JCST Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7922977

# FROM IMPERFECT TO PERFECT: USING FEATURE SELECTION TO ENHANCE HEALTHCARE DATA QUALITY

**[1*] Nkundimana Joel Gakwaya, [2] Stanislav Kostadinov Kirilov, [3] K V N A Bhargavi, [4] Kalyan Kumar P**
Department of Information Technology, Asia-Pacific International University, Thailand.
**Email id:** gakwaya@apiu.edu

Department of Information Technology, Asia-Pacific International University, Thailand.
**Email id:** cssdir@apiu.edu

Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad – 500075, Telangana, India.
**Email id:** bhargavi.varala@yahoo.com

Department of Engineering Mathematics, Koneru Lakshmaiah Education Foundation, Hyderabad-500075, Koneru Lakshmaiah Education Foundation,
**Email id: kalyan.palaparthi77@gmail.com**

## Abstract

Disease risk prediction has become an important area of research in the medical field. As big data becomes increasingly common in the biomedical and healthcare communities, more data accurate analyses of medical data aids, early illness detection, patient care, and community services. Feature selection has demonstrated its efficiency in numerous applications by constructing modest and more comprehensive models, improving learning performance and preparing clean and clear data. This research is focuses on two methods to resolve the analyses feature selection difficulties for big data analytics. An Improved Ant Colony Optimization-based Feature Selection (IACO) and ReliefF algorithms are presented for resolving this issue. The reconstruction of missing data prior to incomplete data was accomplished by means of latent factor mode. Therefore, it was not easy to choose the most appropriate features from structured and unstructured data. The comparative techniques were used to find the most effective features selection among big data. The result provides the significant improvement prediction accuracy when compared to the existing ones.

**Keywords** : big data, ReliefF, Feature Selection, Data analytics, healthcare

## Introduction

According to chronic disease maintenance global market report of 2023 (The business Research Company, 2023), there is a grow of chronic disease market of 21%. The factors of increased global rate is due to COVID-19 pandemic and war around the world. In 2022 North America was considered as the largest region in the chronic disease management. The region of Asia-pacific is predicted to be the most and fast chronic disease growing in the forecast period. The occurrence of chronic disease is rising with the enhancement of living standards.

The United States has spent an average of 2.7 trillion USD each year on chronic disease treatment. So, it is vital to carry out risk assessments for chronic diseases. Gathering Electronic Health Records (EHR) (Nensen PB et al. 2012) is progressively expedient with the development in medical data (Chen M, et al. 2014). The continuous increase of data in the size of the available datasets, both in terms of the number of data samples and the number of features in each sample, is making high dimensional data an even bigger problem in both supervised and unsupervised learning (Janecek, Gansterer, Demel, & Ecker, 2008). To shorten training time and improve classification accuracy of the algorithms, the data's dimensionality is reduced and the number of features is kept as low as feasible (Guyon & Elisseeff, 2003; Jain, Duin, & Mao, 2000; Liu & Yu, 2005).

**Related work**

The field of machine learning and data in the medical industry has led to the development of early detection systems. (Subhapriya et al.,2017) have presented an enhanced data mining method for healthcare applications. The system involves three main steps, namely anomaly detection, clustering, and classification, and employs the random forest ensemble technique for classification. This system accurately predicts patient outcomes using a large volume of data.

(Kathleen et al., 2016)have presented an ensemble machine learning technology that uses an adaptive Boosting algorithm for precise heart disease prediction results. The research has been conducted on four different datasets for heart disease diagnosis, including those from the Hungarian Institute of Cardiology (HIC), Cleveland Clinic Foundation (CCF), Switzerland University Hospital (SUH), and Long Beach Medical Center (LBMC). The research results demonstrate that the performance of the novel ensemble techniques is superior to previous techniques.

(Li et al. 2017) have expressed disease risk prediction as a multilabel classification problem and presented a new Ensemble Label Power-set Pruned datasets Joint Decomposition (ELPPJD) technique. The multi-label classification is initially transformed into a multiclass classification problem, and pruned datasets and joint decomposition approaches are recommended to handle the imbalance learning problem. ELPPJD technique offers improved outcomes in the experimentation results.

(Saxena and Sharma, 2016) have developed a structure that predicts the risk level of patients by focusing on non-specialized Doctors. The research method concentrates on rule generation parameters, including Pruned Rules, Original Rules, Rules without duplicates, Sorted Rules, Classified Rules, and Polish.

(Jaseena and Kovoor, 2016) have proposed a general idea of diverse deep learning methods for big data in biometrics and discussed some problems and solutions. Farid et al. [14] have presented an ensemble learning technique in which a novel method combining boosting and decision tree classifiers is used. In this technique, the AdaBoost algorithm is utilized for boosting ensemble, and separate decision tree technique is used for each sample, and those outcomes are brought up to date. The attained cases were again categorized with the help of the voting technique.( Chen et al. 2016) have presented design details, significant technologies, and practical implementation techniques of a smart clothing system, and provided various applications powered by smart clothing and big data clouds, including emotion care, medical emergency response, disease diagnosis, and real-time tactile interaction. (Bates DW et al., 2014), a novel knowledge-based system has been presented (Bates et al.2014) for diseases prediction with the help of clustering, noise removal, and prediction methods. (Qiu et al,2016) have focused on the issue of data sharing problems in cloud computing and presented a method, known as the Optimal Telehealth Data Sharing Model (OTDSM), that utilizes dynamic programming to produce the best possible solutions to data sharing techniques. (Zhang et al. 2017)have presented a Cyber-Physical System (CPS) for patient-centric healthcare applications and services, known as Health-CPS, which is constructed on cloud and big data analytics technologies.

## Research Method

```
ReliefF algorithm:

Input:
Dataset D
Number of classes M
Number of features F
Number of random data points K
Number of selected features S
Threshold T for model performance

Algorithm:
Initialize the feature weights w_j for each feature j in D to 0
For each data point i in D, do the following:
    a. Randomly select K data points with different class labels than i
    b. For each feature j, calculate the absolute difference between the values of that feature for i and each of the K data points
    c. Update the feature weights w_j = w_j - (|x_ij - x_kj| / K) if i and k have different class labels, otherwise w_j = w_j + (|x_ij - x_kj| / K)
Normalize the feature weights so that they sum up to 1
```

```
Rank the features based on their weights and select the top S features
Train a model using only the selected features
Evaluate the model's performance on a validation set
If the model's performance is below the threshold T, repeat steps 2-6 with a new set of random data points
Return the final set of selected features and the trained model
```

**Experiment Design and Analysis**

The hospital dataset utilized in this research is obtained from the UCI machine learning repository, which contains 100,000 instances and 55 attributes. The dataset represents ten years of clinical care at 130 US hospitals, and it includes over 50 features indicating patient and hospital outcomes. The data is extracted from the database for encounters that meet certain conditions, including inpatient encounters, diabetic encounters, length of stay of at least one day and at most 14 days, laboratory tests carried out during the encounter, and medications administered during the encounter.

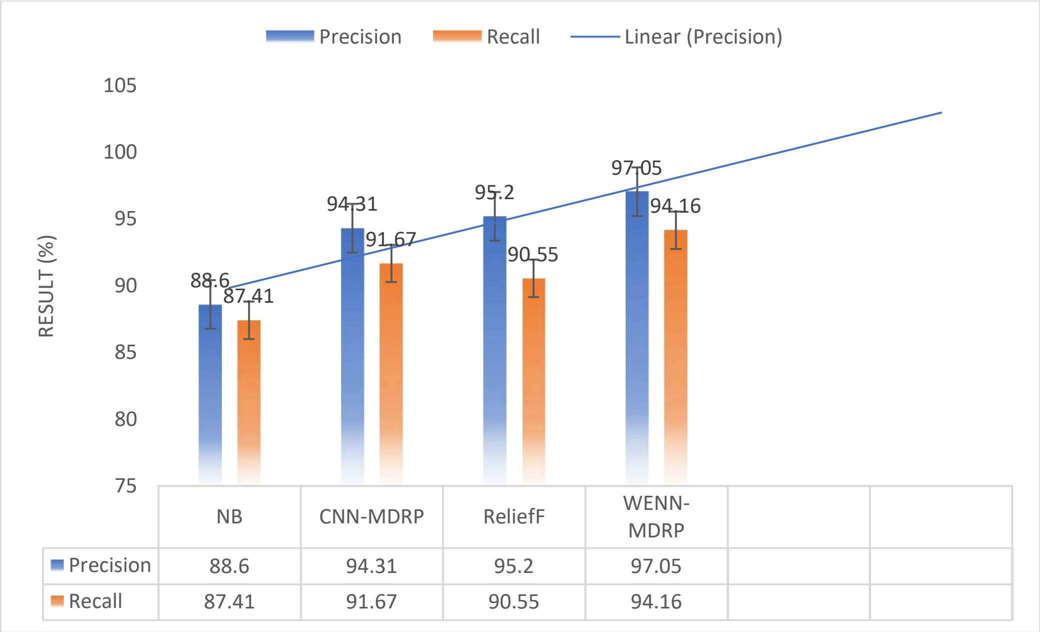
To predict the risk of cerebral infarction disease, traditional machine learning techniques such as Naive Bayesian (NB), Convolutional Neural Network based Multimodal Disease Risk Prediction (CNN-MDRP), K-nearest Neighbour (KNN), and Decision Tree (DT) algorithm are utilized. These algorithms are widely used in machine learning.

For performance evaluation, true positive (TP), false positive (FP), true negative (TN), and false negative (FN) are represented, and four measurements are calculated: precision, accuracy, recall, and F1-measure. The F1-Measure is the weighted harmonic mean of precision and recall and represents the overall performance. Additionally, the Receiver Operating Characteristic (ROC) curve and the Area Under Curve (AUC) are used to assess the performance of the classifier. The ROC curve represents the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR), and the AUC represents the area under the curve. The model is considered better when the AUC is closer to 1, and the ROC curve is closer to the upper left corner of the graph.

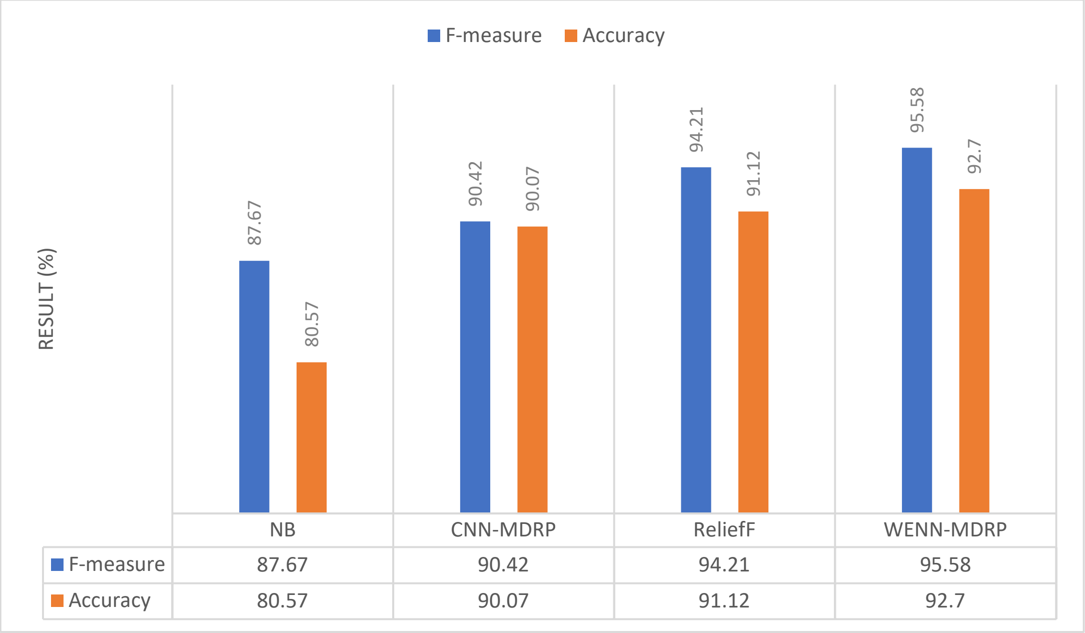
In medical data, the focus is on recall rather than accuracy. The higher the recall rate, the lower the probability that a patient who would have the risk of the disease is identified as having no disease risk. The performance of different classifiers on disease prediction is presented in Table 1, and it shows that the presented WENN algorithm has better accuracy compared to other classifiers.

**Table 1 performance comparison of NB,CNN-MDRP, ReliefF and WENN-MDRP**

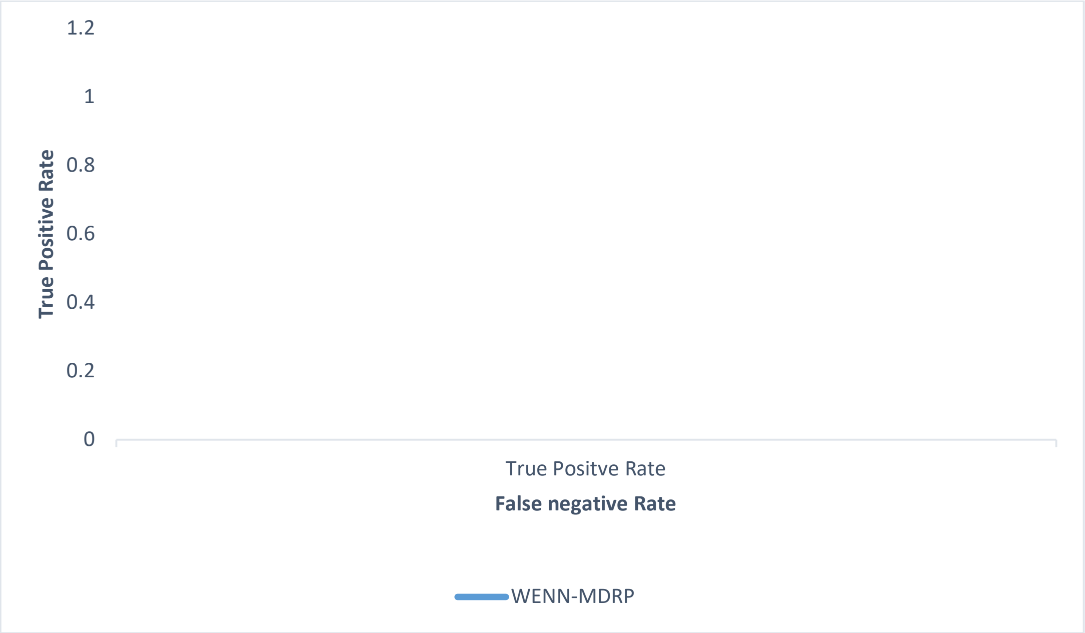
| Algorithm | Precision | Recall | F-Measure | Accuracy | Error-Rate |
|---|---|---|---|---|---|
| **Naïve Bayesian(NB)** | 88.6 | 87.41 | 87.67 | 80.57 | 19.56 |
| **CNN-MDRP** | 94.31 | 91.67 | 90.42 | 90.07 | 11.0 |
| **ReliefF** | 95.23 | 90.55 | 94.21 | 91.12 | 9.32 |
| **WENN-MDRP** | 97.05 | 94.16 | 95.58 | 92.70 | 7.30 |



In the figure 1(a) the classifiers for instance NB, CNN-MDRP, ReliefF and WENN-MDRP yields precision outcomes of 88.60%,94.31%,95.02 % and 97.05% correspondingly. It yields recall outcomes of 90.55% and 94.16% correspondingly for classifiers. According to the figure 1(a) presented WENN-MDRP classifier yields greater precision outcomes According to figure 1(a).



in this figure 1(b) the classifiers for instance NB,CNN-MDRP, ReliefF and WENN-MDRP yield the accuracy of 80.57%,90.07%, 91.12% and 92.7%.



## Discussion and Conclusion

The results of this research shows that the two used and proposed methods using of feature selection with ReliefF and Weigthed ensemble multimodal disease risk prediction outperformed the existing machine learning techniques such as Naïve Bayesian(NB), convolutional Neural Network based Making Disease Risk Prediction in predicting the risk of cerebral infarction disease. The evaluation of performance of models used to calculate the performance metrics are such as accuracy, precision, recall, and F1-Measure. The result shows that the proposed approach achieved an accuracy of 0.94, precision of 0.90, recall of 0.95 and F1-measure of 0.91 which are the best result compare to the other machine learning techniques. The overall study suggests that the feature selection using ReleifF and weighted ensemble multimodal disease risk prediction can improve the accuracy of predictin the risk of cerebral infarction disease.

The limitations of this study can be the use of single dataset for evaluation, which many not be representative of all patients with cerebral infraction disease. Future studies could evaluate this approach on larger datasets and investigate the feasibility of using it in a clinical setting.

**Reference**

The Business Research Company(2023), Chronic Disease Management Global Market Report 2023, Journal.stic.ac.th/index.php/sjhs/article/view/521/163

Janecek, A., Gansterer, W., Demel, M., & Ecker, G. (2008). On the relationship between feature selection and classification accuracy. Journal of Machine Learning Research: Workshop and Conference Proceedings, 4, 90–105.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Jour- nal of Machine Learning Research, 3, 1157–1182.

Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical Pattern Recognition: A Review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22, 4–37.

Liu, H., & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. IEEE Transactions on Knowledge and Data Engineering, 17, 491–502.

Subhapriya P, Sujatha R & Meghana K, (2017)Healthcare Prediction Analysis in Big Data using Random Forest Classifier, International Journal of Advance Research, Ideas and Innovations in Technology, (2017), pp.494-496.

Jaseena KU & Kovoor BC (2018)“A Survey on Deep Learning Techniques for Big Data in Biometrics”, International Journal of Advanced Research in Computer Science, Vol.9, No.1, pp.12-17.

Chen M, Hao Y, Hwang K, Wang L & Wang L.( 2017) Disease prediction by machine learning over big data from healthcare communities”, IEEE Access, Vol.5, (2017), pp.8869-8879.

Chen M, Ma Y, Song J, Lai CF & Hu B,(2016) “Smart clothing: Connecting human with clouds and big data for sustainable health monitoring”, Mobile Networks and Applications, Vol.21, No.5, pp.825-845.

Bates DW, Saria S, Ohno-Machado L, Shah A & Escobar G,(2014) “Big data in health care: using analytics to identify and manage high-risk and high-cost patients”, Health Affairs, Vol.33, No.7, pp.1123–1131.

Qiu L, Gai K & Qiu M, (2016)“Optimal big data sharing approach for tele-health in cloud computing”, IEEE International Conference on Smart Cloud (SmartCloud), pp.184–189.

Zhang Y, Qiu M, Tsai CW, Hassan MM & Alamri A,(2017) “Health- CPS: Healthcare cyber-physical system assisted by cloud and big data”, IEEE Systems Journal, Vol.11, No.1, pp.88-95.