Jurnal Informatika dan Rekayasa Perangkat Lunak ISSN 2656-2855
Vol. 5, No. 2, September 2023, Hal. 130-134 e-ISSN 2685-5518

Analysis of Classification Models for ICU Mortality Prediction using Random Forest and Neural Network

Lymin, Alvin, Bodhi Lhoardi, Darwis, Joseph Siahaan, Abdi Dharma*
Jurusan Informatika, Fakultas Ilmu Komputer,
Universitas Prima Indonesia, Indonesia
Email: abdidharma@unprimdn.ac.id

Abstract

Based on the results of previous studies, research on machine learning for predicting ICU patients is crucial as it can aid doctors in identifying high-risk individuals. A high accuracy in machine learning models is necessary for assisting doctors in making informed decisions. In this study, machine learning models were developed using two models, namely Random Forest and Artificial Neural Network (ANN), to predict patient mortality in the ICU. Patient data was obtained from The Global Open Source Severity of Illness Score (GOSSIS) and underwent preprocessing to address issues of missing values and imbalanced data. The data was then divided into training, validation, and testing sets for model training and evaluation. The results of the study indicate that the Random Forest model performs better with an accuracy of 93% on the testing data compared to the ANN which only achieved an accuracy of 86% on the testing data. Consequently, the Random Forest model can be utilized as a solution for predicting patient mortality in the ICU.

Keywords: *ICU Patient Data Analysis; GOSSIS (The Global Open Source Severity of Illness Score) dataset; ICU Patient Mortality Prediction; Random Forest; Neural Network*

INTRODUCTION

The Intensive Care Unit (ICU) is a hospital room reserved for patients with life-threatening conditions that can result in death, and it requires high-quality services to reduce the risk of mortality (Megawati, 2019). Analyzing patient data in the ICU can be one solution to improve understanding of the factors that contribute to the risk of death and to prepare medical staff in the ICU to provide better services and save patients' lives. Technology can play a role in predicting, identifying, and recognizing various patient conditions to enhance patient care (Wardani et al., 2022).

The MIMIC-III BIDMC dataset records patient care in the ICU, including patient death descriptions, totaling 53,423 data points. This dataset has been studied for predicting ICU mortality caused by various diseases (Xie et al., 2020), kidney failure (Lin et al., 2019), and heart disease (Jayasudha et al., 2021). In addition, research has been conducted to diagnose diseases such as Pulmonary (Mlodzinski et al., 2020) and Blood Infections (Roimi et al., 2020). Many Random Forest (Cohen et al., 2021) and Neural Network models (Zhu et al., 2021), (Yu et al., 2020), (Na Pattalung et al., 2021) have been used to study the dataset and have produced good performance.

On the other hand, similar previous research using the GOSSIS (The Global OpenSource Severity of Illness Score) includes 91,713 data points to study the prediction of Diabetes Mellitus, Heart Disease, and ICU mortality using Random Forest machine learning models. While previous research by (Gaffney, 2021; Li, 2022; Mundra et al., 2022) achieved good accuracy rates using this dataset, there have been only a few studies that directly compare the Random Forest and Artificial Neural Network (ANN) models for ICU mortality prediction. Moreover, previous research that used ANN and Random Forest with GOSSIS dataset were not observing both performances directly and only use ANN for hyper-tuning (Li, 2022).

Therefore, the novelty of this research in comparison with previous research is by aiming to observe the performance of Random Forest and ANN to predict ICU mortality using the GOSSIS dataset. The difference between this research and the previous research is that this research observes two types of models directly, which are a simple Machine Learning model i.e. Random Forest, and a deep learning model i.e. ANN.

METHOD

A. ICU Mortality Prediction

ICU mortality prediction aims to classify the level of mortality in the ICU based on patient data in the ICU. The benefit of this research is as statistics and supporting data for the Decision Support System (DSS) in ICU treatment.

B. Artificial Neural Network

Artificial Neural Network (ANN) is one of the Machine Learning algorithms that mimics human neurons in learning to perform classification or other problem-solving tasks.

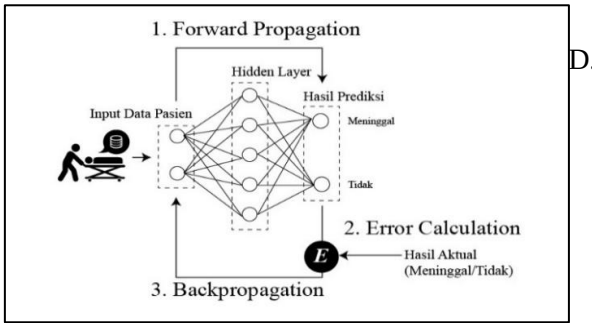

Figure 1. Steps of ANN

The three main steps of ANN as shown in the figure 1 are:

1. Forward Propagation: Starting from the input layer, data is propagated forward to the output layer. With z1(h) known as the vector containing the input's linear product with the weight matrix.
2. Error Calculation: Based on the output results, the error value can be calculated. With the number of errors obtained as the difference between the Target and the Output, the Total Error can be generated.

3. Backpropagation: This step looks for derivatives or gradients of every weight in the network. Using the chain rule commonly applied in calculus.

Finally, the Backpropagation process is continued by calculating the new weight values, namely w1, w2, w3, and w4.

C. *Random Forest*

Random Forest algorithm is a Machine Learning model that uses an ensemble technique and several random decision trees combined at the end of the process to perform classification as shown in the figure 2.

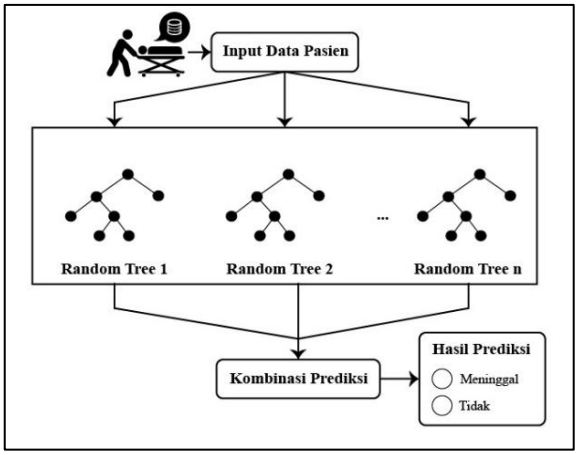

Figure 2. Steps Random Forest algorithm

D. Machine Learning Pipeline

The procedure in this research follows the Machine Learning Pipeline, which is divided into the dataset preparation process and Machine Learning modeling with Random Forest and ANN models, as shown in the figure 3.

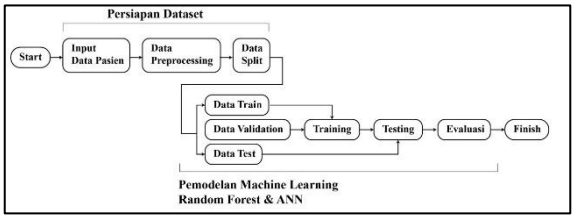

Figure 3. Procedure Machine Learning Pipeline

The steps shown in the figure 3 are:

1. Data Preprocessing: This process cleans the ICU patient dataset so that it can be processed by Machine Learning models. This process consists of the following stages:
2. Remove duplicates: Removing duplicate or identical data in the dataset.
3. Data cleaning: Data cleaning and processing to fix or remove invalid, duplicate, or unnecessary values in a dataset.
4. Handle outliers: Handling extreme or unusual values in the dataset that can affect data analysis.
5. Dealing with missing values: Handling missing values in the dataset, such as by filling in empty values or deleting missing data rows.
6. Data encoding: Converting categorical or text variables in the dataset into numerical

forms that can be processed by Machine Learning models, such as using methods like one-hot encoding or label encoding.
7. Data Train: This data is used to train the models in the Training process.
8. Data Validation: This data is used to validate the Training process.
9. Data Test: This data is used in the Testing process.
10. Training: This process trains the Random Forest and ANN models to predict patient mortality in the ICU.
11. Testing: This process tests the trained model to predict patient mortality in the ICU.

E. Dataset

The data used in this study is the GOSSIS (The Global Open-Source Severity of Illness Score) dataset with 91,713 ICU patient data. The format of the dataset is CSV (Comma Separated Values) as shown in figure 4.

	A	B	C	D	E	F
1	encounter_id	patient_id	hospital_id	hospital_death	age	bmi
2	66154	25312	118	0	68	22.73
3	114252	59342	81	0	77	27.42
4	119783	50777	118	0	25	31.95
5	79267	46918	118	0	81	22.64
6	92056	34377	33	0	19	NA
7	33181	74489	83	0	67	27.56

Figure 4. Dataset

RESULT

The process of creating ANN and Random Forest models with Jupyter Notebook can be done through several stages that will be explained in this chapter.

A. Preprocessing

This stage involves several processes as explained in Methodology, namely as shown in the table 1

Table 1. Preprocessing

Preprocessing Method	Explanation
Remove duplicates	The drop_duplicates function is used to remove duplicated data in the dataset
Data Cleaning	The Dropna function is used on data that can disrupt the data structure and do not provide useful information in the form of patient IDs and admission status.
Handle Outliers	By correcting values that are considered outliers or unreasonable in a data column, namely the pre_icu_los_days <= 0 column is set to 0.
Dealing with missing values	The imputation process is carried out using Iterative Imputer from Sci-Kit Learn and Linear Regression to estimate missing values during the imputation process.
Data encoding	The get_dummies function from the Pandas library is used to perform data encoding in this study.

After the data is processed and cleaned, the Synthetic Minority Oversampling Technique (SMOTE) process is used to overcome imbalanced dataset problems.

B. Data Split

This stage involves dividing the data into Training, Validation, and Testing data with a ratio of 80% for Training data, 10% for Validation data, and 10% for Testing data.

C. Training

This stage involves the process of training the model using the training data. The accuracy results of the training and validation of the ANN and Random Forest models as shown in table 2.

Table 2. Training Result

Model	Training Accuracy	Validation Accuracy
ANN	87,23%	86,37%
Random Forest	92%	91%

D. Testing

This stage involves the process of testing the model using test data. During this stage, the trained model is used to make predictions on the test data, and the results are compared with the actual values, as shown in table 3.

Table 3. Test Result

Model	Test Accuracy
ANN	86%
Random Forest	93%

DISCUSSION

This research utilized the Global Open Source Severity of Illness Score (GOSSIS) dataset to develop prediction models using Artificial Neural Network (ANN) and Random Forest models that can aid doctors in making clinical decisions about patient care in ICU. The preprocessing stage involved handling missing data, label encoding, and Synthetic Minority Oversampling Technique (SMOTE) to handle imbalanced datasets. The results show that the Random Forest model outperformed the ANN model by 7% in predicting ICU patient outcomes with reasonable accuracy. The developed prediction models in this research can improve patient care quality and reduce mortality rates. However, this research has several limitations, such as limited data and feature usage. Further research could develop more accurate and sophisticated prediction models by considering more features and larger datasets.

CONCLUSION

In summary, based on the research results that used the GOSSIS dataset, predictive models using ANN and Random Forest have been developed to assist medical professionals in making clinical decisions regarding patient care in the ICU. The GOSSIS dataset can be used to develop predictive models that can aid doctors in making clinical decisions about patient care. Regarding accuracy, the Random Forest model showed better performance compared to the ANN model, with an accuracy of 93% on testing data. This indicates that the Random Forest model is more effective and can provide more accurate prediction results.

Based on the results of this study, the future works are:

1. Conduct research using other ICU patient datasets, for example, larger and newer datasets, so that the results obtained are more representative and accurate.
2. Try using other algorithms such as Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) to compare performance with the ANN and Random Forest models.
3. Conduct research focused on feature engineering or hyper-tuning to enhance the model's performance in predicting the risk of patient death in the ICU.

REFERENCES

Cohen, S., Dagan, N., Cohen-Inger, N., Ofer, D., & Rokach, L. (2021). ICU Survival Prediction Incorporating Test-Time Augmentation to Improve the Accuracy of Ensemble-Based Models. *IEEE Access*, *9*, 91584–91592. https://doi.org/10.1109/ACCESS.2021.3091622

Gaffney, A. (2021). An Ensemble Learning Algorithm for ICU Patient Mortality Prediction. *Doctoral Dissertation, Dublin, National College of Ireland*.

Jayasudha, B. S. K., Sudha, P. N., Keshav, K., & Ramesh, N. (2021). Ensemble learning as a prerogative method of predicting mortality of patients with cardiovascular diseases. *Proceedings of CONECCT 2021: 7th IEEE International Conference on Electronics, Computing and Communication Technologies*, 1–5. https://doi.org/10.1109/CONECCT52877.2021.9622600

Li, Z. (2022). Study of ICU Mortality Prediction and Analysis based on Random Forest. *Proceedings of the 7th International Conference on Cyber Security and Information Engineering*, 691–695.

Lin, K., Hu, Y., & Kong, G. (2019). Predicting in-hospital mortality of patients with acute kidney injury in the ICU using random forest model. *International Journal of Medical Informatics*, *125*, 55–61. https://doi.org/10.1016/j.ijmedinf.2019.02.002

Megawati, S. W. (2019). Analisis Mortalitas Pasien di Ruang Intensive Care Unit (ICU). *Universitas Bhakti Kencana*, *7*(02), 127–135. https://doi.org/10.33482/medika.v7i02.151.

Mlodzinski, E., Stone, D. J., & Celi, L. A. (2020). Machine Learning for Pulmonary and Critical Care Medicine: A Narrative Review. *Pulmonary Therapy*, *6*(1), 67–77. https://doi.org/10.1007/s41030-020-00110-z

Mundra, S., Vijay, S., Mundra, A., Gupta, P., Goyal, M. K., Kaur, M., Khaitan, S., & Rajpoot, A. K. (2022). Classification of imbalanced medical data: An empirical study of machine learning approaches. *Journal of Intelligent & Fuzzy Systems*,

43(2), 1933–1946. https://doi.org/10.3233/JIFS-219294

Na Pattalung, T., Ingviya, T., & Chaichulee, S. (2021). Feature explanations in recurrent neural networks for predicting risk of mortality in intensive care patients. *Journal of Personalized Medicine*, *11*(9), 934. https://doi.org/10.3390/jpm11090934

Roimi, M., Neuberger, A., Shrot, A., Paul, M., Geffen, Y., & Bar-Lavie, Y. (2020). Early diagnosis of bloodstream infections in the intensive care unit using machine-learning algorithms. *Intensive Care Medicine*, *46*(3), 454–462. https://doi.org/10.1007/s00134-019-05876-8

Wardani, S., Akbar, M. U., Sitanggang, A. H. Y., Tupa, J. B., Pardede, J., & Dharma, A. (2022). Smart Prediction Model For Unplanned Icu Transfer Based On Deep Learning Optimization: An Article Review. *Jurnal Mantik*, *6*(2), 2659–2663.

Xie, F., Chakraborty, B., Hock Ong, M. E., Goldstein, B. A., & Liu, N. (2020). AutoScore: A machine learning–based automatic clinical score generator and its application to mortality prediction using electronic health records. *JMIR Medical Informatics*, *8*(10), 21798. https://doi.org/10.2196/21798

Yu, R., Zheng, Y., Zhang, R., Jiang, Y., & Poon, C. C. Y. (2020). Using a Multi-Task Recurrent Neural Network with Attention Mechanisms to Predict Hospital Mortality of Patients. *IEEE Journal of Biomedical and Health Informatics*, *24*(2), 486–492. https://doi.org/10.1109/JBHI.2019.2916667

Zhu, Y., Zhang, J., Wang, G., Yao, R., Ren, C., Chen, G., Jin, X., Guo, J., Liu, S., Zheng, H., Chen, Y., Guo, Q., Li, L., Du, B., Xi, X., Li, W., Huang, H., Li, Y., & Yu, Q. (2021). Machine Learning Prediction Models for Mechanically Ventilated Patients: Analyses of the MIMIC-III Database. *Frontiers in Medicine*, *8*. https://doi.org/10.3389/fmed.2021.662340