

DOI: <https://doi.org/10.38035/dhps.v2i1>Received: July 17th 2024, Revised: July 25th 2024, Publish: August 02nd 2024<https://creativecommons.org/licenses/by/4.0/>

Feature Optimization for Classification of Medical Records with Evolutionary Algorithm

Yuda Syahidin¹, Ade Irma Suryani², Ika Rahman³

¹Health Information Management, Piksi Ganesha Politechnic, Bandung, Indonesia, yudasy@gmail.com

²Medical Records and Health Information, Piksi Ganesha Polytechnic, Bandung, Indonesia,

³Physiotherapy, Ganesha Polytechnic, Bandung, Indonesia,

Corresponding Author: yudasy@gmail.com¹

Abstract: Feature selection in medical records is necessary because the data usually contains many irrelevant features and noise. Electronic Health Record, abbreviated as EHR, makes it possible to analyze large amounts of medical data. A Genetic Algorithm is widely used for feature selection because it has the ability or potential for global optimization of the selected features. Genetic Algorithm-based methods include many iterations (generations) in the crossover process, and mutation can produce new individuals because the Genetic Algorithm adopts a fitness value to represent how "good" the resulting individual is. The problem with heuristic algorithms is those simple genetic algorithms are not for processing high-dimensional data. Genetic algorithms in solution search techniques always get local optimum solutions which can cause failure to obtain optimal solutions during random searches. From these limitations, developing and improving genetic algorithms for feature selection on clinical data is necessary. First, sort the features based on the feature evaluation criteria to exclude irrelevant features through the fitness process in the evaluation with the accuracy value of the Support Vector Machine calculation. This way reduces the number of features and results in optimal features. Then to get the optimal solution, it is necessary to optimize the subset features that have been selected using a machine learning algorithm that determines the best parameters using a genetic algorithm.

Keywords: Electronic Medical Record, Genetic Algorithm, Feature Optimization, Support Vector Machine

INTRODUCTION

Genetic Algorithms are widely used for feature selection because they have the ability or potential for global optimization of the selected features (Dewan & Sharma, 2015; Ghorbani dkk, 2020; Kalinin dkk, 2018; Pawlovsky & Matsushashi, 2017). Genetic Algorithm-based methods include many iterations (generations) in the process of crossover, and mutation can produce new individuals because the Genetic Algorithm adopts a fitness value to represent how "good" the resulting individual (Duan dkk, 2021; Guo dkk, 2014). Each individual is evaluated through the stages of fitness. Calculation of fitness can be done with the MAE (Mean Absolute Error) technique stages (Guan dkk, 2017; Nguyen dkk, 2021), which is the output of the prediction model with a combination of specification features that act as model inputs (Dewan & Sharma, 2015; Liang dkk, 2020).

Research using genetic algorithms, the search technique carried out, namely sequence searches, can be divided into three categories such as forward search, reverse search and two-way search (Dewan & Sharma, 2015; Ghorbani dkk, 2020; Majidnezhad, 2015; Nguyen dkk, 2021). Time complexity is the optimization chosen based on the lowest, but the resulting subset of features only produces the locally optimal (Sun dkk, 2004), Random search can make the advantages of two methods; namely, The temporal complexity is lower than the global ideal but greater than the sequence search, and the best solution can be calculated. The genetic algorithm is a popular random search technique that has been and continues to be used in a variety of fields. (Guo dkk, 2014; Liang dkk, 2020; Luque dkk, 2011) presents a genetically based algorithm to perform simultaneous parameter selection and optimization for features from clinical data. The ideal global search, sequence search, and random search are the three types of search feature subsets and search algorithms. Global optimum tracking identifies a global subset of the original feature collection that is optimal. So far, only the whole method and the branch-and-bound (BBM) method have been implemented.(Peng dkk, 2021), It can attain global optimality, but is only useful in low-dimensional areas because to its high time complexity (Daniel, 2018; Zhaoke Huang dkk, 2019).

Optimization algorithms such as genetic algorithms have been carried out effectively in medical records to optimize the required parameters. They optimize feature values or weights to improve the quality of features resulting from medical image extraction features to be measured accurately (Duan dkk, 2021). Genetic Algorithm-based methods are used in optimizing pathology clinical data features to evaluate features and reduce error rates in data classification and classification techniques using the Artificial Neural Networks algorithm (Majidnezhad, 2015). Optimization technique by combining genetic algorithms and regularization approaches for feature selection in the formation of learning models for cancer classification data and gene selection, where this selection is for global optimization and regularization methods for local search (Zhengxing Huang dkk, 2018a). A genetic algorithm is used to optimize the hyperparameters of each classification algorithm which, using ensemble learning (Multistage ensemble learning) techniques, will choose the best accuracy value (Dhar, 2021). Research by combining the Genetic algorithm with other optimization algorithms is the GWO (grey wolf optimization) algorithm, where the method is used for feature selection and optimization of the MLP and SVM classification algorithms in clinical data (Multi-Layer Perceptron) (Daniel, 2018; Zhengxing Huang dkk, 2018b; Kamel dkk, 2019).

Based on previous research (Daniel, 2018; Dhar, 2021; Duan dkk, 2021; Ghorbani dkk, 2020; Kamel dkk, 2019; X. Y. Liu dkk, 2018; Majidnezhad, 2015), There are still certain flaws in using genetic algorithms, such as heuristic algorithms, and simple genetic algorithms are not ideal for creating optimal features. If the data sought has a large number of characteristics and data volumes, local optimum solutions are always found in the solution search technique, which can lead to failure to find solutions (Liang dkk, 2020; Zeng dkk, 2021; Zhong dkk, 2021), problems in the accuracy of the accuracy results from combining the GA algorithm with other optimization algorithms as well as testing techniques with the selected classification algorithm. Issues in the field of medical records using genetic algorithms in optimizing still have limitations in producing optimal features and the specified algorithm related to the relatively small number of features and instances due to the random determination of the population still in the local optimum and the time required in the optimization process(Duan dkk, 2021). It is necessary to review the testing of the optimal features, and from the testing stage, it affects the accuracy and time results if the dataset used is changed (Duan dkk, 2021). Another limitation is the problem of determining the optimum solution generated by the genetic algorithm as well as the time performance for selecting optimal features, the dataset size and features with noise (Ghorbani dkk, 2020; Majidnezhad, 2015).

The development of methods to handle feature optimization is quite open to being carried out, one of which is the development of genetic algorithms as criteria for determining features that are expected to produce optimal features in their formation and stages in optimizing the parameters of the SVM algorithm (Kamel dkk, 2019; Le dkk, 2021; Rojas-Dominguez dkk, 2017) namely C (complexity) and Gamma, determined by genetic algorithms and the evaluation

stage of the proposed model.

Medical records contain a variety of data kinds, ranging from structured data like prescription dates and doses to unstructured data like clinical narratives that explain the medical reasons for the medical record paperwork (Alharthi, 2018; Jensen dkk, 2012; Roque dkk, 2011). The relationship between structured and unstructured data can be seen in Figure 1 below.

Intelligent optimizations such as grey wolf optimization (GWO), particle swarm optimization (PSO), and genetic algorithm (GA) have been carried out effectively in medical record data where some optimization for parameters is require (Gong dkk, 2019; Guo dkk, 2014; Kamel dkk, 2019; Nguyen dkk, 2021; Sakri dkk, 2018).

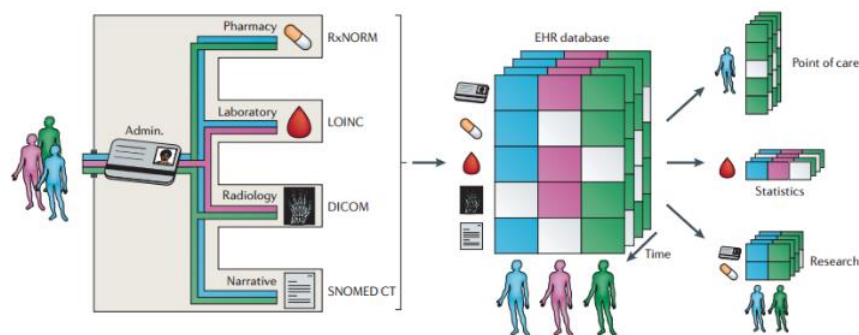


Figure 1. Electronic health record content (Jensen dkk, 2012)

Ghorbani (Ghorbani dkk, 2020) The method proposed in feature selection using a Genetic Algorithm to optimize unbalanced data features and tested with ensemble learning algorithms on private ICU hospital datasets resulted in an accuracy of 82.52%; only in this study there was no optimization of the parameters used in ensemble learning in testing the optimal features. Majidnezhad (Majidnezhad, 2015) proposes optimization using a genetic algorithm (GA) to optimize the features of pathological clinical data to evaluate features and reduce the error rate in data classification and classification techniques using the ANN algorithm. They were used so that it can seem complicated to optimize features. X.Y. Liu(X. Y. Liu dkk, 2018) proposes an optimization technique by combining genetic algorithm and regularization approach for feature selection in the formation of learning models for cancer classification data and gene selection where this selection for global optimization and regularization method used for local search of this study requires high computational time. Dhar (Dhar, 2021) proposes a feature optimization technique with a 2 (two) stage strategy.

The first stage is to determine the optimal features with 8 (eight) Machine Learning algorithms for classification accuracy in fitness values. The second stage combines the results of the optimal features into the learning ensemble and is selected based on the classification weight value. T.Minh.Le (Le dkk, 2021) proposes the PSO (Particle Swarm Intelligence) and GWO (grey wolf optimization) algorithm technique where the method is used to optimize the data features of diabetes medical records and optimize the Multi-Layer Perceptron classification algorithm. Daniel (Daniel, 2018) Research by combining the Genetic Algorithm with other optimization algorithms, namely the GWO (grey wolf optimization) algorithm for optimizing the feature value coefficients of medical record images by updating the position (distance) using a genetic algorithm to produce the optimum work. S.B. Sakri (Sakri dkk, 2018) proposed a breast cancer feature optimization technique with the Particle Swam Optimization algorithm, which then tested the results with a classification algorithm, namely nave Bayes, REPTree and KNN. Kamel (Kamel dkk, 2019) proposes optimization techniques for breast cancer features with Gray Wolf Optimization and improves accuracy with the Support Vector Machine classification algorithm.

METHODS

This study proposes developing a genetic algorithm to optimize medical record features and parameters for the Support Vector Machine algorithm that will be used to test the optimized features. The search process to find the best feature subset in the genetic algorithm is determined

by the fitness value (Chen dkk, 2021; Xiong dkk, 2019) from the calculation of accuracy using the SVM algorithm (Soguero-Ruiz dkk, 2016). The following is a diagram of the feature subset determination process and feature optimization using genetic algorithms and SVM parameter optimization, as shown in Figure 2 the proposed model approach.

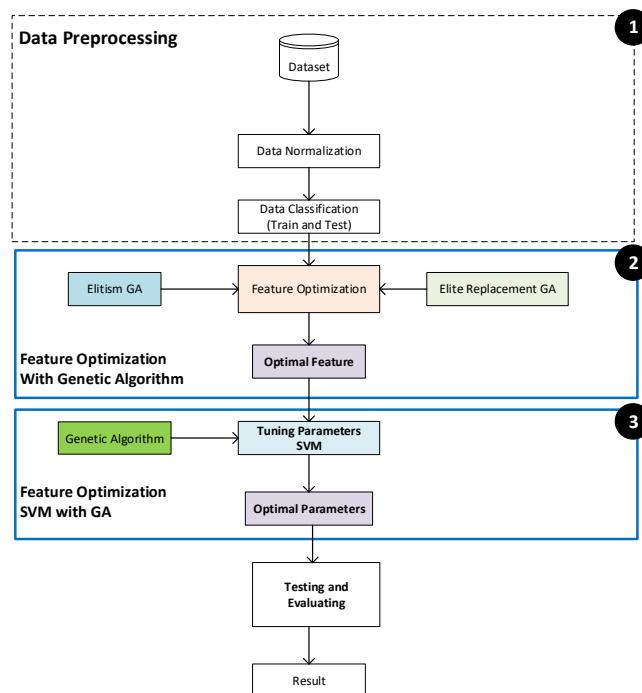


Figure 2. Proposed Method

RESULTS AND DISCUSSION

Based on the proposed model, the following steps can be explained:

Data Preprocessing

This stage must be completed in order to avoid the establishment of ineffective rules and patterns. Use a normalization approach, such as the min-max technique, to eliminate repeated data at this step. The data was then split into two groups: training and testing. The remaining 20% (samples) were chosen at random for the test group, while the remaining 80% (samples) were distributed at random for training. The following is Eqs 1. Min – Max Normalization (Louridi dkk, 2021; Swamynathan, 2017).

$$X[:, i] = \frac{x[:,i] - \min(x[:,i])}{\max(x[:,i]) - \min(x[:,i])} \quad (1)$$

Where X represents a dataset that has a number of data rows and columns of features, while x[:,i] represents feature i.

Feature Optimization with Genetic Algorithms

This stage develops a genetic algorithm for determining population initialization, calculating fitness, cross-over and mutation of each feature (individual). This stage applies a machine learning algorithm, namely the SVM (Support Vector Machine) algorithm, to determine the score of each feature. Population initialization is generated randomly based on the length of the feature to be optimized (Guo dkk, 2014; Sun dkk, 2004)(Sun dkk, 2004). Determination of the weight value in the fitness calculation can be used to determine the critical importance of the feature (Y. Liu dkk, 2011). Standard Deviation can be used to determine the best feature selection (Prasetyowati dkk, 2021). They determine the proposed fitness value by combining the weight value, classification accuracy and standard deviation value.

$$w_i = \frac{F(\text{accuracy}(FS_i))}{F(\text{accuracy}(FS_k))} \quad (2)$$

Where w_i is the weight value of the feature, $F(\text{accuracy}(FS_i))$ is a function to calculate the accuracy of feature selection and $F(\text{accuracy}(FS_k))$ is a function to calculate the accuracy of all features.

$$S = \sqrt{\frac{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n X_i)^2}{n(n-1)}} \quad (3)$$

After obtaining the weight values, accuracy and standard deviation values, to determine the fitness value using a combination of the formulas eq 2 and 3.

$$\text{Fitness}_i = w_i + \text{accuracy}_i + S \quad (4)$$

Where Fitness_i is the fitness value of the population, w_i is the weighted value of the subset features, accuracy_i is the accuracy value of the feature subset and S is the standard deviation value of the data. The crossover process in this study uses the technique of replacing all parents with all offspring if the offspring's fitness value is greater than the parent's fitness (locally elite replacement)(Ashlock, n.d.). The crossover process must ensure that the solution is still in the solution space (Luque dkk, 2011) like eqs 5.

$$\text{Individu}(K) \in S \quad (5)$$

Mutation determination is used because operators in genetic algorithms change certain genes of a chromosome. Mutations used by changing the value of a gene at a certain position (Ahn, 2006). Elite is used to maintain the fitness value (Guan dkk, 2017; Xiong dkk, 2019)The elitist proposed to retain the fitness value is placed in the position of the following population determination as shown in Figure 4. Elite genetic algorithm.

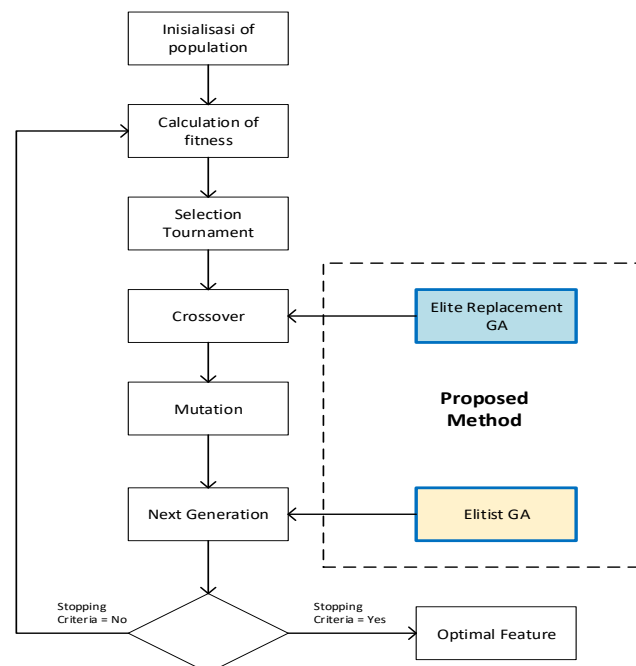


Figure 3. Proposed Elitist

Optimization of SVM Parameters with Genetic Algorithms

This stage is to perform the parameter optimization of the SVM algorithm with a genetic algorithm where the initialization of the population is obtained from the feature selection results and applies the KFold function (Zhenya & Zhang, 2021) in finding the optimal solution to these parameters. This stage will test the gamma parameter because the gamma value affects the learning speed. The higher the gamma

value, the faster the learning and achieving convergence. But if the gamma value is too large, it will cause the optimal gamma accuracy value to be missed. It will cause a decrease in accuracy when testing the C parameter (complexity). This parameter influences minimizing the error value in the classification process (Rojas-Dominguez dkk, 2017). The optimum hyperplane is found using SVM, which maximizes the distance between classes. A hyperplane is a function that can be used to separate groups (Kendale dkk, 2018; Rojas-Dominguez dkk, 2017). Figure 5 Hyperplane, which illustrates the illustration of a hyperplane in SVM (Yala dkk, 2017) as well as figure 6. determination of GA-SVM parameter values.

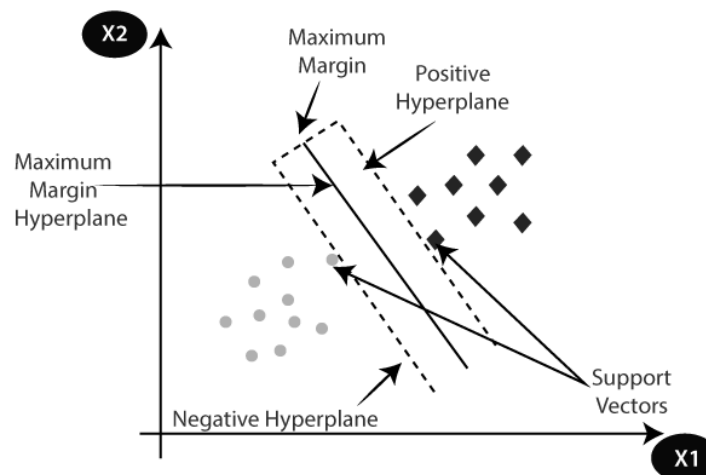


Figure 4. Hyperplane SVM (Yala dkk, 2017)

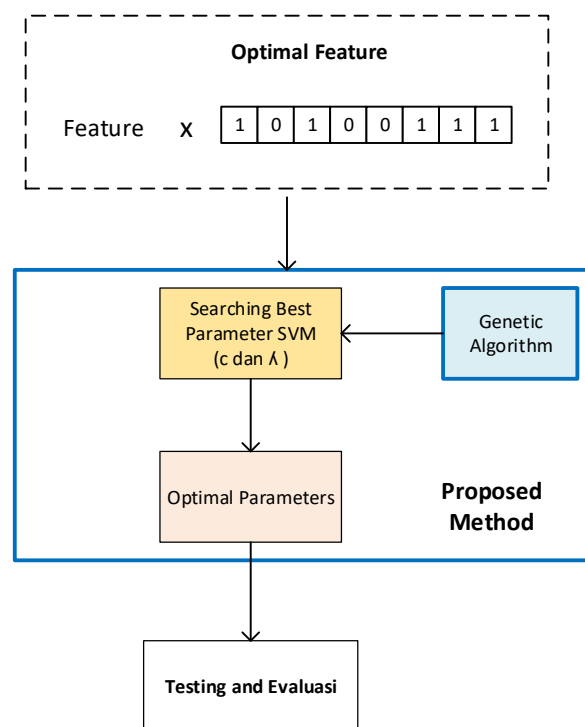


Figure 5. Proposed GA-SVM

The results of the trial process of the feature optimization model should be assessed against a series of criteria to evaluate the performance of the proposed method. Expressed in Eqs. 4 to 8 sets of criteria are used to carry out the evaluation.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$Specificity = \frac{TN}{FP+TN} \quad (6)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (7)$$

$$F_{\text{measure}} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Precision} + \text{Recall}} \quad (8)$$

CONCLUSION

The proposed method aims to produce optimal features of medical record classification with the initial stage of having optimal features from the medical record classification dataset, where the results obtained will be used to go to the next step, namely determining the best parameters c and γ parameters in the Support Vector Machine algorithm which aims to increase the value of accuracy better classified medical records. In the following research stage, data will be compiled, including the classification of medical records used to test the proposed optimal feature method and compared with the results of previous studies (state-of-the-art).

REFERENCE

- Ahn, C. W. (2006). Practical genetic algorithms. In *Studies in Computational Intelligence* (Vol. 18). https://doi.org/10.1007/11543138_2
- Alharthi, H. (2018). Healthcare predictive analytics: An overview with a focus on Saudi Arabia. *Journal of Infection and Public Health*, 11(6), 749–756. <https://doi.org/10.1016/j.jiph.2018.02.005>
- Ashlock, D. (n.d.). Evolutionary Computation for Modeling and Optimization. In *Department of Mathematics and Statistics University of Guelph Guelph, Ontario N1G 2W1 CANADA* (Vol. 59).
- Chen, Y., Wang, Y., Cao, L., & Jin, Q. (2021). CCFS: A Confidence-Based Cost-Effective Feature Selection Scheme for Healthcare Data Classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18(3), 902–911. <https://doi.org/10.1109/TCBB.2019.2903804>
- Daniel, E. (2018). Optimum Wavelet-Based Homomorphic Medical Image Fusion Using Hybrid Genetic-Grey Wolf Optimization Algorithm. *IEEE Sensors Journal*, 18(16), 6804–6811. <https://doi.org/10.1109/JSEN.2018.2822712>
- Dewan, A., & Sharma, M. (2015). Prediction of heart disease using a hybrid technique in data mining classification. *2015 International Conference on Computing for Sustainable Global Development, INDIACom 2015*, 704–706.
- Dhar, J. (2021). Multistage Ensemble Learning Model with Weighted Voting and Genetic Algorithm Optimization Strategy for Detecting Chronic Obstructive Pulmonary Disease. *IEEE Access*, 9, 48640–48657. <https://doi.org/10.1109/ACCESS.2021.3067949>
- Duan, J., Mao, S., Jin, J., Zhou, Z., Chen, L., & Chen, C. L. P. (2021). A Novel GA-Based Optimized Approach for Regional Multimodal Medical Image Fusion with Superpixel Segmentation. *IEEE Access*, 9, 96353–96366. <https://doi.org/10.1109/ACCESS.2021.3094972>
- Ghorbani, R., Ghousi, R., Makui, A., & Atashi, A. (2020). A New Hybrid Predictive Model to Predict the Early Mortality Risk in Intensive Care Units on a Highly Imbalanced Dataset. *IEEE Access*, 8, 141066–141079. <https://doi.org/10.1109/ACCESS.2020.3013320>
- Gong, X., Liu, L., Fong, S., Xu, Q., Wen, T., & Liu, Z. (2019). Comparative research of swarm intelligence clustering algorithms for analyzing medical data. *IEEE Access*, 7, 137560–137569. <https://doi.org/10.1109/ACCESS.2018.2881020>
- Guan, B., Zhang, C., & Ning, J. (2017). Genetic algorithm with a crossover elitist preservation mechanism for protein–ligand docking. *AMB Express*, 7(1). <https://doi.org/10.1186/s13568-017-0476-0>
- Guo, T., Han, L., He, L., & Yang, X. (2014). A GA-based feature selection and parameter optimization for linear support higher-order tensor machine. *Neurocomputing*, 144, 408–416. <https://doi.org/10.1016/j.neucom.2014.05.018>

- Huang, Zhaoke, Yang, C., Zhou, X., & Huang, T. (2019). A Hybrid Feature Selection Method Based on Binary State Transition Algorithm and ReliefF. *IEEE Journal of Biomedical and Health Informatics*, 23(5), 1888–1898. <https://doi.org/10.1109/JBHI.2018.2872811>
- Huang, Zhengxing, Dong, W., Duan, H., & Liu, J. (2018a). A Regularized Deep Learning Approach for Clinical Risk Prediction of Acute Coronary Syndrome Using Electronic Health Records. *IEEE Transactions on Biomedical Engineering*, 65(5), 956–968. <https://doi.org/10.1109/TBME.2017.2731158>
- Huang, Zhengxing, Dong, W., Duan, H., & Liu, J. (2018b). A Regularized Deep Learning Approach for Clinical Risk Prediction of Acute Coronary Syndrome Using Electronic Health Records. *IEEE Transactions on Biomedical Engineering*, 65(5), 956–968. <https://doi.org/10.1109/TBME.2017.2731158>
- Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: Towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6), 395–405. <https://doi.org/10.1038/nrg3208>
- Kalinin, A. A., Higgins, G. A., Reamaroon, N., Soroushmehr, S., Allyn-Feuer, A., Dinov, I. D., Najarian, K., & Athey, B. D. (2018). Deep learning in pharmacogenomics: From gene regulation to patient stratification. In *Pharmacogenomics* (Vol. 19, Nomor 7, hal. 629–650). <https://doi.org/10.2217/pgs-2018-0008>
- Kamel, S. R., Yaghoubzadeh, R., & Kheirabadi, M. (2019). Improving the performance of support-vector machine by selecting the best features by Gray Wolf algorithm to increase the accuracy of diagnosis of breast cancer. *Journal of Big Data*, 6(1). <https://doi.org/10.1186/s40537-019-0247-7>
- Kendale, S., Kulkarni, P., Rosenberg, A. D., & Wang, J. (2018). Supervised Machine-learning Predictive Analytics for Prediction of Postinduction Hypotension. *Anesthesiology*, 129(4), 675–688. <https://doi.org/10.1097/ALN.0000000000002374>
- Le, T. M., Vo, T. M., Pham, T. N., & Dao, S. V. T. (2021). A Novel Wrapper-Based Feature Selection for Early Diabetes Prediction Enhanced with a Metaheuristic. *IEEE Access*, 9, 7869–7884. <https://doi.org/10.1109/ACCESS.2020.3047942>
- Liang, K., Dai, W., & Du, R. (2020). A Feature Selection Method Based on Improved Genetic Algorithm. *2020 Global Reliability and Prognostics and Health Management, PHM-Shanghai 2020*. <https://doi.org/10.1109/PHM-Shanghai49105.2020.9281001>
- Liu, X. Y., Liang, Y., Wang, S., Yang, Z. Y., & Ye, H. S. (2018). A Hybrid Genetic Algorithm with Wrapper-Embedded Approaches for Feature Selection. *IEEE Access*, 6, 22863–22874. <https://doi.org/10.1109/ACCESS.2018.2818682>
- Liu, Y., Wang, G., Chen, H., Dong, H., Zhu, X., & Wang, S. (2011). An improved particle swarm optimization for feature selection. *Journal of Bionic Engineering*, 8(2), 191–200. [https://doi.org/10.1016/S1672-6529\(11\)60020-6](https://doi.org/10.1016/S1672-6529(11)60020-6)
- Louridi, N., Douzi, S., & El Ouahidi, B. (2021). Machine learning-based identification of patients with a cardiovascular defect. *Journal of Big Data*, 8(1). <https://doi.org/10.1186/s40537-021-00524-9>
- Luque, R. M., Elizondo, D., López-Rubio, E., & Palomo, E. J. (2011). GA-based feature selection approach in biometric hand systems. *Proceedings of the International Joint Conference on Neural Networks*, 246–253. <https://doi.org/10.1109/IJCNN.2011.6033228>
- Majidnezhad, V. (2015). A novel hybrid of genetic algorithm and ANN for developing a high efficient method for vocal fold pathology diagnosis. *Eurasip Journal on Audio, Speech, and Music Processing*, 2015(1), 1–11. <https://doi.org/10.1186/s13636-014-0046-1>
- Nguyen, M. H., Le Nguyen, P., Nguyen, K., Le, V. A., Nguyen, T. H., & Ji, Y. (2021). PM2.5 Prediction Using Genetic Algorithm-Based Feature Selection and Encoder-Decoder Model. *IEEE Access*, 9, 57338–57350. <https://doi.org/10.1109/ACCESS.2021.3072280>
- Pawlovsky, A. P., & Matsushashi, H. (2017). The use of a novel genetic algorithm in component selection for a kNN method for breast cancer prognosis. *Pan American Health Care Exchanges, PAHCE, 2017-March*. <https://doi.org/10.1109/GMEPE-PAHCE.2017.7972084>
- Peng, C., Wu, X., Yuan, W., Zhang, X., Zhang, Y., & Li, Y. (2021). MGRFE: Multilayer Recursive Feature Elimination Based on an Embedded Genetic Algorithm for Cancer

- Classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18(2), 621–632. <https://doi.org/10.1109/TCBB.2019.2921961>
- Prasetyowati, M. I., Maulidevi, N. U., & Surendro, K. (2021). Determining threshold value on information gain feature selection to increase speed and prediction accuracy of random forest. *Journal of Big Data*, 8(1). <https://doi.org/10.1186/s40537-021-00472-4>
- Rojas-Dominguez, A., Padierna, L. C., Carpio Valadez, J. M., Puga-Soberanes, H. J., & Fraire, H. J. (2017). Optimal Hyper-Parameter Tuning of SVM Classifiers with Application to Medical Diagnosis. *IEEE Access*, 6, 7164–7176. <https://doi.org/10.1109/ACCESS.2017.2779794>
- Roque, F. S., Jensen, P. B., Schmock, H., Dalgaard, M., Andreatta, M., Hansen, T., S ebye, K., Bredkj er, S., Juul, A., Werge, T., Jensen, L. J., & Brunak, S. (2011). Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Computational Biology*, 7(8). <https://doi.org/10.1371/journal.pcbi.1002141>
- Sakri, S. B., Abdul Rashid, N. B., & Muhammad Zain, Z. (2018). Particle Swarm Optimization Feature Selection for Breast Cancer Recurrence Prediction. *IEEE Access*, 6, 29637–29647. <https://doi.org/10.1109/ACCESS.2018.2843443>
- Soguero-Ruiz, C., Hindberg, K., Rojo-Alvarez, J. L., Skrovseth, S. O., Godtliebsen, F., Mortensen, K., Revhaug, A., Lindsetmo, R. O., Augestad, K. M., & Jenssen, R. (2016). Support Vector Feature Selection for Early Detection of Anastomosis Leakage from Bag-of-Words in Electronic Health Records. *IEEE Journal of Biomedical and Health Informatics*, 20(5), 1404–1415. <https://doi.org/10.1109/JBHI.2014.2361688>
- Sun, Z., Bebis, G., & Miller, R. (2004). Object detection using feature subset selection. *Pattern Recognition*, 37(11), 2165–2176. <https://doi.org/10.1016/j.patcog.2004.03.013>
- Swamynathan, M. (2017). Mastering Machine Learning with Python in Six Steps - review and good into in ML and NN approaches and basics + Python samples --Each topic has two parts: the first part will cover the theoretical concepts and the second part will cover practical impleme. In *Scandinavian Journal of Information Systems* (Vol. 19, Nomor 2). <http://aisel.aisnet.org/sjis%0Ahttp://aisel.aisnet.org/sjis/vol19/iss2/4>
- Xiong, B., Li, Y., Huang, M., Shi, W., Du, M., & Yang, Y. (2019). Feature Selection of Input Variables for Intelligence Joint Moment Prediction Based on Binary Particle Swarm Optimization. *IEEE Access*, 7, 182289–182295. <https://doi.org/10.1109/ACCESS.2019.2959064>
- Yala, A., Barzilay, R., Salama, L., Griffin, M., Sollender, G., Bardia, A., Lehman, C., Buckley, J. M., Coopey, S. B., Polubriaginof, F., Garber, J. E., Smith, B. L., Gadd, M. A., Specht, M. C., Gudewicz, T. M., Guidi, A. J., Taghian, A., & Hughes, K. S. (2017). Using machine learning to parse breast pathology reports. *Breast Cancer Research and Treatment*, 161(2), 203–211. <https://doi.org/10.1007/s10549-016-4035-1>
- Zeng, X., Lin, S., & Liu, C. (2021). *Multi-view Deep Learning Framework for Predicting Patient Expenditure in Healthcare*. <https://doi.org/10.1109/OJCS.2021.3052518>
- Zhenya, Q., & Zhang, Z. (2021). A hybrid cost-sensitive ensemble for heart disease prediction. *BMC Medical Informatics and Decision Making*, 21(1), 1–18. <https://doi.org/10.1186/s12911-021-01436-7>
- Zhong, Z., Yuan, X., Liu, S., Yang, Y., & Liu, F. (2021). Machine learning prediction models for prognosis of critically ill patients after open-heart surgery. *Scientific Reports*, 11(1), 1–10. <https://doi.org/10.1038/s41598-021-83020-7>