Utilization of data mining classification techniques to identify the effect of Madden-Julian Oscillation on increasing sea wave height over East Java Waters

F Alfahmi1,2,3, O S Hakim2, R C Dewi1 and A Khaerima4
1 Department of Geophysics and Meteorology, IPB University, Indonesia
2 Juanda Meteorological Station, BMKG, Indonesia
3 Department of Mathematics Faculty of Mathematics, Computation, and Data Science, Sepuluh Nopember Technical College, Indonesia
4 Bogor Climatological Station, BMKG, Indonesia

E-mail: furqon.alfahmi@bmkg.go.id

Abstract. East Java BPBD data recorded 18 marine accidents in 2018, which increased by 1 event compared to the previous year. It is interesting to study the waters around East Java which are divided into 9 regions. The wind is a major factor in the high wave generation, but the contribution of weather phenomena triggered by the marine environment is important to identify. Phenomena such as Madden-Julian Oscillation (MJO) has a cycle through the Indonesia territory, becomes a factor that should be suspected. MJO identification uses the Real-Time Multivariate MJO (RMM)-1 and RMM-2 index, which can be combined with the wind speed data using data mining classification techniques to get the thresholds value of wave height data obtained from the analysis of Windwave-05 model. The classification is helped by WEKA's machine learning algorithm, by determining 4 selected classification algorithms including Naïve Bayes, J48, JRip, and Multi-Class Classifier. The data validation using the K-fold cross-validation method with a number of folds is 10 units. The accuracy value of the best algorithm obtained in each waters region ranges from 63.02% to 84.50%. The overall accuracy value increases by 0.24% to 4.41% compared to only using wind factors, except for the Waters of Bawean Island and Masulembu Islands.

Keywords: MJO, wave height, data mining classification techniques

1. Introduction
Arin Jaya Motorboat was sailed from Gowa Gowa Island and sank in Giliyang Waters, Sumenep, Madura, East Java on June 17, 2019. The cause of this tragedy was bad weather and high waves after the motorboat sailed 20 minutes. This boat carried dozens of passengers which causing dozens of lifeless [1]. Data from East Java BPBD recorded 18 sea accidents throughout 2018 in East Java Waters, which increased by 1 incidence compared to the previous year. On May 27, 2018, there were also dozens of fishermen were drown when they sailed after their ship was wrecked hitting by waves at Puger Beach, Selayan, Jember, with one of the 23 fishermen was found to be lifeless [2].

A wave is a movement of up and down water in a direction perpendicular to the sea level that forms a sinusoidal curve [3]. The World Meteorological Organization (WMO) classifies ocean waves

To whom any correspondence should be addressed (furqon.alfahmi@bmkg.go.id)
based on the height shown in table 1 [4]. Ocean waves generally arise by the influence of the wind, although there are still other factors that can cause waves in the sea such as seismic activity on the seabed (earthquake), volcanic eruptions, and attraction forces of celestial bodies (moon and sun) [5]. The wind speed value directly proportional to the wave height, which when the westerly season with strong wind speed, wave height higher than when the eastern season with weak wind speed [6].

Table 1. Douglas sea scale of wave height

| Code | Wave Height (m) | Characteristic Name |
|------|-----------------|---------------------|
| 0    | 0.0             | Calm (Glassy)       |
| 1    | 0.0 – 0.1       | Calm (Rippled)      |
| 2    | 0.1 – 0.5       | Smooth              |
| 3    | 0.5 – 1.25      | Slight              |
| 4    | 1.25 – 2.5      | Moderate            |
| 5    | 2.5 – 4.0       | Rough               |
| 6    | 4.0 – 6.0       | Very Rough          |
| 7    | 6.0 – 9.0       | High                |
| 8    | 9.0 – 14.0      | Very High           |
| 9    | > 14.0          | Phenomenal          |

The interaction of the ocean and atmosphere can include microphysical and macro. Examples of the ocean and atmospheric interactions on the global scale are El-Nino Southern Oscillation (ENSO) events in the Pacific Ocean and Dipole Mode in the Indian Ocean [7]. Madden Julian Oscillation (MJO) is a periodic intra-seasonal oscillation between 30 - 90 days which influences the movement of convection areas from the Indian Ocean to the east through the province of West Sumatra and results in high rainfall in the area passed [7]. The MJO activities are closely related to increased convection in tropical regions that can interact with the atmosphere on a global scale [8,9].

The Bureau of Meteorology (Australian Meteorological Agency) monitors MJO using the Real-time Multivariate MJO Index (RMM1 and RMM2) [10]. This index involves wind variables with a height of 200 mb and 850 mb, as well as using Outgoing Longwave Radiation (OLR) data [10]. RMM1 and RMM2 are mathematical models in a combination of the number of clouds and winds with upper and lower altitudes from the atmosphere to measure the strength and location or phase of MJO [11]. The 8 phases correlated with the geographical location of the negative outgoing longwave radiation (OLR) anomaly associated with the active phase of the MJO as shown in figure 1 [12,13].

In the analysis to find out the relationship between response variables and predictors, data mining classification techniques can be used. Data mining or often referred to as knowledge discovery in databases (KDD) is an activity that includes gathering, using historical data to find order, pattern or relationship in large data sizes [14]. Classification is processing to find models or functions that explain and characterize concepts or data classes, for specific purposes [15]. Machine Learning can be used to help implement data mining classification techniques, which is one of the popular learning machines that will be used in this research is WEKA. Machine learning learns how machines or computers can learn from experience or how to program machines for learning, which requires data to be learned so it is usually also called learning from data [16]. Act as a rapid machine learning models, efficient to wave height forecast system [17].

The background of this study was that there were still many marine accidents caused by bad weather or high waves in East Java waters. Recently, Meteorological, Climatological, and Geophysical Berau (BMKG) have predicted the wave height model using a numerical model through only inputting the effect of winds. Interaction with another phenomenon not clearly described in this model. This research was lead to predicts wave height using data mining classification techniques.
With data mining classification techniques, the impact of high waves clearly can be calculated and understood. This study is to be able to become a proposed a predictor model for BMKG and for a more accurate predictor that can be used as a reference to provide early warnings to reduce the number of accidents at sea.

Figure 2. The distribution of wave height data based on the classification of the Douglas sea scale.

2. Methods
The data used in this study were RMM1, RMM2 and wind speed as predictor variables. The response variable uses significant wave height data. RMM1 and RMM2 data were obtained from BOM, while wind speed and wave height data were obtained from Windwave-05 model analysis at 9 points representing each East Java waters region. The 9 waters sample points are; Eastern of Java Sea (113.572600° E and 5.204826° S), Bawean Island Waters (112.720700° E and 5.950492° S), Masalembu Islands Waters (114.767800° E and 5.383968° S), Kangean Islands Waters (115.537600° E and 7.202166° S), Waters north of East Java (112.000000° E and 6.534514° S), Gresik Waters (112.721200° E and 6.856208° S), Madura Strait (113.440600° E and 7.448720° S), South East Java Waters (112.683600° E and 8.716989° S), and the southern Indian Ocean of East Java (112.731200° E and 9.682922° S).

In the analysis of the MJO phase pattern temporally, daily RMM1 and RMM2 data were used from June 1st, 1974 to February 1st, 2019. Every MJO phase was challenged within 1 month of its appearance. Furthermore, in constructing wave height classification algorithms, RMM1 and RMM2 data only followed random hourly Windwave-05 output data of 1271 days from February 1st, 2013 to January 1st, 2018, but only maximum daily data was taken. The wave height data distribution is shown in figure 2. The data validation technique used is k-fold cross-validation with the number of folds is 10. In this technique the data collection is divided into several groups, a classifier is studied using k-1 fold, and the error value is calculated by testing the classifier in the group remaining [18].

The wave height classification was tested using Naïve Bayes, J48, JRip, and Multi-Class algorithm Classifiers. Naïve Bayes classifiers assume that the effects of attribute values on a given class do not depend on other attribute values [19]. J48 implements the C4.5 Quinlan algorithm to produce a decision tree pruned or unpruned C4.5 [20]. JRip class implements student propositional rules, Repetitive Pruning to Produce Error Reduction (RIPPER), proposed as an optimized IREP version
The Multi-Class is a meta-classifier for handling multi-class datasets with a 2-class classifier [21]. The value used as a reference for accuracy is Percent Correct [22]. The accuracy of the best algorithm obtained from each waters region is then deducted by the accuracy of the same algorithm by only using predictor variables only wind speed. If the greater positive value obtained, then the effect of MJO on wave height is greater too. Conversely, if the negative value of the difference is obtained, then RMM1 and RMM2 are not the supporting factors of wave height.

3. Results and Discussion

3.1. MJO Phase Pattern Temporally

MJO moved from west to east with 8 phases. Each phase described the location of an active MJO. Phase 3 and 4 were located in the west and east Indonesia area. Based on figure 3, the research found that that during data observation, MJO possible occurred in every month. The difference between often appear phases and rarely appear phases, only ranges from 3.15% in March to 12.70% in February. Each phase had the highest MJO occurrence percentage value at 1 or 2 months. Phase 1 possessed 5 months as the lowest percentage value which becoming the most compared to the others. If the occurrence of MJO was accumulated throughout the year, then the difference was only 1.97% between Phase 2 which was often appears phase and Phase 8 which was rarely appears.

Phase 4 and phase 5 which show the appearance of the MJO in the Maritime Continent or Indonesia territories only had the highest percentage occurrence value at 1 month, which is phase 4 in April and phases 5 in October. However, both phases did not have the lowest percentage month. In accumulation throughout the year, phase 4 has the fourth-highest percentage value with a value of 12.39% and phase 5 has the third-highest percentage value with a value of 12.99%.

![Figure 3. The chance of MJO occurrence on a monthly average.](image)

3.2. Wave Height Classification based on Wind Speed, RMM1 and RMM2

The accuracy of the best algorithm obtained by each waters region was ranged from 63.02% to 84.50% as shown in figure 4. The highest accuracy value was obtained by the Multi-Class algorithm on Gresik Waters, while the lowest was obtained by the JRip algorithm on the Masalembu Islands Waters. On the other hand, the Naïve Bayes and J48 algorithms were both the best with the most
number, namely in 3 waters territorials. However, Naïve Bayes produced the higher accuracy values compared to J48.

Based on table 2, it can be concluded that the Naïve Bayes algorithm was suitable for waters that had wave height variation characteristics from Smooth to Very Rough. On the other hand, the JRip algorithm corresponded to the less variation in wave height from Slight to Very Rough as shown in table 3. The least variation was shown in table 4 in the Multi Class Algorithm, with variations also in the low wave height of Calm to Slight. The non-uniform variations between waters in accordance with the J48 Algorithm were shown in table 5. The waters of the Kangean Islands was varied from Smooth to Very Rough, The northern waters of East Java was also varied from Smooth to Rough, while the Madura Strait was varied from Smooth to Moderate.

![Figure 4. The accuracy of the algorithm classifiers obtained by each water region.](image)

| Eastern of Java Sea | Smooth | Slight | Moderate | Rough | Very Rough |
|---------------------|--------|--------|----------|-------|------------|
| Wind Velocity (kt)  | 4.9 ± 1.2 | 7.9 ± 2.3 | 13.8 ± 2.0 | 18.4 ± 1.7 | 23.6 ± 1.2 |
| RMM1                | -0.07 ± 1.04 | 0.13 ± 0.94 | -0.09 ± 0.93 | -0.11 ± 0.99 | 0.60 ± 0.97 |
| RMM2                | -0.32 ± 1.06 | 0.08 ± 0.98 | -0.14 ± 0.89 | 0.03 ± 0.83 | 0.49 ± 1.04 |

| Bawean Island Waters | Smooth | Slight | Moderate | Rough | Very Rough |
|-----------------------|--------|--------|----------|-------|------------|
| Wind Velocity (kt)    | 5.2 ± 1.6 | 8.4 ± 2.6 | 14.0 ± 2.3 | 18.5 ± 2.0 | 23.0 ± 1.6 |
| RMM1                  | 0.05 ± 0.82 | 0.13 ± 1.03 | -0.11 ± 0.89 | -0.14 ± 1.06 | 0.69 ± 1.01 |
| RMM2                  | -0.08 ± 0.97 | 0.07 ± 1.00 | -0.19 ± 0.89 | 0.15 ± 0.93 | 0.74 ± 1.02 |

| Southern Waters of East Java | Smooth | Slight | Moderate | Rough | Very Rough |
|-----------------------------|--------|--------|----------|-------|------------|
| Wind Velocity (kt)          | 3.2 ± 0.00 | 8.1 ± 3.4 | 11.5 ± 4.3 | 15.4 ± 4.1 | 18.0 ± 3.4 |
| RMM1                        | -0.73 ± 0.00 | -0.57 ± 0.94 | 0.03 ± 0.92 | -0.02 ± 1.03 | 0.40 ± 0.89 |
| RMM2                        | -1.17 ± 0.00 | -0.07 ± 1.18 | -0.06 ± 0.90 | -0.09 ± 0.94 | 0.97 ± 0.94 |
Table 3. Location for the best Algorithm (JRip Algorithm)

| Masalembu Islands Waters | Slight | Moderate | Rough | Very Rough |
|-------------------------|--------|----------|-------|------------|
| Wind Velocity (kt)      | ≤ 12.0 | > 12.0 and < 16.3 | ≥ 16.3 | ≥ 21.5     |
| RMM1                   | ≤ -1.01|          |       |            |
| RMM2                   | ≤ 0.43 |          |       |            |
| The Indian Ocean of southern East Java | Slight | Moderate | Rough | Very Rough |
| Wind Velocity (kt)      | ≤ 7.0  | > 12.0 and < 12.9 | ≥ 12.9 | ≥ 14.2     |
| RMM1                   | ≤ -1.03| ≤ -1.00  | ≤ 0.44 | ≥ -0.37    |
| RMM2                   | ≤ -1.13| ≤ -0.38  | ≥ -0.01| ≥ 0.63     |

Table 4. Location for the best Algorithm (Multi-Class Algorithm)

| Gresik Waters | Coefficients |
|---------------|--------------|
|               | Calm | Smooth | Slight |
| Intercept     | -1.5001 | -2.7463 | 11.8967 |
| Wind Velocity (kt) | 0.466 | 0.1159 | -0.6706 |
| RMM1          | 0.2182 | 0.0089 | 0.1088  |
| RMM2          | -0.0329| 0.2679 | -0.5571 |

Table 5. Location for the best Algorithm (J48 Algorithm)

| Northern Waters of East Java | Madura Strait | Kangean Islands Waters |
|-----------------------------|--------------|------------------------|
| Wind <= 11.5                | Wind <= 11.0: Smooth | Wind <= 11.6 |
| Wind <= 9.2                 | Wind > 11.0 | RMM1 <= 1.91: Slight |
| Wind <= 6.5                 | RMM2 <= 0.85 | RMM1 > 1.91: Moderate |
| | Wind <= 16.8: Slight | Wind > 11.6 |
| | Wind <= 16.8: Slight | Wind > 11.6 |
| Wind > 6.5                  | RMM1 <= 0.22 | Wind > 11.6 |
| RMM2 <= 1.49: Smooth        | Wind <= 17.9: Slight | RMM2 <= 1.06 |
| | Wind <= 17.9 | RMM1 <= 0.56 |
| | RMM1 <= -0.32 | RMM2 <= -0.48 |
| RMM2 > 1.49                | Wind > 17.9 | RMM1 <= -0.32 |
| Wind <= 7.2: Slight         | RMM1 <= 0.08 | RMM2 <= -1.34 |
| Wind > 7.2: Slight          | RMM2 <= 0.24 | RMM1 <= -0.07 |
| Wind > 9.2: Slight          | RMM1 <= -0.07 | RMM2 > -0.48 |
| Wind > 11.5                 | RMM2 <= -0.48 | RMM1 > 0.56: Rough |
| Wind <= 16.0                | RMM1 > 0.22 | RMM1 > 0.08: Rough |
| Wind <= 14.73: Slight       | Wind <= 18.1 | RMM1 > 0.56: Rough |
| Wind <= 14.73: Slight       | Wind <= 16.8: Smooth | Rough |
| RMM2 <= 0.86                | Wind <= 20.3 | RMM1 <= 0.08 |
| | Smooth | Wind > 20.5 |
| | RMM1 <= 0.95 | RMM1 <= 0.08: Rough |
| | RMM1 <= -0.95 | RMM1 <= -0.30: Rough |
| | RMM1 > 0.95: Slight | RMM1 > 0.08: Rough |
| | RMM2 <= 1.06: | RMM2 <= 1.06: Rough |
### Northern Waters of East Java

|          | RMM2 > 1.06: Rough | Wind > 22.1 | RMM1 <= 0.49 | Wind <= 24.5: Rough | Wind > 24.5: Moderate |
|----------|--------------------|-------------|--------------|---------------------|----------------------|
|          | Smooth             | Slight      | Rough        | Smooth              | Slight               |
|          | RMM2 > -1.01:      | Wind > 18.1 | RMM1 <= 1.61 | RMM1 <= 0.97:       |                      |
|          |                    |              |              |                     | RMM1 > 1.61: Smooth  |
|          |                    |              |              |                     | RMM2 > 0.85          |
|          |                    |              |              |                     | Wind <= 21: Smooth   |
|          |                    |              |              |                     | Wind > 21: Slight    |

### Madura Strait

|          | Wind <= 20.7: Rough | Wind > 20.7: Very |
|----------|--------------------|-------------------|
|          | Smooth             | Rough             |

### Kangean Islands Waters

|          | RMM2 > 0.85        |
|----------|-------------------|

Based on 3 waters which were in accordance with the Naïve Bayes algorithm, it was suitable for waters with a fairly long wave height variation characteristic. Compared to Naïve Bayes, the JRip algorithm possessed a shorter range of wave height variations. The opposite pattern was shown in Gresik waters which was the only one according to the Multi-Class algorithm, but the shortest and lowest value variation. The pattern of variations in wave height characteristics, were not the same as shown in the J48 algorithm.

#### 3.3. Effect of MJO on Wave Height

The accuracy of the best classification algorithm between the use of predictor variables RMM1, RMM2 and wind speed on the use of predictor variables only wind speed, produced a positive value in 7 waters region as shown in figure 5. Positive difference values indicated that the addition of predictor variables was RMM1 and RMM2 have increased the value of accuracy. Madura Strait waters had the greatest positive accuracy difference compared to other waters. However, the accuracy of this waters region by calculating 3 predictor variables was only 66.56%. Gresik waters that had the highest accuracy value only have a difference of +0.63, which indicates the dominance of wind speed factors in these waters.

The decrease in the value of accuracy was only shown in the Masalembu Islands and Bawean Island waters. The wave height in these two water areas can be declared not affected by the MJO. This makes the Masalembu Islands Waters the lowest in both the accuracy and difference values. It is interesting to learn more about the wave height in the waters of the Masalembu Islands.

However, with 7 of the 9 territorial waters that had a positive difference in value, the majority of wave heights in East Java were influenced by MJO. Only local waters that have a small coverage area and are close to the shoreline are likely not affected by the MJO phenomenon.
Figure 5. The difference value between the results of the best classification algorithm by calculating the velocity wind and MJO factors to only the velocity wind factor.

4. Conclusion
The Madden Julian Oscillation (MJO) appearances in the territory of Indonesia most often occurred during the monsoonal rain type transition, precisely in April and October. The Gresik Waters obtained the highest accuracy value with Multi-Class, but the variations range was shortest and the normal wave height was the lowest. The Naïve Bayes and JRip had the most suitability in 3 waters territorial, but better accuracy and more uniform wave height variation characteristics were shown in Naïve Bayes. The characteristics of uniform wave height variations were also shown in the JRip, but the accuracy was the lowest. Thus among other algorithms, Naïve Bayes was the best of the classifier to wave height based on the influence of the wind speed and the MJO factors. The negative difference results have made the Waters of Bawean Island and Masalembu Islands was not affected by the MJO phenomenon. The highest MJO contribution was obtained in the Madura Strait region, but the accuracy was low. Therefore it was necessary to study more deeply, related to weather phenomena or factor others that contribute to wave heights in East Java Waters.

5. References
[1] Arnani M 2019 Kronologi hingga jumlah korban ini 5 fakta kapal tenggelam di Sumenep https://regional.kompas.com/read/2019/06/18/17350231/kronologi-hingga-jumlah-korban-ini-5-fakta-kapal-tenggelam-di-sumenep accessed Juli 27th 2019
[2] Sugiarto B 2018 Kapal nelayan karam dihantam ombak, puluhan tenggelam 1 tewas https://www.inews.id/daerah/jatim/kapal-nelayan-karam-dihantam-ombak-puluhan-tenggelam-1-tewas accessed July 27th 2019
[3] Hidayat J J, Yusuf M and Indrayanti 2013 Dinamika penjalaran gelombang menggunakan model CMS-wave di Pulau Parang Kepulauan Karimunjawa J. Oceanografi 2 pp 255-264
[4] World Meteorological Organization 1998 Guide to wave and forecasting WMO-No 702 second edition (Geneva: Secretariat of WMO)
[5] Nining S N 2000 Oceanografi Fisis (Bandung, ID: ITB)
[6] Sugianto D N 2010 Model Distribusi Data Kecepatan Angin dan Pemanfaatannya dalam Peramalan Gelombang di Perairan Laut Paciran Jawa Timur Jurnal Ilmu Kelautan 15 pp 143-152
[7] Aldrian E 2008 Meteorologi Laut Indonesia (Jakarta: BMG)
[8] Zhang C and Ling J 2012 Potential Vorticity of the Madden–Julian Oscillation J. Atmos. Sci. 69 pp 65–78

[9] Zhang C, Gottschalk J, Maloney E D, Moncrieff M W, Vitart F, Waliser D E, Wang B and Wheeler M C 2013 Cracking the MJO nut. Geophys. Res. Lett. 40 pp 1223–1230

[10] Evana L, Effendy S, and Hermawan E 2008 Pengembangan model prediksi madden Julian Oscillation (MJO) berbasis pada hasil analisis data real-time multivariate MJO (RMM1 DAN RMM2) Jurnal Agromet 22 pp 144-159

[11] Bureau of Meteorology 2019 Madden-Julian Oscillation http://www.bom.gov.au/climate/mjo/ accessed Juli 27th 2019

[12] Donald A M 2004 The Application of real-time multivariate Madden-Julian rainfall forecasting in the mid-latitudes Dissertation of Climatology University of Southern Queensland

[13] Wheeler M C and Hendon H H 2004 An all-season real-time multivariate MJO index: development of an index for monitoring and prediction American Meteorological Society Journal 132

[14] Mujiasih S 2011 Utilization of Data Mining for Weather Forecasting

[15] Defiyanti S and Jajuli M 2015 Integrasi metode klasifikasi dan clustering dalam data mining Konferensi Nasional Informatika

[16] Ethem A 2010 Introduction to machine learning second edition (London, UK: MIT Press)

[17] James S C, Zhang Y and O’Donncha F 2017 A machine learning framework to forecast wave conditions Preprint submitted to Coastal Engineering

[18] Han J, Kamber M and Pei J 2012 Data Mining Concepts and Techniques (USA: Elsevier)

[19] Rodriguez J D and Perez A 2010 Sensitivity analysis of k-fold cross-validation in prediction error estimation IEEE Transactions on Pattern Analysis and Machine Intelligence 32

[20] Ross J 1992 Quinlan: Learning with continuous classes 5th Australian Joint Conference on Artificial Intelligence (Singapore, SG) pp 343-348

[21] George-Nektarios T 2013 Weka Classifiers Summary

[22] Adiyasa A and Kiki 2017 Evaluasi dan verifikasi cuaca publik (Jakarta, ID: BMKG)