e-ISSN: 2963-2129, p-ISSN: 2962-0562
DOI: https://doi.org/10.38035/ijphs.v1i3
Received: 29 July 2023, Revised: 21 August 2023, Publish: 14 September 2023

International Journal of Psychology and Health Science (IJPHS)
GREENATION Publisher
+62 811 7404 455 http://greenpub.org editor@greenpub.org
Jl. Kapt. A. Hasan, Telanaipura Jambi - Indonesia. Unipark Condominium, Block D 3-9 Selangor, Malaysia. Singapore 655B Jurong West ST 61#12-542 Singapore 642655.

Hybrid Methods for Feature Selection Algorithms in the Field of Medical Records

Yuda Syahidin[1], Ade Irma Suryani[2]
[1] Manajemen Informasi Kesehatan, Piksi Ganesha Politechnic, Bandung, Indonesia.
Email:yudasy@gmail.com
[2] Rekam Medis dan Informasi Kesehatan Piksi Ganesha Politechnic, Bandung, Indonesia.
Email: adeirmasuryani20@gmail.com

Corresponding Author: yudasy@gmail.com [1]

Abstract: Big data growth in the healthcare community, accurate analysis of medical data supports early disease detection, patient care and community services. However, the accuracy of the analysis decreases when the quality of the medical data is incomplete. Feature selection is a process that selects a subset of features that are relevant for a predictive modeling problem. This method can identify and remove unnecessary, irrelevant, and redundant attributes from the dataset, which do not contribute to the accuracy of the model or reduce the accuracy of the model. The challenge in research in the field of medical records is in structured and unstructured data which results in a method being needed to assist the algorithm in selecting good features. This paper provides an overview of the proposed hybrid method that can be used for feature selection algorithms in the field of medical records.

Keywords: Predictive, Selection Feature, Hybrid Method

INTRODUCTION

Patient health is a matter of priority to be prioritized and medical experts are constantly trying to apply new technologies and achieve significant results. Use of medical record data that allows for the analysis of large amounts of medical data that can be used in areas of research such as clinical decision support, information extraction, phenotyping, disease inference, and personalized healthcare. Predictive analytics is one of the important areas of clinical science to offer improved care to patients. In recent years, most of the methods used to evaluate Electronic Health Record data which contains a lot of information about patient health, to obtain information by leveraging predictive analytics can use machine learning techniques as well as statistical techniques such as clinical trial results. Predictive algorithm performance depends on data representation and feature selection. The challenge in the field of health data is detecting patterns that produce predictive models for clinical decision support and there is still a large portion of health data that is still not utilized. Scalability and efficiency in analyzing

medical record data is still lacking in generating Electronic Health Reports and there is still a lack of decision-making systems in health because there is still unstructured health data.

In addition, there is an approach in selecting features by means of filters and wrapper methods such as in carrying out search strategies, steps to determine feature quality, and evaluating features. All the features in the dataset play as important factors in determining the model hypothesis to make predictions. The filter method is a selection method that is independent of the machine learning method and is a selection method based on the relationship between variables and machine learning algorithms. Research on feature selection has been carried out using the wrapper technique to reduce the features selected by Yang, et a, wrapper technique in feature selection to predict diabetes, Le, et al, cluster ranking technique in feature selection by Anwar et al, and using genetic algorithms in feature selection to help predict mortality models by Ghorbani et al.

However, from these various feature selection approaches, there are still challenges in terms of high-dimensional data problems, irrelevant data reduction techniques, and thresholds in feature selection, a more in-depth study is needed, and continuous research is needed to combine features. Based on this, it is necessary to develop methods for feature selection algorithms to contribute to feature selection of medical record data that has structured and unstructured properties.

LITERATURE REVIEW

Electronic Health Records

Electronic Health Record (EHR) is a person's health information that is stored digitally and is instantly and securely available to authorized users. EHRs contain patient diagnoses, medications, vital signs, treatment plans, progress notes, radiological images, and test results. Classification of records exists as unstructured and structured EHR data. Unstructured EHR data is written based on the clinical context that describes the patient's condition and is most useful for clinical documentation.

EHR data consists of various types of data, from structured information such as prescription drug data consisting of dates and doses to unstructured data such as clinical narratives that describe the medical reasons behind medical record documents. The relationship between structured data and unstructured data can be seen in Figure 1 below.

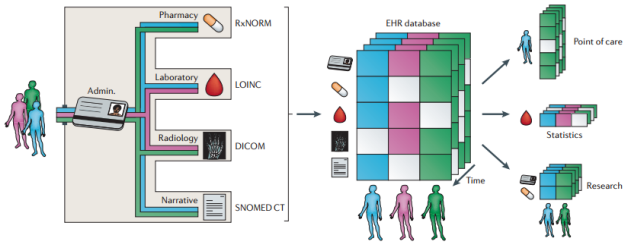

Figure 1. Electronic health record content

A patient's EHR can be viewed as a storehouse of information regarding his health status in a computer-readable form. When connected to the health care system it produces various types of patient-related data. Patient data is stored in a database and can be viewed in a format that suits the needs and authority of certain user groups. The term Electronic Health Record (EHR), or Electronic Health Record, refers to the collection of patient health information in a digital format. EHRs can be categorized in terms of functionality: (i) basic EHR without clinical records, (ii) basic EHR with clinical records and (iii) comprehensive

systems. EHRs, even in their simplest form, provide researchers with a rich collection of data. Data can be shared across networks and can include, as previously described, a wide variety of information. EHR is primarily designed for internal hospital administration tasks and many different schemes exist in different structures.

Feature Selection

Feature selection is a critical process in machine learning, designed to remove irrelevant, redundant, and noisy features and retain a small fraction of features from the main feature space. Thus, effective feature selection can help reduce computational complexity, improve model accuracy, and improve model interpretability. In machine learning and data science more generally, feature selection (also known as variable/feature selection, attribute selection or subset selection) is the process by which data is automatically or manually selected for a relevant subset of features for use in building a machine learning model. In fact, it is one of the core concepts in machine learning that has a huge impact on model performance and is key to creating reliable machine learning models.

Feature selection can be described in two steps as follows:

a) Combination of search techniques to generate new feature subsets.

b) Take measurements of features to evaluate or judge how well different subsets of features are.

Feature selection methods can be divided into three categories:

a) *Filter Methods*

Rely on the characteristics of the features without using any machine learning algorithms. The filter method selects features from the dataset independently for the machine learning algorithm. This method only depends on the characteristics of the variables, so that features are filtered from the data before learning begins.

b) *Wrapper Methods*

Based on consideration of selecting a feature set as a search problem, then using predictive machine learning algorithms to select the best feature subset. This method trains a new model on each feature subset with the aim of generating the best performing feature subset for a given machine learning algorithm.

c) *Embedded Methods*

The Embedded method considers the interaction of features and models. This method performs feature selection as part of the model construction process.

Research in Prediction Models in The Fields Of Medical Records

Basically, in the prediction model with a machine learning approach in data processing activities, feature engineering and feature selection and feature extraction techniques are carried out.

Research that utilizes feature selection by eliminating data dimensions is carried out in research for patient readmission to the hospital. Models that use Genetic Algorithm as a feature selection method and ensemble models based on a combination of Stacking and Boosting. This process selects an optimal subset of the relevant features for use in predictive model development.

Research that approaches the feature extraction technique for a predictive model of the risk of depression through national health data and normalization techniques for clinical time series data. Feature reduction technique approach for predictive models of cardiac risk Disease risk prediction method by classification and selection of features in handling imbalanced data (JICFS) is proposed to handle the problem of hospital readmission. Research using the Random Forest (RF) method with feature selection to develop a predictive model in the field of medical

records with clustered and longitudinal data and predictive models based on medical record data applying machine learning techniques with LSTM (Long Short Term Memory networks) algorithms to predict patient mortality. The Neural Network model is used to assist in making a prediction model for kidney disease.

Research for predictive models with various health datasets with outlier techniques to examine features of diabetes and hypertension as well as research that performs predictive models on structured and unstructured data on disease.

STATE-OF-THE-ART

Feature Selection – Genetic Algorithm

The hybrid model uses a Genetic Algorithm as a feature selection method and an ensemble model based on a combination of Stacking and Boosting. This process selects an optimal subset of the relevant features for use in predictive model development.

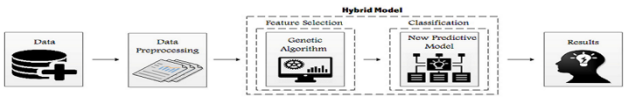

Figure 2. Model Framework *New Hybrid*

Feature Selection - Wrapper

This research uses the wrapper method to select features to remove features that have nothing to do with the diabetes dataset and this method helps optimize the number of attributes for the Multilayer Perceptron algorithm. Below is picture 3 of the Wrapper Method and picture 4 of the framework for early diabetes prediction.

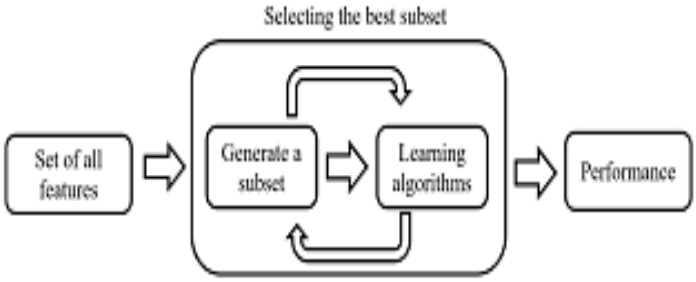

Figure 3. Metode Wrapper

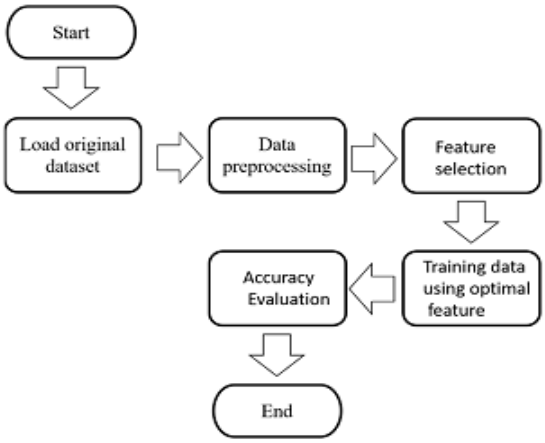

Figure 4. Framework untuk *early diabetes Prediction*.

Feature Selection – Imbalanced Data Classification

This study uses filter techniques in selecting features for unbalanced data. The results of this feature selection will be used for the prediction model for hospital readmission. Below is figure 5. Framework Imbalanced data classification

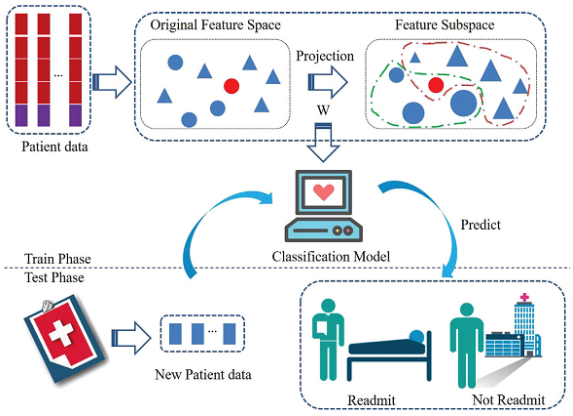

Figure 5. Framework *Imbalanced data classification*

Feature Selection - Classification models for heart disease prediction

This study uses the dimension reduction method and finds features that have a correlation with heart disease by applying feature selection techniques. Combining PCA (principal component analysis) techniques can help with data dimension problems. Below is a picture of 6 approaches to feature selection.

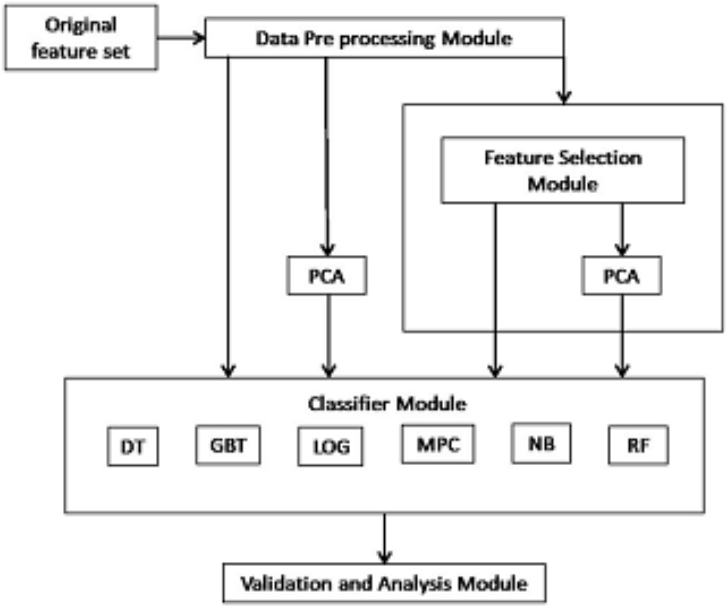

Figure 6. Approach to Feature Selection

METHODS

In selecting features for the dataset that will be modeled against machine learning algorithms based on the feature selection technique used, namely filter, wrapper and embedded techniques. The following is a picture of the approach method in feature selection.

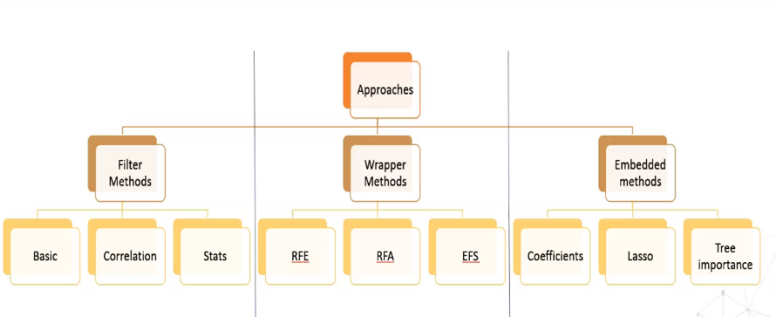

Figure 7. Feature Selection Approach Method.

The Hybrid method in feature selection depends on the combination of the selected approaches. The approach chosen in this research is wrapper and embedded. This hybrid method can use the Feature Schuffling, Recursive Feature elimination and Recursive Feature Addition algorithm approaches.

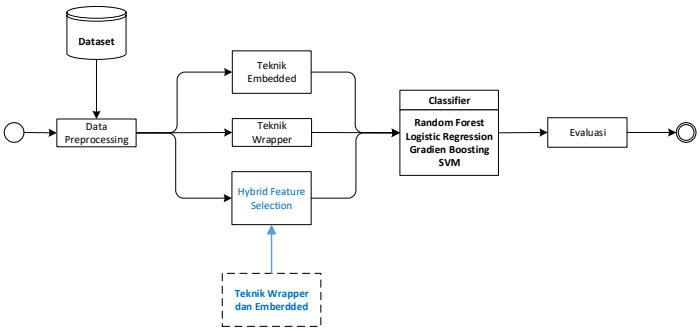

Figure 7. The proposed Hybrid Approach method.

RESULT AND DISCUSSION

As part of this study, initial trials were conducted on a collection of datasets to test the proposed Hybrid Feature Selection (HBFS) model. The required dataset is a dataset that fits the criteria in the Health sector. In this study, the Hybrid model uses cardiac datasets from Cleveland, Hungary, Statlog. The dataset is taken from the UCI Machine Learning Repository. Meanwhile, the dataset has 76 attributes/features (including predictive labels), but only 14 features are used in each experiment. The following is Table 1 Summary of Heart Disease Dataset.

Table 1. Summary Dataset UCI Repository

Dataset	Number of instances	Number of features	Number of classes
Cleveland Heart Disease	303	13	5
Hungarian Heart Disease	294	13	5
Statlog Heart Disease	270	13	2

Initial experiments weighted the features of each dataset. The features that have been given weight, the next stage these features are sorted based on the value of the largest to the smallest weight. After conducting the experiment, according to the stages of the research designed, the results are obtained in the form of a comparison without hybrid with hybrid. From the experimental results it can be seen that the comparison of the accuracy values of the datasets tested is based on the weight of the features that are relevant to the target (label). Following are the comparison tables of the embedded technique (LR Coefficient, NCAFW) and the Hybrid Feature Selection technique. The following is Table 2 Comparison of Election Results.

Table 2. Comparison of LR, NCAFW and Hybrid FS models against UCI ML Disease Datasets

Dataset	Algoritma Machine Learning	Performance Metrics		
		LR Coef	NCAFW	HBFS
Cleveland	RF	75	**80.35**	**91.66**
	LR	73.21	**82.5**	**88.09**
	SVM	67.85	**85.71**	**88.09**
	GBT	69.64	**83.92**	**89.28**
Hungarian	RF	61.81	**83.63**	**94.33**
	LR	72.72	**81.81**	**92.45**
	SVM	69.09	**80.09**	**89.18**
	GBT	67.27	**83.63**	**86.48**
Statlog	RF	85.71	**90.47**	**95.23**
	LR	71.42	**85.51**	**88.09**
	SVM	76.19	**89.41**	**92.85**
	GBT	80.95	**85.71**	**90.47**

CONCLUSION

Research in determining the attributes in the dataset with the feature selection method at this time has already been done by making a very good contribution and from the results of the research there are still opportunities for further research. The approach through previous research can be seen as an opportunity for improvement or improvement of methods that can contribute to the feature selection technique (feature section). By looking at the methods in feature selection, namely filtering, wrapper and embedding, a hybrid method can be used by combining the methods and using Feature Schuffling, Recursive Feature Elimination and Recursive Feature Addition techniques.

This study aims to contribute to the hybrid method for feature selection algorithms in the field of medical record data or medical records obtained due to problems and challenges to structured and unstructured medical record data so that it can help improve its accuracy and performance in a given prediction model.

REFERENCES

B. Shickel, P. J. Tighe, A. Bihorac, dan P. Rashidi, "Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis," *arXiv*, vol. 22, no. 5, hal. 1589–1604, 2017.

A. Panesar, *Machine Learning and AI for Healthcare*. .

P. B. Jensen, L. J. Jensen, dan S. Brunak, "Mining electronic health records: Towards better research applications and clinical care," *Nat. Rev. Genet.*, vol. 13, no. 6, hal. 395–405, 2012, doi: 10.1038/nrg3208.

C. Xiao, E. Choi, dan J. Sun, "Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review," *Journal of the American Medical Informatics Association*, vol. 25, no. 10. hal. 1419–1428, 2018, doi: 10.1093/jamia/ocy068.

I. D. Dinov, *Data science and predictive analytics: Biomedical and health applications using R*. 2018.

Y. Bengio, A. Courville, dan P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, hal. 1798–1828, 2013, doi: 10.1109/TPAMI.2013.50.

N. Polyzotis, S. Roy, S. E. Whang, dan M. Zinkevich, "Data lifecycle challenges in production

machine learning: A survey," *SIGMOD Rec.*, vol. 47, no. 2, hal. 17–28, 2018, doi: 10.1145/3299887.3299891.
B. A. Goldstein, A. M. Navar, M. J. Pencina, dan J. P. A. Ioannidis, "Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review," *J. Am. Med. Informatics Assoc.*, vol. 24, no. 1, hal. 198–208, 2017, doi: 10.1093/jamia/ocw042.
N. G. Weiskopf, G. Hripcsak, S. Swaminathan, dan C. Weng, "Defining and measuring completeness of electronic health records for secondary use," *J. Biomed. Inform.*, vol. 46, no. 5, hal. 830–836, 2013, doi: 10.1016/j.jbi.2013.06.010.
J. Latif, C. Xiao, S. Tu, S. U. Rehman, A. Imran, dan A. Bilal, "Implementation and Use of Disease Diagnosis Systems for Electronic Medical Records Based on Machine Learning: A Complete Review," *IEEE Access*, vol. 8, hal. 150489–150513, 2020, doi: 10.1109/ACCESS.2020.3016782.
K. Yan dan D. Zhang, "Feature selection and analysis on correlated gas sensor data with recursive feature elimination," *Sensors Actuators, B Chem.*, vol. 212, hal. 353–363, 2015, doi: 10.1016/j.snb.2015.02.025.
T. M. Le, T. M. Vo, T. N. Pham, dan S. V. T. Dao, "A Novel Wrapper-Based Feature Selection for Early Diabetes Prediction Enhanced with a Metaheuristic," *IEEE Access*, vol. 9, hal. 7869–7884, 2021, doi: 10.1109/ACCESS.2020.3047942.
A. U. Haq, D. Zhang, H. Peng, dan S. U. Rahman, "Combining Multiple Feature-Ranking Techniques and Clustering of Variables for Feature Selection," *IEEE Access*, vol. 7, hal. 151482–151492, 2019, doi: 10.1109/ACCESS.2019.2947701.
R. Ghorbani, R. Ghousi, A. Makui, dan A. Atashi, "A New Hybrid Predictive Model to Predict the Early Mortality Risk in Intensive Care Units on a Highly Imbalanced Dataset," *IEEE Access*, vol. 8, hal. 141066–141079, 2020, doi: 10.1109/ACCESS.2020.3013320.
T. Poongodi, D. Sumathi, P. Suresh, dan B. Balusamy, "Deep learning techniques for electronic health record (EHR) analysis," *Stud. Comput. Intell.*, vol. 903, hal. 73–103, 2021, doi: 10.1007/978-981-15-5495-7_5.
A. Zheng dan A. Casari, *Feature Engineering for Machine Learning*. .
K. J. Max Kuhn, *Feature Engineering and Selection: A Practical Approach for Predictive Models*. (Chapman & Hall/CRC Data Science Series) 1st Edition, 2019.
K. Yu dan X. Xie, "Predicting Hospital Readmission: A Joint Ensemble-Learning Model," *IEEE J. Biomed. Heal. Informatics*, vol. 24, no. 2, hal. 447–456, 2020, doi: 10.1109/JBHI.2019.2938995.
H. Wang, Z. Cui, Y. Chen, M. Avidan, A. Ben Abdallah, dan A. Kronzer, "Predicting Hospital Readmission via Cost-Sensitive Deep Learning," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 15, no. 6, hal. 1968–1978, 2018, doi: 10.1109/TCBB.2018.2827029.
H. Wang, Z. Cui, Y. Chen, M. Avidan, A. Ben Abdallah, dan A. Kronzer, "Predicting Hospital Readmission via Cost-Sensitive Deep Learning," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 15, no. 6, hal. 1968–1978, 2018, doi: 10.1109/TCBB.2018.2827029.
J. W. Baek dan K. Chung, "Context Deep Neural Network Model for Predicting Depression Risk Using Multiple Regression," *IEEE Access*, vol. 8, hal. 18171–18181, 2020, doi: 10.1109/ACCESS.2020.2968393.
T. Ruan *dkk.*, "Representation learning for clinical time series prediction tasks in electronic health records," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. Suppl 8, hal. 1–14, 2019, doi: 10.1186/s12911-019-0985-7.
M. Jamshidi *dkk.*, "Artificial Intelligence and COVID-19: Deep Learning Approaches for Diagnosis and Treatment," *IEEE Access*, vol. 8, no. December 2019, hal. 109581–109595, 2020, doi: 10.1109/ACCESS.2020.3001973.
C. Zhou, Y. Jia, dan M. Motani, "Optimizing Autoencoders for Learning Deep Representations

from Health Data," *IEEE J. Biomed. Heal. Informatics*, vol. 23, no. 1, hal. 103–111, 2019, doi: 10.1109/JBHI.2018.2856820.

X. Shi *dkk.*, "Multiple Disease Risk Assessment with Uniform Model Based on Medical Clinical Notes," *IEEE Access*, vol. 4, hal. 7074–7083, 2016, doi: 10.1109/ACCESS.2016.2614541.

S. J. Pasha dan E. S. Mohamed, "Novel Feature Reduction (NFR) model with machine learning and data mining algorithms for effective disease risk prediction," *IEEE Access*, vol. 8, hal. 184087–184108, 2020, doi: 10.1109/ACCESS.2020.3028714.

Q. Zhenya dan Z. Zhang, "A hybrid cost-sensitive ensemble for heart disease prediction," *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, hal. 1–18, 2021, doi: 10.1186/s12911-021-01436-7.

B. Wang *dkk.*, "A Multi-Task Neural Network Architecture for Renal Dysfunction Prediction in Heart Failure Patients with Electronic Health Records," *IEEE Access*, vol. 7, hal. 178392–178400, 2019, doi: 10.1109/ACCESS.2019.2956859.

G. Du, J. Zhang, Z. Luo, F. Ma, L. Ma, dan S. Li, "Knowledge-Based Systems Joint imbalanced classification and feature selection for hospital readmissions," *Knowledge-Based Syst.*, vol. 200, hal. 106020, 2020, doi: 10.1016/j.knosys.2020.106020.

M. Khalilia, S. Chakraborty, dan M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," 2011.

Y. Maali, O. Perez-concha, E. Coiera, D. Roffe, R. O. Day, dan B. Gallego, "Predicting 7-day, 30-day and 60-day allcause unplanned readmission: a case study of a Sydney hospital," hal. 1–11, 2018, doi: 10.1186/s12911-017-0580-8.

S. B. Golas *dkk.*, "A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: A retrospective analysis of electronic medical records data," *BMC Med. Inform. Decis. Mak.*, vol. 18, no. 1, 2018, doi: 10.1186/s12911-018-0620-z.

J. L. Speiser, "A random forest method with feature selection for developing medical prediction models with clustered and longitudinal data," *J. Biomed. Inform.*, vol. 117, no. October 2020, hal. 103763, 2021, doi: 10.1016/j.jbi.2021.103763.

T. Wanyan *dkk.*, "Relational Learning Improves Prediction of Mortality in COVID-19 in the Intensive Care Unit," *IEEE Trans. Big Data*, no. December 2020, 2020, doi: 10.1109/TBDATA.2020.3048644.

G. Kong, K. Lin, dan Y. Hu, "Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU," vol. 2, hal. 1–10, 2020.

Y. Ren, H. Fei, X. Liang, D. Ji, dan M. Cheng, "A hybrid neural network model for predicting kidney disease in hypertension patients based on electronic health records," vol. 19, no. Suppl 2, 2019, doi: 10.1186/s12911-019-0765-4.

N. L. Fitriyani, M. Syafrudin, G. Alfian, dan J. Rhee, "Development of Disease Prediction Model Based on Ensemble Learning Approach for Diabetes and Hypertension," *IEEE Access*, vol. 7, hal. 144777–144789, 2019, doi: 10.1109/ACCESS.2019.2945129.

Y. Sun dan D. Zhang, "Diagnosis and Analysis of Diabetic Retinopathy Based on Electronic Health Records," *IEEE Access*, vol. 7, hal. 86115–86120, 2019, doi: 10.1109/ACCESS.2019.2918625.

Y. Hao, M. Usama, J. Yang, M. S. Hossain, dan A. Ghoneim, "Recurrent convolutional neural network based multimodal disease risk prediction," *Futur. Gener. Comput. Syst.*, vol. 92, hal. 76–83, 2019, doi: 10.1016/j.future.2018.09.031.

A. Hajjam, E. Hassani, E. Andr, dan A. K. G, "Classification models for heart disease prediction using feature selection and PCA," vol. 19, 2020, doi: 10.1016/j.imu.2020.100330.

S. Galli, *Python Feature Engineering Cookbook*. Packt, 2020.