Azerbaijan Journal of High Performance Computing, Vol 6, Issue 2, 2023, pp. 135-152
https://doi.org/10.32010/26166127.2023.6.2.135.152

Predicting the Status of Thyroid and Cardiovascular Patients According to Their Electronic Records Using Temporal Elements Based on the Combination of Shuffled Frog Leaping Algorithm (SFLA) and Deep Learning

Amirhossein Jalilzadeh Afshari

Azad University of Zanjan, Zanjan, Iran, amir.jalilzadeafshar@gmail.com

*Correspondence: Amirhossein Jalilzadeh Afshari, Azad University of Zanjan, Zanjan, Iran, amir.jalilzadeafshar@gmail.com

Abstract

Health and treatment are two of the most important application fields of information technology, in which the problem of predicting a disease is highly important. The physician makes such predictions based on the clinical condition of the patient and the level of facilities and advances in medical knowledge for the patient—information technology benefits from multiple methods to help this field. Accordingly, the patient information storage system, drug information, treatment and surgery systems, treatment follow-up systems, remote treatment systems, etc., aim to facilitate the treatment process. The patient can receive the best services within the shortest time due to these systems and information availability. The doctor can provide services to his patient anywhere in the world. This paper provided a model to predict the condition of patients based on their electronic records using temporal elements based on combining the shuffled frog leaping algorithm (SFLA) and deep learning. Accordingly, the evolutionary shuffled frog leaping algorithm (SFLA) and deep learning were used for preprocessing, feature selection, and classification. Two datasets of cardiovascular and thyroid diseases were utilized in the simulation section to ensure the efficiency of the proposed method. Based on this simulation, the proposed method indicated improvement compared to similar methods in the evaluated datasets. In the cardiovascular diseases dataset, this improvement was recorded as 1.4% and 3.2% compared to the author's previous and updated similar methods, respectively.

Keyword: Prediction of Patients' Conditions, Electronic File, Shuffled Frog Leaping Algorithm (SFLA), Deep Learning.

1. Introduction

According to WHO, health care quality is defined as a level of health services provided for individuals and populations that enhances the likelihood of optimal health outcomes.

Due to the advancement of digital technology, most tasks and work in the healthcare sector are being digitized and well-organized, which may dramatically improve the quality of healthcare services compared to the traditional approach. The Electronic Health Record (EHR) system appears at the forefront of implementation in healthcare institutions to enhance healthcare measure quality. EHR systems enable data-based clinical decision-making to improve the quality of healthcare. According to Gattiti et al., the accurate adoption of EHR systems can improve healthcare quality by increasing patients' safety and ensuring effective, efficient, timely, fair, and patient-centered care. Despite the advantages of EHR systems, problems or unintended outcomes prevent the acceptance and successful utilization of EHR systems in healthcare settings. Some of the most common of these factors are as follows: Physician burnout, failure of expectations, EHR market saturation, absences of innovation, data obfuscation, interoperability, privacy in sharing data, process prolonging until the completion of tasks, interruption in accomplishing tasks and solutions at the point of care, and non-coordination between technology and clinical context (Woldemariam, M. T., & Jimma, W., 2023). Disease prediction can benefit stakeholders such as the government and health insurance companies. It is capable of identifying patients at risk of diseases or health conditions. Subsequently, doctors can take appropriate measures to prevent or minimize the risk and, in turn, improve the quality of care and avoid potential hospital admissions. Also, considering the recent advancements in data analysis tools and techniques, disease risk prediction would be able to use large amounts of semantic information such as demographics, clinical diagnoses and measurements, health behaviors, laboratory results, prescriptions, and use of care measures (Hossain et al.; S., 2019).

Electronic health data are computerized medical records of patients that contain information about healthcare institutions. These data refer to diseases or conditions of the patient and are recorded in electronic systems with the initial goal of providing relevant health care and services. Administrative Healthcare Data, Administrative Claims Data, Computerized Claims Data, Digital Health Records, or Electronic Health Records are all used to describe electronic health data. Electronic health data are rapidly used for modeling and decision-making in the health care research section. This data type is used beyond keeping records in the health care research section. For example, they are used to analyze healthcare utilization, monitor the hospital care network's effectiveness, and develop predictive models for disease prediction (Lu et al., S., 2023, April).

Data mining is a systematic method to extract valuable and meaningful patterns from big data. It is a process that discovers unknown patterns and trends in data stores. Such information is mainly used to make prediction models. Prediction models play a substantial role in the healthcare sector. Different approaches and models help reduce human efforts to observe and quantify the relationships among the various features, patterns, colored graphs of healthcare datasets, etc. (Mulla, F. D., & Jayakumar, N., 2018, November). Machine-learning and deep-learning approaches have recently been used in data-driven healthcare research. Many supervised machine-learning algorithms have been utilized for

risk assessment by disease risk prediction models.

Similarly, using deep learning methods has brought remarkable advances in health informatics. Such models can efficiently capture and record complicated relationships between high-dimensional features through hierarchical levels of manipulation when used to train a prediction model. For instance, the convolutional neural network performs exceptionally well in visual medical image analysis. In addition, recurrent neural networks provide exceptional accuracy in language processing through the recurrent neural network architecture (Lu, H., & Uddin, S., 2023, April).

The accuracy and reliability of risk assessment models mainly depend on predictors and methods of development, validation, calibration, and clinical application. Administrative data are limited due to not having clinical specificity in choosing a proper set of predictors. The utilization of machine learning methods in medicine has also developed for laboratory conditions and results with recent multi-billion dollar investments in electronic medical records and their ever-increasing use and application in healthcare systems. Thus, an increase has emerged in the development of highly sophisticated prediction models using EMR over the last few years (Mahmoudi, E., Kamdar N. et al., 2020). This paper focused on predicting the health status of patients according to their electronic records using temporal elements based on the evolutionary shuffled frog leaping algorithm (SFLA) and deep learning technique.

2. Relevant literature

Mahmoudi et al. provided a model using the data mining of electronic health records for heart failure subtyping in 2023. Their paper was focused on evaluating whether text mining of electronic health record data can be used to improve register-based heart failure (HF) subtyping or not. The EHR data of 43,405 individuals were extracted from two Finnish hospital biobanks for mentioning the unstructured text of the Eruption Fraction (EF), and two 100-subject groups were randomly chosen versus the clinical evaluation. The structured laboratory data were then included for classification based on the HF subtype (Vuori, M. A., Kiiskinen, T., et al., 2023).

Amirhossein Jalilzadeh Afshari, M.S.S. (2018) proposed a model to predict the condition of patients according to their electronic records using time elements based on combining the artificial bee colony (ABC) algorithm and support vector machine. According to them, the issue of predicting diseases is one of the most critical issues today in the healthcare field, which is highly important. The physician makes this prediction based on the patient's clinical situation. Their research concluded that the best system for accurately diagnosing the disease can be developed based on the electronic file of the patient's conditions. They accomplished the prediction of the patient's condition according to their electronic files using time elements based on the combination of the artificial bee colony (ABC) algorithm and support vector machine. In their procedure, the artificial bee colony (ABC) algorithm was used for preprocessing and feature selection, followed by applying the decision support vector for classification. In the central part, two datasets were used to simulate

the proposed method to ensure its efficiency. The proposed method was compared with similar methods based on this simulation. Accordingly, the proposed method recorded a more appropriate improvement than those approached within the mentioned datasets. Their proposed method recorded 4% and 0.1% improvement rates in the heart and thyroid disease datasets, respectively [7].

Getzen et al. provided an exploratory model for equitable health in 2023 to assess the effect of missing data in electronic health records. According to them, electronic health records are gathered as a routine process of providing health care measures with a high potential to be used for improving patient health outcomes. These records contain multiple years of health data that can be used to predict risks, diagnose diseases, and evaluate treatments. However, they need a standardized and consistent format among institutions, especially in the United States, and can present significant analytical challenges. They encompass multi-scale data from heterogeneous domains and include structured and unstructured data. These data are gathered for individual patients at irregular intervals and with different frequencies. Besides analytical challenges, EHRs can reflect disparity; i.e., patients from different groups would have different amounts of data in their health records. Many of these issues can contribute to gathering biased data. As a result, the data from underserved groups may contain less information partly due to more sporadic care, which can be regarded as a missing data problem. There needs to be a framework for introducing missing values for the EHR data in this complicated form. Also, more research must be conducted to assess the effect of missing data in the EHR. In their work, they first introduced a term to define the three levels of EHR data. Then, they suggested a new framework to simulate real scenarios of missing data in the EHR to sufficiently evaluate their impact on predictive modeling. They combined a medical knowledge graph in the model to find and record dependencies between medical events to develop a more realistic missing data framework. They realized in the ICUs that missing data had a more significant negative impact on the performance of disease prediction models in groups tending to have less access to health care or seeking less health care. They also found that the effect of missing data on disease prediction models is more potent when using the knowledge graph framework for introducing actual missing values rather than eliminating random events (Getzen, E., Ungar, L., Mowery, D., Jiang, X., & Long, Q., 2023).

Mukherjee, P., Humbert-Droz, M., Chen, J. H., & Gevaert, O. (2023) proposed the SCOPE model to predict future diagnoses in office visits using electronic health records. They suggested an interpretable and scalable model to predict probable diagnoses in an encounter based on previous diagnoses and laboratory results. This model is designed to assist physicians in interacting with electronic health records. To do so, the EHR data of 2,701,522 patients at the Stanford Healthcare Center were collected and identified from January 2008 to December 2016. A population-based sample of 524,198 patients with multiple encounters with at least one recurrent diagnosis code was chosen. Then, a calibrated model was developed to predict ICD-10 diagnosis codes in an encounter based on previous diagnoses and laboratory results by utilizing a multi-label modeling

strategy based on binary communication. Logistic regression and random forests were examined as base classifiers, and multiple time windows were tested to gather previous diagnoses and laboratory results. This modeling approach was compared with the deep learning method based on the recurrent neural network. The best model utilized random forests as the base classifier and integrated demographic characteristics, diagnosis codes, and laboratory results. The best model had been calibrated, and its performance was comparable to or better than existing methods concerning different criteria, including the average AUROC of 0.904 in 583 diseases. The average AUROC with the best model was equal to 0.796 when predicting a patient's first occurrence of a disease label. The modeling approach performed better in terms of the AUROC ($p < 0.001$) compared to the tested deep learning method; however, it showed a lower performance regarding the AUPRC ($p < 0.001$). The interpretation of the model revealed that the model uses meaningful features and highlights many exciting relationships between diagnoses and laboratory results. The paper concluded that the multi-label model performs better than the RNN-based deep learning model, providing simplicity and potentially superior interpretability. While this model has been trained and validated on data obtained from a single institution, its simplicity, interpretability, and functionality make it a promising deployment candidate.

In 2023, Mohapatra et al. suggested a prediction model for heart diseases based on stacking classifiers. Cardiovascular diseases or heart diseases are known as one of the most substantial causes of death around the world. As estimated, about 1 out of every 4 deaths are caused by cardiovascular diseases, which are extensively classified as different types of abnormal heart diseases. However, diagnosing cardiovascular diseases is a time-consuming process in which the data obtained from different clinical trials are manually analyzed. Therefore, new approaches need to be developed to automate the detection of such abnormalities in human cardiac conditions to ultimately provide physicians with faster analysis by reducing diagnosis time and increasing results. Electronic health records are often utilized to discover valuable data patterns that help improve the prediction of machine-learning algorithms. In particular, machine learning helps solve problems like prediction in various domains, such as healthcare. Given the abundance of available clinical data, it seems necessary to use such information for the betterment of humanity. To this end, they presented a prediction model to predict heart diseases based on stacking different classifiers in two levels (basic and meta-level). Various heterogeneous learners are combined to generate robust model results. This model achieved a 92% precision rate in prediction with a 92.6% accuracy score, 92.6% sensitivity, and 91% specificity. The model performance was evaluated using different criteria, including accuracy, precision, recall, F1 scores, and area under the ROC curve values (Mohapatra, S., Maneesha, S., Patra, P. K., & Mohanty, S., 2023).

Essam et al. presented a model in 2023 for identifying heart disease risk factors according to electronic health records using advanced NLP and deep learning techniques. According to them, heart disease has still saved its place as the leading cause of death of people despite recent advances in the context of prediction and prevention. Thus,

identifying risk factors is a significant step in diagnosing and preventing heart disease. Automated detection of heart disease risk factors in clinical notes can contribute to modeling disease progression and clinical decision-making. Many studies have attempted to identify risk factors for heart diseases; however, none have identified all risk factors. These studies have proposed hybrid systems that combine knowledge-based and data-driven techniques based on dictionaries, rules, and methods of machine learning, which require significant human efforts. The National Informatics for Integrating Biology and the Bedside (i2b2) proposed a clinical natural language processing (NLP) challenge in 2014 with a track (Track2) focused on identifying risk factors for cardiovascular diseases risk factors in clinical notes over time. Clinical narratives provide a wealth of information that can be extracted using NLP and deep learning techniques. This paper was designed to improve the previous work in this field as a part of the i2b2 2014 challenge by identifying tags and features related to disease diagnosis, risk factors, and drugs by presenting advanced techniques of using stacked word embeddings. The i2b2 heart disease risk factors challenge dataset obtained using the stacked embeddings approach, which combines different embeddings, has remarkably improved. The model of this paper achieved an F1 score of 93.66% by utilizing BERT and character embedding (CHARACTER-BERT Embedding). The proposed model benefits from substantial results compared to other models and systems we have developed for the i2b2 2014 challenge (Houssein, E. H., Mohamed, R. E., & Ali, A. A., 2023).

Liang et al. suggested a disease prediction model based on integrating the data of several types of China's electronic health records. They state that disease prediction using various healthcare data to help doctors diagnose diseases has recently become a more prominent research topic. Their paper proposed a disease prediction model that combines different types of encrypted representations of Chinese EHRs. The model framework uses a Multi-head self-attention mechanism that combines textual and numeric features to improve the text representation. The BiLSTM-CRF and TextCNN models are used to extract entities and obtain representations from them. Text representations and the entities contained therein are combined to formulate the representations of electronic health records. The experimental results on electronic health record data gathered from a Class B general hospital in Gansu Province, China, indicated that their model has an F1 score of 91.92, which is better than the previous basic methods (Liang, Z., Zhang, Z., Chen, H., & Zhang, Z., 2022).

In 2022, Rakhmetulayeva, S., & Kulbayeva, A. (2022) presented a model for disease prediction using machine-learning algorithms based on electronic health record reports. They believed that the number of tasks assigned to predict the occurrence of infectious diseases is on a rapid growth path because of the availability of statistical data that supports the relevant analysis. Their paper described the current leading solutions to make short-term and long-term disease predictions. Also, the limitations and practical applications of these solutions have been mentioned. The paper gave much attention to the Naive Bayes classification, logistic regression, artificial neural network algorithm, and

k-means artificial neural networks as model analysis methods based on machine learning. The article provided an overview of two popular machine-learning algorithms for disease prediction. It used standard datasets for various diseases, including fungal infections, allergies, GERD, chronic cholestasis, stomach ulcer disease, diabetes, bronchial asthma, migraine, paralysis (cerebral hemorrhage), etc.

Zhao et al. provided a disease progression prediction model based on data from electronic health records in 2022. They suggested that electronic health records encompass patients' diagnostic, hospitalization, and medication records, in which a tremendous amount of structured time series data is at reach. Significant advances have occurred in electronic health record analysis and mortality prediction research. However, available electronic health records data have a sporadic and irregular nature, which prevents the accomplishment of scientific research and practical applications based on time-series electronic health records data. This paper generated several models based on deep neural networks to evaluate the prediction of patients' mortality. First, an attention mechanism was introduced to extend the factoring machine model, which dynamically learns the weights of different feature combinations to obtain some interpretability in the model. Second, the bidirectional gated regression unit (BiGRU) was applied to simultaneously capture the long-term dependencies in the forward and backward directions. Third, the BiGRU-AFM model was proposed and extensively examined in data mining based on electronic health records. The principal, second-order, and higher-order features are utilized to achieve a complete combination of features and a comprehensive emotional interaction in electronic health records. In particular, the attention-based FM (AFM) part presupposes combinations of low-order features, and the BiGRU section records higher-order feature interaction vectors. Their joint output appeared to make highly expressive vectors to predict the patients' mortality. Finally, a series of experiments were conducted on the public electronic health record dataset, in which experimental results demonstrated that the proposed BiLSTM-FM model performed better than the advanced basic models and gained about 97.9% in the widely used metric of the area under the curve (Zhao, F., Yu, X., Zhang, J., Li, X., & Li, R., 2022, December).

2. The analysis of the proposed method

Predicting the condition of patients in medicine and health has become highly important and essential. To achieve this goal, access to adequate information on the patient's complete health records is essential for accurate prediction. The first innovation to implement this section in the current research was using electronic health records (EHR). Utilizing this system provides sufficient information to the doctor or intelligent system. It has proven much better than the traditional manual systems available in many medical fields. A growing body of literature uses data obtained from EHRs to shape and design medical and health studies. Also, there is a growing but much smaller body of literature on essential and intrinsic data quality (DQ) problems in the EHRs as a source of research data, which is involved with the broad range of non-accidental human errors in different dimensions.

Some frameworks, such as those introduced by Weiskopf, N. G., & Weng, C. (2013) and Kahn, M. G., et al. (2016), may be used to classify the EHR DQ dimensions and help identify appropriate strategies for reduction. Selecting them appears essential in the EHR research. Process mining in healthcare can be challenging due to highly different care patterns among patients, healthcare professionals, and organizations. Thus, relying on this approach to complete those event reports with a time stamp can enhance measurable DQ requirements. Systematic event mapping, reconstruction, and analysis techniques are essential, just like other forms of data mining, because they are necessary for transparency about data cleaning and control procedures. DQ problems may happen during the life cycle of EHR data, from the design time of the original EHR application and database and its application in practice to data extraction for research, technologies, and methods used in them.

Another innovation of this research came from using the evolutionary shuffled frog leaping algorithm (SFLA) for "feature selection" to reduce the dimension and improve the quality of the available data. Also, a standard classification was needed in the final part of the current research for an optimal outcome to predict the patient's health status, which was met by the deep learning technique in the proposed system. The block diagram of the proposed method is illustrated in Figure 1.

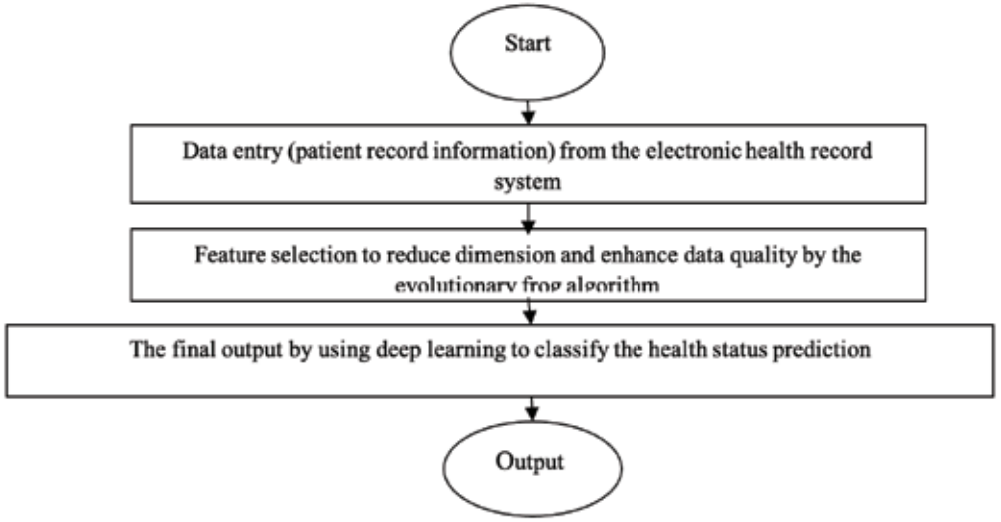


Fig.1: The block diagram of the proposed method

The general approach of the proposed method consists of the following steps (Weiskopf, N. G., & Weng, C., 2013):

o Preprocessing and selection of proper features

o We are identifying the DQ dimensions, detecting possible sources of information

related to the DQ, preparing a list of potential DQ problems, establishing links between these problems and experiments, data marking, and reducing the DQ problem if possible.

o Studying: The analysis of the results of the "Do" phase.

o Action: Taking measures to improve the future DQ.

The goal here is to specify data with non-acceptable quality as "bad" data, i.e., unusable. Incomplete but acceptable data are determined as "moderate" data, meaning that they can be utilized in some circumstances or experiments. Other data are unspecified or "good" and available for all purposes. The use of the proposed method consists of three major stages. The first stage involves identifying the archive of DQ problems for the research. The known problems are moved to this archive in advance. This archive will be completed with other issues specific to the research questions (the output of stage 1 includes the three dictionary entities: dimension, levels, and sources, as well as the very archive of the DQ issues and the list of tests for this research). The second stage presents the research data through analysis and classification tools. The third stage involves preparing a report about what has happened (Weiskopf, N. G., & Weng, C., 2013).

3. The analysis of simulation results of the proposed method

The simulation results of the proposed method were examined in this section, and the results of different evaluation stages were represented in Matlab and Weka software in the form of graphs and tables. Moreover, two datasets were utilized to assess the proposed method. The Heart Disease and Thyroid Disease datasets obtained from the UCI databases were used in the evaluations. The Heart Disease dataset included 303 samples with 75 features for each sample. Due to using electronic health records, the features are in two different sections, including patients' personal information and clinical and disease information. The dataset had missing data (values), just like the real-world datasets. The thyroid Disease dataset included 7200 samples in 10 different sets with 21 features, of which a set with 2800 samples was used in the evaluation. This set also had missing data, and the so-called used sets were not clean. The profile tables of the two datasets are shown in Figure 2.

Data Set Characteristics:	Multivariate	**Number of Instances:**	303	**Area:**	Life
Attribute Characteristics:	Categorical, Integer, Real	**Number of Attributes:**	75	**Date Donated**	1988-07-01
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	981302

A. The Heart Disease dataset

Data Set Characteristics:	Multivariate, Domain-Theory	**Number of Instances:**	7200	**Area:**	Life
Attribute Characteristics:	Categorical, Real	**Number of Attributes:**	21	**Date Donated**	1987-01-01
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	192054

B. The Thyroid Disease dataset

Fig. 2: The specifications of the two datasets

4. The parameters used to select the feature of the proposed method

The shuffled frog leaping algorithm (SFLA) was used in the proposed method to improve the process of discovering similarities and help the final clustering in extracting the impact of nodes and the proximity of influential factors to each other and weighting these factors as well as to select proper parameters for determining the most critical parameters, reducing the conclusion time, and enhancing the quality. The parameters considered for this algorithm are given in Table 1.

Table 1: The parameters used in the shuffled frog leaping algorithm

Row	Parameter Name	Description	Value
1	nVar	Number of decision-making variables	Equal to the number of features
2	VarMin	The lower limit of the variable values	-10
3	VarMax	The higher limit of the variable values	10
4	MaxIt	Maximum iteration of the algorithm	300
5	nPop	The number of frogs	50
6	nPopMemeplex	Memeplex size	10
7	nMemeplex	Number of Memeplex	5
8	alpha	Alpha value	3
9	beta	Beta value	5
10	sigma	Sigma value	2

The fixed values used in the algorithm have been obtained from repeated iterations and represent the best output with these values. An instance of the result of this algorithm is depicted in Figure 3.

As illustrated in Figure 3, the value of the best cost is high at the beginning of implementing the algorithm, and the algorithm tries to minimize and incline it towards zero, which approaches this value by advancing in the number of iterations. However, choosing the maximum number of iterations is essential in evolutionary and optimization algorithms such as the frog leaping algorithm since a small number may result in an improper best cost, and a large number may enhance the response time. Thus, choosing the correct value seems very important, and usually, this value is obtained by repeating the execution and observing the result. The algorithm has been implemented with different iterations in the proposed method aimed at achieving optimal results. The results are shown in Figure 4.

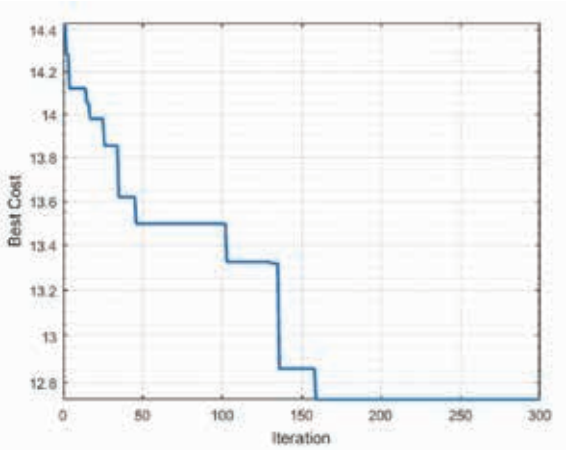


Fig. 3: The result of implementing the shuffled frog leaping algorithm

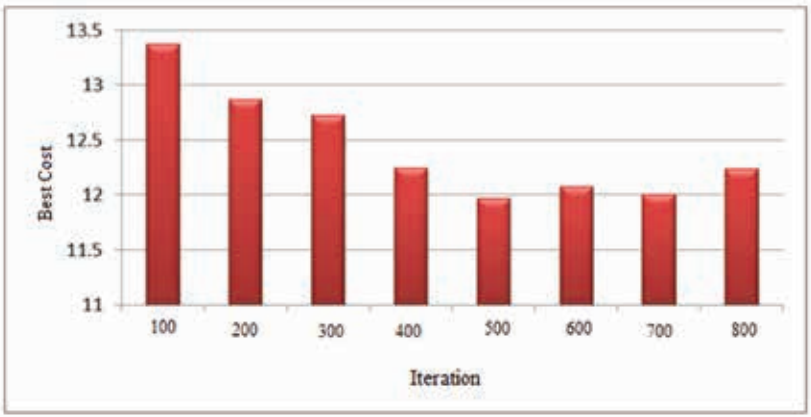


Fig. 4: Comparing the value of the best cost in different implementations of the frog leaping algorithm

As illustrated in Figure 4, the results in different implementations and the best cost value follow a decreasing trend. This reduction continued until repetition 500, and the repetitions increased to 100 units at each stage. This decrease in value has been slight in rounds 400 and 500 and increased after that. Hence, the value of 500 has been used as the maximum number of iterations of the frog leaping algorithm in the proposed method.

There are three different cost functions to choose from according to the problem in the frog leaping algorithm. These three functions are given in Table 2.

Table 2: Different functions to be used in the frog leaping algorithm

Row	Function Name	Calculation procedure

1	Rosenbrock	z=sum((1-x(1:n-1)).^2)+100*sum((x(2:n)-x(1:n-1).^2).^2);
2	Ackley	z=20*(1-exp(-0.2*sqrt(mean(x.^2))))+exp(1)-exp(mean(cos(2*pi*x)));
3	Sphere	z=sum(x.^2);

As shown in Table 2, there are three functions to choose for calculating the best cost in the algorithm. Accordingly, the algorithm was implemented with identical conditions and 300 repetitions with all three functions, and the results of the best final cost are provided in Figure 5.

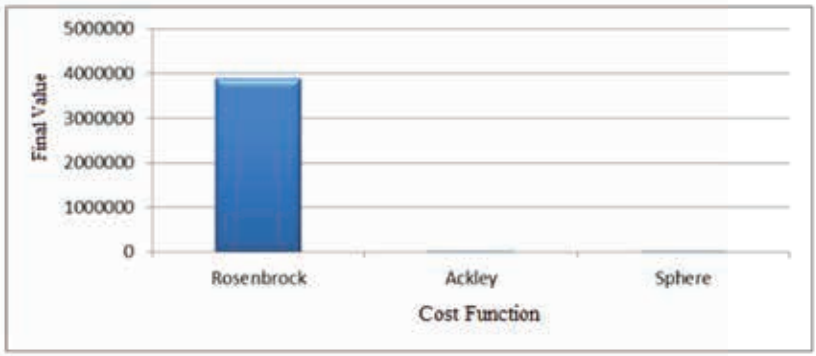


Fig. 5: The comparison of the functions that can be used in the frog leaping algorithm in terms of the value of the best final cost

As shown in Figure 5, the Rosenbrock and the Ackley functions have obtained the worst and best values at the end of the execution and after 300 iterations, respectively. The diagram of the cost function value change in the three compared functions is illustrated in Figure 6.

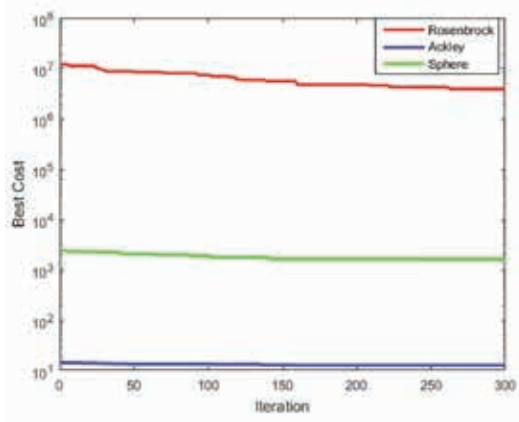


Fig. 6: The comparison of the value of the cost function at different iterations of the frog leaping algorithm with three usable functions

As shown in Figure 6, the Rosenbrock and the Ackley functions have shown the worst and best values at the end of the execution and after 300 iterations, respectively. According to these results, the Ackley function was used in the route selection section of the proposed method. The dataset following feature selection by adding the SF is shown below.

5. Comparing the proposed method with similar methods

The obtained collection was then evaluated and compared with different classification methods to determine the strength and improvement of the proposed method compared to the known and widely used methods. The criteria used for the evaluation are given in Table 3. The validation test by the K-Fold method with K=10 was used in all experiments. In this type of validation, the data is divided into K subsets. One of these K subsets is used for validation each time, and the other K-1 will be used for training. The K-Nearest Neighbor classification algorithms, J48, SMO, Decision Table, and the algorithm based on the Bayes theory were employed for evaluation.

Table 3: Introducing the evaluated criteria in the proposed method

Criterion	Description
Accuracy	The closeness of the agreement between the average value obtained from many test results and the accepted reference value is also called "average accuracy."
TP Rate	It represents the number of records whose real category is positive, and the classification algorithm has also recognized their category as positive.
FP Rate	It represents the number of records whose actual category is negative, and the classification algorithm has mistakenly recognized their category as positive.
Recall	A general parameter is used to evaluate the usefulness of the proposed algorithm and acts as the following equation (Th is the set of members inside each class, and Tr is the set of members inside the algorithm). $Recell= (T_h \cap T_r)/T_h$
Precision	A general parameter is used to measure the usefulness of the proposed algorithm and is obtained from the following relation. $Precision= (T_h \cap T_r)/T_r$
F-Measure	This criterion is obtained by calculating the average of the correlation between the two criteria of usefulness and utility, which is obtained by using Recall(R) and Precision(P) parameters in the form of the following equation. $F=2PR/P+R=2/(1/R+1/P)$

6. The first evaluation of the validity criteria results on the thyroid disease and heart disease datasets

As seen in Figure 7, in the heart disease dataset, followed by applying the DQ proposed in the previous chapter, all methods have demonstrated more acceptable and favorable results compared to the same set before this DQ application. Before applying the data quality improvement on the dataset, the support vector classification method did not have a good performance, and 67% of accuracy was in the third category of compared algorithms, or the proposed deep learning method with 89.3% has been in a place two algorithms above the rule Bayes algorithm method. After applying the proposed method, obtaining a new dataset, and repeating the experiments on this dataset, the recorded results indicated performance improvement and optimization. The deep learning method ranked first with an accuracy of 95.8%. It is worth noting that the results of all evaluated algorithms improved after applying the proposed DQ in the previous chapter. Figure 7 shows the results of the accuracy evaluation of these methods on the thyroid disease dataset.

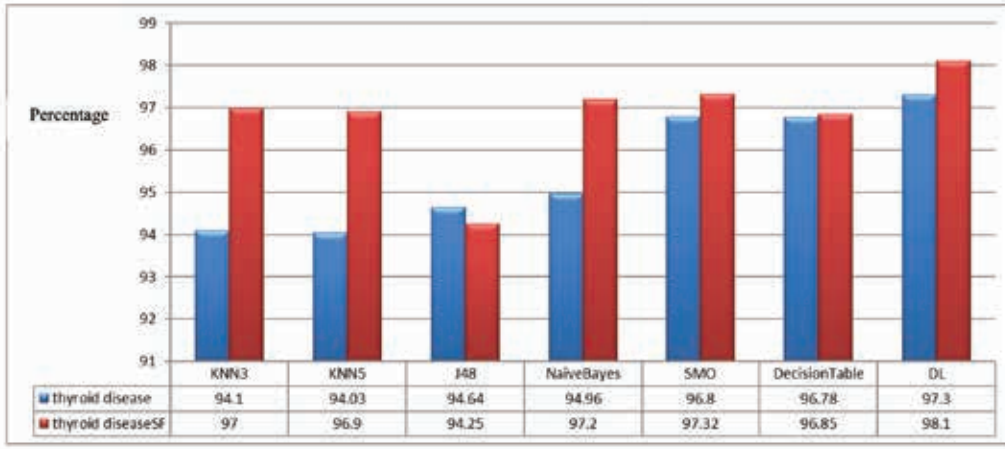
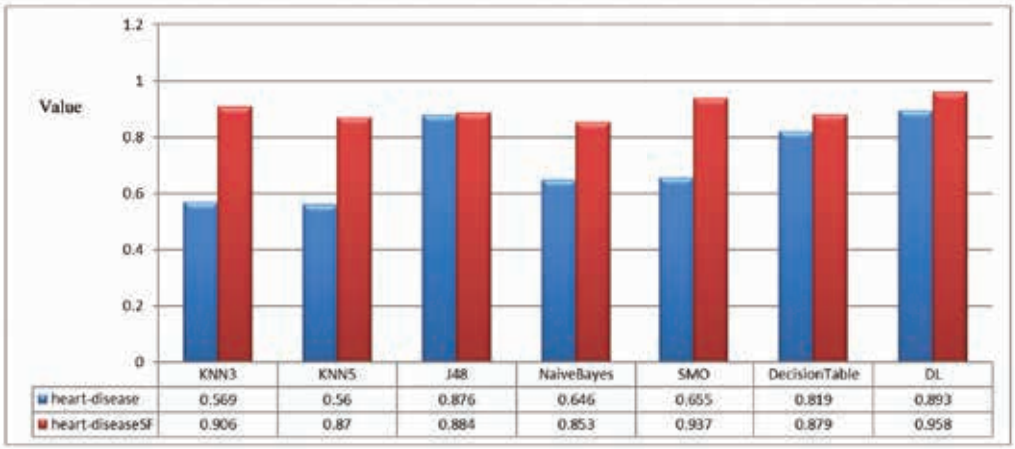


Fig. 7: Comparing the accuracy of the proposed method with other methods in the dataset of thyroid diseases

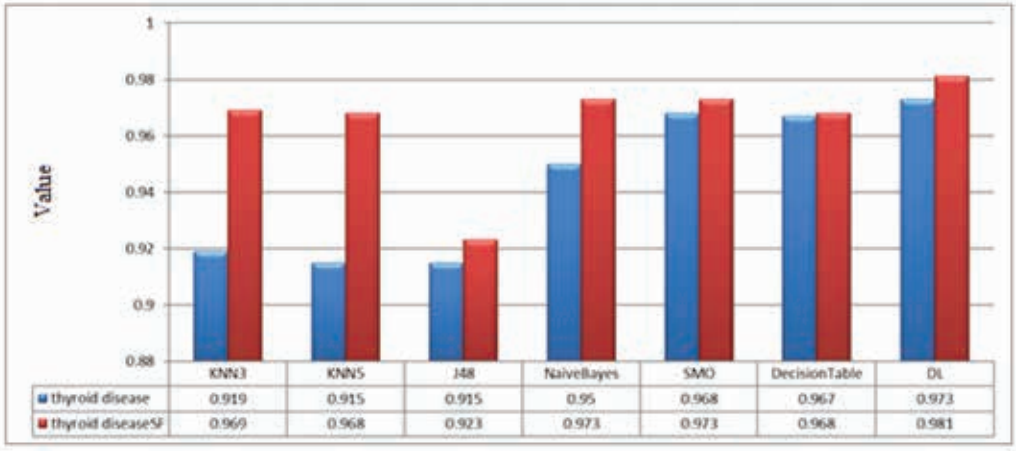
As shown in Figure 7, the evaluated methods have had a lower performance than the proposed method. However, it should be noted that the results of the Bayesian method are very close to those of the proposed method. After applying the DQ steps to the data, all methods have shown improvement and optimization. The deep learning method has performed better than other methods before and after applying the DQ in this dataset. The proposed method and other similar methods were evaluated on two datasets with the F-Measure to enhance the reliability of the evaluation process. The results of this evaluation of the heart disease dataset are illustrated in Figure 8.



	KNN3	KNN5	J48	NaiveBayes	SMO	DecisionTable	DL
heart-disease	0.569	0.56	0.876	0.646	0.655	0.819	0.893
heart-diseaseSF	0.906	0.87	0.884	0.853	0.937	0.879	0.958

Fig. 8: The comparison of the F-Measure of the proposed method and other methods in the heart disease dataset

As shown in Figure 8, all methods have demonstrated more acceptable and favorable results after applying the DQ proposed in the previous chapter compared to the same set before applying these measures in the heart disease dataset. Before applying the data quality improvement on the dataset, the proposed classification method had an F-Measure value of 0.893 and has been two algorithms above the rule Bayes algorithm method. After applying the proposed method, obtaining a new dataset, and repeating the tests on this dataset, the recorded results indicated performance improvement and optimization. The deep learning method ranked first with an F-measure value of 0.958. It is worth noting that the results of all the evaluated algorithms improved after applying the proposed DQ in the previous chapter. Figure 9 shows the results of the F-Measure evaluation of the methods on the thyroid disease dataset.



	KNN3	KNN5	J48	NaiveBayes	SMO	DecisionTable	DL
thyroid disease	0.919	0.915	0.915	0.95	0.968	0.967	0.973
thyroid diseaseSF	0.969	0.968	0.923	0.973	0.973	0.968	0.981

Fig. 9: The comparison of the F-Measure of the proposed method and other methods in the thyroid disease dataset

As shown in Figure 9, the evaluated methods had a lower performance than the proposed method. However, it should be noted that the results of the Bayesian method were very close to the proposed method. All methods demonstrated improvement and optimization after applying the DQ steps of the data. The deep learning decision method has performed better than other methods before and after applying the DQ in this dataset. The proposed method was compared with the results of papers (Mohapatra, S., Maneesha, S., Patra, P. K., & Mohanty, S., 2023) and [7] in the final evaluation. In the article [7], the author has performed the proposed method with the Artificial Bee Colony (ABC) and Decision Support Vector Machine (DSVM) algorithms. In contrast, the article (Mohapatra, S., Maneesha, S., Patra, P. K., & Mohanty, S., 2023) has only worked on the heart dataset, and this evaluation has been made merely on this set. The evaluation of the proposed method is depicted in Figure 10.

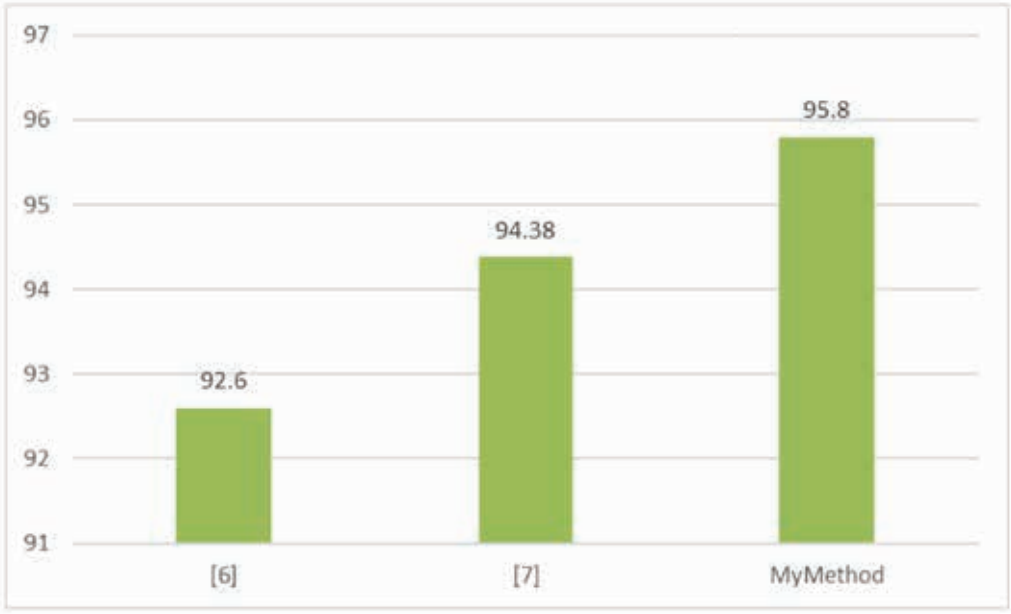


Fig. 10: The evaluation of the proposed method and similar papers in terms of accuracy criterion

As shown in Figure 10, the proposed method recorded an accuracy value of 95.8%, which has recorded an improvement of 1.4 compared to the same method of the present author, with an accuracy of 94.38%. Also, it shows an improvement of 3.2% compared to the similar work of the paper (Mohapatra, S., Maneesha, S., Patra, P. K., & Mohanty, S., 2023), which has a value of 92.6%.

Conclusion and summary

This research provided a method to create a proper platform for analyzing and using time data, which was used to analyze the data, followed by making the necessary

changes to advance and enhance the data quality based on the analysis. Accordingly, the goal was set to enhance the quality and optimize the data in future measures. The evolutionary shuffled frog leaping algorithm (SFLA) was used for preprocessing and feature selection in the proposed method, followed by the deep learning technique for classification. Accordingly, all the requested results and information were examined, and then the method was simulated. Two datasets were used in the simulation to ensure the method's performance. Based on this simulation, the proposed method improved the evaluated datasets compared to similar methods. The improvement rates were recorded as 1.4% and 3.2% in the heart disease dataset compared to the author's previous and updated methods, respectively.

For future work, there is scope to manage an application for collaboration of institutions that acquire and process such data and can perform clinical-level ML computations to solve real-time problems. This program leads to faster decision-making. The same method can be used to diagnose other chronic diseases as well. Implementing the proposed method in the real world can improve performance, including future work on the proposed method aimed at finding and fixing its shortcomings. Moreover, using other optimization and evolutionary algorithms rather than the frog optimization algorithm may bring different results, which need to be evaluated and suggested if they would be better.

References

Amirhossein Jalilzadeh Afshari, M.S.S. (2018). Predicting the condition of patients from electronic records using temporal elements based on the combination of bee colony algorithm and support vector machine. Master thesis, Azad University

Getzen, E., Ungar, L., Mowery, D., Jiang, X., & Long, Q. (2023). Mining for equitable health: Assessing the impact of missing data in electronic health records. *Journal of Biomedical Informatics, 139,* 104269.

Hossain, M. E., Khan, A., Moni, M. A., & Uddin, S. (2019). Use of electronic health data for disease prediction: A comprehensive literature review. *IEEE/ACM transactions on computational biology and bioinformatics, 18*(2), 745-758.

Houssein, E. H., Mohamed, R. E., & Ali, A. A. (2023). Heart disease risk factors detection from electronic health records using advanced NLP and deep learning techniques. *Scientific Reports, 13*(1), 7173.

Kahn, M. G., et al. (2016). A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *Egems, 4*(1).

Liang, Z., Zhang, Z., Chen, H., & Zhang, Z. (2022). Disease prediction based on multi-type data fusion from Chinese electronic health record. *Math. Biosci. Eng, 19*(12), 13732-13746.

Lu, H., & Uddin, S. (2023, April). Disease Prediction Using Graph Machine Learning Based on Electronic Health Data: A Review of Approaches and Trends. In *Healthcare* (Vol. 11, No. 7, p. 1031). MDPI.

Mahmoudi, E., Kamdar, N., et al. (2020). Use of electronic medical records in development and validation of risk prediction models of hospital readmission: systematic review. *bmj, 369*.

Mohapatra, S., Maneesha, S., Patra, P. K., & Mohanty, S. (2023). Heart Diseases Prediction based on Stacking Classifiers Model. *Procedia Computer Science, 218,* 1621-1630.

Mukherjee, P., Humbert-Droz, M., Chen, J. H., & Gevaert, O. (2023). SCOPE: predicting future diagnoses in office visits using electronic health records. *Scientific Reports, 13*(1), 11005.

Mulla, F. D., & Jayakumar, N. (2018, November). A Review of Data Mining & Machine Learning approaches for identifying Risk Factor contributing to likelihood of cardiovascular diseases. In *2018 3rd International Conference on Inventive Computation Technologies (ICICT)* (pp. 631-635). IEEE.

Rakhmetulayeva, S., & Kulbayeva, A. (2022). Building Disease Prediction Model Using Machine Learning Algorithms on Electronic Health Records' Logs. DTESI.

Vuori, M. A., Kiiskinen, T., et al. (2023). Use of electronic health record data mining for heart failure subtyping. *BMC Research Notes, 16*(1), 208.7.

Weiskopf, N. G., & Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association, 20*(1), 144-151.

Woldemariam, M. T., & Jimma, W. (2023). Adoption of electronic health record systems to enhance the quality of healthcare in low-income countries: a systematic review. *BMJ Health & Care Informatics, 30*(1).

Zhao, F., Yu, X., Zhang, J., Li, X., & Li, R. (2022, December). A Disease Progression Prediction Model Based on EHR data. In *2022 IEEE 24th Int Conf on High Performance Computing & Communications; 8th Int Conf on Data Science & Systems; 20th Int Conf on Smart City; 8th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)* (pp. 1625-1632). IEEE.

Submitted 12.09.2023
Accepted 02.11.2023