Deep Learning approach for predicting the therapeutic usage of Jamu

S H Wijaya\textsuperscript{1*}, M Saumnuari\textsuperscript{1}, A K Nasution\textsuperscript{1}, D A Ramadhan\textsuperscript{1} and L S Hasibuan\textsuperscript{1}

\textsuperscript{1}Department of Computer Science, Faculty of Mathematics and Natural Sciences, IPB, University, Bogor 16680, West Java, Indonesia

*Email: sony@apps.ipb.ac.id

Abstract. As a country with the largest number of medicinal plants in the world, Indonesia uses medicinal plants as a composition of herbal medicine. The ingredient of herbal medicine is generally made based on experiences and hereditary. This research aims to build a scientific background of Jamu through analysis of the relationship between medicinal plants used as the composition of Jamu and its therapeutic usage. Deep Learning was chosen as a classifier because it shows good effectiveness in generating predictive models in many studies. As a comparison, we also applied the Random Forest and Support Vector Machine as classifiers and examined the classifier performances while predicting the therapeutic usage of Jamu. To handle the imbalanced data between efficacy classes, the Synthetic Minority Oversampling Technique was applied before model generation. The result shows that the highest accuracy for Deep Learning is 88.74\%, relatively higher than Random Forest and Support Vector Machine, which obtain accuracy values of 78.84\% and 77.60\%, respectively. Variable importance of the best prediction model using Deep Learning identified 105 medicinal plants, and 39 of them were selected as potential plants for 14 therapeutic usages.

1. Introduction

The number of plant species worldwide is around 40,000. In Indonesia, there are approximately 30,000 plant species, and an estimated 1,000 of them are medicinal plants [1]. Indonesia, as the country with the most significant number of medicinal plants in the world, uses these plants as a composition of herbal medicines known as Jamu [2]. Traditionally, the formulations of Jamu are made from generation to generation. Nevertheless, the formulation and production of herbal medicine have been carried out on an industrial scale. There were 1,247 herbal medicine factories in Indonesia that produce many herbal formulas with various therapeutic usages and under the supervision of the Indonesia National Agency of Drug and Control [3].

The systematization of Jamu has been examined in the last decade because of its benefits. Afendi \textit{et al.} (2013) examined the effect of medicinal plants by utilizing Partial Least Squares with Discriminant Analysis (PLS-DA) [4]. The efficacy prediction of herbal medicine from this study is 71.6\%. Wijaya \textit{et al.} (2014) conducted a study on the relationship between plants and diseases using a supervised learning method based on DPClusO [5]. They identified 135 potential plants from 11 disease classes, in which 63 plants have a specific efficacy, and another 72 plants have more than one therapeutic usage.

Supervised learning is a machine learning task by building a learning model to make a prediction. Currently, one of the popular methods for making the prediction model is the Deep Neural Network (DNN). DNN architecture has been used in several studies, including in the omics research. Li \textit{et al.}
(2018) applied DNN to predict protein interactions [6]. The predictions were performed using *Escherichia coli*, *Drosophila*, and *Caenorhabditis elegans* data sets and obtained the predictive accuracy of 95.95%, 98.39%, and 98.67%, respectively. In addition, Farhan (2018) performed the classification of herbal medicines based on metabolites using the DNN method, and the result was compared with the Random Forest [7]. The result showed that DNN achieved the highest accuracy compared to Random Forest with an accuracy of 86.57%. These studies indicate that DNN is an excellent method for making a prediction model of protein interactions using primary sequences and also therapeutic usage of herbal medicines based on compounds that exist in the medicinal plants. Therefore, we can exploit the advantages of DNN for predicting the efficacy of Jamu.

The number of Jamu formulas produced and distributed for the public is affected by its efficacy. For general diseases, such as the digestive system disease, the number of Jamu formulas available on the market is more numerous compared to Jamu formulas for a specific disease such as cancer. As a consequence, this situation resulted in imbalanced data of Jamu formulas in each disease class. Imbalanced data can affect the prediction model. Therefore, it is necessary to utilize sampling techniques to overcome imbalanced data [8]. This study aims to build a scientific background of Jamu through analysis of the relationship between medicinal plants used as the composition of Jamu and the therapeutic usage of Jamu using Deep Learning. The systematization of Jamu is needed to give a better understanding of Jamu. Also, the results can be used to predict the efficacy of new herbal formulations or to determine the substitution of medicinal plants for a particular disease.

2. Material and methods

2.1. Datasets

The data used in this study were taken from Wijaya *et al.* (2014), consisted of 3 138 Jamu formulas, 463 plants, and 116 lists of diseases that had been grouped into 18 disease classes [5]. The disease classes are as follows: Blood and lymph diseases (E1), Cancers (E2), The digestive system (E3), Ear, nose, and throat (E4), Diseases of the eye (E5), Female-specific diseases (E6), Glands and hormones (E7), The heart and blood vessels (E8), Diseases of the immune system (E9), Male-specific diseases (E10), Muscle and bone (E11), Neonatal diseases (E12), The nervous system (E13), Nutritional and metabolic diseases (E14), Respiratory diseases (E15), Skin and connective tissue (E16), The urinary system (E17), and Mental and behavioral disorders (E18).

Jamu data has been registered at the Indonesia National Agency of Drug and Control, and the professional doctor determined the efficacy of each herbal medicine formally using the International Classification of Diseases ver. 10 (ICD-10, http://apps.who.int/classifications/icd10/). The list of diseases based on ICD-10 then grouped into disease classes based on the National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/books/NBK22183/). Data is represented as binary vectors, which indicate whether a medicinal plant is contained in the Jamu formula or not.

2.2. Methods

This study mainly consists of three steps, namely data pre-processing, prediction of Jamu efficacies based on medicinal plants used as Jamu ingredients, and identification of potential medicinal plants. Deep Learning was used as a classifier for predicting Jamu efficacies and also for identifying the potential medicinal plants for each disease class.

2.2.1. Data pre-processing. At this stage, the preparation of data for predicting the therapeutic usage of Jamu was carried out by eliminating Jamu formulas with certain conditions. Jamu formula with multiple therapeutic usages and also has a very general efficacy were excluded from our analysis. Additionally, we created two datasets, namely dataset I and dataset II, to evaluate the performance of classifiers while making the model for efficacy predictions. Dataset I was formed by only eliminating Jamu formulas with certain conditions as we mentioned before, whereas dataset II was created as the same as dataset I.
but with some additional pre-process such as feature selection, and oversampling. In terms of oversampling, we applied the Synthetic Minority Over-sampling Technique (SMOTE) [8].

2.2.2. Prediction of Jamu efficacies. The prediction of Jamu efficacy using dataset I and II was started by tuning parameters to identify the best parameter values. In addition, we applied k-fold cross-validation to evaluate the machine learning model on a limited data sample and also examined the accuracy of different k-folds [9]. The choice of k depends on best-resulted accuracy. The evaluation was done by comparing the accuracy obtained by Deep Learning with other classifiers, i.e., Random Forest (RF) and Support Vector Machine (SVM). The best model from the Deep Learning method was then utilized to identify potential medicinal plants for each therapeutic usage.

2.2.3. Identification of potential medicinal plants for each disease class. The process of selecting potential medicinal plants used the variable importance in Deep Learning. In this case, the variable represents medicinal plants, and the number of selected potential medicinal plants was determined based on Wijaya et al. (2014). The efficacy of potential medicinal plants that had been identified for each disease class was further validated using open-access databases and literature.

3. Results and discussion

3.1. Data preparation
Pre-processing of Jamu data was applied to Jamu formulas that have a general symptom and diverse efficacy such as fever and Jamu formulas with two or more efficacies. Jamu formula for fever was removed because this symptom was very general and can be found in many disease classes, while Jamu formula that classified into more than one disease class was removed because the prediction model was only for a specific disease. The number of eliminated data was 123 Jamu formulas and only 3 015 data left for further analysis. Disease class related to ear, nose, and throat (E4), disease of the eye (E5), glands and hormones (E7), and neonatal disease (E12) were also excluded in the analysis because they did not have enough samples for making the prediction model. Therefore, there were 3 012 Jamu formulas and 14 disease classes used as dataset I. Because the number of Jamu formulas for each disease class relatively diverse, we also examined the effect of sampling technique using SMOTE. The parameters used in oversampling were ‘auto’ and ‘kind=SVM’. In this case, we obtained 8 474 new samples. Moreover, the number of plants as features used in the dataset I relatively large, so that we did feature selection using the ExtraTreesClassifier algorithm [10] to obtain smaller features, which is 105 plants. Therefore, dataset II consisted of 8 474 samples, 105 plants, and 14 disease classes. The distribution of herbal medicine data in each class of disease from the dataset I and II can be seen in Figures 1a and 1b.

(a) Figure 1. The distribution of dataset I and dataset II for each disease class.
3.2. Deep Learning method for predicting Jamu efficacies

The prediction of Jamu efficacy aims to get Jamu formula that has a specific effect against certain disease classes. The method used in this study is Deep Learning, while the comparative methods are Random Forest and Support Vector Machine.

The Deep Learning method was applied using the Keras package on the Jupyter Notebook. The architecture of the Deep Learning method using dataset I and II can be seen in Figure 2. Initially, tuning parameters for Deep Learning were done, including the activation function, dropout value, and the number of \( k \) in the \( k \)-fold cross-validation. The activation function used in the input and hidden layer was ReLU because it provided a faster learning process and better accuracy compared to Linear, Sigmoid, and Tanh activation functions [11]. Additionally, we applied the Softmax activation function in the output layer because the output model consisted of multi-class, and this function can determine the probability value for each class [11]. The dropout value was used in the hidden layer to reduce overfitting. In this case, we examined four different dropout values (0.15, 0.25, 0.40, and 0.50), and the best accuracy was obtained if we set 0.15 as dropout value. We applied StratifiedKFold for \( k \)-fold cross-validation for \( k=4 \) to 10, and the best accuracy was obtained when we set \( k=10 \). The use of \( k=10 \) has computational advantages and a more feasible learning procedure [12]. In the tuning parameters, we also examined the number of hidden layers and epochs. The determination of the number of hidden layers and epochs is in accordance with the resulting accuracy value. The best accuracy was obtained if we set the number of hidden layers and epoch equal to 4 and 500, respectively.

![Figure 2. Deep Learning Architecture. In term of the dataset I and II, the number of nodes as input layer are 463 and 105, respectively.](image)

As a comparison, we chose RF and SVM as classifiers. Random Forest is a method that consists of structured trees and classifies by looking at the distribution of the most data from the input data on the tree [13], whereas Support Vector Machine is a discriminative classifier defined by a separating hyperplane [14]. RF and SVM were done by utilizing the Sklearn package. Initially, we applied a tuning parameter for both classifiers. In the case of RF, the highest accuracy was obtained if we set the number of trees (\( n\_estimators \)) and the number of \( k \) in \( k \)-fold cross-validation using Repeated StratifiedKFold equal to 1000 and 10, respectively. In the case of SVM, we examined three different kernels, i.e. RBF, Polynomial, and Sigmoid kernels. The highest accuracy was obtained if we used RBF kernel with \( k \) in the \( k \)-fold cross validation equal to 10 [15].

A comparison of classification models using three different classifiers and two datasets is shown in Figure 3. Overall, the best prediction accuracy is obtained by a combination of Deep Learning and dataset II, with 88.74\% accuracy. In addition, Deep Learning outperforms Random Forest and Support Vector Machine in dataset II. Otherwise, Random Forest outperforms Deep Learning and Support Vector Machine with 66.68\% accuracy in dataset I. If we compared the best accuracy from both datasets, utilization of feature selection and sampling methods can increase the accuracy by 30.38\% for Deep Learning, 12.16\% for Random Forest and 14.53\% for Support Vector Machine.
3.3. Identification of potential medicinal plants for each disease class

Potential medicinal plants for each disease class were obtained based on variable importance from the best model in the Deep Learning method using the KerasRegressor and PermutationImportance packages. Based on Wijaya et al. (2014), the maximum number of efficacies for each medicinal plant is seven therapeutic usages [15]. Therefore, we applied this threshold to determine the potential plant for each disease class. Out of 105 important features used in dataset II, we selected 39 plants as a candidate for the potential plants. The distribution of selected plants for each disease class is shown in Table 1.

**Table 1. The distribution of 39 potential plants in each disease class.**

| No. | Plant Species              | Weight | 1 | 2 | 3 | 6 | 8 | 9 | 10 | 11 | 13 | 14 | 15 | 16 | 17 | 18 |
|-----|---------------------------|--------|---|---|---|---|---|---|----|----|----|----|----|----|----|----|
| 1   | Guazuma ulmifolia         | 0.495  | 1 | 0 | 1 | 1 | 1 | 0 | 0  | 1  | 0  | 1  | 0  | 0  | 0  | 0  |
| 2   | Serenoa repens            | 0.398  | 0 | 0 | 0 | 0 | 0 | 0 | 1  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |
| 3   | Cocos nucifera            | 0.365  | 0 | 0 | 1 | 1 | 0 | 0 | 0  | 1  | 1  | 1  | 1  | 1  | 0  | 0  |
| 4   | Curcuma zedoaria          | 0.349  | 1 | 1 | 1 | 1 | 0 | 0 | 0  | 1  | 0  | 1  | 0  | 0  | 1  | 0  |
| 5   | Soya max                  | 0.275  | 1 | 0 | 1 | 1 | 1 | 0 | 1  | 1  | 0  | 1  | 0  | 0  | 0  | 0  |
| 6   | Curcuma heynaneana        | 0.223  | 1 | 0 | 1 | 1 | 0 | 0 | 0  | 1  | 0  | 1  | 0  | 1  | 0  | 1  |
| 7   | Apium graveolens          | 0.189  | 1 | 0 | 0 | 0 | 1 | 0 | 0  | 1  | 0  | 1  | 0  | 0  | 0  | 1  |
| 8   | Graptothyllum pictum      | 0.179  | 1 | 0 | 0 | 1 | 0 | 0 | 0  | 1  | 0  | 0  | 1  | 1  | 0  | 0  |
| 9   | Citrus aurantium          | 0.172  | 1 | 0 | 1 | 1 | 0 | 0 | 0  | 1  | 0  | 1  | 1  | 1  | 0  | 0  |
| 10  | Eurycoma longifolia       | 0.165  | 1 | 0 | 0 | 1 | 0 | 0 | 0  | 1  | 0  | 1  | 0  | 0  | 0  | 0  |
| 11  | Valeriana javanica        | 0.155  | 0 | 0 | 1 | 1 | 1 | 0 | 0  | 1  | 1  | 0  | 1  | 0  | 0  | 1  |
| 12  | Aloe vera                 | 0.132  | 0 | 0 | 1 | 1 | 0 | 0 | 0  | 1  | 0  | 1  | 1  | 1  | 0  | 0  |
| 13  | Murraya paniculata        | 0.128  | 1 | 0 | 0 | 1 | 0 | 0 | 0  | 1  | 0  | 1  | 0  | 1  | 0  | 0  |
| 14  | Syzygium polyanthum       | 0.121  | 1 | 0 | 0 | 0 | 1 | 0 | 0  | 1  | 1  | 1  | 0  | 0  | 1  | 0  |
| 15  | Vetiveria zizanioides     | 0.121  | 1 | 0 | 1 | 1 | 0 | 0 | 0  | 1  | 0  | 1  | 1  | 1  | 0  | 0  |
| 16  | Galla lusitania           | 0.104  | 1 | 0 | 1 | 1 | 0 | 0 | 1  | 0  | 1  | 1  | 0  | 1  | 0  | 0  |
| 17  | Clausena anisum-olens     | 0.094  | 0 | 0 | 1 | 1 | 0 | 0 | 0  | 1  | 1  | 1  | 1  | 0  | 0  | 0  |
| 18  | Quercus lusitanica        | 0.090  | 0 | 0 | 0 | 1 | 0 | 0 | 0  | 1  | 0  | 1  | 0  | 1  | 0  | 0  |
| 19  | Trigonella foenum-graecum | 0.074  | 1 | 0 | 0 | 1 | 0 | 0 | 0  | 1  | 0  | 1  | 1  | 1  | 0  | 0  |
| 20  | Sida rhombifolia          | 0.073  | 1 | 0 | 0 | 0 | 0 | 0 | 0  | 1  | 0  | 0  | 0  | 0  | 0  | 0  |
| 21  | Phaleria papuana          | 0.071  | 1 | 0 | 1 | 1 | 1 | 0 | 0  | 1  | 0  | 1  | 1  | 0  | 0  | 0  |
| 22  | Pogostemon cablin         | 0.071  | 1 | 0 | 1 | 0 | 0 | 0 | 0  | 0  | 0  | 0  | 0  | 1  | 1  | 0  |
| 23  | Symlocos odoratissima     | 0.063  | 0 | 0 | 1 | 1 | 0 | 0 | 0  | 0  | 0  | 0  | 1  | 1  | 1  | 0  |
| 24  | Psidium guajava           | 0.060  | 1 | 0 | 1 | 1 | 0 | 0 | 0  | 1  | 0  | 1  | 1  | 0  | 0  | 0  |
| 25  | Sauropus androgynus       | 0.059  | 0 | 0 | 0 | 1 | 0 | 0 | 0  | 1  | 0  | 0  | 0  | 0  | 0  | 0  |
Based on Table 1, the validation of some potential plants was carried out by published literature, such as *Curcuma zedoaria* for disease class E2 (Cancers), *Echinacea purpurea* for disease class E9 (Diseases of the immune system), *Sida Rhombifolia* for disease class E11 (Muscle and bone), *Citrus hystrix* for disease class E15 (Respiratory diseases), and *Pogostemon cablin* for disease class E16 (Skin and connective tissue). The results of the validation from these plants are as follows:

- *Curcuma zedoaria* belongs to the *Zingiberaceae* family. It has an active compound Isocurcumenol, which has a function to inhibit cancer cells [16].
- Parker (2006) states that the *Echinacea purpurea* can stimulate the body's immunity in humans [17].
- *Sida Rhombifolia* has properties to reduce uric acid levels because they contain flavonoids in leaf extracts. The content of flavonoids has an inhibitory effect of xanthine oxidase so that it can reduce excess uric acid production. These plants can help remove the excess uric acid in the blood, so it does not continue to accumulate in the body [18].
- The *Citrus hystrix* has a synonym, *Citrus aurantifolia*. Al-Snafi (2016) states that these plants traditionally have properties to treat sinusitis, bronchitis, and asthma [19].
- Pujiarti et al. (2012) state that oil in the *Pogostemon cablin* can be used to treat skin diseases. Oil from these plants can help stimulate the growth of new skin cells, help in the healing process of wounds, and treat dermatitis [20].

### 4. Conclusion

The prediction of Jamu efficacies based on medicinal plants used as Jamu ingredients using the Deep Learning method obtained the best accuracy on dataset II. The Deep Learning method produced an accuracy value of 88.74%, increased 30.38% compared to the same method without pre-processing, feature selection, and oversampling techniques. Moreover, Random Forest obtained slightly higher accuracy while using dataset II compared to Deep Learning and Support Vector Machine. Based on the best prediction model formed using Deep Learning, we identified 39 of 105 potential medicinal plants from 14 disease classes, and some of them had been validated by published literature. The prediction of Jamu efficacies can be extended to the molecular level by considering compounds that exist in each medicinal plant [21].
Acknowledgments
This research is sponsored by the Ministry of Research, Technology, and Higher Education of Republic of Indonesia through Penelitian Dasar Unggulan Perguruan Tinggi (PDUPT), IPB University, with contract number 4178/IT3.L1/PN/2019.

References
[1] Widyastuti U, Suharsono G, Ogasawara N, Kanaya S. 2009. Sistematisasi produk metabolit sekunder alami Indonesia sebagai bahan obat herbal menggunakan pendekatan metabolomik. Direktorat Jenderal Pendidikan Tinggi.
[2] Hanafi M, Nina A, Fadia Z, Nurbaeti. 2006. Indonesian country report on traditional medicine. New Delhi, India.
[3] Indonesian Food Technologist C. 2018. Seminar nasional dan pameran industri Jamu [internet]. [accessed 2018 Oct 25]. Available online: http://www.ift.or.id/2014/04/seminar-nasional-dan-pameran-industri.html.
[4] Afendi FM, Darusman LK, Hirai A, Amin MA, Takahashi H, Nakamura K, Kanaya S. 2013. Efficacy prediction of Jamu formulations by PLS modeling. Current Computer-Aided Drug Design. 46-59.
[5] Wijaya SH, Husnawati H, Afendi FM, Batubara I, Darusman LK, Ul-Amin MA, Sato T, Ono N, Sugiiura T, Kanaya S. 2014. Supervised clustering based on DPClusO: prediction of plant-disease relations using jamu formulas of KNApSAcK database. BioMed Research International. doi:10.1155/2014/831751.
[6] Li H, Gong X, Yu H, Zhou C. 2018. Deep neural network based predictions of protein interactions using primary sequences. Molecules.
[7] Farhan MI. 2018. Deep Neural Network for classification of herbal medicine's formula based on metabolite [Thesis]. Bogor (ID): Institut Pertanian Bogor.
[8] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. 2002. SMOTE: synthetic minority over-sampling technique. J Artificial Intelligence Research. 16:321-357.
[9] Jung Y. 2016. A K-fold Averaging Cross-validation Procedure. California (US): University of California.
[10] Geurts P, Wehenkel L, Ernst D. 2006. Extremely randomized trees. Article in Machine Learning.
[11] James G, Witten D, Hastie T, Tibshirani R. 2017. An Introduction to Statistical Learning. London (UK): Springer. page 176-186.
[12] Goel E, Abhilasha E. 2017. Random forest: a review. Computer Science & Engineering & GZSCCET Bhatinda. doi:10.23956/ijarcse/V711/01113.
[13] Syafrullah SC. 2015. Indonesian sidaguri (Sida Rhombifolia L.) as antigout and inhibition kinetics of flavonoids. J Majority Volume 4 Nomor 1.
[14] Al-Snafi AE. 2016. Nutritional value and pharmacological importance of citrus species grown in Iraq. Department of Pharmacology, College of Medicine, Thi qar University, Iraq.
[20] Pujiarti R, Ohtani Y, Widowati TB, Wahyudi W, Kasmudjo K, Herath NK, Wang CN. 2012. Effect of Melaleuca leucadendron, Cananga odorata and Pogostemon cablin oil odors on human physiological responses. Wood Research Journal. 2012;3(2):100-5.

[21] Wijaya SH, Batubara I, Nishioka T, Altaf-Ul-Amin M, Kanaya S. 2017. Metabolomic Studies of Indonesian Jamu Medicines: Prediction of Jamu Efficacy and Identification of Important Metabolites. Molecular informatics. 2017 Dec;36(12):1700050.