona, flu, allergy, cold) are diseases related to the lungs. In the case of diseases related to the lungs, when lung function analyzes the bio-current pattern, the numerical value of menopausal blood appears higher than that of other intestinal blood. When the virus encounters the lungs, it easily creates inflammation. At this time, the function of the lungs is overworked because they fight against the immune components of the lungs. Therefore, when the function of the menopausal blood is lowered to achieve the equilibrium state of each intestinal blood, the body's immune system is activated, and the strength to fight the disease is gained.

5. Conclusion

This paper proposes a software prediction model based on symptom data to judge four diseases (corona, flu, allergy, cold). While the coronavirus has suffered great damage and the with corona continues, the model proposed in this paper applies a decision tree model to determine the above diseases. Through this study, it is expected that the public and people in areas where medical access is difficult will be able to more easily determine whether or not they are infected with the corona without a test tool.

As a result of using the model, it will be possible to take quick action, such as recommending a PCR test for those who are judged to be infected with the corona.

In the future, we expect to continuously collect data collection, add symptom data, and continuously train the model to improve performance. If reliable data are collected for other diseases besides the coronavirus, the model can be expanded by applying them as well. Afterward, web pages and applications will be developed to provide the service of the proposed model.

Accordingly, preventive intervention against the spread of contagious diseases is expected to be possible by determining whether there is a coronavirus infection.

Acknowledgement

This research was supported by Korea Creative Content Agency(KOCCA) grant funded by the Ministry of Culture, Sports and Tourism(MCST) in 2023(Project Name: Development of AI-based user interactive multi-modal interactive storytelling 3D scene authoring technology, Project Number: RS-2023-00227917, Contribution Rate: 50%) and Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education in 2023(Project Name: The Faultless Codinglization through Automatic Refactoring based on NLP BERT Model, Project Number: 2021R1I1A3050407, Contribution Rate: 50%)

References

[1] Global COVID-19 Tracker, https://www.kff.org/coronavirus-covid-19/issue-brief/global-covid-19-tracker/

[2] T. G. Yu, H. J. Lee, "A Study on the Establishment of a New Quarantine System in the COVID-19 Era." *International Journal of Internet, Broadcasting and Communication*, 15(1), pp275–280. (2023) DOI: https://doi.org/10.7236/IJIBC.2023.15.1.275

[3] B. Hu, H. Guo, P. Zhou et al. Characteristics of SARS-CoV-2 and COVID-19. Nat Rev Microbiol 19, pp141–154 (2021). DOI: https://doi.org/10.1038/s41579-020-00459-7

[4] Coronavirus disease (COVID-19): Similarities and differences between COVID-19 and Influenza https://www.who.int/news-room/questions-and-answers/item/coronavirus-disease-covid-19-similarities-and-differences-with-influenza

[5] S. Yoo, Y. Kim, J. Kim, R. Kim, "Comparative Prediction Prototype Model of COVID19 and Pandemic Influenza", *Proceedings of KCS 2020, Korean Information Science Society*, pp. 1295-1297, (2020)
[6] M. Jin, S. Yeom, "DDoS-based Analysis of Decision Tree Considering Number of Traffic Attributes" *Journal of Information and Communications Society of Korea*, Vol.25, (2021)
[7] S. Yeom, S. Park, K. Shin, "Analysis of DDoS Attacks Based on Decision Tree Considering Importance", *Proceedings of the 2021 Spring Conference of the Korea Information and Communications Society*
[8] H. Jeong, H. Kim, S. Park, et al., "Prediction of rental car traffic accident severity using Naive Bayes big data classifier", *Journal of the Korea ITS Society*, Vol.16 No.4, (2017)
[9] Y. Shuai, Y. Zheng and H. Huang, "Hybrid Software Obsolescence Evaluation Model Based on PCA-SVM-GridSearchCV", Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS vol. 2018-Novem, pp. 449-453, 2019.
[10] E. Choi, N. Park, "Development and Application of Machine Learning Training Program Based on K-NN Algorithm Understanding", *Journal of The Korean Association of Information Education* Vol. 25, no. 1, (2021)
[11] M. Kim. "Lane Color and Shape Recognition Using Multilayer Perceptron and Partial Projection." *Domestic master's thesis Pusan National University Graduate School*, 2014. Busan
[12] M. Oh. “The Relationship between the Quality of Educational Services and Repurchase Intention of Sujichim and Seogeum Therapy.” *Master's thesis*, 2017.
[13] S. Ahn, "Another understanding of the electrical characteristics of acupuncture points and meridians," Korean Journal of Acupuncture 25 no.2, pp. 33-41, (2008)
[14] Y. Jin, et al. "The Classification Model of Human Characteristics based on Human Bioelectric Patterns." *The IIBC 2022 O2O Integrated Conference*, Nov. 17, (2022)
[15] Covid19 Prediction Using Symptoms, https://github.com/GuduruAishwarya/Covid19_Prediction_Using_Symptoms