OPEN ACCESS

162

ISSN 2356-5462
http://socj.telkomuniversity.ac.id/ijoict/

Intl. Journal on ICT
Vol. 9, No. 2, Dec 2023. pp. 162-176
doi: doi.org/10.21108/ijoict.v9i2.868

Performance of Time-Based Feature Expansion in Developing ANN Classification Prediction Models on Time Series Data

Sri Suryani Prasetiyowati [1], Arnasli Yahya [2], Aniq Atiqi Rochmawati [3]

[1,3]*School of Computing, Telkom University*
Bandung, Indonesia

[2]*Department of Algebra and Geometry, Budapest University of Technology and Economics*
Hungary

* aniqatiqi@telkomuniversity.ac.id

Abstract

The prediction problem in most research is the main goal, to estimate future events related to the field under study. Research on classification that involves the prediction process in it, with spatial-time data and influenced by many features, such as the problem of disease spread, climate change, regional planning, environment, economic growth, requires methods that can predict while solving the problem of features and time. To obtain a time-based classification prediction model using many features, this research uses machine learning methods, one of which is Artificial Neural Network (ANN). The scenario carried out is to develop a $t + r$ classification prediction model by expanding features based on the time $t - r$ of the previous period. The performance of feature expansion in the development of ANN classification prediction models is determined based on the optimal accuracy value of the combination of $t - r$ classification prediction models for the previous time period. By implementing the model on the data, it is found that the performance of time-based feature expansion in ANN classification ranges from 3.5% to 11%. While the optimal accuracy value is obtained from the feature expansion scenario of 3 to 5 time periods earlier.

Keywords: Prediction, Classification, Time-Based, Feature Expansion, ANN

I. INTRODUCTION

FORECASTING has always been a benchmark in decision-making and planning in various fields. The uncertainty of future events is interesting and challenging to study, both individually and in groups by minimizing risk and maximizing utility [1]. Prediction is one of the methods in the field of statistics used to

Received on 19 Nov 2023, Revised on 16 Dec 2023, Accepted and Published on 31 Dec 2023.

study future behaviour. Predictive modelling solutions include data mining technology that analysis historical and current data to generate future models. Machine learning involves computational methods that learn from complex data to build a variety of prediction, classification, and evaluation models [2].

Research involving future predictions is being involved in a variety of areas such as pandemic diseases such as Covid-19 [3] and DBD [4], climate [5], rainfall [6]–[8], air quality index [2], [9], [10], social, environmental, urban planning, and others with time-based spatial data [11]. Predictive models are not only limited to numerical predictions, but also can be extended to classification results. To solve some problems e.g. the spread of diseases, floods, air pollution, unemployment, forest fires, disasters and development of a region for some time to come, some practical information are needed in the form of level classifications.

The use of machine learning methods and regression analysis for prediction of spatial-time data classification has already been explored in some research e.g., [11]–[15] [3] [4] [5] [6]. However, in this research the classification process and the classification prediction process for the future use different methods. The classification process uses machine learning, the prediction process uses a linear regression model with time as the independent variable. For instance, the study in [13] uses ANN and linear regression to predict climate for the next 100 years, and [14] uses Random Forest and linear regression in cases of Covid-19 for 6 days ahead, whereas [11] applies ANN, wavelet method and linear regression for predicting rainfall. Based on these research studies, ANN emerge as the only deep learning method among various other techniques employed. According to [16], [17], Deep Learning is a sub-part of Machine Learning that can represent data on learning that has multiple layers, making it more meaningful.

Based on the research above, there is still the possibility of developing a predictive classification model based on spatial-time data. Consequently, a specific scenario is required to integrate the prediction process with the classification process. The research idea to be developed is to expand features based on previous time to find a classification model some time before. The model is used to predict future classifications. The classification method used is ANN, where ANN is one of the classification methods in Deep Learning which is a continuation of research from the previous period. In this research, we develop a time-based ANN architecture, a time-based feature expansion matrix, a time-based feature expansion ANN model, and measure the performance of the model when implemented on several datasets. The contribution of this research is the development of a time-based classification prediction model using one of the ANN classification methods.

II. LITERATURE REVIEW

A. Prediction

Predictive models are essential in various applications fields such as health, business, climate, transportation [18]. The predictive problem in most research is the primary objective; it can predict future events related to the field being studied. Most of the predictive methods are found in statistical approaches, and typically, predictions are based on time and space. Predictions can also be implemented with machine learning approaches, to solve a variety of structured and unstructured data problems, not limited to regression, clustering, classification, and prediction [19].

B. Artificial Neural Network (ANN)

ANN consists of cells which connected to a system is a computational model based on the biological nerve structure of the brain, where each link has a numerical weight, which concludes the importance of the link. Perceptron is the most important and widely developed model of neural tissue studied in the last six decades [20]. Perceptrons are the simplest form of nerve tissue consisting of a single neuron with an adjustable weight and bias (can change values). Perceptron receives input signals from examples of training data, with given weights, and combined in a linear equation called activation. The single-layer perceptron, introduced by Rosenblatt in 1958, is one of the earliest and simplest models of neural networks [20].

To overcome the limitations that cannot be separated linearly, one can extend SLP to layered structures, such as MLP [20]. Perceptron multilayer (MLP) backpropagation, also called multi-layer advanced feed nerve network. MLP is based on supervised procedures, i.e. a network builds models based on examples in data with known outputs. The MLP consists of three layers (input, hidden, and output) with neurons and processing units. All neurons of one layer are fully connected to the neurons in the adjacent layer. This connection is represented as the weight (connection intensity) in the computational process. The number of neurons in the input layer depends on the number of free variables in the model, whereas the number of neurons in the output layer is equal to the number and number of dependent variables, with the number or number of output neurons being single or multiple [21].

Multilayer perceptrons are the most commonly used type of nerve tissue. Signals are transmitted in a one-way network that is from input to output, there is a loop, where the output of each neuron does not affect the neuron itself. Such architecture is called a forward feed, which is not directly connected to the environment called a hidden layer. The introduction of several layers is determined by the need to increase the complexity of the decision area [22]. The steps in MLP are forward passes, loss crossings, and backward passes. In forward pass, the input is forwarded to the model by multiplying each by weight and adding bias. The output of the j neuron on the hidden layer can be described by the following formula:

$$z_j = \emptyset_h \left(\sum_{i=1}^{n} w_{ij} x_i + b_j \right) \tag{1}$$

As for the value of z_j used to find the output model. The output of the k-neuron on the output layer can be described by the formula as follows:

$$y_k = \emptyset_o \left(\sum_{j=1}^{m} v_{jk} z_j + b_k \right) \tag{2}$$

Based on equations (1) and (2), Fig. 1. is an architecture of the MLP.

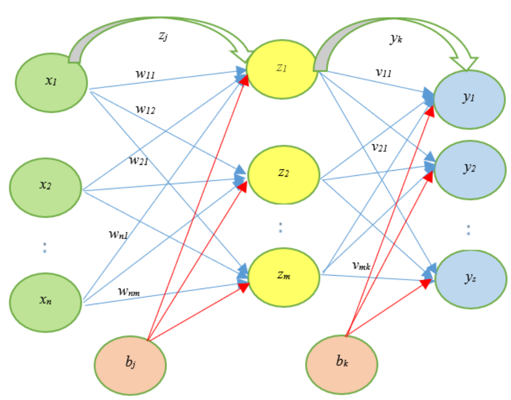

Fig. 1. The architecture on the MLP.

x_i : neuron i^{th} on input layer, $i = 1, 2, 3, \ldots, n$
z_j : neuron j^{th} pada hidden layer, $j = 1, 2, 3, \ldots, m$

y_k : neuon k^{th} pada output layer, $k = 1, 2, 3, \ldots, s$
w_{ij} : weight of neuron i^{th} on the input layer and neuron j^{th} on hidden layer
v_{jk} : weight of neuron j^{th} on a hidden layer and neuron k^{th} on output layer
b_j : bias on z_j
b_k : bias on y_k
$\emptyset_h$: activation function on the hidden layer
$\emptyset_o$: activation function on the output layer

C. *Feature Expansion*

Feature engineering techniques such as feature extraction, feature selection, and feature extension are often applied to classifications [23]. Feature extraction is the process of selecting the best feature subset from the overall feature set, whereas feature extension combines additional features from input data, which combines different relationships between the original features of two objects. The objective is to expand the original vector or form a new feature, which is associated with the distance of each data sample to the number of centroids found by the clustering algorithm. The feature extension approach to the problem of one-dimensional (1-D) time series data classification is the lowest proposed by [24]. The feature extended includes temporal characteristics, frequencies, and statistics. The results showed that the proposed extension of the feature could improve the accuracy of classification, compared to using conventional machine learning. Feature expansion allows classification to consider several dimensions that are inadequate for low-dimensional data. Feature expansion works by taking a feature on the original data and doing something with or on the feature, then adding additional dimensions, to see if there is an improvement in the accuracy of the hyperplane resulting [25]. In feature expansion, it is possible to use linear classification on some data by creating new features in a new dimension.

III. RESEARCH METHOD

The research phase of the development of the prediction model of Artificial Neural Network with time-based feature expansion begins the process of developing the design of the feature extension based on previous $t - r$ time and development of ANN model Time-Based Feature Expansion. After that, an experiment is carried out which is the implementation of the ANN time-based features expansion model on some data.

A. Data Matrix Design with Feature Expansion

The Feature Expansion Process is based on the modeling of predictions of classification based on time. The idea of this study is to determine a predictive model of classification for future $t + r$ using the best result of a combination of previous $t - r$ models. Based on this idea, a classification model with a y_t class target can be developed based on the previous $t - r$ features. Table I shows the spatial-time matrix and Table II shows the expansion of the matrices for designing a model for predicting future $t + r$ classifications based on previous $t - r$ features.

TABLE I

DATA SET MATRIX

Class Target	y_{t-r}	y_{t-r-1}	$\cdots$	y_{t-2}	y_{t-1}	y_t
	$x_{1(t-r)}$	$x_{1(t-r-1)}$		$x_{1(t-2)}$	$x_{1(t-1)}$	$x_{1(t)}$
	$x_{2(t-r)}$	$x_{2(t-r-1)}$		$x_{2(t-2)}$	$x_{2(t-1)}$	$x_{2(t)}$
Features	$\ldots$	$\ldots$		$\ldots$	$\ldots$	$\ldots$
	$x_{(n-1)(t-r)}$	$x_{(n-1)(t-r-1)}$		$x_{(n-1)(t-2)}$	$x_{(n-1)(t-1)}$	$x_{(n-1)(t)}$
	$x_{n\,(t-r)}$	$x_{n(t-r-1)}$		$x_{n\,(t-2)}$	$x_{n(t-1)}$	$x_{n(t)}$

TABLE II
TIME-BASED FEATURE EXPANSION DATA SET MATRIX

Time Prediction	$t+r$	…	$t+r-1$	$t+1$
Class	y_t		y_t	y_t
	$x_{1(t-1)}$		$x_{1(t-r)}$	$x_{1(t-r-1)}$
	…		…	…
	$x_{n(t-1)}$		$x_{n(t-r)}$	$x_{n(t-r-1)}$
	$x_{1(t-2)}$		$x_{1(t-r-1)}$	
Features	…		…	
	$x_{n(t-2)}$		$x_{n(t-r-1)}$	
	…			
	$x_{1(t-r)}$			
	…			
	$x_{n(t-r)}$			

B. *Design of Feature Expansion Model For ANN*

This model is an ANN development by expanding the input layer based on time. It is expected to be used to predict the classification model in the future. This research will develop ANN Multilayer Perceptron with time-based input layer expansion. Based on equations (1), (2), and the design of the data matrix in Table II, then the process of development of the architecture and model of the ANN Time-Based Feature Expansion is described in Fig. 2 and the equations (3) to (13) of the MPL Architecture on the image, developed on the basis of Fig. 2 with an expansion on the input layer part based on $t-r$ previous time.

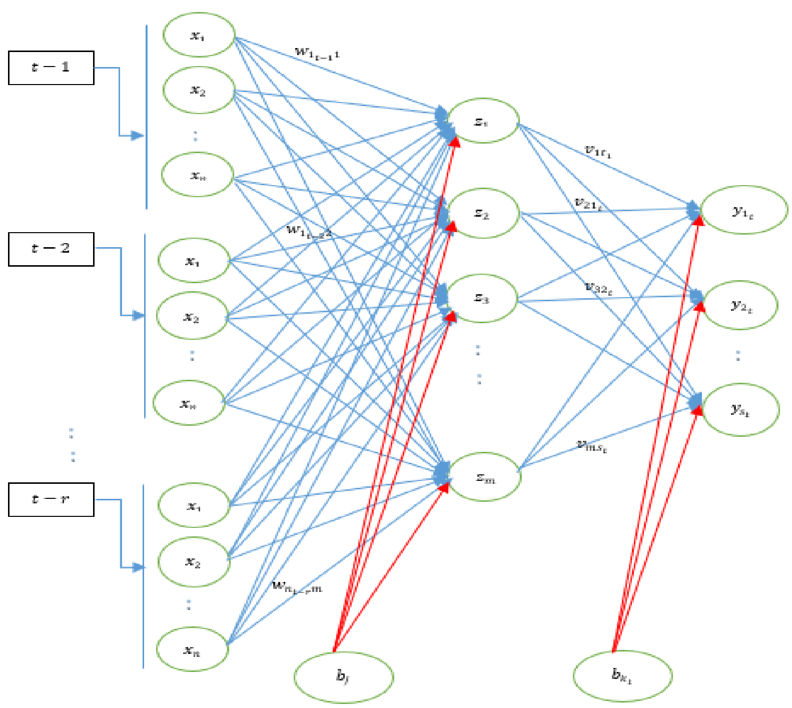

Fig 2. Network structure of MLP with Feature Expansion

Based on the architecture in Fig. 2, on the MLP ANN Time-Based Feature Expansion, the stages performed the extension process to all stages of forward pass, loss calculation, and backward pass. Analog to equations (1) and (2), the input layer is expanded by time, by multiplying each by the weight and added by the bias, then the output of the *j*-neuron on the hidden layer can be developed into the formula as follows.

$$z_j = \emptyset_h\left(\sum_{i=1}^{n} w_{i_{t-r}j} x_{i_{t-r}} + b_j\right) \tag{3}$$

The first layer input at the time before the $t-r$ period is $x_{(t-r)i}$ meets a strong stationary process, where the joint distribution of x_{t_1} , x_{t_2} , …, x_{t_n} with $x_{t_{1-r}}$, $x_{t_{2-r}}$, …, $x_{t_{n-r}}$ is the same for all t and $t-r$. The value of z_j is used to find the output model. The output of the s-th neuron on the outputs layer can be described by the following formula:

$$y_{k_t} = \emptyset_o\left(\sum_{j=1}^{m} v_{jk_t} \emptyset_h\left(\sum_{i=1}^{n} w_{i_{t-r}j} x_{i_{t-r}} + b_j\right) + b_{k_t}\right) \tag{4}$$

Then the weight is updated for each neuron continuously on the loss calculation, therefore the output value approaches the target value. Evaluation of error values is done by calculating the gradient of the loss function. Equation (5) is an error function calculated with the concept of the square number of errors.

$$E = \sum_{k=1}^{s} \frac{1}{2}\left(c_{k_t} - y_{k_t}\right)^2 \tag{5}$$

Using the equation (4), then the error function can be written as follows

$$E = \sum_{k=1}^{s} \frac{1}{2}\left(c_{k_t} - \emptyset_h\left(\sum_{j=1}^{m} v_{jk_t} \emptyset_h\left(\sum_{i=1}^{n} w_{i_{t-r}j} x_{i_{t-r}} + b_j\right) + b_{k_t}\right)\right)^2 \tag{6}$$

E : The error function
c_{t_k} : The k^{th} target at t period
y_{t_k} : The k^{th} output at t period

Further, on the backward pass, the weight correction is performed on both the hidden layer and the output layer, using errors obtained from the second stage with backpropagation. The requirement for the use of backpropagation is that the activation function must be nonlinear and deductible. To meet these requirements, the following logistical or sigmoid functions are used:

$$\varphi(z) = \frac{1}{1+e^{-z}} \tag{7}$$

When the function φ(z) is the derivative form of z, then obtained

$$\frac{d\varphi(z)}{dz} = \varphi(z)\left(1-\varphi(z)\right) \tag{8}$$

Then the magnitude of the change in weight is

$$\Delta v_{jt_k} = -\eta \frac{\partial E}{\partial v_{jk_t}} \tag{9}$$

Δv_{jt_k} : The change in weight
η : The learning rate
$\frac{\partial E}{\partial v_{jk_t}}$: The partial derivative of the error function with respect to v_{jk_t}

Using a chain rule, the partial resultant of an error function against the weight can be written as follows:

$$\frac{\partial E}{\partial v_{jk_t}} = \frac{\partial E}{\partial u^o_{k_t}} \frac{\partial u^o_{k_t}}{\partial v_{jk_t}} \tag{10}$$

By using the equation $u^o_{t_k} = \sum_{j=1}^{m} v_{jk_t} z_j + b_{k_t}$ and $\frac{\partial u^o_{k_t}}{\partial v_{jk_t}} = \frac{\partial\left(\sum_{j=1}^{m} v_{jk_t} z_j + b_{k_t}\right)}{\partial v_{jk_t}} = z_j$

Then $\frac{\partial E}{\partial u^o_{k_t}}$ can be written as

$$\frac{\partial E}{\partial u^o_{k_t}} = -\phi_o\left(u^o_{k_t}\right)\left(c_{k_t} - y_{k_t}\right) \tag{11}$$

By using the equation (11), then the magnitude of the weight changes as follows

$$\Delta v_{jk_t} = \eta \phi'_o\left(u^o_{k_t}\right)\left(c_{k_t} - y_{k_t}\right) z_j \tag{12}$$

The new weight update can be written by

$$v^{r+1}_{jk_t} = v^r_{jk_t} + \Delta v_{jk_t} = v^r_{jk_t} + \eta \phi'_o\left(u^o_{k_t}\right)\left(c_{k_t} - y_{k_t}\right) z_j \tag{13}$$

Where ϕ'_o is a derivative of the activation function, and it becomes a requirement when using backpropagation.

C. *Experiments*

In the experiment, the ANN time-based feature expansion method was implemented on data sets of rainfall, air quality index, and number of Covid-19 cases. Table III and Table IV describe the description and features of the three data sets. Meanwhile, Fig. 3. is a distribution plot of several features from the Covid-19 data set.

TABLE III

DATA SET DESCRIPTION

Data Set	Record	Features	Model Period
Rainfall	4032	6	Day and Month
Air Quality Index	270	6	Day
Covid-19 Cases	2700	22	Month

TABLE IV

FEATURES DESCRIPTION

Data Set	Features
Rainfall	The average temperature (°C), the humidity percentage (%), rainfall (mm), sunlight (hours), average wind speed (m/s), and wind direction (°)
Air Quality Index	Pollutant content of SO_2, NO_2, CO, PM10, PM25, and O_3
Covid-19 Cases	Number of male, number of female, amount of rainfall (%), solar radiation (^{0}C), the average temperature (^{0}C), the maximum temperature (^{0}C), the minimum temperature (^{0}C), number of people not in school, number of people in elementary school, number of people with elementary school graduates, number of people with junior high school graduates, number of people with high school graduates, number of people with 1 year diplomas, number of 2 year diploma graduates, number of 3 year diploma graduates, number of Undergraduate, number of master graduates, number of doctoral graduates, proportion of the first dose of vaccine (%),the proportion of the second dose of vaccine (%), proportion of the third dose of vaccine (%),proportion of adherence to wearing a mask (%),proportion of adherence to physical distancing (%)

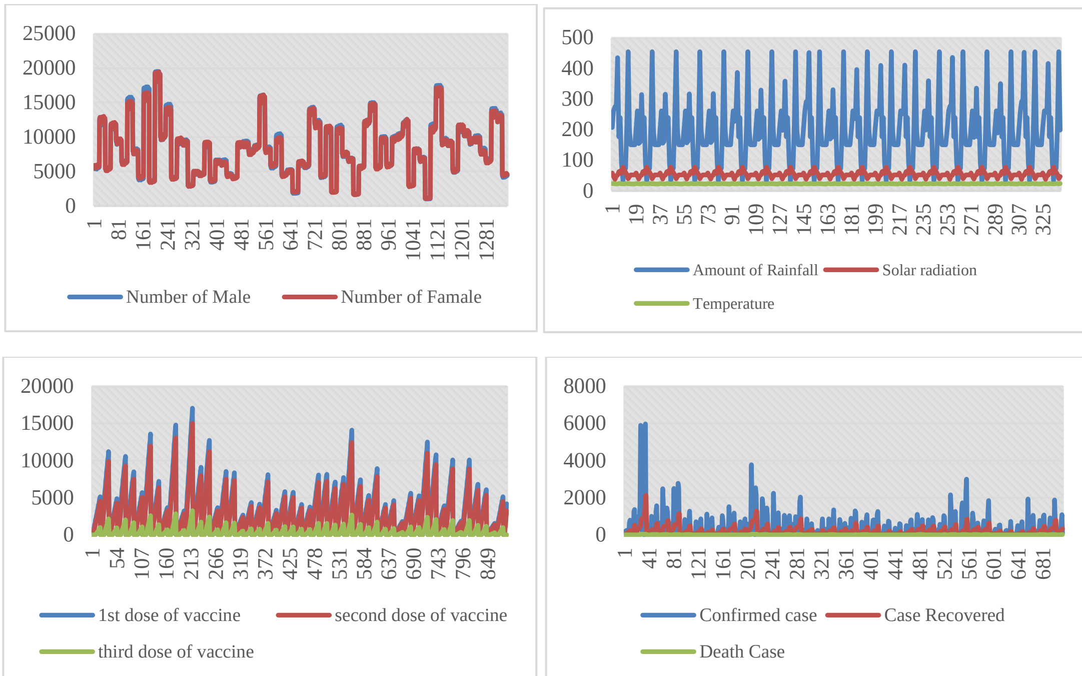

Fig 3. Distribution of Covid-19 Data Features

Table V, Table VI, and Table VII explain the target class labeling for each data. The rainfall data set consists of 6 classes, the Covid-19 data set consists of 3 classes, and the Air Quality Index data set consists of 5 classes.

TABLE V
RAINFALL DATA SET CLASS LABELING

Class	Range	Label
Cloudy	RR < 0	0
Light Rain	0 ≤ RR < 20	1
Moderate Rain	20 ≤ RR < 50	2
Heavy Rain	50 ≤ RR < 100	3
Very Heavy Rain	100 ≤ RR < 150	4
Extreme Rain	IR ≥ 150	5

TABLE VI
COVID-19 DATA SET CLASS LABELING

Class	Range	Label
Low	Cases < 218	0
Medium	218 ≤ Cases < 419	1
High	Cases ≥ 419	2

TABLE VII
AIR QUALITY INDEX DATA SET CLASS LABELING

Class	Range	Label
Good	0 ≤ ISPU < 51	0
Medium	51 ≤ ISPU < 101	1
Unhealthy	101 ≤ ISPU < 200	2
Very Unhealthy	200 ≤ ISPU < 300	3
Harmful	300 ≤ ISPU	4

After the target class labelling process of each data, a feature extension is carried out to develop a classification model based on the previous period of time, using the model design in Table IV. Then the ANN Time-Based Feature Expansion classification process is implemented with the equation (3) to (13). From the combination of classification models that have already been selected, the best features with the best accuracy have been chosen as the model for prediction of future $t + r$ classifications.

D) Evaluation

Evaluation of the performance of implementing time-based feature expansion in ANN for developing classification prediction models, based on classification accuracy values. This accuracy is a measure that describes the system's performance in producing correct predictions. The calculation of classification accuracy in this study uses a multiclass confusion matrix because the number of target classes is more than two. The multi-class confusion matrix is described in Table VIII, which has dimensions NxN, where N is the number of different class labels $C_1, C_2, \dots, C_N$. The analysis is only focused on certain classes and can be carried out based on the characterization described in Table VIII. Equation (18) is a formula for calculating classification accuracy based on a multiclass confusion matrix [26].

$$Accuracy = \frac{\sum_{i=1}^{N} TP(C_i)}{\sum_{i=1}^{N}\sum_{j=1}^{N} c_{i,j}} x100\% \quad (18)$$

TABLE VII

MULTICLASS CLASSIFICATION PROBLEM CONFUSION MATRIX

		Predicted Class			
		C_1	C_2	...	C_N
True Class	C_1	$C_{1,1}$	FP	...	$C_{1,N}$
	C_2	FN	TP	...	FN
	...	...	...	...	
	C_N	$C_{N,1}$	FP	...	$C_{N,N}$

IV. RESULTS AND DISCUSSION

A. Results

In this section, the results of the implementation and testing of the ANN Time-Based Feature Expansion classification model will be described based on the $t - r$ time of the previous three data sets.

TABLE VIII

PERFORMANCE OF THE RAINFALL CLASSIFICATION PREDICTION MODEL FOR THE PREVIOUS 2-6 MONTHS PERIOD

Model Prediction	Model Combination	Target Prediction (month)	Accuracy (%)
2 months before	A	03/22	83.51
	B	02/22	75.00
	C	01/22	86.55
	D	12/21	70.39
	E	12/21, 01/22, 02/22, 03/22	72.38
3 months before	A	03/22	79.37
	B	02/22	84.49
	C	01/22	85.66
	D	12/21	73.06
	E	12/21, 01/22, 02/22, 03/22	76.09

Model Prediction	Model Combination	Target Prediction (month)	Accuracy (%)
4 months before	A	03/22	78.49
	B	02/22	82.55
	C	01/22	84.70
	D	12/21	75.60
	E	12/21, 01/22,02/22, 03/22	76.91
5 months before	A	03/22	78.49
	B	02/22	86.73
	C	01/22	89.17
	D	12/21	80.35
	E	12/21, 01/22,02/22, 03/22	80.60
6 months before	A	03/22	81.39
	B	02/22	84.78
	C	01/22	87.91
	D	12/21	78.45
	E	12/21, 01/22,02/22, 03/22	82.11

TABLE IX

PERFORMANCE OF THE PREDICTION MODEL FOR CLASSIFICATION OF COVID-19 IN THE PREVIOUS 2-5 MONTHS PERIOD

Model Prediction	Model Combination	Target Prediction (month)	Accuracy (%)
2 months before	A	01/22	81
	B	02/22	70
	C	03/22	68
	D	04/22	73
3 months before	A	01/22	72
	B	02/22	73
	C	03/22	70
	D	04/22	92
4 months before	A	01/22	83
	B	02/22	71
	C	03/22	69
	D	04/22	90
5 months before	A	01/22	77
	B	02/22	71
	C	03/22	72
	D	04/22	63

TABLE IX

PERFORMANCE OF THE RAINFALL CLASSIFICATION PREDICTION MODEL FOR THE PREVIOUS 3-7 DAYS PERIOD

Model Prediction	Model Combination	Target Prediction (date/month/year)	Accuracy (%)
3 days before	A	30/03/22	78.04
	B	29/03/22	75.36
	C	28/03/22	80.03
	D	27/03/21	74.97
4 days before	A	30/03/22	78.13
	B	29/03/22	75.43
	C	28/03/22	79.91
	D	27/03/21	74.94
5 days before	A	30/03/22	77.94
	B	29/03/22	75.33
	C	28/03/22	80.64
	D	27/03/21	74.41

Model Prediction	Model Combination	Target Prediction (date/month/year)	Accuracy (%)
6 days before	A	30/03/22	78.62
	B	29/03/22	75.88
	C	28/03/22	81.10
	D	27/03/21	74.21
7 days before	A	30/03/22	77.99
	B	29/03/22	75.13
	C	28/03/22	80.14
	D	27/03/21	75.10

TABLE XI

PERFORMANCE OF THE AIR QUALITY INDEX CLASSIFICATION PREDICTION MODEL IN THE PREVIOUS 5 AND 7 DAY PERIOD

Model Prediction	Model Combination	Target Prediction (date/month/year)	Accuracy (%)
5 days before	A	24/06/20	59.75
	B	25/06/20	54.06
	C	26/06/20	25.60
	D	27/06/20	26.01
7 days before	A	24/06/20	57.47
	B	25/06/20	55.74
	C	26/06/20	28.16
	D	27/06/20	29.31

Based on Table VIII to Table XI, we can obtain models with the best accuracy as predictive classification models for some future time ahead. Table XII contains models used for classification predictions with a significant feature combination.

TABLE XII

OPTIMAL ACCURACY VALUE OF CLASSIFICATION PREDICTION MODEL

Data Set	Prediction Period	The Best Model	Accuracy (%)
Rainfall	Months	2C	86.55
		3C	85.66
		4C	84.70
		5C	89.17
		6C	87.91
	Days	3C	80.03
		4C	79.91
		5C	80.64
		6C	81.10
		7C	80.14
Air Quality Index	Days	5A	59.75
		7A	57.47
Covid-19	Months	2A	81
		3D	92
		4D	90
		5A	77

B. Discussion

This section discusses and analyzes the implementation of the ANN time-based Feature Expansion development. The data used are the data set of rainfall, air quality index, and the number of cases of covid-19 that have been shown in the table and pictures in the results and evaluation section. The tables and images that have already been displayed in the outcomes and evaluations section show that ANN's time-based feature expansion method can be implemented to build a predictive model of classification based on time. Table VIII and Table X show that the accuracy patterns of feature selection results in the ANN Time-Based Feature Expansion method show an increase and a decrease in the combination of models in each period of time. For the monthly period, the prediction accuracy range is from 9.1% to 16.16%. In contrast, for the daily period the predicted accurate range ranges from 4.97% to 6.89%. In Table XII it is shown that the time-based feature expansion of the ANN method increases the accuracy of 4.47% in the month prediction model and 1.09% in the daily predicting model. Similarly, in Table IX, the precision patterns of the Covid-19 data set, the month-based prediction precision range is 13% to 22%, with the impact of the time-based feature extension of 11% from the two-month model to the three-month model. Meanwhile, the larger monthly model drops not significantly.

The implementation of the ANN Time-Based Feature Expansion for the Air Quality Index data set is shown in Table XI. The accuracy range for each daily prediction period was 29.31% to 34.15%, with the highest accuracy only reaching 59.75%. Similarly, the impact of the feature expansion based on the daily time of the air quality index has been reduced by 2.28%. This is influenced by the amount of data that is too small to make a time-based feature expansion [27]. The accuracy of the prediction model using the ANN Time-Based Features Expansion combines all features of all years and records, never exceeding the most accurate model. It is consistent with [28] which states that the larger number of feature dimensions will lead to imbalances in the classification results. Based on the results, it can be demonstrated that the combination scenario of all features and the addition of records has no significant effect on increasing the accuracy.

The highest accuracy value of each model presented in Table XII indicates that the addition of features affects the optimum precision of each data set. On data sets of rainfall based on monthly classification forecast scenarios, the optimal precision is obtained when a feature extension was performed five years earlier. When a feature was extended two, three and four years earlier, the precision drops steadily to 1.85% and rises quite significantly by 4.47% when the amount of feature expansion is increased to five years before. The optimum accuracy in Table XII is 89.17% higher than the accuracy without the feature extended, which is 86.65% [29]. As for the development of the daily classification prediction model, the optimum accuracy is obtained when the feature extension was performed six days earlier, but the increase is not too significant, namely 0.46%. In the daily Air Quality Index data set classification forecast model, optimum precision is achieved 59.75% of the previous five days' feature expansion. The optimal precision of the monthly classification model prediction on the Covid-19 data set is achievable when the features are expanded three or four months earlier, with a fairly significant increase of 9% to 11%. Based on the size of the optimal exactitude improvement, the greatest improvement occurs in the COVID-19 data sets, which can reach 11%. The optimal accurate value in Table XII is 92% higher than the non-functional precision value of 71.21% [30]. Thus, on big data with a lot of features by doing an extension of $t - r$ features previously, then you will get a choice of combination of many features with a high accuracy.

V. CONCLUSION

Time-based feature expansion method can be used to build a classification prediction model for the future. The classification prediction model is obtained by building a classification model through the expansion of the previous time $t - r$ features with the target class $t + r$. Time-Based Feature Expansion implemented on Rainfall and Covid-19 data sets with ANN classification is proven to increase the accuracy value of the model. While the performance of Time-Based Feature Expansion on data sets with too little data, features and time periods, has not been optimal in increasing the accuracy value. By implementing the model on the data, it is found that the performance of time-based feature expansion in ANN classification ranges from 3.5% to 11%. The optimal accuracy value is obtained from the feature expansion scenario of 3 to 5 time periods earlier. Future work focuses on the application of Time-Based Feature Expansion to other classification methods, performance comparison with the LSTM method, and implementation of data sets that have been used by other studies.

ACKNOWLEDGMENT

The authors wish to thank Telkom University for the support of funds and infrastructure so that this research can be completed.

REFERENCES

[1] F. Petropoulos *et al.*, “Forecasting: theory and practice,” *International Journal of Forecasting*, vol. 38, no. 3. Elsevier B.V., pp. 705–871, 2022, doi: 10.1016/j.ijforecast.2021.11.001.

[2] Y. C. Liang, Y. Maimury, A. H. L. Chen, and J. R. C. Juarez, “Machine learning-based prediction of air quality,” *Appl. Sci.*, vol. 10, no. 24, pp. 1–17, 2020, doi: 10.3390/app10249151.

[3] C. Anastassopoulou, L. Russo, A. Tsakris, and C. Siettos, “Data-based analysis, modelling and forecasting of the COVID-19 outbreak,” *PLoS One*, vol. 15, no. 3, 2020, doi: 10.1371/journal.pone.0230405.

[4] S. S. Prasetiyowati and Y. Sibaroni, “Prediction of DHF disease spreading patterns using inverse distances weighted (IDW), ordinary and universal kriging,” *J. Phys. Conf. Ser.*, vol. 971, no. 1, pp. 0–15, 2018, doi: 10.1088/1742-6596/971/1/012010.

[5] D. Jang, “An Application of ANN Ensemble for Estimating of Precipitation Using Regional Climate Models,” *Adv. Civ. Eng.*, vol. 2021, 2021, doi: 10.1155/2021/7363471.

[6] A. N. Sihananto and W. F. Mahmudy, “Rainfall Forecasting Using Backpropagation Neural Network,” *J. Inf. Technol. Comput. Sci.*, vol. 2, no. 2, pp. 66–76, 2017.

[7] W. M. Ridwan, M. Sapitang, A. Aziz, K. F. Kushiar, A. N. Ahmed, and A. El-Shafie, “Rainfall forecasting model using machine learning methods: Case study Terengganu, Malaysia,” *Ain Shams Eng. J.*, vol. 12, no. 2, pp. 1651–1663, 2021, doi: 10.1016/j.asej.2020.09.011.

[8] A. Y. Barrera-Animas, L. O. Oyedele, M. Bilal, T. D. Akinosho, J. M. D. Delgado, and L. A. Akanbi, “Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting,” *Mach. Learn. with Appl.*, vol. 7, p. 100204, 2022, doi: 10.1016/j.mlwa.2021.100204.

[9] A. J. Lepperød, A. Jacobsen, and L. Ntnu, “Air Quality Prediction with Machine Learning.” 2019.

[10] S. M. S. Nagendra and M. Khare, “Artificial neural network approach for modelling nitrogen dioxide dispersion from vehicular exhaust emissions,” *Ecol. Modell.*, vol. 190, no. 1–2, pp. 99–115, 2006, doi: 10.1016/j.ecolmodel.2005.01.062.

[11] A. S. Raj, R. J. Usha, S. Akshaya, K. Saranya, and D. Shyamilee, “Artificial Neural Network Model for Rainfall Data Analysis During 2004-2017 in Tamil Nadu, India – Prevailing Pattern Evaluation on Climate Change,” *J. Model Based Res.*, vol. 1, no. 2, pp. 34–47, 2020, doi: 10.14302/issn.2643-

2811.jmbr-20-3402.

[12] F. MinminMiao, HongZeng, AiminWang, ChangsenZhao, “Discriminative spatial-frequency-temporal feature extraction and classification of motor imagery EEG: An sparse regression and Weighted Naïve Bayesian Classifier-based approach,” *elsevier*, vol. 278, pp. 13–24, 2017.

[13] M. Akhter and M. A. Ahanger, “Climate modelling using ANN,” *Int. J. Hydrol. Sci. Technol.*, vol. 9, no. 3, pp. 251–265, 2019, doi: 10.1504/IJHST.2019.102316.

[14] C. M. YEŞİLKANAT, “Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm,” *Chaos, Solitons and Fractals*, vol. 140, 2020, doi: 10.1016/j.chaos.2020.110210.

[15] J. I. abMunag. V.N.K.Prasadb, S.Nickolasa, and G.R.Gangadharan, “Representational primitives using trend based global features for time series classification,” *elsevier*, vol. 167, 2021, doi: https://doi.org/10.1016/j.eswa.2020.114376.

[16] L. Deng and D. Yu, “Deep learning: Methods and applications,” *Found. Trends Signal Process.*, vol. 7, no. 3–4, pp. 197–387, 2013, doi: 10.1561/2000000039.

[17] H. Wang, R. Czerminski, and A. C. Jamieson, “Neural Networks and Deep Learning,” *Mach. Age Cust. Insight*, pp. 91–101, 2021, doi: 10.1108/978-1-83909-694-520211010.

[18] M. Robnik-Šikonja, “Explanation of prediction models with explain prediction,” *Inform.*, vol. 42, no. 1, pp. 13–22, 2018.

[19] B. Nikparvar and J. C. Thill, “Machine learning of spatial data,” *ISPRS Int. J. Geo-Information*, vol. 10, no. 9, pp. 1–28, 2021, doi: 10.3390/ijgi10090600.

[20] K. L. Du, C. S. Leung, W. H. Mow, and M. N. S. Swamy, “Perceptron: Learning, Generalization, Model Selection, Fault Tolerance, and Role in the Deep Learning Era,” *Mathematics*, vol. 10, no. 24. MDPI, 2022, doi: 10.3390/math10244730.

[21] Y. S. Park and S. Lek, *Artificial Neural Networks: Multilayer Perceptron for Ecological Modeling*, vol. 28. Elsevier, 2016.

[22] V. E. Balas, N. E. Mastorakis, M.-C. Popescu, and V. E. Balas, “Multilayer perceptron and neural networks Analysis and Synthesis of perturbed Duffing oscillators’ View project Education View project Multilayer Perceptron and Neural Networks.” 2009, [Online]. Available: https://www.researchgate.net/publication/228340819.

[23] A. K. Sanjiv Storcheus Dmitry; Rostamizadeh, “A Survey of Modern Questions and Challenges in Feature Extraction,” *1st Int. “Feature Extr. Mod. Quest. Challenges,”* vol. 44, pp. 1–18, 2015.

[24] Daeun Jung; Jungjin Lee; Hyunggon Park, “Feature expansion of single dimensional time series data for machine learning classification,” in *IEEE Xplore*, 2021, p. 181, doi: 10.1109/ICUFN49451.2021.9528690.

[25] J. Eden, “Expand Your Horizons,” 2021.

[26] I. Markoulidakis, I. Rallis, I. Georgoulas, G. Kopsiaftis, A. Doulamis and N. Doulamis, “Multiclass Confusion Matrix Reduction Method and Its Application on Net Promoter Score Classification Problem,” Technologies, 2021, 9, 81, doi: 10.3390/technologies9040081

[27] R. F. Ramadhani, S. S. Prasetiyowati and Y. Sibaroni, "Performance Analysis of Air Pollution Classification Prediction Map with Decision Tree and ANN," vol. 3, no. 4, pp. 536–543, 2024, doi: 10.47065/josyc.v3i4.2117.

[28] D. A. Petrusevich, "Features addition and dimensionality reduction in classification," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 919, no. 4, 2020, doi: 10.1088/1757-899X/919/4/042018.

[29] A. Gumilar, S. S. Prasetiyowati, and Y. Sibaroni, "Performance Analysis of Hybrid Machine Learning Methods on," vol. 5, no. 158, pp. 481–490, 2022.

[30] C. Natalia, A. A. Rohmawati, and S. S. Prasetyowati, "Classification of COVID-19 Monthly Cases Using Artificial Neural Network (ANN) Method," vol. 4, no. 1, pp. 33–41, 2022, doi: 10.47065/josh.v4i1.2236.