Comparing Different Supervised Machine Learning Accuracy on Analyzing COVID-19 Data using ANOVA Test

Nurrahma
School of Electrical Engineering and Informatics
Bandung Institute of Technology
Bandung, Indonesia
rahma.rara06@gmail.com

Rahadian Yusuf
School of Electrical Engineering and Informatics
Bandung Institute of Technology
Bandung, Indonesia
yrahadian@itb.ac.id

2020 6th International Conference on Interactive Digital Media (ICIDM) | 978-1-7281-4928-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICIDM51048.2020.9339676

***Abstract*—COVID-19 belongs to the genus of Coronaviridae, has emerged and affected the whole world. A virus without a vaccine creates unexpected havoc in human life, financial and economic systems of every country around the world. This disease carries a heavy workload for doctors and health workers. This workload can be reduced by machine learning and the development of computer-aided diagnostic systems. Any scientific study of this disease will help manage this disease as soon as possible. In this experiment, support vector machine, decision tree, and neural network classifiers have been built for COVID-19 dataset and the accuracies among those classifiers have been analyzed by ANOVA test. First, we performed data preprocessing to handle the missing values and categorical data. Second, we performed feature selection using chi-squared statistics and mutual-information statistics methods. Third, we split the data into 4 train-sets and test-sets for performing 4-fold cross validation. And lastly, we built SVM, DT, and NN models and started the experiment. In the experiment without feature selection, accuracy results among SVM, DT, and NN is stastistically significant (reject the null hypothesis). Otherwise, in the experiment with feature selection, accuracy results among SVM, DT, and NN is not stastistically significant (accept the null hypothesis). Both experiments are tested by ANOVA test. In this experiments, neural network classifiers (both without and with feature selection) have better accuracy result (98,16% and 97,08%) than SVM (96,63% and 96,72%) and DT (98,08% and 96,89%) for the dataset used in this experiment.**

Keywords—COVID-19, supervised machine learning, support vector machine, decision tree, neural network, accuracy, ANOVA test

I. INTRODUCTION

COVID-19 is the disease caused by a new coronavirus called SARS-CoV-2. COVID-19 epidemic occurred in Wuhan, China in December 2019. WHO declared the outbreak as an emergency and pandemic for public health on January 30, 2020. COVID-19's symptoms are fever, dry cough, fatigue and other symptoms that are less common and may affect some patients. The way to protect others and ourselves are by taking some precautions, such as physical distancing, wearing a mask, keeping the room well ventilated, avoiding crowds and close contact, regularly cleaning our hands, and coughing into a bent elbow or tissue [1].

The number of worldwide COVID-19 cases are 52.568.539 with 1.291.849 death cases, as of November 12, 2020 [2]. The increase in the number of COVID-19 cases infected rapidly exceeds the amount of medical resources available in hospitals. It imposes a substantial burden on the health care system [3]. Due to the limited availability of resources at hospitals, the time delay for the results of the medical tests, and the rapid increasing number of cases to test for COVID-19 day by day, it is not possible to test due to the time and cost factors [4]. Hence, it is required to build an automatic detection system to predict whether a person has COVID-19 or not. One of the tools that can be used is machine learning.

Machine learning refers to concepts and results from a variety of fields, including statistics, artificial intelligence, philosophy, information theory, biology, cognitive science, computational complexity, and control theory. Machine learning can learn from the past and detect useful patterns from large, unstructured, and complex data sets using a variety of statistics, probabilistic, and optimization methods [5]. Machine learning has a wide range of applications, including computer vision, speech recognition, bio-surveillance, robot and automation control, and empirical science experiments (e.g., spam filtering, fraud detection, topic identification, and predictive analytics) [6]. These applications can be implemented by both supervised and unsupervised machine learning methods. Supervised machine learning is based on training a data sample from a data source with the correct classification or label already assigned. In unsupervised machine learning, training data is not labeled. The classifier is designed to identify hidden patterns in unlabeled input data [7].

There are many classifiers nowadays and each classifier has different accuracy on different datasets. It makes great concern on how to choose the best classifier for a particular dataset. It would be preferable to compare the performance of classifiers on the same dataset. It is important to make an appropriate comparison of all available classifiers and choose the best among those classifiers [8]. One way to compare classifiers is to use the analysis of variance (ANOVA) test to compare the accuracy results of each classifier [9].

ANOVA test was developed by statistician Ronald Fisher. ANOVA is a statistical technique to analyze the differences among group means in a sample. Frequently, ANOVA test is used to test equality among several means by comparing variance among groups relative to variance within groups [10]. Almost every statistical package provides ANOVA test, which makes it accessible to researchers in all experimental sciences. It is easy to enter data sets and run a simple ANOVA test [11]. The statistically significant result, when the probability (p-value) was less than the predetermined threshold (level of significance), justified the null hypothesis being rejected [12].

Thus, this paper aims to simulate the different machine learning methods on COVID-19 dataset. The scope of this experiment is primarily on building several supervised machine learning classifiers on COVID-19 dataset and analyzing the accuracies among those classifiers. The classifiers used in this experiment are Support Vector Machine (SVM), Decision Tree (DT), and Neural Network (NN). Each classifier will be given the same dataset without a selection feature and with a selection feature. Feature selection will be carried out statistically, in order to make it easier for machine learning to classify. The accuracy performance of each classifier will be compared statistically. ANOVA test will be used in this experiment to determine whether any of those classifier's accuracy are statistically significant from each other. Therefore, we can determine which machine learning classifiers are good in this case.

II. Related Work

A. Narin, C. Kaya, and Z. Pamuk [13] have developed an automatic detection system as an alternative diagnosis option of COVID-19. In this research, the authors proposed three different Convolutional Neural Network (CNN) based models (ResNet50, Inception-V3, and Inception-ResNetV2) for the detection of coronavirus pneumonia infected patients using chest X-Ray radiographs. The authors discussed the performance accuracy among those CNN models. Performance results show that the ResNet50 pre-trained model yielded the highest accuracy among five models for used three different datasets (Dataset-1: 96.1%, Dataset-2: 99.5% and Dataset-3: 99.7%).

Y. Ünal and M. N. Dudak [14] have applied naive bayes, k-nearest neighbor, support vector machine, decision tree on COVID-19 Mexico Patient Health dataset. The dataset containing data related to the COVID-19 outbreak affecting the worldwide. The accuracy results from those classifiers show that the support vector machine algorithm performs 100% accuracy which is the best classifier for that dataset.

D. M. Matta and M. K. Saraf [15] conducted a systematic literature review to identify suitable algorithms for prediction of COVID-19 in patients. To evaluate the accuracy of machine learning classifiers, each algorithm is trained with record sets of varying numbers of patients. The trained algorithms were assessed using accuracy performance metric. After result analysis, Random Forest (99,44%) showed better prediction accuracy in comparison with both SVM (98,33%) and Artificial Neural Networks (99,25%). The trained algorithms were also assessed to find the features that affect the prediction of COVID-19 in patients.

K. B. Prakash, S. S. Imambi, M. Ismail, T. P. Kumar, and Y. V. R. N. Pawan [16] built different prediction machine learning algorithms, computed and evaluated their performances. The results show that the Random Forest Regressor and Random Forest Classifier has better results than Decision Tree, Gaussian Naive Bayes, Multilinear Regression, Logistic Regression, SGB Regression, SVM, and KNN+NCA in terms of Coefficient of Determination and Accuracy.

III. Experiment Dataset and Methods

A. Dataset

The dataset used in this experiment is obtained from the Kaggle site. The dataset's name is "Symptoms and COVID Presence". This dataset consists of 5434 cases which are formed of 20 features and 1 label (classified into 2 classes, "Yes" and "No") as shown in Table I. This dataset was recorded worldwide between April 17, 2020 and August 29, 2020 [17].

TABLE I. Symptoms and Covid Presence Dataset Features

Feature No.	Feature Name	Feature Type	Description
1	Breathing Problem	Boolean	Yes = Positive COVID-19, No = Negative COVID-19
2	Fever	Boolean	Yes = Positive COVID-19, No = Negative COVID-19
3	Dry Cough	Boolean	Yes = Positive COVID-19, No = Negative COVID-19
4	Sore Throat	Boolean	Yes = Positive COVID-19, No = Negative COVID-19
5	Running Nose	Boolean	Yes = Positive COVID-19, No = Negative COVID-19
6	Asthma	Boolean	Yes = Positive COVID-19, No = Negative COVID-19
7	Chronic Lung Disease	Boolean	Yes = Positive COVID-19, No = Negative COVID-19
8	Headache	Boolean	Yes = Positive COVID-19, No = Negative COVID-19
9	Heart Disease	Boolean	Yes = Positive COVID-19, No = Negative COVID-19
10	Diabetes	Boolean	Yes = Positive COVID-19, No = Negative COVID-19
11	Hypertension	Boolean	Yes = Positive COVID-19, No = Negative COVID-19
12	Fatigue	Boolean	Yes = Positive COVID-19, No = Negative COVID-19
13	Gastrointestinal	Boolean	Yes = Positive COVID-19, No = Negative COVID-19
14	Abroad Travel	Boolean	Yes = Positive COVID-19, No = Negative COVID-19
15	Contact with COVID Patient	Boolean	Yes = Positive COVID-19, No = Negative COVID-19
16	Attended Large Gathering	Boolean	Yes = Positive COVID-19, No = Negative COVID-19
17	Visited Public Exposed Place	Boolean	Yes = Positive COVID-19, No = Negative COVID-19
18	Family Working in Public Exposed Places	Boolean	Yes = Positive COVID-19, No = Negative COVID-19
19	Wearing Masks	Boolean	Yes = Positive COVID-19, No = Negative COVID-19
20	Sanitization from Market	Boolean	Yes = Positive COVID-19, No = Negative COVID-19

B. Methods

1) Data Preprocessing: Data preprocessing is an important process in the development of machine learning models. The dataset is often loosely controlled with missing values, out-of-range values, etc. This kind of data can mislead the experimental result. In this experiment, the dataset features are categorical data and has no missing values. The categorical data have been handled by using LabelEncoder in python scikit-learn library.

TABLE II. FEATURES SELECTION RESULTS

Feature No.	Feature Name	Chi-squared	Mutual Information
1	Breathing Problem	275,30	0,10
2	Fever	11,67	0,06
3	Dry Cough	193,23	0,10
4	Sore Throat	275,26	0,11
5	Running Nose	0,15	0,00
6	Asthma	17,98	0,01
7	Cronic Lung Disease	9,66	0,01
8	Headache	1,06	0,00
9	Heart Disease	1,39	0,00
10	Diabetes	3,22	0,01
11	Hyper Tension	23,67	0,01
12	Fatigue	4,32	0,00
13	Gastrointestinal	0,00	0,00
14	Abroad Travel	439,37	0,13
15	Contact with COVID Patient	265,09	0,06
16	Attended Large Gathering	332,08	0,09
17	Visited Public Exposed Place	26,02	0,02
18	Family Working in Public Exposed Places	54,42	0,01
19	Wearing Masks	0,00	0,00
20	Sanitization from Market	0,00	0,00

2) Feature Selection: Feature selection is the process of identifying and selecting the sub-set of input features that are most relevant to the target label. Chi-squared statistics and mutual information statistics are two popular feature selection methods that can be used in categorical data [18].

The python scikit-learn library provides an implementation of the chi-squared test in the chi2() function and mutual information in the mutual_classif_info() function. These functions can be used in a feature selection strategy, such as selecting the top *k* most relevant features (the highest values) via the SelectKBest class. The result of features selection using chi-squared and mutual information can be seen in Table II.

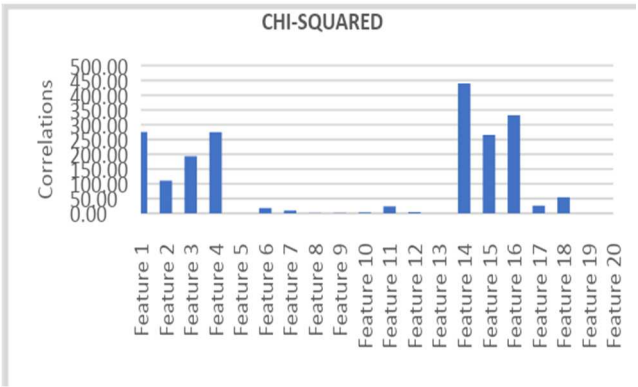

Fig. 1. Chi-squared results.

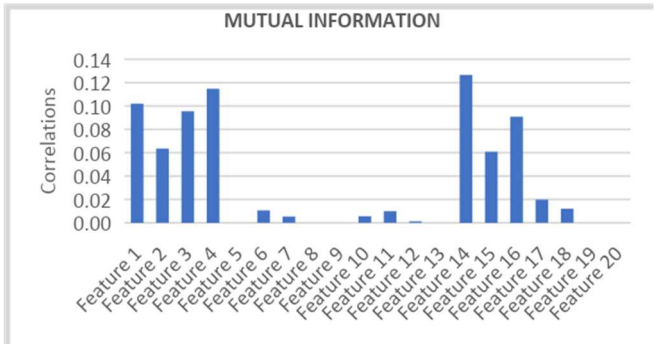

Fig. 2. Mutual information results.

As the feature selection results shown in Fig 1. and Fig. 2, feature 1, feature 2, feature 3, feature 4, feature 14, feature 15, and feature 16 are the top 7 features that are most relevant with COVID-19. So in this experiment, we used these 7 features.

3) Splitting Dataset: The dataset is randomly divided into train-set and test-set for training the data using 4-fold cross validation. The illustration can be seen in Fig. 3.

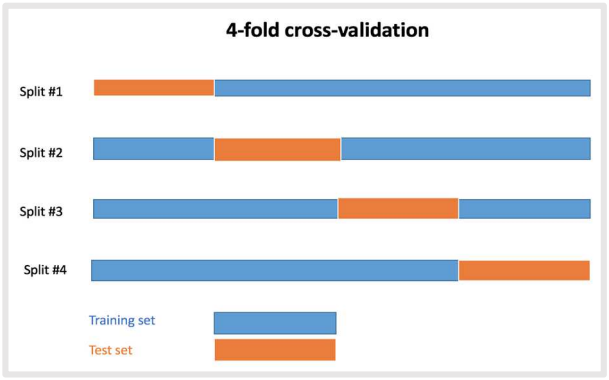

Fig. 3. 4-Fold cross validation illustration [19].

4) Building the Machine Learning Models: Models are built using python scikit-learn and keras library. Changes made to the algorithms can affect the results.

a) Support Vector Machine (SVM): SVM is used when the data has exactly two classes. The base of the SVM algorithm is using precision to generalize the errors. The algorithm divides the data into classes by finding the best [20] hyperplane so that all samples belonging to one class will be categorized on one side and the rest on the other side [21]. In SVM, the predictor variable is called an attribute and the transformed attribute is called a feature. Selecting the most suitable representative data is called feature selection. A set of features describing one case is called a vector [15]. Here is the SVM model used in this experiments:

```
model = svm.SVC(kernel = 'linear', verbose = True)
```

b) Decision Tree: One of the most accurate methods used for classification purposes in data mining is the decision tree. Decision tree often utilized for classification, clustering, prediction models, and to create subgroups within the related research areas of a problem [14]. The decision tree uses regression or classification to predict response to data. Regression is used when the data is continuous and classification is used when the features are grouped. A

decision tree is made of the root node, branches, and leaf nodes. To evaluate the data, follow the path from the root node to reach a leaf node [22]. Here is the DT model used in this experiments:

```
model = DecisionTreeClassifier()
```

c) Neural Network: Neural network is an algorithm originated from imitating the neural system of the human brain. Neurons are the basic unit of the neural networks. A neuron is said to perform functions on input and produces an output. Neurons combined are called neural networks. Training of the data is started once the neural networks are formed. The neural network has 3 layer architecture, that are input layer, hidden layer, and output layer [15]. Here is the NN model used in this experiments:

```
def NN_model():
    model = Sequential()
    model.add(Dense(128, activation = 'relu', input_dim
= 20))
    model.add(Dropout(0.5))
    model.add(Dense(64, activation = 'relu'))
    model.add(Dropout(0.5))
    model.add(Dense(32, activation = 'relu'))
    model.add(Dropout(0.5))
    model.add(Dense(16, activation = 'relu'))
    model.add(Dropout(0.5))
    model.add(Dense(8, activation = 'relu'))
    model.add(Dropout(0.5))
    model.add(Dense(1, activation = 'sigmoid'))
    return model
```

Compiling NN Model:

```
model.compile(loss = 'binary_crossentropy', optimizer =
'adam', metrics = ['accuracy'])
```

IV. Results and Analysis

This section presents the results and analysis for this experiment. In this experiment, 3 supervised machine learning methods are built to predict COVID-19. Data divided into train-set (75%) and test-set (25%). The train-set is used to build the classifier and the test-set is used to validate the classifier. In this paper, we visualize the data as a 95%-100% scale chart due to the slight differences between each category, so it will be very difficult to see if using a 0%-100% scale.

A. SVM Results – With and Without Feature Selection

Table III. represents the accuracy for every fold with and without feature selection achieved by SVM model. The average training and testing time in SVM without feature selection are around 0,570 seconds and 0,231 seconds each fold. The average training and testing time in SVM with feature selection are around 0,253 seconds and 0,226 seconds each fold. The accuracy results of SVM can be clearly identified from the chart in Fig. 4.

TABLE III. SVM Accuracy Results

Fold	All Features (Without Feature Selection)	7 Features (With Feature Selection)
Fold-1	96,86 %	**97,06 %**
Fold-2	96,98 %	96,98 %
Fold-3	96,69 %	96,81 %
Fold-4	95,98 %	96,02 %
AVG	96,63 %	**96,72 %**

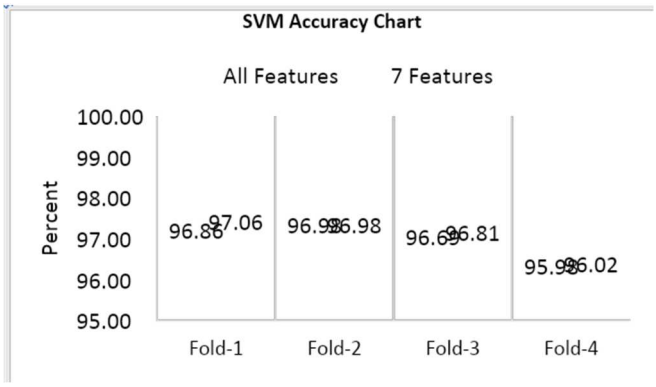

Fig. 4. SVM accuracy chart.

B. DT Results – With and Without Feature Selection

Table IV. represents the accuracy for every fold with and without feature selection achieved by decision tree model. The average training and testing time in DT without feature selection are around 0,125 seconds and 0,111 seconds each fold. The average training and testing time in DT with feature selection are around 0,104 seconds and 0,030 seconds each fold. The accuracy results of the decision tree can be clearly identified from the chart in Fig. 5.

TABLE IV. Decision Tree Accuracy Results

Fold	All Features (Without Feature Selection)	7 Features (With Feature Selection)
Fold-1	98,36 %	97,28 %
Fold-2	97,77 %	96,98 %
Fold-3	**98,38 %**	96,76 %
Fold-4	97,82 %	96,54 %
AVG	**98,08 %**	96,89 %

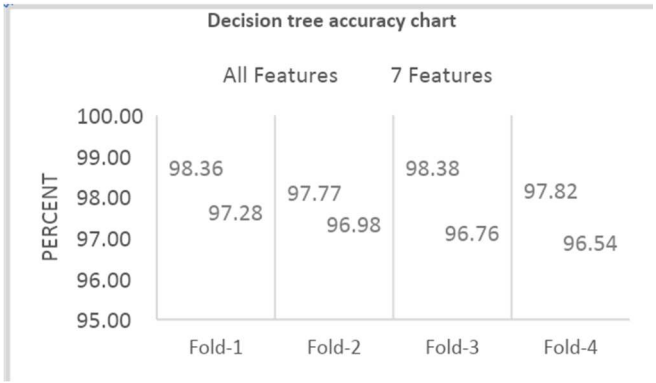

Fig. 5. Decision tree accuracy chart.

C. NN Results – With and Without Feature Selection

In neural network, the accuracy results will not always have the same results every time it is run, in contrast to SVM and DT which always give the same results. So, in this experiment, we ran the neural network methods 20 times for every fold without and with feature selection. The overall average of NN results are 98,01% (without feature selection) and 96,54% (with feature selection). In this experiment, we select the best accuracy among those results to compare with other methods. The best accuracy is in the sixth built. Table V represents the best accuracy among those results. The average training and testing time in NN without feature

selection are around 76,25 seconds and 0,265 seconds each fold. The average training and testing time in NN with feature selection are around 88,50 seconds and 0,357 seconds each fold. The accuracy results of neural network can be clearly identified from the chart in Fig. 6.

TABLE V. NEURAL NETWORK ACCURACY RESULTS

Fold	All Features (Without Feature Selection)	7 Features (With Feature Selection)
Fold-1	98,28 %	97,28 %
Fold-2	97,82 %	96,98 %
Fold-3	**98,50 %**	97,18 %
Fold-4	98,04 %	96,88 %
AVG	**98,16 %**	97,08 %

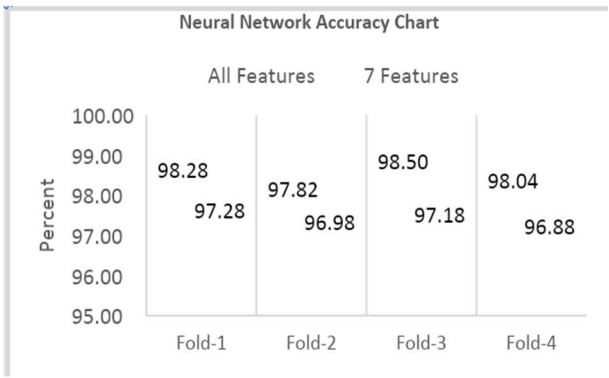

Fig. 6. Neural network accuracy chart.

D. *Comparison Between SVM, DT, and NN Results – without Feature Selection*

Table VI represents the accuracy results achieved by each method for every fold without feature selection. The accuracy results achieved by each method for every fold without feature selection can be clearly identified from the chart in Fig. 7.

TABLE VI. ACCURACY RESULTS – WITHOUT FEATURE SELECTION

Fold	Methods Name		
	SVM	DT	NN
Fold-1	96,86 %	98,36 %	98,28 %
Fold-2	96,98 %	97,77 %	97,82 %
Fold-3	96,96 %	98,38 %	**98,50 %**
Fold-4	95,98 %	97,82 %	98,04 %
AVG	96,63 %	98,08 %	**98,16 %**

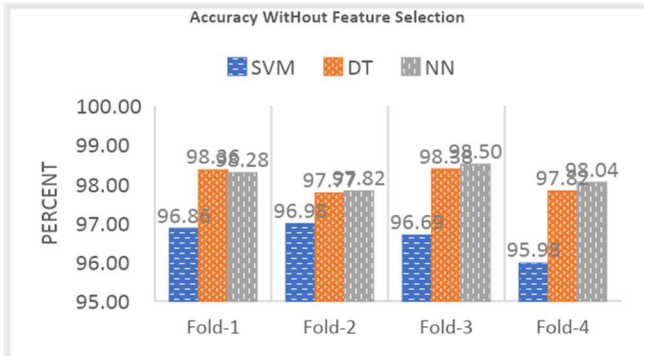

Fig. 7. Accuracy results – without feature selection.

From the data in Table VI, we performed an ANOVA test using the Data Analysis feature in Excel and got the following results that can be seen in Table VII and Table VIII. On this test, we determined the threshold (alpha) = 0,05. From the ANOVA test result, we can see that the p-value is less than the alpha (p-value < 0,05). It means the accuracy results between SVM, DT, and NN without feature selection is stastistically significant (reject the null hypothesis). This means that each classifier has a different ability to recognize features for classification. NN has the highest accuracy result. This is probably because NN has the best ability to recognize features.

TABLE VII. SUMMARY OF ACCURACY RESULTS - WITHOUT FEATURE SELECTION

Groups	Count	Sum	Average	Variance
SVM	4	386,5045	96,6261	0,2027
DT	4	392,3198	98,0799	0,1115
NN	4	392,6388	**98,1597**	0,0888

TABLE VIII. ANOVA TEST ON ACCURACY RESULTS - WTHOUT FEATURE SELECTION

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	5,9624	2	2,9812	22,1865	**0,0003**	4,2565
Within Groups	1,2093	9	0,1344			
Total	7,1717	11				

E. *Comparison Between SVM, DT, and NN Results – with Feature Selection*

Table IX represents the accuracy results achieved by each method for every fold with feature selection. The accuracy results achieved by each method for every fold with feature selection can be clearly identified from the chart in Fig. 8.

TABLE IX. ACCURACY RESULTS – WITH FEATURE SELECTION

Fold	Methods Name		
	SVM	DT	NN
Fold-1	97,06 %	**97,28 %**	**97,28 %**
Fold-2	96,98 %	96,98 %	96,98 %
Fold-3	96,81 %	96,76 %	97,18 %
Fold-4	96,02 %	96,54 %	96,88 %
AVG	96,72 %	96,89 %	**97,08 %**

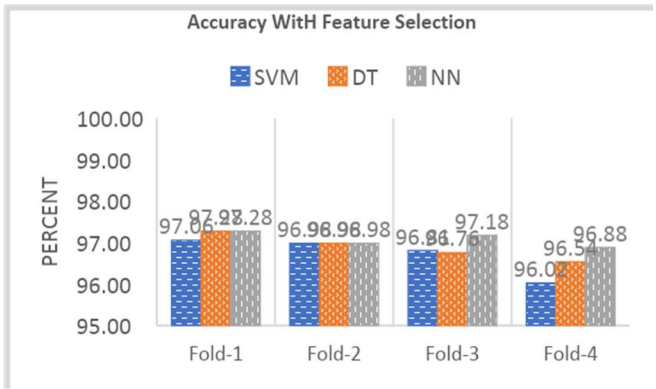

Fig. 8. Accuracy results – with feature selection.

From the data in Table IX, we performed an ANOVA test using the Data Analysis feature in Excel and got the following results that can be seen in Table X and Table XI. On this test, we determined the threshold (alpha) = 0,05. From the ANOVA test result, we can see that the p-value is greater than the alpha (p-value > 0,05). It means the accuracy results between SVM, DT, and NN with feature selection is not stastistically significant (accept the null hypothesis). This means that each classifier relatively has the same ability on classification with feature selection. This may be due to the statistical feature selection we have done. This makes machine learning easier to classify the input with relevant features we have selected.

TABLE X. SUMMARY OF ACCURACY RESULTS – WITH FEATURE SELECTION

Groups	Count	Sum	Average	Variance
SVM	4	386,8724	96,7181	0,2245
DT	4	387,5596	96,8899	0,0990
NN	4	388,3204	**97,0801**	0,0323

TABLE XI. ANOVA TEST ON ACCURACY RESULTS – WITH FEATURE SELECTION

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	0,2623	2	0,13115	1,10587	**0,37202**	4,25649
Within Groups	1,0673	9	0,11859			
Total	1,3296	11				

V. CONCLUSION AND FUTURE WORKS

In this experiment, SVM, DT, and NN classifiers (with and without feature selection) were built. These classifiers were applied on a dataset named "Symptoms and COVID Presence" from Kaggle site. In the case of this dataset, SVM and DT seem less able to get the main features of classification compared to NN. NN has better accuracy performance than SVM and DT (both without features selection and with features selection). NN has the longest training time. However, the testing time of NN is not much different than SVM and DT. The accuracy results among those classifiers were also tested using ANOVA test. In the experiment without feature selection, the accuracy results between SVM, DT, and NN is stastistically significant. Whereas, in the experiment with feature selection, accuracy results between SVM, DT, and NN with feature selection is not stastistically significant.

For future work, it is recommended to work on different machine learning methods that could resolve problems with better outcomes than NN, SVM, and DT. In addition, other datasets both in the medical and non-medical fields can be used for further experiments.

REFERENCES

[1] WHO, 'Coronavirus disease (COVID-19)', 2020. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19. [Accessed: 02-Nov-2020].

[2] Worldmeters, 'Corona Virus Update (Live)', 2020. [Online]. Available: https://www.worldometers.info/coronavirus/. [Accessed: 12-Nov-2020].

[3] K. Roosa et al., 'Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th , 2020', Infect. Dis. Model., vol. 5, pp. 256–263, 2020.

[4] H. S. Maghdid, K. Z. Ghafoor, A. S. Sadiq, K. Curran, and K. Rabie, 'A Novel AI-enabled Framework to Diagnose Coronavirus COVID-19 using Smartphone Embedded Sensors : Design Study', pp. 1–5, 2020.

[5] T. M. Mitchell, Machine Learning. McGraw-Hill Science/Engineering/Math, 1997.

[6] K. Das and R. N. Behera, 'A Survey on Machine Learning : Concept , Algorithms and Applications', Int. J. Innov. Res. Comput. Commun. Eng., vol. 5, pp. 1301–1309, 2017.

[7] R. Sathya and A. Abraham, 'Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification', vol. 2, no. 2, pp. 34–38, 2013.

[8] N. Karankar, P. Shukla, and N. Agrawal, 'Comparative Study of Various Machine Learning Classifiers on Medical Data', Int. Conf. Commun. Syst. Netw. Technol., 2017.

[9] S. L. Salzberg, 'On Comparing Classifiers : Pitfalls to Avoid and a Recommended Approach', vol. 327, pp. 317–327, 1997.

[10] R. A. Fisher, 'Statistical Methods for Research Workers', Biol. Monogr. MANUALS, 1934.

[11] M. G. Larson, 'Analysis of Variance', pp. 115–121, 2008.

[12] Wikipedia, 'Analysis of Variance', 2020. [Online]. Available: https://en.wikipedia.org/wiki/Analysis_of_variance. [Accessed: 01-Nov-2020].

[13] A. Narin, C. Kaya, and Z. Pamuk, 'Automatic Detection of Coronavirus Disease (COVID-19) Using X-ray Images and Deep Convolutional Neural Networks', 2020.

[14] Y. Ünal and M. N. Dudak, 'Classification of Covid-19 Dataset with Some Machine Learning Methods', J. Amasya Univ. Inst. Sci. Technol., 2020.

[15] D. M. Matta and M. K. Saraf, 'Prediction of COVID-19 using Machine Learning Techniques', Blekinge Institute of Technology, 2020.

[16] K. B. Prakash, S. S. Imambi, M. Ismail, T. P. Kumar, and Y. V. R. N. Pawan, 'Analysis , Prediction and Evaluation of COVID-19 Datasets', Int. J. Emerg. Trends Eng. Res., vol. 8, 2020.

[17] Kaggle.com, 'Symptomps and COVID Presence', 2020. [Online]. Available: https://www.kaggle.com/hemanthhari/symptoms-and-covid-presence/metadata. [Accessed: 04-Oct-2020].

[18] J. Brownlee, 'How to Choose a Feature Selection Method For Machine Learning'. [Online]. Available: https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/. [Accessed: 01-Nov-2020].

[19] T. Moudiki, '4-Fold Cross-Validation'. [Online]. Available: https://www.r-bloggers.com/2020/04/linear-model-xgboost-and-randomforest-cross-validation-using-crossvalcrossval_ml/. [Accessed: 01-Nov-2020].

[20] J. R. Quinlan, 'Induction of Decision Trees', pp. 81–106, 2007.

[21] V. N. Vapnik, 'The Nature of Statistical Learning Theory'. Springer, New York, 1995.

[22] T. Mythili, D. Mukherji, N. Padalia, and A. Naidu, 'A Heart Disease Prediction Model using SVM-Decision Trees-Logistic A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL)', Int. J. Comput. Appl. Technol., vol. 68, pp. 10–15, 2013.