Informatics in Medicine Unlocked 23 (2021) 100561

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: http://www.elsevier.com/locate/imu

Medical decision-making based on the exploration of a personalized medicine dataset

Hafid Kadi [a,b,*], Mohammed Rebbah [a], Boudjelal Meftah [a], Olivier Lézoray [b]

[a] *Department of Computer Science, University Mustapha Stambouli, Mascara, Algeria*
[b] *Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, Caen, France*

ARTICLE INFO

Keywords:
Personalized medicine
Medical decision
Data representation
Dimension reduction
Data visualization
Classification

ABSTRACT

The emergence of personalized medicine and its exceptional advancements reveal new needs regarding the availability of adequate medical decision-making models. Considering detailed data on this medicine, the creation of a medical decision-making system may encounter many inhibitory factors, such as data representation, data reduction, data classification, and overall processing complexity. To address these challenges, this paper aims to create a useful model that can classify new patient data using efficient computations by choosing the best data processing series. Our methodology represents data with a recent model in the first task. During the second task, we continue with distance matrix production. The third task aims to reduce the dimensions of the last matrix. The fourth task applies a classification according to the results of reduced dimensionality. We have tested several distance measurements, dimensionality reduction methods, and classification techniques to achieve maximum performance. The evaluation results of the proposed model have shown excellent performance. Its F-measure can achieve an impressive rating with several classifiers (F-measure = 0.917, F-measure = 0.923, F-measure = 0.987 by 3-NN, random forest (RF) and support vector machine (SVM) classifiers). In addition to these performance measures, the computation time is also taken into account to choose among the proposed model's derived methods (time = 2 ms, time = 76 ms, time = 118 ms for the 3-NN, RF, and SVM classifiers, respectively). According to the performance and processing time criteria, we defined three use-case scenarios. However, we recommend using the RF classifier for the data reduced by the t-distributed stochastic neighbor embedding (TSNE) technique in practical cases to compromise performance and speed criteria.

1. Introduction

A medical decision, regardless of its nature, about the prognosis or diagnosis of a patient's response to prescription treatments or therapeutic plans, is a difficult task. Personalized medicine designs a medicine centered on an attitude adapted to patient profiles. These profiles can include information about behavioral preferences, environmental factors, key biomarkers, treatment history, demographics, genetic composition, and other useful information. This information, such as body temperature and blood pressure, is observed and organized by medical events. The electronic storage form of these profiles is referred to as Electronic Health Records (EHR). EHR offers a data source with a very detailed level of data. Data can be structural, nonstructural or semistructural with different qualities, such as nominal, numerical, Boolean, date-time, image, text, video, sound, and Extensible Markup Language (XML). In addition, numerical and nominal medical events can be presented as time series or nominal sequences. Therefore, there is a considerable need for models to explore these datasets and facilitate decision-making. To facilitate medical decision-making, one can consider computer-aided procedures. The latter usually rely on classification techniques that operate on training data, the quality of both are important to achieve excellent results. The qualities and types of the considered data and the ways of representing and preprocessing data can be factors of decisive influence. As a sample, we can quote the example of structural data processing and the completion of missing values, as well as coping with both loss of information and data during the transformation process [16].

However, choosing the best classification techniques regarding a specific medical problem is also difficult if one wants to conceive an efficient computer-aided medical decision-making system. Different exploration results for the same data set may be returned due to each technique's computation strategies. The computation time of medical

[*] Corresponding author. Department of Computer Science, University Mustapha Stambouli, Mascara, Algeria.
E-mail address: kadi.hafid@univ-mascara.dz (H. Kadi).

https://doi.org/10.1016/j.imu.2021.100561
Received 16 December 2020; Received in revised form 25 March 2021; Accepted 25 March 2021
Available online 31 March 2021

decisions is another crucial element of these systems, especially for urgent cases. The processed data volumes, data dimension reduction techniques, and complexity of the classifiers utilized are the main factors affecting the processing time taken to perform an automated medical decision. This paper proposes a model for automated medical decision-making based on personalized medicine datasets. The model is intended to classify new patient data as predictive and preventive measures for a given disease and estimate the appropriateness of treatment for a given patient to avoid adverse effects. To achieve this process, our model combines three main steps. The first step produces a data representation by selecting, transforming, and coding structural, temporal, and nontemporal data. We treat all types of data and their heterogeneity. These data can be simple data, such as age, or time series, such as periodic temperature tests. These data can also be Boolean data “0/1” to indicate the presence and absence of a given event. Nominal data sequences, such as pain sites or organ colors, are also processed during treatments. Minimizing the loss of data and information is another aspect treated during the representation, which generates an important number of attributes. The result will be a very detailed symbolic representation. Via the latter, the second step calculates the distance between all patients according to a determined measure; the result will be a numerical matrix. The dataset sometimes contains thousands of patients, which makes classification treatments computationally demanding. The third step reduces the previous numerical matrix dimensionality, using dimensionality reduction to simplify the calculations and reduce the processing time. The reduction is performed to obtain three dimensions. This choice is motivated by the visualization and interpretation needs of the classification results. The last step classifies the resulting 3D data and orients medical decisions by defining patient categories to be examined. We have identified the best combination of dimensionality reduction techniques with the best performing classifier. We implemented data mining techniques that produce an efficient automated medical decision-making model using these three steps.

In the contribution plan, we targeted structural data by maximizing the types of data applied (four types) and minimizing as much as possible the loss of data and information carried on the time series during the choice of model representation. We tested three distances between patients and four-dimensionality reduction techniques on their results. Four different classifiers are tested, and global evaluations consider both performance and computation time constraints.

According to the considered classification constraints, our work expresses three scenarios for applying the most appropriate treatment series. A final compromise between the computation time and the achieved performance allowed us to determine a preferred processing suite for practical application.

This paper is organized as follows: The data dimensionality reduction methods employed in related works are reviewed in Section 2. We detail the principle of our classification model in Section 3. Section 4 describes the experimentation and the evaluation of our proposal using a dataset for the Alzheimer’s Disease Neuroimaging Initiative.[1] The results and the performance achieved are also discussed. Section 5 globally summarizes our work and indicates the final choices on reduction and classification techniques in a real medical decision-making application.

2. Related works

The selected criteria can vary from one to several during a medical decision process, explaining this process treatment as a single or multiobjective optimization problem [38]. Papers [39–42] are interesting works that discuss medical decision-making, patient prioritization, and telemedicine in general. More than confidence, some cases require a fast decision, especially in urgent cases. This necessity requires dimension reduction techniques to simplify the calculations of the different classification techniques involved and decrease time.

2.1. Data dimensionality reduction

Personalized medicine data exploration involves working with large and multidimensional data. Many works use dimensionality reduction and visualization to understand, analyze, and treat such datasets. One of the famous dimensionality reduction techniques is principal component analysis (PCA) [1]. The PCA technique is based on transforming variables in a linear space and maximizing the variance to extract new characteristics in a reduced space that is referred to as principal components. In a study of Parkinson’s disease, Shukla et al. [2] employed the PCA technique for data dimensionality reduction in the dysphonia features’ classification process. In another study on the disease “systemic lupus erythematosus” [3], researchers applied PCA followed by visualization and interpretation of the results. Kernel PCA (KPCA) is another technique for reducing and extracting characteristics [1]. To solve the problem of linear data inseparability, the KPCA technique was applied with other techniques on subsets of gene expression from several diseases [4]. In a disease classification application, the proposed model uses the KPCA technique to reduce the data dimensionality and applies the least squares support vector machine (LSSVM) technique for classification [5]. Multidimensional scaling (MDS) is another technique for data reduction and visualization [6]. In an analysis realized by Vital et al. [7], the MDS technique is employed to reduce dimensionality and visualize, analyze, and report the results. T-distributed stochastic neighbor embedding (TSNE) is another dimension reduction and data visualization technique that is based on the distribution of data and their similarities to represent them in a new space of two or three dimensions. The new distribution more closely represents the data that have a high probability of similarity. Thus, the most dissimilar data are represented farther from each other [8]. After the data representation and similarity calculation, Zhang et al. [9] utilized the TNSE technique to reduce dimensionality and applied a clustering technique to define Parkinson’s disease subtypes. Workman et al. [10] investigated the TSNE reduction technique for patients’ temporal clinical data in a classification and analysis process that is intended to evaluate the effectiveness of deep learning on temporal data. This short overview shows that many studies have considered data dimensionality reduction techniques in EHR mining and classification.

The main finding of these studies reveals an interest in the direct application of a single data reduction technology without comparison with others. This finding raises a question about the compatibility of these reduction techniques and the knowledge behind the data.

2.2. Classification techniques

Sometimes exploration works use different classification techniques to compare the results and evaluate performance. For example, the work of Mohammad et al. [11] applied seven classification algorithms for heart disease prediction: k-nearest neighbor (k-NN), decision tree, naïve Bayes (NB), logistic regression, support vector machine (SVM), neural network and vote (hybrid technique with NB and logistic regression). Other approaches [12,13] predict coronary artery and heart diseases by the following classification techniques: multilayer ANN, SVM, NB, and decision trees (C4.5). In İlkim et al. [14] and Vital et al. [15], six classification techniques are considered. The first technique applies the NB

[1] Data utilized in preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The investigators within the ADNI contributed to the design and implementation of the ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be obtained at:http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

classifier, decision trees (Classification and Regression Tree (CART), C4.5, C5.0, C5.0 boosted), and random forest (RF) algorithms to analyze the effect of rheumatic fever on heart disease in childhood [14]. The second technique applies the alternating decision tree (AD tree), decision trees (C4.5), NB, BayesNet, K-Star, and RF to predict cancer disease and analyze the performance of the dataset [15].

The classifier selection behaviors applied for these latter approaches [11–15] vary among unjustified, popular and recommended approaches. This observation expresses no determined form of choice, and every professional can adopt any subjective measure. Additionally, the tested classifier number for each approach can vary between two and seven, a finding that may suggest exaggeration or insufficiency around the evaluation of the approaches.

In addition to the previous reviews, most of the cited approaches treat only one or two data types. The missing values are removed or supplemented, which leads to a case of data loss. The studies do not specify the transformations applied to the data during the representation, which probably generates a loss of information. In addition, the studies do not adopt the time series and nominal sequence forms for the data and keep only the numerical or nominal forms of certain events.

3. Method

We represent the set of patients by $P = \{P_1, P_2 \ldots, P_n\}$ and the total number of patients by n. E is the notation of the medical events set, where $E = \{E_1, E_2 \ldots, E_m\}$ and m is the total number of medical events.

Our classification model (Fig. 1) has four sequential tasks.

a) The model treats structural data and their numeric, date-time, nominal, and Boolean types to represent them more conveniently,
b) Distances between patients are computed with the Jaccard distance,
c) A dimensionality reduction of the distance matrix is produced on the output of the previous task, and
d) The classification is performed on the data in the obtained new embedding space.

We detail each of these steps in the sequel.

More than the vision of the best processing suite exploration on personalized medicine datasets, the first task applies a recent and promising data representation model that will be applied for the first time in our classification process.

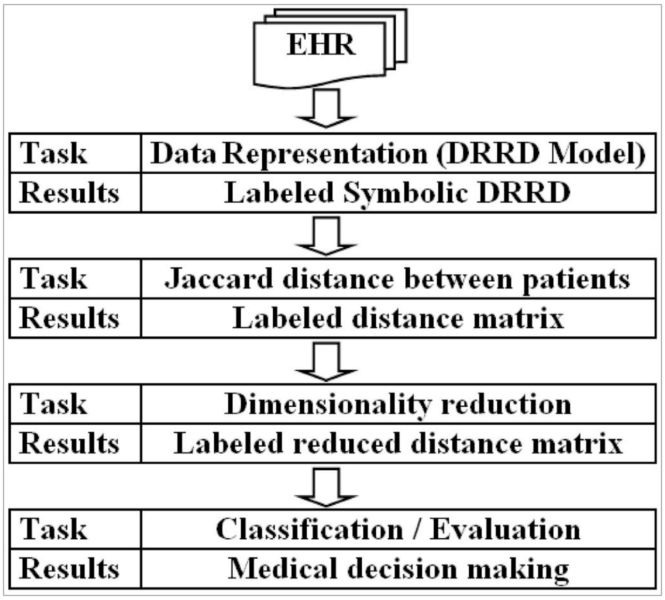

Fig. 1. Treatment process.

3.1. Data representation

This task includes preprocessing, transformation, and representation of the data. Due to the many data types and data time series encountered in the EHR, we use the Data Representation model per Region and Dispersion (DRRD) model that was recently proposed by Kadi et al. [16] to generate the symbolic representation of these data. We briefly expose its principle in the sequel.

The first phase in the DRRD model is numeric data representation by region (DRR), which is based on numeric data clustering. We notify this first phase by DRR. DRR transforms date-time type events into numeric data by calculating the age of observations. Next, it partitions numeric data from each event and uses the clusters as membership regions of observations. A table will represent each numeric event. Linearization by the join of all produced representation tables generates a single global representation table. Based on notification and marking operations, the DRR phase will generate the following three representations: by a real value, binary, and by a symbol.

The second phase attempts to mimic the first phase process plan, but this time the Boolean and nominal type events are considered. This phase begins with the transformation of data from Boolean events into nominal data. Each value of "0" will be replaced by "F", and each value of "1" will be replaced by "Y". For each event, E_i, a list L_i is created, where each list must include only the various values observed for the corresponding event. Subsequently, each list L_i corresponding to the event E_i will be transformed into the Table T_i, and each value in this list will generate a column in the associated table. This phase represents the list L_i's dispersion, which is why we refer to it as Data Representation by Dispersion (DRD). With the same linearization mechanism as the first phase of DRR, the DRRD model generates a single global representation table for all nominal events. The notification and marking operations also generate three representations: by a real value, binary, and by a symbol.

The DRRD model assembles representations of the same type generated by the DRR and DRD phases to form a single global representation of type by a real value, binary, or by a symbol as needed.

In this paper, we use the symbolic representation of the model DRRD (SDRRD). The produced SDRRD table (Equation (1)) includes $n = |P|$ rows according to the number of patients and q columns according to the new representation of the $m = |E|$ dataset events, such that $q \geq |E|$.

$$\forall s_{ij} \in \text{SDRRD} \quad (0 < i < |P| \ and \ 0 < j < q \ ) \Rightarrow \begin{cases} s_{ij} = null \\ Or \\ s_{ij} \ is \ \ a \ \ symbol \end{cases} \tag{1}$$

The DRRD process parameters will be saved for subsequent use to represent the data of new patients. Fig. 2 resumes the whole data representation task.

3.2. Distance between patients (distance matrix generation)

The SDRRD representation resulting from the previous task constitutes the essential point and main entry of the current task. The SDRRD matrix rows have varying sizes in terms of the expressed symbol number due to the variability among patients. This variability depends on the

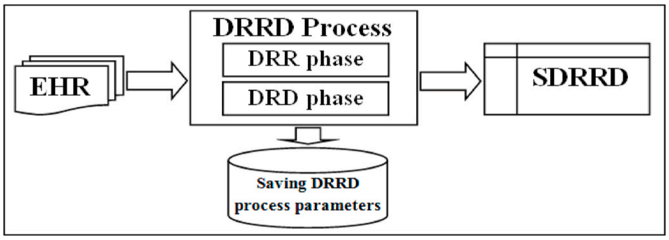

Fig. 2. Data representation process.

captured events and the saved time-series length. As an example, a fever as an event E_f can occur for only some patients P_i, and the number of times N_f of when its measurements are taken can also vary between two patients. Therefore, the symbolic representation will generate symbolic sequences of E_f for those patients only and with varying lengths N_f. The remaining patients will have no representation for E_f. Generally, this variability is the main cause of missing values present in the SDRRD representation matrix.

To compare the patients' representations and avoid searching for a method for missing data completion, we decided to calculate the distance matrix between all the patients (DMP) according to a distance that considers this problem. The Jaccard distance was chosen for this task. For all patients P_i and P_j, we compute the Jaccard index $J(P_i, Pj)$ according to Equation (2). This index computes the percentage of common attributes compared to all the attributes of patients P_i and P_j.

$$J\left(P_i, P_j\right) = \left|P_i \cap P_j\right| / \left|P_i \cup P_j\right| \tag{2}$$

The compliment of this latter index gives the Jaccard distance $DJ(P_i, Pj)$:

$$DJ\left(P_i, P_j\right) = 1 - J\left(P_i, P_j\right) \tag{3}$$

The generated DMP matrix is a symmetric matrix with:

$$\forall \quad \mathrm{v_{ij}} \in \mathrm{DMP} \quad (0 < i < |P| \ and \ 0 < j < |P| \ ) \Rightarrow \begin{cases} \mathrm{v_{ij}} = 0 \ if \quad i = j \\ Or \\ \mathrm{v_{ij}} \geq 0 \ if \quad i \neq j \end{cases} \tag{4}$$

The resulting DMP matrix provides strong basis support for comparison among patients. This matrix constitutes a tool for measuring the homogeneity of the captured observations expression among patients. Surrounded by its characteristics, such as the data volume and the main variables in expressivity value terms, this matrix requires specialized processing with the following task.

3.3. Dimensionality reduction

Sometimes the EHRs include the data of thousands of patients, which generates an extensive DMP matrix (n X n). The complexity of the classification technique to be applied in the following task and the large size of the DMP matrix can result in certain computational difficulties. In addition to the computation time diminution, the dimension reduction objective in our approach is also considered for data visualization, and the relevant variable exploitation induces a learning improvement of our techniques during the next task. The current task consists of reducing the DMP matrix to only three dimensions (3D). The result of this reduction is reduced DMP (RDMP). Intuitively, this dimensionality reduction will give for each patient its three most similar dimensions. Saving the advanced parameters of the reduction is an inevitable requirement to apply them in future patients' future processing. The data preparation for the next task of classification requires inserting the fourth column in the RDMP matrix, which carries the labeling information for each patient. This approach will be employed for classifier learning only. Filling in the labeling information based on the dataset EHR produces a labeled RDMP (LRDMP) matrix (refer to Fig. 3).

It is enough to apply only one reduction technique to the DMP matrix in our model. However, we have considered several reduction techniques for comparison and analysis and selected the best technique. Based on the study of comparison realized by Shaeela et al. [17] on dimension reduction techniques, we choose four techniques: PCA, KPCA, MDS, and TSNE. Each technique has a strategy that gives us the following four reduced matrices: RPCA (reduction of PCA), RKPCA (reduction of KPCA), RMDS (reduction of MDS), and RTSNE (reduction of TSNE). Each strategy has several rows equal to the number of patients and three columns (3D reduction). Labeling the reduction of each technique produces the reduced matrices labeled LRPCA (Labeled RPCA), LRKPCA (Labeled RKPCA), LRMDS (Labeled RMDS), and LRTSNE (Labeled RTSNE).

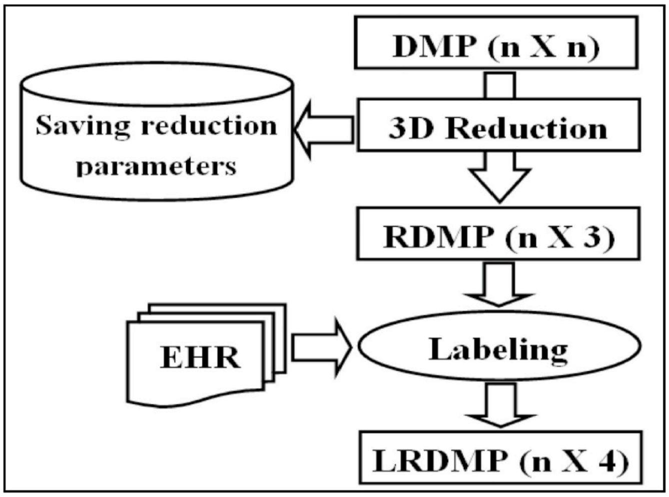

Fig. 3. Dimensionality reduction process.

In addition, reducing the dimensionality to a 3D space enables the visualization of all the EHRs at once, which can be of interest for interactive data exploration.

3.4. *Classification*

The classification of a new patient NP requires its symbolic representation SDRRD, which is referred to as NSDRRD. In this case, we use the parameters of the DRRD process that are saved during the first task. Let SDRRD(NP) be the function that returns the symbolic representation of the patient NP according to the DRRD model where:

$$\forall P_i \in P \ \mathrm{SDRRD}_i = \mathrm{SDRRD}(P_i) \quad \wedge \quad \mathrm{NSDRRD} = \mathrm{SDRRD(NP)} \tag{5}$$

Subsequently, the Jaccard distance is calculated from this NP to all patients using the SDRRD representations (Equation (6)). Of course, this distance will generate a new line DMP_{NP} with |P| columns.

$$\begin{aligned} &\forall P_i \in P \quad \mathrm{DMP}_{NP}[i] = \mathrm{DJ}(NP, P_i) \quad \text{where} \\ &\mathrm{DJ}(NP, P_i) = \mathrm{DJ}(\mathrm{NSDRRD}, \mathrm{SDRRD}_i) \wedge \mathrm{SDRRD}_i = \mathrm{SDRRD}(P_i) \end{aligned} \tag{6}$$

To reduce the dimension of this new line DMP_{NP}, we reuse the reduction parameters generated and saved during the third task concerned with dimension reduction (Equation (7)). The term "new patient RDMP (NRDMP)" will reduce this new patient line.

$$\begin{aligned} &\mathrm{NRDMP} = \mathrm{3DREDUCTION}\ (\mathrm{DMP_{NP}}) \quad \text{where} \\ &\forall P_i \ \in \ P \ \mathrm{RDMP}_i \ = \mathrm{3DREDUCTION} \ \ (\mathrm{DMP}_i) \end{aligned} \tag{7}$$

DMP_i is the line number i in the DMP matrix, and 3DREDUCTION (DMP_i) is the function that returns the reduction of this line according to the deployed technique.

Based on learning from the LRDMP matrix data, our process applies a classification technique CT on the new reduced line NRDMP to find the new patient category NPC.

$$\mathrm{NPC} = \mathrm{CLASSIFY(CT, NRDMP, LRDMP)}. \tag{8}$$

The calculation strategies adopted by the classification techniques and the performances obtained differ among techniques. To have an exhaustive evaluation, we will apply several classification techniques on the four reductions labeled LRPCA, LRKPCA, LRMDS, LRTSNE and apply several classification techniques on the labeled DMP (LDMP) matrix without any dimensionality reduction. Based on studies [18–21], with a reasonable number and by a compromise among popularity, accuracy, and recommendation of techniques, we choose the following four classifiers to apply:

- **NB:** It is a probabilistic model of supervised classification based on Bayes' theorem and assumes independence between two attributes. This conditional model computes the posterior probability of the class categories for the input observations, and the observations are classified according to the class with the maximum posterior probability [22,23].
- **SVM:** It is a supervised learning method based on a transformation using kernels and data separation by margin maximization to produce separating hyperplanes. The simplicity of using the SVM and its theoretical foundations justify its usefulness in several domains [24].
- **KNN:** KNN uses the labeled learning set to classify a given example. This algorithm uses a distance to find the first K examples closest to this entry. By a majority vote, the most present class among this close K will be assigned to this example [25,26]. For our classification, we use k = 3 and notify this technique by (3-NN).
- **RF:** It is a classification and regression technique in its origin that uses a combination of random decision trees. Similar to the bagging technique, the RF uses average aggregation for regression and majority voting for classification. This technique's principal idea is to train decision trees on different data subsets and randomly chosen variables [27,28].

New patients can be considered 'out-of-sample' examples that do not belong to the initial training set. Their embedding coordinates in the dimensionality reduction are calculated by projection using the PCA and KPCA techniques. For the MDS technique, we use the out-of-sample extension proposed by Bengio et al. [29], which considers a normalized kernel. We consider the out-of-sample algorithm proposed in the paper of Gisbrecht et al. [30] for the kernel TSNE technique as an extension to embed new patients.

4. Experimental results

The Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu) is the source that provides the dataset employed in this experimentation. The ADNI was launched in 2003 as a public-private partnership led by Principal Investigator Michael W. Weiner, MD. The primary goal of the ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessments can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD) [31].

The dataset, especially the ADNIMERGE_May15 table, contains 90 attributes. Among these attributes we can cite the following attributes: AGE (patient age), FDG (average FDG-PET of angular, temporal, and posterior cingulate), PTETHCAT (ethnicity), PTEDUCAT (education), COLPROT (study protocol of data collection), Hippocampus_bl (UCSF hippocampus at baseline), APOE4 (apolipoprotein epsilon 4), PET (average PIB SUVR of frontal cortex; anterior cingulate; precuneus cortex; and parietal cortex), AV45 (average AV45 SUVR of frontal; anterior cingulate; precuneus; and parietal cortex relative to the cerebellum), CDRSB (clinical dementia rating scale – sum of boxes), ADAS11 (ADAS-Cog-with 11 tasks), GDP (average PIB SUVR of frontal cortex; anterior cingulate; precuneus cortex; and parietal cortex), MMSE (Mini-Mental State Examination), RAVLT_immediate (Rey Auditory Verbal Learning Test immediate) [31] and others. Patients are categorized into five classes: Cognitively Normal (CN), Alzheimer's disease (AD), Early Mild Cognitive Impairment (EMCI), Late Mild Cognitive Impairment (LMCI), and Significant Memory Concern (SMC). Based on these classes, we applied data from 500 patients randomly selected (100 patients per class) to evaluate our model. Table 1 describes the statistics of the data line number in this dataset, where each line can contain up to 90 values according to the number of attributes.

Table 1
Dataset statistics.

Classes	AD	CN	EMCI	LMCI	SMC	ALL classes
No. of lines	456	1096	684	900	270	3406

We have eliminated some attributes that have no medical significance from this set of attributes and can bias the results, such as the patient ID. Only 87 attributes remain after this elimination. The DRRD process directly triggers the transformation operation of date-time and Boolean data, and each line in this step corresponds to a single observation (i.e., a single value). Table 2 summarizes the result statistics after this transformation.

When the two treatments of the DRR and DRD representations terminate, the DRRD process generates the final SDRRD symbolic representation by assembling the results. The symbolic table SDRRD represents the 500 patients in 1393 columns.

The second task's direct result is a symmetric matrix DMP composed of five hundred rows and five hundred columns.

Fig. 4 shows the 3D display of the four-dimensionality reduction methods that we have considered. To show more detail about the distribution of all categories, we have colored each category. The AD, CN, EMCI, LMCI, and SMC classes are colored red, cyan, green, magenta, and orange, respectively.

For the classification validation, we perform cross-validation ten times. Subsequently, we use the last test results for the reduction visualization of new patients as a demonstration. First letter abbreviations of class names are used to avoid color overlap. The test examples of the Alzheimer's disease, CN, EMCI, LMCI, and SMC classes are presented by A, C, E, L, and S, respectively. Fig. 5 visualizes these results.

To evaluate our classification model on the data without reduction and with all the reduction techniques, we calculate the F-measure (FM) by Equation (9):

$$\mathrm{FM_T} = \frac{2\mathrm{TP_T}}{2\mathrm{TP_T} + \mathrm{FP_T} + \mathrm{FN_T}}.\ \mathrm{T} \in \left\{\mathrm{NB},\ \mathrm{SVM},\ \mathrm{3NN},\ \mathrm{RF}.\right. \tag{9}$$

where $\mathrm{TP_T}$ are all test patients of the category considered positive that were classified as positive during the T technique assessment. The $\mathrm{FP_T}$ are all the test patients of the category considered negative that were classified as positive during the evaluation by the T technique assessment. $\mathrm{FN_T}$ are all test patients of the category considered positive that were classified as negative during evaluation by the T technique assessment.

Table 3 shows the classification evaluation results for all categories. The classifications on the data reduced by the TSNE technique have 19 better cases than the PCA (0 cases), KPCA (1 case), and MDS (0 cases) techniques. Subsequently, and against the 9 cases for the classification without LDMP reduction, we have 11 cases for the classification with TSNE reduction.

Globally, for the five categories AD, CN, EMCI, LMCI, and SMC, and corresponding to each classifier NB, SVM, 3-NN, and RF, we calculate the global FM by Equation (10).

$$\mathrm{FM_T} = \underset{c\in C}{\mathrm{AVG}}\left(\mathrm{FM}_{\mathrm{T}}^{\mathrm{c}}\right) \quad \text{where}$$
$$\mathrm{C} = \left\{\mathrm{AD},\ \mathrm{CN},\ \mathrm{EMCI},\ \mathrm{LMCI},\ \mathrm{SMC}\right\} \wedge \mathrm{T} \in \left\{\mathrm{NB},\ \mathrm{SVM},\ \mathrm{3NN},\ \mathrm{RF}.\right. \tag{10}$$

where $\mathrm{FM}_{\mathrm{T}}^{\mathrm{c}}$ is the FM evaluation of classifier T corresponding to class c.

Table 4 includes these results organized according to the cases with

Table 2
Observation statistics after the transformation step.

No. of patients	No. Of Numerical observations	No. of Nominal observation	All observations
500	75,428	12,188	87,616

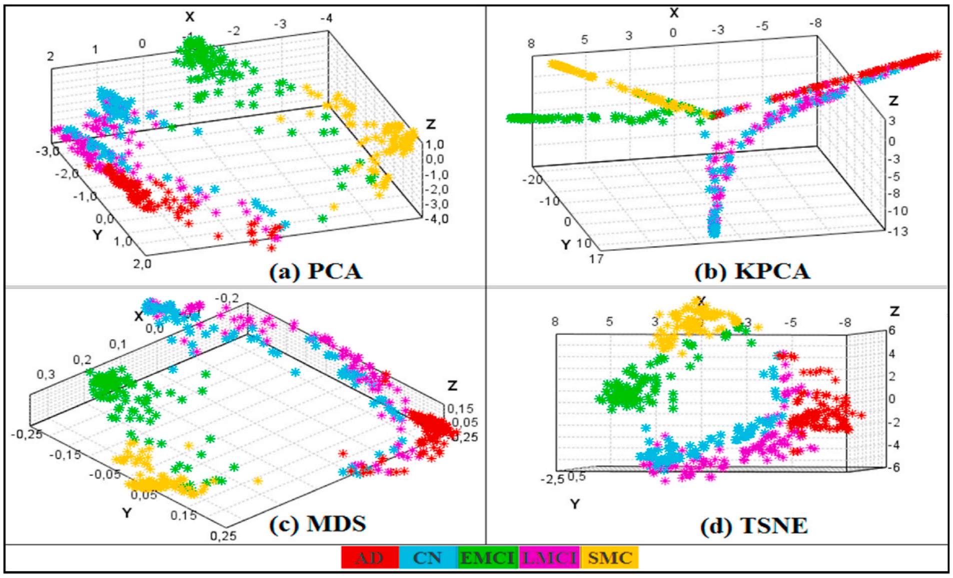

Fig. 4. 3D reductions visualization. (a) PCA, (b) KPCA, (c) MDS, and (d) TSNE.

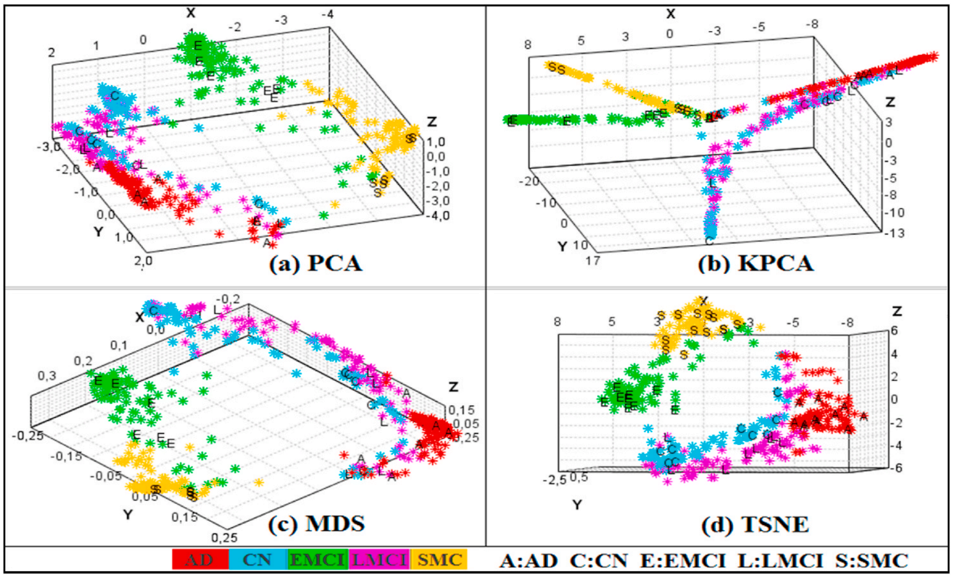

Fig. 5. 3D reduction visualization of the last test fold. (a) PCA, (b) KPCA, (c) MDS, and (d) TSNE.

and without reduction.

For the data reduction results in the last table (Table 4), the TSNE technique's classification generated the four best averages. The technique generates 3 better F-measure averages with the classifiers NB (FM = 0.801), 3-NN (FM = 0.917) and RF (FM = 0.923) than the evaluation without reduction, which returned only one case with the SVM classifier (FM = 0.987).

To study the chosen distance impact on our model result, we try to test another distance instead of the Jaccard distance (J. dist). We will repeat the entire evaluation of our model by calculating the Hamming distance [32] (H. dist) and Levenshtein distance [33] (L. dist) between the patients during the second task. Table 5 shows the FM results of this evaluation using data without reduction LRDMP and with reduction LRTSNE.

For the statistics on cases without reduction, the classifications based on the Jaccard distance yielded better results (3 cases) than the Hamming distance (1 case) and the Levenshtein distance (0 cases). The SVM technique with the Jaccard distance always returns the best performance (FM = 0.987).

As previously mentioned, dimensionality reduction is utilized to minimize computation times to direct our evaluation of the classification time study. For the case comparison with reduction and without reduction, we evaluate only the elapsed time by the classifications applied to the matrices labeled LDMP and LRTSNE. For each technique, this computation time is calculated by the average classification time of the five categories. Table 6 displays the time percentages of the classifications with reduction versus without reduction.

The SVM classifier returned the classifications' maximum performance on the data without reduction (FM = 0.987) in 118 ms. On the other hand, the two classifiers 3-NN and RF achieve excellent performances on TSNE reduction (FM> = 0.917) with the superiority of RF but with a short time and superiority of 3-NN (Time. 3-NN: 2 ms, RF: 76 ms).

Table 3
F-measure results for all classes.

Class	Classifier	FM				
		Without reduction	With reduction			
		LDMP	LRPCA	LRKPCA	LRMDS	LRTSNE
AD	**NB**	0.814	0.82	0.702	0.814	**0.841**
	SVM	**0.985**	0.822	0.711	0.82	**0.823**
	3-NN	0.908	0.849	0.792	0.884	**0.913**
	RF	0.919	0.905	0.839	0.906	**0.937**
CN	**NB**	0.694	0.578	0.59	0.579	**0.728**
	SVM	**0.983**	0.633	0.577	0.615	**0.715**
	3-NN	0.882	0.724	0.545	0.763	**0.894**
	RF	0.869	0.737	0.636	0.854	**0.886**
EMCI	**NB**	**0.917**	0.905	0.877	0.916	**0.914**
	SVM	**1**	0.915	0.901	0.921	**0.931**
	3-NN	**0.976**	0.923	0.873	0.96	**0.974**
	RF	0.944	0.93	0.894	0.956	**0.979**
LMCI	**NB**	**0.633**	0.524	0.461	0.504	**0.603**
	SVM	**0.968**	0.543	0.466	0.533	**0.604**
	3-NN	0.813	0.599	0.437	0.659	**0.819**
	RF	0.809	0.675	0.541	0.792	**0.827**
SMC	**NB**	**0.934**	0.927	**0.936**	0.923	0.921
	SVM	**1**	0.939	0.928	0.936	**0.944**
	3-NN	0.986	0.934	0.883	0.96	**0.987**
	RF	0.961	0.941	0.913	0.956	**0.987**
No. best classifications	**Reductions cases**	/	0	1	0	**19**
	LDMP vs. LRTSNE	9	/	/	/	**11**

Table 4
F-measure global results.

AVG all Class		FM				
		Without reduction	With reduction			
		LDMP	LRPCA	LRKPCA	LRMDS	LRTSNE
NB		0.798	0.751	0.713	0.747	**0.801**
SVM		**0.987**	0.77	0.717	0.765	**0.803**
3-NN		0.913	0.806	0.706	0.845	**0.917**
RF		0.9	0.838	0.765	0.893	**0.923**
No. of best classifications	**Reduction cases**	/	0	0	0	**4**
	LDMP vs. LRTSNE	1	/	/	/	**3**

Table 5
Classification evaluation comparison according to the chosen distances.

AVG all Class	FM					
	LDMP based on			LRTSNE based on		
	J. dist	H. dist	L. dist	J. dist	H. dist	L. dist
NB	**0.798**	0.77	0.765	**0.801**	0.793	0.779
SVM	**0.987**	0.957	0.955	**0.803**	0.773	0.776
3-NN	**0.913**	0.886	0.89	**0.917**	0.896	0.902
RF	0.9	**0.913**	0.907	**0.923**	0.907	0.906
No. of best classifications	**3**	1	0	**4**	0	0

Table 6
Percentage of classification time elapsed on the LDMP and LRTSNE matrices in milliseconds for class AD.

Classification techniques	Classification times (Milliseconds)		
	On LDMP (T1)	On LRTSNE (T2)	% (T2/T1)
NB	109	1	0.90
SVM	118	32	27.11
3-NN	53	2	3.77
RF	443	76	17.15

5. Discussion and evaluation

5.1. Results evaluation

Automated medical decision-making has to cope with many difficult challenges, as the results are related to public health and individual cases of patients and people. One of the challenges facing health professionals and practitioners is adopting well-performing models. Our contribution generates many results, and their analysis allows us to argue our choices and discuss priorities in the use of different techniques.

The excellent visualized data separation in Figs. 4 and 5 expresses the successful choice of the applied data representation model and excellent data quality.

Table 3 shows that the TSNE technique's classification occupies almost all the results and statistically shows complete dominance. On the other hand, the SVM classifier has shown impressive results and almost perfect performance for data without reduction. Simultaneously, the RF technique collected all the best classification performances on the data reduced by TSNE, except with the CN category.

For the global evaluation by the average, Table 4 confirms the excellent classification evaluation of the data reduced by the TSNE technique for the global evaluation by the average.

Tables 3 and 4 show that the relevant variables exploited for the representation in a 3D space by the TSNE technique are more meaningful and preferable for use by classifiers, especially with 3-NN and RF and techniques that adopt majority voting overall. The full set of LDMP matrix variables remained more useful for transformation and separation by margin maximization with the SVM classifier. These results justify our SVM classifier choice for a classification scenario without reduction and the choice of the TSNE technique and the 3-NN or RF classifier for the classification scenario with a reduction.

Subsequently, the performances obtained according to the tests of the three distance measurements in Table 5 clearly show that the Jaccard distance is the most efficient with the data reduction by the TSNE technique. For better readability, Fig. 6 presents the FM curves as a function of the three distances following the LRTSNE technique.

Fig. 6 concludes that the Jaccard distance is the best and most suitable and that the 3-NN and RF classifiers are the best performers with the reduced data.

The analysis of Table 6 shows that the elapsed time for classification on the LRTSNE matrix is short and sometimes negligible compared to the case without LDMP reduction, which is consistent with the benefit of data reduction and our model's overall requirements.

From this analysis and based on the usage needs, we define three application scenarios for our model. Data representation is a common task between them. Simultaneously, the first scenario is involved if the time factor is significant, which invokes the 3-NN classifier. However, if the time is less critical, we use the RF classifier in a second scenario. Of course, the chosen classifier will be applied to the data matrix reduced by the TSNE technique for these two scenarios. Conversely, if the time factor does not have any importance against FM performance, the third scenario applies the SVM classifier on the data without reduction. According to the Jaccard distance, the distance matrix production among the patients is another common task for these three scenarios.

We have attained the challenge of choosing the most suitable representation model to maximize the processed data set and minimize data and information loss. We have successfully passed the challenge of choosing the similarity distance, reduction plan and classifiers to apply. The three selected use scenarios confirm our success against the problem of choosing the best processing series.

For Alzheimer's disease described by the chosen dataset, the resulting performances demonstrate the importance of these diagnoses for judging patients' medical statuses. These diagnoses may include elements that are weakly associated with this disease, but the resulting performance indicates that some of them are strongly correlated to the different categories if it is not the complete set.

5.2. Comparison

To judge our model and discuss its characteristics compared to current research, we have examined previous works [34–37] that address the same classification problem based on patient and participant data. Globally, Table 7 summarizes the four targeted approaches and our contribution. The points adopted in this table are the characteristics considered almost typical among all the state-of-the-art works.

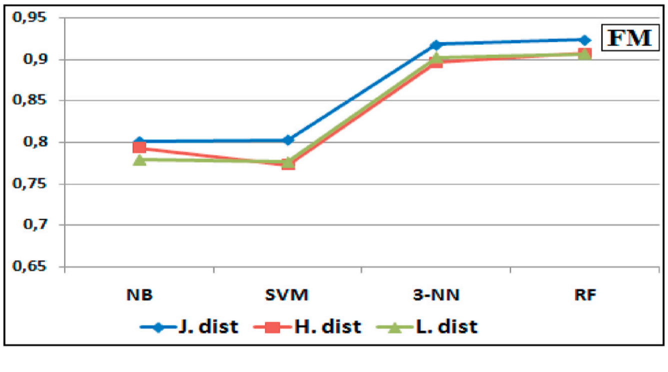

Fig. 6. FM comparison of classifications on LRTSNE data according to the distances chosen.

Joloudari et al. [34] proposed a work to predict the patients' state and the possibility of their liver disease suffering. Their process utilized data for 583 patients collected from 3 data sources, described 14 attributes and tested five classifiers for comparison. With an attribute selection strategy, the particle swarm optimization (PSO)-SVM is the best classifier that achieves the performance FM = 0.958. During processing, missing values, numeric and nominal data types are all treated, but Boolean, date, and time-series data are not treated.

Terrada et al. [35] proposed an automatic process to boost atherosclerosis diagnosis. Data from 835 patients were employed, including 29 attributes, and seven classifiers were tested for preference. This model applied an attribute selection strategy, and an ANN classifier generated a maximum performance of FM = 0.98. Except for numeric, nominal, and Boolean data, this process does not operate on the date data type, time series, and missing values.

The approach of Carvalho et al. [36] is a dynamic decision model that is based on supervised learning. For the experimentation, a data set of 319 patients was employed, but this approach does not specify the adopted number of attributes for evaluating the eight tested classifiers. The best performance obtained was FM = 0.95 by the A1DE classifier within 145 min. In addition to disregarding time series, Boolean and date data are not considered in this approach.

For 349 concerned patients, Lu et al. [37] classified cancer patients into two categories: ovarian cancer and benign ovarian tumor. For the relevant attribute selection among the 49 classifiers employed, the results of three classifiers are compared. The logistic regression classifier (log reg) returned the best performance FM = 0.97. This model processes only missing values and numeric and nominal data and disregards Boolean, date, and time-series data.

Compared to our contribution, we utilized the richest dataset in terms of attribute number (87 attributes) without taking into account the approach [36], which does not include this detail. Our model classifies patient data into five distinct categories and tests four classifiers against other approaches. The TSNE technique adopted for data reduction allowed us to minimize the calculation time up to 32 ms. Although we did not use the same dataset as the other approaches, our dataset has a special quality (considered types and structures), but the excellent performance obtained indicates that our proposal succeeds in terms of selecting the most appropriate processing series. The complete variety of data types processed, the consideration of time series and missing values, and processing performance and elapsed time of treatment are all factors that highlight the exceptional characteristics of our model and its advantages compared to other works. This modular rating allows us to consider our model to be more reliable and more effective.

6. Conclusion

Automatic medical decision-making based on a personalized medicine dataset is the main purpose of this work. Our proposal applies a treatment series to achieve this goal. Applied tasks process structured data that have several types. They also consider time series and missing values. We applied a current data representation model named DRRD. We tested the following three distance measures to calculate the patients' similarity: Jaccard distance, Hamming distance, and Levenshtein distance. Subsequently, we tested the following four reduction techniques: PCA, KPCA, MDS, and TSNE. We examined four classifiers—NB, SVM, 3-NN, and RF—to categorize patients. By experimentation on an Alzheimer's disease dataset, we defined three use scenarios that were formed according to both performance requirements and processing time. Our model evaluation reached FM = 0.987 according to the third scenario and 2 ms of the first scenario's elapsed time. By considering a compromise between performance and computation time (FM = 0.923, time = 76 ms), we recommend the second scenario of the following tasks: SDRRD, Jaccard distance, TSNE, and RF. Compared to other works, our proposal satisfied more characteristic factors and

Table 7
Proposed approach versus current research.

Approach		[34]	[35]	[36]	[37]	Our Model
Patients		583	835	319	349	500
Attributes		14	29	Undefined	49	87
Classifiers		5	7	8	3	4
Best Classifier		PSO-SVM	ANN	A1DE	Log Reg	SVM
Best FM		0,958	0,98	0,95	0,97	0,987
Output categories		2 Liver disease: Yes/No	2 Atherosclerosis: Yes/No	3 Diagnosis: D/AD/MCI	2 Ovarian Cancer/ Benign Ovarian Tumors	5 AD/CN/EMCI/LMCI/SMC
Strategy		Feature selection	Feature selection	Not applicable	Feature selection	Dimension reduction
Elapsed time		Undefined	Undefined	145 min	Undefined	32 millisec
Time series		No	No	No	No	Yes
Missing values		Yes	No	Yes	Yes	Yes
Data type	**Numeric**	Yes	Yes	Yes	Yes	Yes
	Nominal	Yes	Yes	Yes	Yes	Yes
	Boolean	No	Yes	No	No	Yes
	Date	No	No	No	No	Yes

demonstrated wider potential.

For future works and more than the real practical aspect of this proposal, we aim to endow our model with other modules for the most significant diagnoses and drug prescriptions. This extension aims to build a medical decision-making system to control patients' classification, data analysis and personalized treatments without side effects.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was completed as part of the Hubert Curien Partnership (PHC) TASSILI cooperation program between France and Algeria under the project code 19MDU212.

Data collection and sharing for this project was funded by the ADNI (National Institutes of Health Grant U01 AG024904) and Department of Defense (DOD) ADNI (DOD award number W81XWH-12-2-0012). The ADNI was funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and by generous contributions from the following organizations: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd. and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research provided funds to support ADNI clinical sites in Canada. Private sector contributions were facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization was the Northern California Institute for Research and Education, and the study was coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data were disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

[1] Bernhard S, Alexander S, Klaus-Robert M. Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput 1998;10(5):1299–319.

[2] Shukla AK, Singh P, Vardhan M. Medical diagnosis of Parkinson disease driven by multiple preprocessing technique with scarce lee silverman voice treatment data. Engineering vibration, communication and information processing. Lecture Notes in Electrical Engineering 2019;478:407–21.

[3] Raymond WD, Eilertsen GØ, Nossent J. Principal component analysis reveals disconnect between regulatory cytokines and disease activity in. Systemic Lupus Erythematosus. Cytokine. 2019;114:67–73.

[4] Lu H, Meng Y, Yan K, Gao Z. Kernel principal component analysis combining rotation forest method for linearly inseparable data. Cognit Syst Res 2019;53: 111–22.

[5] Jiang JL, Li SY, Liao ML, Jiang Y. Application in disease classification based on KPCA-IBA-LSSVM 2019;154:109–16.

[6] Li T, Yin Q, Song R, Gao M, Chen Y. Multidimensional scaling method for prediction of lysine glycation sites. Computing 2019;101:705–24.

[7] Vital TP, Kumer KD, Sri HVB, Krishna MM. Analysis of cancer data set with statistical and unsupervised machine learning methods. Smart intelligent computing and applications. Smart Innovation, Systems and Technologies 2019; 104:267–76.

[8] Maaten LVD, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res 2008;9: 2579–605.

[9] Zhang X, Chou J, Liang J, Xiao C, Zhao Y, Sarva H, Henchcliffe C, Wang F. Data-driven subtyping of Parkinson's disease using longitudinal clinical Records: a cohort study. Sci Rep 2019;9:797.

[10] Workman TE, Hirezi M, Trujillo-Rivera E, Patel AK, Heneghan JA, Bost JE, Zeng-Treitler Q, Pollack M. A novel deep learning pipeline to analyze temporal clinical data. In: 2018 IEEE international conference on big data (big data). Seattle, WA, USA; 2018. p. 2879–83.

[11] Mohammad SA, Yin KC, Kasturi DV. Identification of significant features and data mining techniques in predicting heart disease. Telematics Inf 2019;36:82–93.

[12] Ayatollahi H, Gholamhosseini L, Salehi M. Predicting coronary artery disease: a comparison between two data mining algorithms. BMC Publ Health 2019;19:448.

[13] Gultepe Y, Rashed S. The use of data mining techniques in heart disease prediction. Int J Comput Sci Mobile Comput 2019;8(4):136–41.

[14] İlkim EE, Nurdan E, Yusuf İA, Yalçın Ö, Çiğdem E. The analysis of the effects of acute rheumatic fever in childhood on cardiac disease with data mining. Int J Med Inf 2019;123:68–75.

[15] Vital TP, Krishna MM, Narayana GVL, Suneel P, Ramarao P. Empirical analysis on cancer dataset with machine learning algorithms. Soft computing in data analytics. Advances in Intelligent Systems and Computing 2019;758:789–801.

[16] Kadi H, Rebbah M, Meftah B, Lezoray O. A data representation model for personalized medicine. Int J Healthc Inf Syst Inf. (in press).

[17] Shaeela A, Muhammad KH, Ramzan T. Overview and comparative study of dimensionality reduction techniques for high dimensional data. Inf Fusion 2020; 59:44–58.

[18] Pak I, Teh PL. Machine learning classifiers: evaluation of the performance in online reviews. Indian Journal of Science and Technology 2016;9(45).

[19] Babar AH, Mahoto NA. Comparative analysis of classification models for Healthcare data analysis. Int J Comput Inf Technol 2018;7(4):170–5.

[20] Sudhir MG, Ankit D, Preetesh P. A study of some data mining classification techniques. International Research Journal of Engineering and Technology 2017;4 (4):3112–5.

[21] Paul Y, Kumar N. A comparative study of famous classification techniques and data mining tools. In: Proceedings of ICRIC 2019. Lecture notes in electrical engineering. Cham, Switzerland: Springer; 2020. p. 627–44.

[22] Aggarwal G, Vig R. Acoustic methodologies for classifying gender and emotions using machine learning algorithms. In: Amity international conference on artificial intelligence. Dubai: United Arab Emirates; 2019. p. 672–7.

[23] Salmi N, Rustam Z. Naïve Bayes classifier models for predicting the colon cancer. IOP Conf Ser Mater Sci Eng 2019;546(5).

[24] Noble WS. What is a support vector machine? Nat Biotechnol 2006;24:1565–7.

[25] Al Bataineh A. A comparative analysis of nonlinear machine learning algorithms for breast cancer detection. International Journal of Machine Learning and Computing 2019;9:248–54.

[26] Sarkar M, Leong TY. Application of K-nearest neighbors algorithm on breast cancer diagnosis problem. In: Proceedings/AMIA … Annual symposium. AMIA Symposium; 2000. p. 759–63.

[27] Ishwaran H, Lu M. Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. Stat Med 2018;38(4):558–82.

[28] Kavzoglu T. Object-oriented random forest for high resolution land cover mapping using quickbird-2 imagery. In: Pijush S, Sanjiban SR, Valentina EB, editors. Handbook of neural computation. London, UK: Academic Press; 2017. p. 607–19.

[29] Bengio Y, Paiement JF, Vincent P, Delalleau O, Le Roux N, Ouimet M. Out-of-Sample extensions for LLE, isomap, MDS, eigenmaps, and spectral clustering. In: Proceedings of the 16th international conference on neural information and processing systems; MA, United States; 2003. p. 177–84.

[30] Gisbrecht A, Schulz A, Hammer B. Parametric nonlinear dimensionality reduction using kernel t-SNE. Neurocomputing 2015;147:71–82.

[31] ADNI [Internet]. Alzheimer's Disease Neuroimaging Initiative. [accessed 2019]. Available from:: http://adni.loni.usc.edu/.

[32] Yang H, Wang Y. A LBP-based face recognition method with hamming distance constraint. In: Proceedings of fourth international conference on image and graphics (ICIG 2007); sichuan; 2007. p. 645–9.

[33] Krishna NS, Behara AB, Edward C. A novel approach for the structural comparison of origin-destination matrices: Levenshtein distance. Transport Res C Emerg Technol 2020;111:513–30.

[34] Joloudari JH, Saadatfar H, Dehzangi A, Shamshirband S. Computer-aided decision-making for predicting liver disease using PSO-based optimized SVM with feature selection. Informatics in Medicine Unlocked 2019;17.

[35] Terrada O, Cherradi B, Raihani A, Bouattane O. A novel medical diagnosis support system for predicting patients with atherosclerosis diseases. Informatics in Medicine Unlocked 2020;21.

[36] Carvalho CM, Seixas FL, Conci A, Débora C, Muchaluat-Saade, Laks J, Boechat Y. A dynamic decision model for diagnosis of dementia, Alzheimer's disease and Mild Cognitive Impairment. Comput Biol Med 2020;126.

[37] Lu M, Fan Z, Xu B, Chen L, Zheng X, Li J, Znati T, Mi Q, Jiang J. Using machine learning to predict ovarian cancer. Int J Med Inf 2020:141.

[38] Tariq I, AlSattar HA, Zaidan AA, Zaidan BB, Abu Bakar MR, Mohammed RT, Albahri OS, Alsalem MA, Albahri AS. MOGSABAT: a metaheuristic hybrid algorithm for solving multi-objective optimisation problems. Neural Comput Appl 2020;32:3101–15.

[39] Mohammed KI, Jaafar J, Zaidan AA, Albahri OS, Zaidan BB, Abdulkareem KH, Jasim AN, Shareef AH, Baqer MJ, Albahri AS, Alsalem MA, Alamoodi AH. A uniform intelligent prioritisation for solving diverse and big data generated from multiple chronic diseases patients based on hybrid decision-making and voting method. IEEE 2020;8:91521–30.

[40] Mohammed KI, Zaidan AA, Zaidan BB, Albahri OS, Albahri AS, Alsalem MA, Mohsin AH. Novel technique for reorganisation of opinion order to interval levels for solving several instances representing prioritisation in patients with multiple chronic diseases. Comput Methods Progr Biomed 2020:185.

[41] Almahdi EM, Zaidan AA, Zaidan BB, Alsalem MA, Albahri OS, Albahri AS. Mobile-based patient monitoring systems: a prioritisation framework using multi-criteria decision-making techniques. J Med Syst 2019;43:219.

[42] Albahri AS, Alwan JK, Taha ZK, Ismail SF, Hamid RA, Zaidan AA, Albahri OS, Zaidan BB, Alamoodi AH, Alsalem MA. IoT-based telemedicine for disease prevention and health promotion: state-of-the-Art. J Netw Comput Appl 2021:173.