Kritik Artikel Ilmiah "Indonesian Seismic Mitigation using Earthquake Predicted Artificial Intelligence Model"

Alden Luthfi - 2206028932 - MPPI B

Judul Penelitian

Jika ditinjau dari substansinya, judul dari artikel ini sudah cukup bagus, hanya saja, judul penelitian ini sebaiknya lebih merincikan bahwa metodologi penelitian ini hanya menggunakan kecerdasan buatan untuk mendasari pengembangan model yang dibuat. Jika ditinjau dari segi bahasa, judul artikel ini masih tidak sesuai dengan kaidah kebahasaan bahasa Inggris. Anak Kalimat dari judul artikel ini mengunakan struktur yang diawali objek dan diakhiri subjek sehingga kata *Predicted* seharusnya diganti dengan bentuk *-ing* menjadi *Predicting*. Kalimat juga bisa disusun menjadi kalimat yang lebih benar secara gramatikal. Contoh judul artikel yang telah diperbaiki adalah *Developing an Artificial Intelligence-Based Model for Earthquake Prediction and Mitigation in Indonesia*.

Isu Sentral Penelitian

Isu sentral dari artikel ini adalah pengembangan suatu model kecerdasan buatan untuk memprediksi gempa bumi lebih dini untuk meminimalisir kerugian yang terjadi akibat gempat tersebut. Isu sentral ini secara eksplisit telah tertulis pada artikel ini. Pada bagian kajian literatur, rtikel ini sudah menyebutkan penelitian-penelitian terdahulu yang dilakukan di Cina, India, Kosta Rika, dan Amerika Serikat. Masing-masing penelitian menggunakan algoritma kecerdasan buatan yang beragam. Meskipun demikian, artikel ini tidak menyebutkan alasan mengapa algoritma *Random Forest Regression* lebih dipilih ketimbang semua algoritma yang sudah dipakai pada penelitian-penelitian sebelumnya. Artikel ini langsung menjelaskan bagaimana model ini menggunakan *Random Forest Regression* tanpa menjelaskan keunggulan algoritma tersebut diatas algoritma lain.

Identifikasi Masalah Penelitian

Proses identifikasi masalah pada artikel ini dicantumkan secara singkat dan jelas. Identifikasi masalah secara eksplisit tercantum di bagian pendahuluan. Meskipun demikian, Artikel ini tidak cukup mengatasi keterbatasan dari penelitian ini. Sebagai contoh, dataset yang digunakan untuk pelatihan dan pengujian model tidak dibahas secara mendalam, termasuk potensi bias atau keterbatasan dalam kualitas data. Selain itu, artikel ini tidak mengakui keterbatasan dari algoritma *Random Forest Regression* atau sumber potensial kesalahan dalam prediksi gempa bumi. Diskusi yang lebih komprehensif tentang keterbatasan-keterbatasan ini akan memberikan pandangan yang lebih seimbang tentang temuan penelitian ini.

Metodologi Penelitian

Metodologi penelitian pada artikel ini adalah pengembangan model prediksi gempa bumi berbasis kecerdasan buatan menggunakan algoritma *Random Forest Regression*. *Dataset* yang digunakan untuk pelatihan dan pengujian model terdiri dari catatan gempa bumi di Indonesia dari Januari 1900 hingga Januari 2022. Atribut yang digunakan dalam model mencakup tanggal, lokasi, lintang, bujur, hiposenter, magnitudo, kedalaman gempa bumi, percepatan puncak tanah maksimum atau *Peak Ground Acceleration* (PGA), intensitas Mercalli yang dimodifikasi atau Modified Mercalli Intensity (MMI), dan kelas situs. Model ini dilatih menggunakan 80% dari data dan diuji menggunakan 20% sisanya. Akurasi model *Random Forest Regression* dievaluasi, dan metrik kinerja tingkat kesalahan, *Mean Absolute Error* (MAE) dan *Mean Square Error* (MSE). Salah satu kelemahan dalam penulisan artikel ini adalah dengan banyaknya istilah, penulis artikel cenderung sering menyingkat istilah-istilah yang digunakan. Bahkan, istilah *Random Forest Regression* langsung disingkat menjadi RF tanpa penjelasan sebelumnya. Kenyamanan pembaca menjadi terganggu karena harus menghafal semua istilah-istilah yang digunakan para artikel ini.

Penarikan Kesimpulan dan Pengumpulan data

Walaupun tahapan-tahapan pengumpulan dan pengolahan data yang digunakan pada artikel ini sudah terperinci, artikel ini kurang menjelaskan bagaimana hasil yang didapatkan berhubungan satu sama lain. Sebagai contoh, pengumpulan data dimulai dari mencari distribusi kerusakan terhadap bangunan, kemudian algoritma *Random Forest Regression* digunakan untuk menganalisis tren terjadinya gempa. Artikel ini tidak memberikan penjelasan yang jelas tentang bagaimana data dikumpulkan, diproses, dan dianalisis. Selain itu, langkah-langkah dan prosedur khusus yang diikuti dalam pengembangan model prediksi gempa bumi berbasis kecerdasan buatan menggunakan algoritma regresi hutan acak tidak cukup dijelaskan. Kurangnya detail metodologis ini membuat pembaca sulit untuk menilai validitas dan reproduktibilitas dari penelitian ini.

Interpretasi Data

Meskipun dalam artikel inidisebutkan bahwa model yang dikembangkan mencapai akurasi sebesar 98,8% dan tingkat kesalahan yang rendah, namun kurangnya evaluasi komprehensif dan perbandingan dengan metode yang sudah ada. Artikel ini tidak memberikan informasi tentang bagaimana kinerja model dievaluasi, seperti melalui validasi silang atau pengujian pada dataset independen. Tanpa evaluasi tersebut, menjadi sulit untuk menentukan kekokohan dan kemampuan umum dari model yang diajukan. Selain itu, Meskipun artikel ini menyebutkan tujuan untuk memberikan masukan bagi strategi mitigasi gempa bumi di Indonesia, diskusi mengenai implementasi praktis dari model yang dikembangkan terbatas, seperti pertimbangan seperti integrasi model ke dalam sistem pemantauan gempa bumi yang sudah ada.

Daftar Pustaka

Tidak terdapat masalah yang signifikan dalam daftar referensi artikel ini, artikel ini telah menyantumkan referensi dan sudah jelas dimana referensi tersebut digunakan.

Referensi

U. Wijaya, Kusrini and A. H. Muhammad, "Indonesian Seismic Mitigation using Earthquake Predicted Artificial Intelligence Model," 2022 5th International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, Indonesia, 2022, pp. 349-354, doi: 10.1109/ICOIACT55506.2022.9972091.

Indonesian Seismic Mitigation using Earthquake Predicted Artificial Intelligence Model

Usman Wijaya
Information Technology Postgraduate Program
Universitas Amikom Yogyakarta
Yogyakarta, Indonesia
usman.wijaya@ukrida.ac.id

Kusrini*
Information Technology Postgraduate Program
Universitas Amikom Yogyakarta
Yogyakarta, Indonesia
kusrini@amikom.ac.id

Alva Hendi Muhammad
Information Technology Postgraduate Program
Universitas Amikom Yogyakarta
Yogyakarta, Indonesia
alva@amikom.ac.id

2022 5th International Conference on Information and Communications Technology (ICOIACT) | 978-1-6654-5140-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICOIACT55506.2022.9972091

***Abstract*—One way to reduce injuries during an earthquake is to determine the probability of an earthquake as early as possible so that rescue mitigation may be carried out. An artificial intelligence-based prediction model using a random forest regression algorithm was developed using a dataset of earthquake records in Indonesia from January 1900 to January 2022. The attributes used were date, location, latitude, longitude, hypocenter, magnitude, depth of the earthquake, maximum peak ground acceleration, modified Mercalli intensity, and site class. The label used to determine earthquake predictions was modified Mercalli intensity. The model was split into 80% training data and 20% testing data. The accuracy of the random forest regression model was 98.8%, and the error rate performance metrics mean absolute error and mean square error produce error values of 5.4% and 4.6%, respectively. This Artificial Intelligence-based earthquake prediction model is expected to provide input on the development of earthquake mitigation strategies in Indonesia. It is expected that there will be no more earthquake fatalities in the future.**

***Keywords*—mitigation, artificial intelligence, earthquake, prediction, accuracy**

I. Introduction

Earthquakes are natural disasters that often occur in Indonesia. The earthquake that occurred in Indonesia was categorized as a tectonic earthquake. Tectonic earthquakes can occur due to the movement of plates in the earth's bowels. Plates move convergent and divergent. If the plates move convergent, there will be subduction. Meanwhile, if it moves divergent, it will result in faults [1]. Geographically, Indonesia was flanked by the Eurasian Plate and the Indo-Australian plate, so almost all parts of Indonesia are included in the high-risk category, prone to earthquakes, and earthquakes can come anytime and anywhere.

From the historical records in the last 20 years, there have been powerful earthquakes in Indonesia, including the 2004 Aceh Earthquake, 2005 Nias Earthquake, 2006 Jogja Earthquake, 2006 Padang Earthquake 2009, Palu Earthquake 2018, Lombok Earthquake 2018 [2]. In addition, in mid-January 2022, an earthquake occurred near the Sunda Megathrust in Banten.

Earthquakes are a frightening specter for the Indonesian state, where Indonesia is in the ring of fire as a country with high earthquake risk. Various efforts have been made to reduce victims due to earthquakes, including regularly updating the earthquake map of Indonesia, where the earthquake map is vital to provide information to structural engineers in determining the design parameters of the earthquake resistant building.

Earthquake prediction can be performed in various ways, including statistical data from ground motion records [3]. The Lembang Fault and the Sunda Megathrust, both located in the Sumatra Strait, are studied by seismologists as potential and probable events of faults and megathrusts that might trigger earthquakes. Monitoring of the Lembang Fault and the Sunda Megathrust also help experts in predicting future earthquakes [4]. In addition to using ground motion data, earthquake predictions can also use meteorological and geophysical data based on wind speed and air temperature data as parameters for signs of an upcoming earthquake. The data was evaluated using artificial intelligence (AI) techniques using machine learning (ML) and deep learning (DL) algorithms such as super vector regression (SVR), multi-layer perceptron (MLP) neural networks, and artificial neural networks (ANN) [5].

Buildings are generally designed for a 50-year earthquake return period with a probability of exceeding the design earthquake magnitude for a 50-year-old building is 2%, so it is necessary to know the intensity and level of damage to structures due to earthquakes within the 50 years [6]. The level of damage to buildings was measured by the Modified Mercalli Intensity (MMI) scale. The MMI scale has a strong relationship with earthquake events. A technical team from the Indonesian Agency for Meteorological, Climatology, and Geophysics (BMKG) will perform a survey and report damage to structures that occur after an earthquake. BMKG has carried out this activity for a long time, so this MMI data can be used as the main parameter in predicting earthquake events in the future. In determining the level of damage on the MMI scale, ten categories of damage levels are used, starting from a scale of 1 to 10 [7].

In this study, earthquake prediction using the ML algorithm uses an MMI scale building damage dataset from the recording of earthquake events in the last 100 years, earthquake magnitude, peak ground acceleration (PGA), earthquake hypocenter, coordinates of the location of the earthquake, and the area affected by the earthquake. The dataset was taken from earthquake recording data belonging to the BMKG. With this research, it is expected that seismic events will be predicted earlier in the future, allowing mitigation and evacuation to take place immediately, ensuring that no more people are killed or injured as a result of the earthquake.

II. Literature Study

Research on earthquake prediction has been carried out since the 1970s, but this research is still developing and being carried out until now. Earthquakes are difficult to predict precisely when and where, but earthquakes will occur. Scientists continuously develop these parameters of time and place of occurrence to predict an earthquake's event accurately.

Jain et al., in 2021, predict earthquake events based on the parameters of the distance to the earthquake center, the date and place of the earthquake, and the time of the quake with several machine learning algorithms, namely RF, MLP, and SVR. Of the several algorithms used, MLP Regressor produces the best predictions compared to other algorithms because the error value was the smallest and the accuracy was highest up to above 90 percent. The Root Mean Square Error (RMSE) method checks the performance and error values of each prediction result of each algorithm [8].

DLEP, a deep learning model with Convolutional Neural Networks (CNN) feature extraction, was developed by combining explicit and implicit features to produce very high accuracy earthquake prediction models. The DLEP predicted model was based on earthquake recording datasets in various big cities and countries, including Sichuan, China, Xinjiang, Tibet, Japan, the Philippines, Chicago, and Los Angeles, USA [9].

Shodiq et al., predicted aftershocks five days after the first earthquake based on earthquake events in Indonesia from 1917 to 2017 obtained from the BMKG catalog data and the United States Geological Survey (USGS). The attributes used were date, location, latitude, longitude, magnitude, and depth of seismic. Because the data used consists of 82580 datasets, to improve the accuracy of the data, they were grouped based on the magnitude data (high magnitude and moderate magnitude). The data were grouped by region, namely region 0 to region 5, east, southeast, west, southwest, north, and south regions. The ANN algorithm was used to predict aftershocks. The result was that the prediction accuracy is only high in high magnitude data ($M > 6$), which was 72 percent. Meanwhile, when the magnitude was moderate to low, the accuracy drops to 56 percent [10].

According to Debnath et al., to find the most suitable algorithm out of seven algorithms (Bayes net, RF, simple logistics, random tree, LMT, ZeroR, and logistic regression) to predict earthquakes in India. The logistic tree model algorithm and simple logistic classifier produce India's most accurate earthquake prediction. None of the seven algorithms used can make predictions with a similar level of accuracy for each class and region [11].

Yousefzadeh et al., predicted earthquakes with the highest accuracy using a feed-forward deep neural network (DNN) architecture, compared to a decision tree (DT) and a super vector machine (SVM). The accuracy of DNN and SVM was high for moderate magnitude values, while DT can predict more accurately for large and low magnitude values. Many other researchers have made earthquake predictions based on magnitude parameters themselves, while predictions based on fault points have not been studied [12].

Calderon and Silva predict earthquake risk based on social conditions of the community, significantly Costa Rica (in the form of population growth and population), construction area, PGA, and building height in Costa Rica. The increase in population, population density within the city, building height, construction implementation instructions affect earthquake mitigation in Costa Rica. These attributes were searched for relationships and correlations to produce earthquake risk predictions until 2030. The prediction results are expected to reduce losses and damage to buildings. There were no research attributes derived from historical earthquake events in magnitude and hypocenter data in previous years. Earthquake predictions were studied only based on the community's social conditions, resulting in an increase in earthquake risk in 2030, which will increase to 18 percent and even 30 percent in 2055 [13].

From the study literature presented above, several significant types of research can be used as a reference for conducting research in earthquake prediction and forecasting, including research on earthquake prediction that can be carried out using other algorithms. Most of the researchers in the literature study further used magnitude data as labels. Other attributes such as hypocenter, PGA, and MMI can be used as labels in predicting earthquake events. Damage to buildings based on the MMI level can be categorized into ten groups, from level 1-10 respectively, namely no damage to buildings for grades 1-3, light architectural damage for grades 4-5, moderate architectural damage for grades 6, heavy architectural damage for grades 7, and heavy damage to the collapse of building structural elements for grades 8-10 [14]. In this study, MMI will be used as labels to predict earthquake events with datasets from BMKG in the last 100 years, i.e., datasets from 1900 to 2022. The algorithm used was RF regression.

III. METHODOLOGY

Prediction of earthquake events was carried out by collecting datasets of earthquake events, earthquake magnitude, earthquake location, earthquake hypocenter, earthquake depth, Peak Ground Acceleration Maximum (PGA_{Max}) data, site class, and level of damage to buildings over more than 100 years. The dataset was taken from BMKG data from January 1900 to January 2022. PGA_{Max} is the maximum ground acceleration data when an earthquake occurs for measuring the intensity of ground motion when an earthquake occurs [15]. Site class was the type and level of soil hardness where an earthquake occurs, usually denoted as A, B, C, D, and E for bedrock, rock, hard layer of soil, the medium layer of soil, and soft soil, respectively [16]. The level of damage to buildings was measured on an MMI scale of 1-10 based on earthquake intensity data [17]. The data obtained visualized the relationship between each attribute to find the strongest data correlation. The information was then built into a predictive model of earthquake events based on the level of building damage caused by earthquakes in events for more than 100 years using a random forest regression algorithm. Prediction value was checked for metrics performance with Mean Absolute Error (MAE) and Mean Square Error (MSE) error values.

The data obtained from BMKG was unclean data. The model prediction was acceptable if it cleans well to determine a prediction model. A data cleansing technique was required for this. Exploratory data analysis in managing missing values and data visualization was crucial in data cleansing. The Equation (1) was used to fill in the missing value of one of the variables, PGA_{Max} [18].

$$MMI = 3.333x \log .PGA \quad (1)$$

After the cleaning process was completed, the next step was the data transformation divided into two stages, namely the conversion of categorical data and numerical data. Modifying categorical data may generally use a label encoder or one hot encoder. The label encoder was carried out in this study, namely changing the location and site class attribute data. Previously, both data were of string data type or object was altered to integer data type. Then the numerical data transformation was in the shape of a standard scaler. The final stage before building the model was splitting the data.

Training data and test data were divided into 80 percent for training data and 20 percent for test data. In this study, the model was built using the RF regression algorithm. The flow chart in Figure 1 describes the research stages.

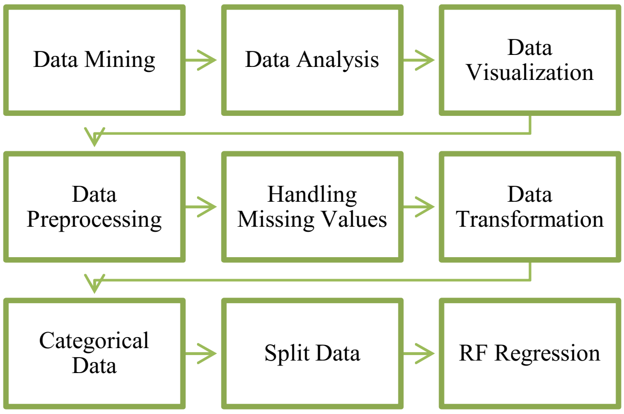

Fig. 1. Flowchart research methodology

After the RF regression prediction model was trained, the accuracy and performance of the test model were tested and validated with the MAE and MSE error values using Equation (1) and (2).

A. MAE

MAE is a method to measure the error rate of a prediction model with regression. The MAE value shows the error rate and absolute error of a prediction model. The MAE error metrics value can be solved by Equation (2) [19].

$$MAE=\frac{\sum_{i=1}^{n}|Y_i-\hat{Y}_i|}{n} \quad (2)$$

B. MSE

MSE results from the square of a metrics measurement to assess the error rate against the actual value. MSE value can be obtained by Equation (3) [20].

$$MSE=\frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{n} \quad (3)$$

Where the value of Y is the ground truth label and $\hat{Y}$ is the predicted value while n is the amount of data.

IV. Results and Discussion

Prediction of earthquake events is urgently needed at this time. AI technology is constantly evolving. This research was one of the AI technology developments in predicting earthquake events to determine earthquake mitigation plans in Indonesia in the future. Predictions were made using the RF Regression algorithm using a dataset from BMKG in the formation of 2149 datasets of records of past earthquakes from January 1900 to January 2022.

A. Data Interpretation

The relationship between variables can be studied by visualizing each variable. This study predicts earthquake events based on the level of building damage. In earthquake engineering, the level of damage to buildings was measured in MMI units on a scale of 1-10. Based on BMKG records for the past 100 years, there have been earthquakes, with the most significant scale of damage being the MMI level 10 scale of 0.2 percent to the MMI level 1 damage scale of 24.9 percent. Consecutively the classification of damage levels from level 1 to 10 can be seen in Figure 2.

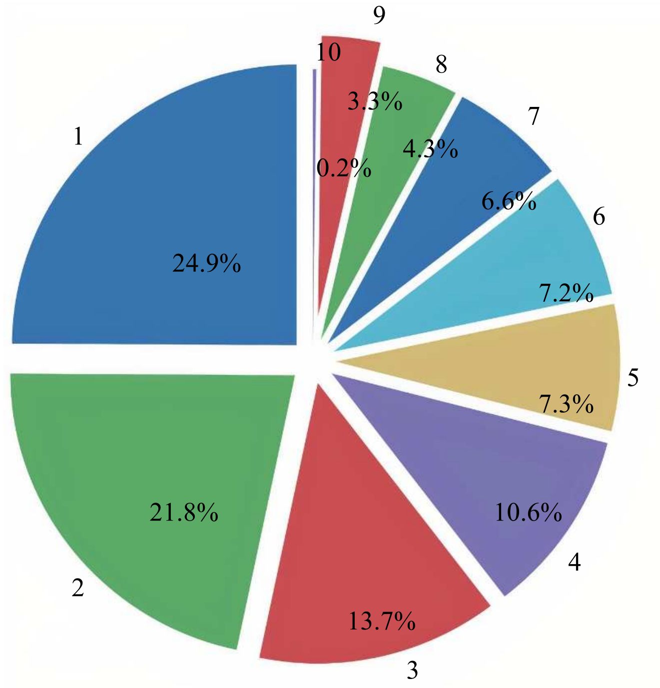

Fig. 2. Class of building damage level

The category of building damage index MMI level 1 showed light damage with the highest number of occurrences and the most frequent events. Index one sometimes cannot be felt by humans; seismographs can only detect index one. While the most extensive MMI index, level 10, occurred once with a sporadic intensity of 0.2 percent in 120 years, which occurred in 1994, while another devastating earthquake with an index level 9 in 120 years had occurred 40 times where 2 out of 40 This time it happened in 2004, i.e., the earthquake accompanied by the Aceh tsunami and in 2006 the Yogyakarta earthquake.

The maximum peak ground acceleration of PGA_{Max} indicates the propagation of earthquake waves from the bottom of the earth, which was called the primary wave. Earthquake waves reach the ground soil surface, called secondary waves, so that PGA can detect it in Gal and seismic magnitude units. Theoretically, the strain energy released after the earthquake will break the rocks in the earth's crust; the rock breaking causes friction, vibration, vertical and horizontal pull in the earth's bowels. The energy will be released to the ground surface as a secondary wave. The greater the energy felt at the ground surface, the greater the PGA value would increase and was directly proportional to the increase in the seismic magnitude value, causing the level of building damage to improve can be seen in Figure 3.

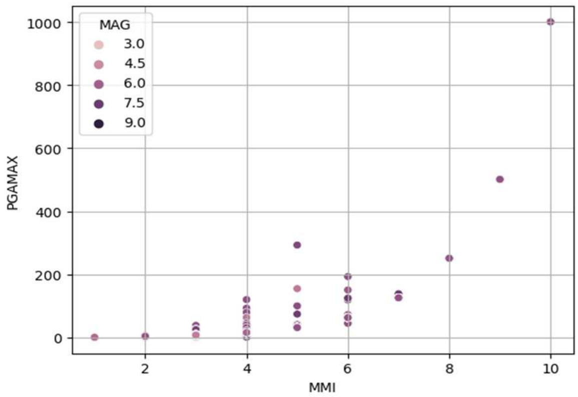

Fig. 3. Relation of PGA_{Max} to building damage level

The response of the soil can reach inelastic in high PGA, so a high level of damping was required. Otherwise, it will cause severe damage to the building. In this case, the impact often felt was on low-rise buildings, where low-level structures will produce a short period for the earthquake vibration, which was very damaging to buildings. In addition, low-level buildings in residential houses ignore the calculation of earthquake-resistant structures, mainly public housing with Economic value considerations often not engineered. Vice versa, a small PGA showed the elastic response of the soil, and the ground acceleration was slight. The PGA that occurs produces random and high irregularities at high MMI values . Nevertheless, the frequency of infrequent occurrences was inversely proportional to low and medium MMI values, i.e., the MMI range of 3 to 6. Based on the statistical data in Figure 3 PGA, earthquake magnitude and MMI have a linear relationship, so the mechanism of earthquake activity in the future has great potential to be predicted, mainly by utilizing AI technology and machine learning algorithms.

The relationship between earthquake magnitude and MMI was very close in an earthquake event. This relationship can be seen in Figure 4.

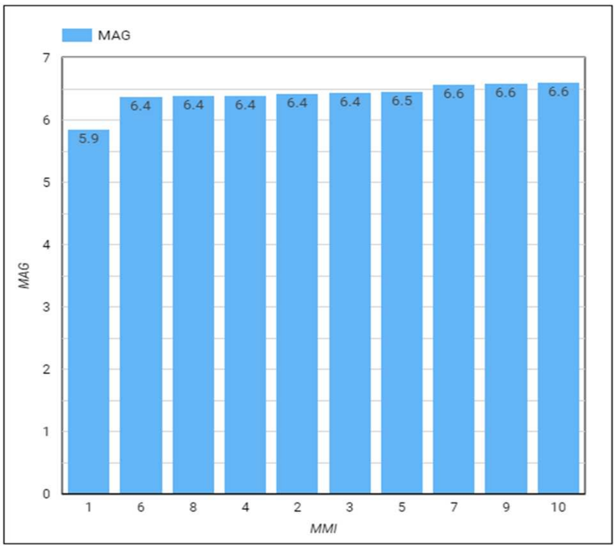

Fig. 4. Relation of magnitude to building damage level

Generally, the level of damage to the MMI building index will increase when the magnitude increases. Figure 4 shows the relationship between MMI and earthquake magnitude increases with the increasing MMI level.

The distance of the earthquake hypocenter affects the earthquake record. The rock medium absorbs the space, the less earthquake energy to the ground surface. Figure 5 shows a study of building damage at the hypocenter distance due to exceeding the degree of damage observed in historical earthquake records in the past.

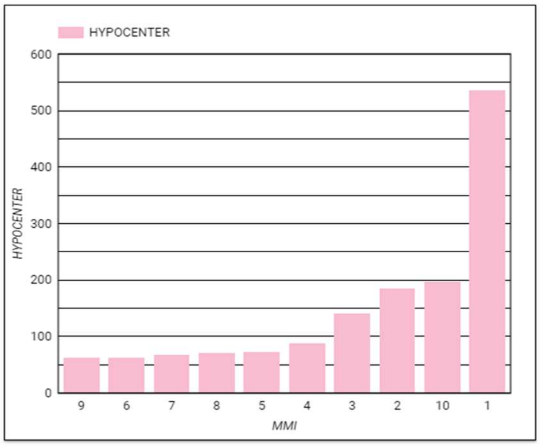

Fig. 5. Relation of hypocenter to building damage level

Figure 5 demonstrates that the shallower earthquake data (hypocenter below 100 km) resulted in extreme damage with MMI levels 7-9. However, there was an anomaly at MMI level 10 where severe building damage occurs in earthquakes at depths above 100 km. This anomaly could happen because the level of damage to the MMI scale building was influenced by the hypocenter and other factors, such as the magnitude and PGA factors, as seen in Figure 3.

B. Data Correlation

Correlation between variables using the heatmap function from a seaborn visualization library displays the correlation dashboard for each variable in the color scheme from the correlation coefficient data -1 to 1, as shown in Figure 6.

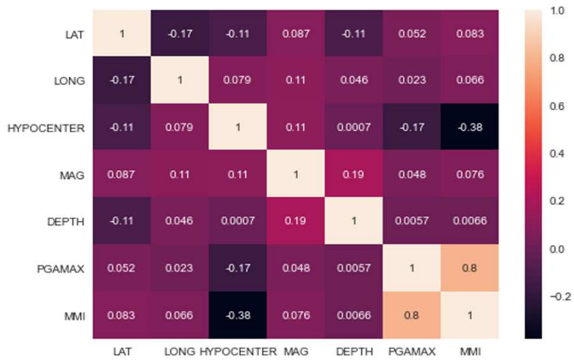

Fig. 6. Independent and dependent variable data correlation

Figure 6 indicates a strong relationship between variables, and there was strong multicollinearity on the independent variables MMI and PGA_{Max}, which was above 0.7 [21]. The Variance Inflation Factor (VIF) value above 10 shows multicollinearity potential [22]. VIF values of 11.9 and 15.2 for PGA_{Max} and MMI, respectively, indicate the possibility for multicollinearity in this study. The standard deviation values of the two variables, 1.13 for PGA_{Max} and 2.03 for MMI, confirm the validity of multicollinearity in both variables. Both values were more than 1, indicating that multicollinearity exists in the two independent variables [23]. Multicollinearity can occur when one independent variable's computation has relied on the values of other independent variables in the model, such as when filling in missing values on the PGA_{Max} variable using Equation (1).

The MMI variable and the hypocenter have a negative association, which indicates that when the MMI variable grows, so does the hypocenter value. The hypocenter, on the other hand, will drop. The MMI variable has a positive correlation with the magnitude, which implies that if the MMI variable grows, the magnitude will increase.

C. *Indonesian Seismic Mapping for Future Mitigation*

The distribution of earthquake events was based on the earthquake magnitude values recorded from January 1900 to January 2022. Figure 7 illustrates the degree of building damage based on the MMI level in Indonesia from January 1900 to January 2022.

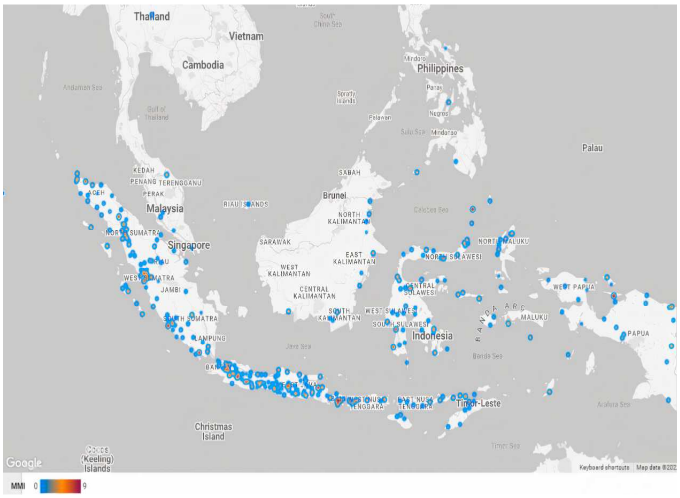

Fig. 7. MMI map based on Indonesian seismic historical data (year of 1900-2022)

The high MMI level in Figure 7 is as outlined in the large magnitude values M>6. There were in the Indian-Australian plate area under the island of Sumatra, i.e., the Mentawai-Pagai Megathrust, Megathrust Enggano, below the Java Island across the Sunda Megathrust, West Central Java Megathrust, under the island of Bali, i.e., the Sumba Megathrust and between the islands of Java and the island of Sulawesi, i.e., the Sulawesi Megathrust, while around the island of Papua. In the design code for earthquake-resistant building structures regulated in SNI 1726: 2019, it has been established that a map of high-risk earthquakes in the megathrust path is designated as Seismic Design Categories D, E, and F. The D-F category is a high-risk earthquake area category [24], earthquake mitigation efforts can be enhanced by earthquake-resistant building design guidelines, including prediction of earthquake events with machine learning algorithms using the statistics of earthquake records mapped in Figures 7.

D. *Earthquake Prediction Using RF Regression*

The dataset consisted of 2149 historical earthquake records from January 1900 to January 2022 with the attributes of date, location, station, latitude, longitude, hypocenter, magnitude, depth, PGA_{Max}, MMI, and SC. Removing the station variable had no impact on the analysis outcomes. However, the dataset of the top five data can be seen in Table I.

TABLE I. EARTHQUAKE DATASET

Date	*Loc*	*Lat*	*Lo ng*	*Hypo*	*Ma g*	*Depth*	*PG A*	*M MI*	*SC*
2018-09-28	Tnt	0.3 3	119 .89	43.8	7.4	11.0	138 .8	7.0	C
2018-09-28	Pos	- 1.4 2	120 .65	138.3	7.4	11.0	124 .9	6.0	D
2018-09-28	Pwt	0.4 7	121 .94	218.2	7.4	11.0	4.9 6	3.0	Na N
2018-09-28	Lu w	- 2.5 5	120 .32	244.2	7.4	11.0	25. 79	4.0	D
2018-09-28	Grt	0.9 8	122 .36	283.3	7.4	11.0	2.5 0	3.0	Na N

From the dataset in Table I, some data were empty and needed to handle missing values. Variables such as hypocenter, magnitude, depth, PGA, MMI, and SC need serious handling of the empty data. Some empty data are handled with the statistical mean, modes, drop row-column, and interpolation functions. The empty data in the PGA was dealt with using the procedure in Equation (1).

There are string or object data types in the Date, Location, and SC variables of the ten variables used for modeling. These three variables need to be preprocessed with data transformation using the label encoder function in the Scikit Learn library. The results of data transformation and handling of missing values can be seen in Figure 8.

	DATE	LOC	LAT	LONG	HYPOCENTER	MAG	DEPTH	PGAMAX	MMI	SC
0	6	576	-6.866	107.162	232.42	6.6	40.0	4.169900	2	3
1	6	577	-6.034	105.942	161.28	6.6	40.0	9.955800	3	3
2	6	567	-6.566	106.404	163.28	6.6	40.0	17.588100	3	3
3	6	812	6.831	105.891	102.24	6.6	40.0	93.107800	4	3
4	6	575	6.729	105.872	105.42	6.6	40.0	193.392200	6	3
...	...	...	...	...	...	...	...	...	...	...
1954	293	323	-3.511	102.056	45.33	6.5	52.0	63.121893	6	3
1955	5	96	-6.500	107.100	53.33	6.5	52.0	125.953436	7	3
1956	5	689	-7.000	106.900	49.33	6.5	52.0	125.953436	7	3
1957	5	542	-7.264	107.480	51.33	6.5	52.0	125.953436	7	3
1958	5	63	-6.399	106.069	55.33	6.5	52.0	125.953436	7	3

Fig. 8. Results of handling missing values and label encoder

After the data transformation, the next step was the data splitting process. The dataset was divided into 80 percent training data and 20 percent testing data. The date, location, latitude, longitude, hypocenter, magnitude, depth, PGA, and SC variables were set as features, and the MMI variables were labeled. The prediction model was built using the RF regression algorithm. From the RF regression model, an accuracy of 98.8 percent has been obtained. In comparison, the metrics performance of the MAE and MSE error values obtained from Equation (2) and (3) were 5.4 percent and 4.6 percent, respectively. The comparison value of the actual and predicted values of the RF regression model can be seen in Figure 9.

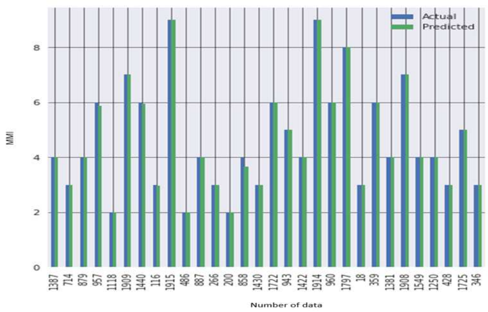

Fig. 9. Actual and prediction RF regression model

Prediction of earthquake events will determine the probabilistic value of earthquakes in the future. This research is expected to add a reference to update the probabilistic seismic coefficients in the provisions for designing earthquake-resistant building structures in the future.

V. CONCLUSIONS

Prediction of earthquake events with the RF regression model using a dataset recording earthquake events in Indonesia for more than 100 years, starting with data interpretation by finding the distribution of damage to buildings with the MMI index value. The damage distribution to high-level structures is scarce, below 10 percent. The distribution most often occurs at low MMI index values. Likewise, infrequent occurrences were also recorded for high magnitude values for high MMI indexes.

There is an indication that the trend of earthquake events from the relationship between the frequency of occurrence and the earthquake's magnitude has an inverse relationship. Earthquakes that have a large size will have a minor frequency of occurrence and rarely occur and or vice versa. Each region has a different level of activity and possible earthquake magnitude. The frequency and occurrence of earthquakes are very uncertain; therefore, earthquake predictions can use statistics to form probabilistic earthquake parameters in the maximum ground acceleration in a certain period, for example, 50 years, 100 years, and so on. The earthquake map and SNI 1726: 2019 refer to the recorded earthquake events in the last 100 years. Thus, earthquake mitigation in designing earthquake-resistant building structures is right on target.

The RF regression algorithm can make predictions accurately and has excellent performance metrics. The developed model can be used and developed into an earthquake event prediction application based on the value of the level of damage to the MMI index.

This research still has many shortcomings and needs improvement in terms of datasets. The variables used as dependent variables or labels can use other variables such as PGA or SC values. In further research, it is recommended to pay attention to the quality of the building in determining the MMI level or index because the earthquake mitigation strategy is not only influenced by the parameters of magnitude, location, hypocenter, and site class. In terms of the algorithms used, many can still be developed for similar research, for example, using other machine learning algorithms or even using other modeling techniques such as deep learning.

ACKNOWLEDGMENT

The authors thank all parties that provide support for this research study. The Ministry of Education, Culture, Research, and Technology in National Competitive Research for Master Thesis Research Scheme supports funding for this research, Master of Information Technology Universitas Amikom, Yogyakarta, the authors' institution, and BMKG who supplied the dataset to the authors.

REFERENCES

[1] T. Tavio and U. Wijaya, *Desain Rekayasa Gempa Berbasis Kinerja*. Yogyakarta: Andi Offset, 2018.

[2] U. Wijaya, R. Soegiarso, Tavio, and A. Wijaya, “Numerical Study of Potential Indonesian Rubber for Elastomeric Base Isolators in Highly-Seismic Zones,” *J. Phys. Conf. Ser.*, vol. 1477, no. 5, 2020.

[3] M. H. Al Banna *et al.*, “Application of Artificial Intelligence in Predicting Earthquakes: State-of-the-Art and Future Challenges,” *IEEE Access*, vol. 8, pp. 192880–192923, 2020.

[4] M. R. Daryono, D. H. Natawidjaja, B. Sapiie, and P. Cummins, “Earthquake Geology of the Lembang Fault, West Java, Indonesia,” *Tectonophysics*, vol. 751, no. July 2018, pp. 180–191, 2019.

[5] P. Hajikhodaverdikhan, M. Nazari, M. Mohsenizadeh, S. Shamshirband, and K. W. Chau, “Earthquake prediction with meteorological data by particle filter-based support vector regression,” *Eng. Appl. Comput. Fluid Mech.*, vol. 12, no. 1, pp. 679–688, 2018.

[6] ASCE/SEI 41-17, “Seismic evaluation and retrofit of existing buildings,” *Am. Soc. Civ. Eng.*, 2017.

[7] D. Contreras, S. Wilkinson, Y. D. Aktas, L. Fallou, R. Bossu, and M. Landès, “Intensity-Based Sentiment and Topic Analysis. The Case of the 2020 Aegean Earthquake,” *Front. Built Environ.*, vol. 8, no. March, 2022.

[8] R. Jain, A. Nayyar, S. Arora, and A. Gupta, “A comprehensive analysis and prediction of earthquake magnitude based on position and depth parameters using machine and deep learning models,” *Multimed. Tools Appl.*, vol. 80, no. 18, pp. 28419–28438, 2021.

[9] R. Li, X. Lu, S. Li, H. Yang, J. Qiu, and L. Zhang, “DLEP: A Deep Learning Model for Earthquake Prediction,” *Proc. Int. Jt. Conf. Neural Networks*, 2020.

[10] M. N. Shodiq, D. H. Kusuma, M. G. Rifqi, A. R. Barakbah, and T. Harsono, “Neural network for earthquake prediction based on automatic clustering in indonesia,” *Int. J. Informatics Vis.*, vol. 2, no. 1, pp. 37–43, 2018.

[11] P. Debnath *et al.*, “Analysis of earthquake forecasting in India using supervised machine learning classifiers,” *Sustain.*, vol. 13, no. 2, pp. 1–13, 2021.

[12] M. Yousefzadeh, S. A. Hosseini, and M. Farnaghi, “Spatiotemporally explicit earthquake prediction using deep neural network,” *Soil Dyn. Earthq. Eng.*, vol. 144, no. August 2020, p. 106663, 2021.

[13] A. Calderón and V. Silva, “Exposure forecasting for seismic risk estimation: Application to Costa Rica,” *Earthq. Spectra*, vol. 37, no. 3, pp. 1806–1826, 2021.

[14] P. Sbarra, P. Tosi, V. De Rubeis, and D. Sorrentino, “Quantification of earthquake diagnostic effects to assess low macroseismic intensities,” *Nat. Hazards*, vol. 104, no. 3, pp. 1957–1973, 2020.

[15] ASCE/SEI 7-16, *Minimum design loads for buildings and other structures*, no. 7 98. ASCE, 2016.

[16] International Code Council, *International Building Code*. International Code Council, 2018.

[17] D. J. Dowrick, G. T. Hancox, N. D. Perrin, and G. D. Dellow, “The modified mercalli intensity scale - Revisions arising from New Zealand experience,” *Bull. New Zeal. Soc. Earthq. Eng.*, vol. 41, no. 3, pp. 193–201, 2008.

[18] X. Tian, Z. Wen, W. Zhang, and J. Yuan, “New Ground Motion to Intensity Conversion Equations for China,” *Shock Vib.*, vol. 2021, 2021.

[19] B. Santosa and A. Umam, *Data Mining and Big Data Analytics*, 2nd ed. Penebar Media Pustaka, 2018.

[20] H. Pham, “A new criterion for model selection,” *Mathematics*, vol. 7, no. 12, pp. 1–12, 2019.

[21] and M. H. R. Kristina P. Vatcheva, MinJae Lee, Joseph B. McCormick, “Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies,” *Physiol. Behav.*, vol. 176, no. 10, pp. 139–148, 2017.

[22] J. I. Daoud, “Multicollinearity and Regression Analysis,” *J. Phys. Conf. Ser.*, vol. 949, no. 1, 2018.

[23] N. Shrestha, “Detecting Multicollinearity in Regression Analysis,” *Am. J. Appl. Math. Stat.*, vol. 8, no. 2, pp. 39–42, 2020.

[24] Badan Standardisasi Nasional, “SNI 1726-2019 tentang Tata cara perencanaan ketahanan gempa untuk struktur bangunan gedung dan nongedung,” *Tata Cara Perenc. Ketahanan Gempa Untuk Strukt. Bangunan Gedung dan Non Gedung*, no. 8, p. 254, 2019.