




ORIGINAL

Advanced Ensemble Machine Learning Techniques for Optimizing Diabetes Mellitus Prognostication: a Detailed Examination of Hospital Data

Técnicas Avanzadas de Aprendizaje Automático para Optimizar el Pronóstico de la Diabetes Mellitus: un Examen Detallado de los Datos Hospitalarios

Najah Al-shanableh¹  , Mazen Alzyoud¹  , Raya Yousef Al-husban²  , Nail M. Alshanableh³  , Ashraf Al-Oun¹  , Mohammad Subhi Al-Batah⁴  , Mowafaq Salem Alzboono⁴  

¹Computer Science Department, Al al-Bayt University. Mafraq, Jordan.

²Faculty of Nursing, Zarqa University. Zarqa, Jordan.

³Vascular Surgery Unit Jordanian Royal Medical Services.

⁴Department of Computer Science, Faculty of Science and Information Technology, Jadara University. 21110 Irbid, Jordan.

Cite as: Al-shanableh N, Alzyoud M, Al-husban RY, Alshanableh NM, Al-Oun A, Al-Batah MS, et al. Advanced Ensemble Machine Learning Techniques for Optimizing Diabetes Mellitus Prognostication: a Detailed Examination of Hospital Data. Data and Metadata. 2024; 3:363. <https://doi.org/10.56294/dm2024.363>

Submitted: 21-01-2024

Revised: 29-05-2024

Accepted: 09-09-2024

Published: 10-09-2024

Editor: Adrián Alejandro Vitón Castillo 

Corresponding Author: Raya Yousef Al-husban 

ABSTRACT

Diabetes is a chronic disease that affects millions of people worldwide. Early diagnosis and effective management are crucial for reducing its complications. Diabetes is the fourth-highest cause of mortality due to its association with various comorbidities, including heart disease, nerve damage, blood vessel damage, and blindness. The potential of machine learning algorithms in predicting Diabetes and related conditions is significant, and mining diabetes data is an efficient method for extracting new insights.

The primary objective of this study is to develop an enhanced ensemble model to predict Diabetes with improved accuracy by leveraging various machine learning algorithms.

This study tested several popular machine learning algorithms commonly used in diabetes prediction, including Naive Bayes (NB), Generalized Linear Model (GLM), Logistic Regression (LR), Fast Large Margin (FLM), Deep Learning (DL), Decision Tree (DT), Random Forest (RF), Gradient Boosted Trees (GBT), and Support Vector Machine (SVM). The performance of these algorithms was compared, and two different ensemble techniques—stacking and voting—were used to build a more accurate predictive model.

The top three algorithms based on accuracy were Deep Learning, Naive Bayes, and Gradient Boosted Trees. The machine learning algorithms revealed that individuals with Diabetes are significantly affected by the number of chronic conditions they have, as well as their gender and age. The ensemble models, particularly the stacking method, provided higher accuracy than individual algorithms. The stacking ensemble model achieved a slightly better accuracy of 99,94 % compared to 99,34 % for the voting method.

Building an ensemble model significantly increased the accuracy of predicting Diabetes and related conditions. The stacking ensemble model, in particular, demonstrated superior performance, highlighting the importance of combining multiple machine learning approaches to enhance predictive accuracy.

Keywords: Machine Learning Algorithms; Ensemble Models; Diabetes Prediction; Data Mining; Predictive Accuracy; Health Informatics.

RESUMEN

La diabetes es una enfermedad crónica que afecta a millones de personas en todo el mundo. El diagnóstico temprano y el tratamiento eficaz son cruciales para reducir sus complicaciones. La diabetes es la cuarta

causa de mortalidad debido a su asociación con diversas comorbilidades, como enfermedades cardíacas, daño a los nervios, daño a los vasos sanguíneos y ceguera. El potencial de los algoritmos de aprendizaje automático para predecir la diabetes y las afecciones relacionadas es significativo, y la extracción de datos sobre la diabetes es un método eficiente para extraer nuevos conocimientos.

El objetivo principal de este estudio es desarrollar un modelo de conjunto mejorado para predecir la diabetes con mayor precisión aprovechando varios algoritmos de aprendizaje automático.

Este estudio probó varios algoritmos de aprendizaje automático populares que se utilizan comúnmente en la predicción de la diabetes, incluidos Naive Bayes (NB), Generalized Linear Model (GLM), Logistic Regression (LR), Fast Large Margin (FLM), Deep Learning (DL), Decision Tree (DT), Random Forest (RF), Gradient Boosted Trees (GBT) y Support Vector Machine (SVM). Se comparó el rendimiento de estos algoritmos y se utilizaron dos técnicas de conjunto diferentes (apilamiento y votación) para construir un modelo predictivo más preciso. Los tres mejores algoritmos en cuanto a precisión fueron Deep Learning, Naive Bayes y Gradient Boosted Trees. Los algoritmos de aprendizaje automático revelaron que las personas con diabetes se ven significativamente afectadas por la cantidad de enfermedades crónicas que tienen, así como por su género y edad. Los modelos de conjunto, en particular el método de apilamiento, proporcionaron una mayor precisión que los algoritmos individuales. El modelo de conjunto de apilamiento logró una precisión ligeramente mejor del 99,94 % en comparación con el 99,34 % del método de votación.

La construcción de un modelo de conjunto aumentó significativamente la precisión de la predicción de la diabetes y las enfermedades relacionadas. El modelo de conjunto de apilamiento, en particular, demostró un rendimiento superior, lo que destaca la importancia de combinar múltiples enfoques de aprendizaje automático para mejorar la precisión predictiva.

Palabras clave: Algoritmos de Aprendizaje Automático; Modelos de Conjunto; Predicción de Diabetes; Minería de Datos; Precisión Predictiva; Informática Sanitaria.

INTRODUCTION

Recently, Machine Learning has gained a reputation as a powerful technique for data predictions and classifications in many disciplines. More specifically, machine learning is used on medical data sets because it can explore hidden patterns in patients' records. Also, it can efficiently address the nature of the medical data set, which is heterogeneous, widely spread, scattered, and requires organization and integration because of the different ways it was collected.^(2,19) Using mathematical analysis to explore patterns from a data set is not a good choice because the relationships are too complex. Therefore, Machine Learning has become more popular in the healthcare industry in many fields, such as medical research, pharmaceuticals, and hospital management.⁽¹⁵⁾

Machine learning in health care provides information to help physicians make decisions, prescribe medication, and determine when to stop the medication.^(3,15,13) It also allows the healthcare industry to decrease the cost of drugs and manage the healthcare process and customer relationships. In addition, doctors gain experience in treatment methods and increase patient response time.⁽¹⁷⁾ The increasing availability of large datasets and improved computing power have made it possible to build complex models for disease prediction with high accuracy.⁽¹⁵⁾ Machine Learning methods used in the healthcare domain are prediction, classification, clustering, and association. These techniques are essential in supporting disease and detecting fraud.^(5,3)

Successful implementation of Machine Learning depends on the quality of the research data. Significant issues in Machine Learning include noise, uncertainty, and incompleteness of data. Data preparation comprises more than half of every machine-learning process.⁽¹⁸⁾ Data preprocessing is the longest and most challenging part of Machine Learning. Data preparation includes collecting needed data or selecting from previously collected data. It also consists of the integration and transformation of data, followed by data cleaning and reduction, if necessary, as well as data set integration, Selection, and transformation.⁽¹⁸⁾ Data from different sources could come in various formats, and attributes may be named differently. Successful data preprocessing improves the quality of the data and the quality of the Machine Learning process, which leads to more benefits from the data.⁽⁸⁾

The machine learning process consists of five stages, with flexible iteration to previous steps, if necessary. These five stages are as follows:

- Selection of data.
- Data preparation and preprocessing include several operations: data cleaning, integration, reduction, and handling outliers.
- Data transformation.
- Machine Learning algorithm application analyzes the data, identifies patterns, and represents knowledge.

- Interpretation of Machine Learning results and evaluation.

Diabetes mellitus (DM)

A chronic condition caused by high blood sugar levels (glucose) is known as Diabetes Mellitus. This problem occurs when a body cannot produce or consume insulin.⁽¹⁾ Then, factors like dietary control and lifestyle of the individuals are considered the main issues for progressive DM.^(4,21) Also, stroke, heart disease, loss of vision, limb removal, and kidney disease are several health complications caused by DM. To minimize these complications, early diagnosis, and effective healthcare are essential.⁽¹⁴⁾ It will also result in cardiovascular disease, nephropathy, neuropathy, foot problems, retinopathy, skin complications, and dental problems, and these are the severe health issues caused by DM.^(6,7)

Early detection and prompt diagnosis can prevent the further progressive nature of the disease, so early detection of disease and DM risk factors detection is also crucial.^(5,15) Then, by knowing the risk factors, it will be easy for doctors to detect the disease level of the individuals, thereby providing appropriate healthcare for reducing the risks. To detect and handle those with DM disease, regular checkups, and monitoring are also crucial. Maintaining blood sugar levels and a healthy lifestyle are essential to minimize the risks of the condition. Regular blood sugar levels and prompt healthcare monitoring are necessary for this disease.⁽¹⁶⁾

ML techniques effectively detect DM-related factors as they determine the relation between the data and new data. DM-related factors can be analyzed using ML, and hidden data in clinical records can be identified.^(3,19,20)

Machine learning for disease prediction

A wide range of data sources can be analyzed via ML. Thus, it is effective in detecting DM risk factors, and it is advantageous.^(29,30,31) Such data sources contain all details of the patients, including the patient demographics, clinical information, and laboratory test results, especially because they contain lifestyle patterns and eating habits. Thus, ML helps in detecting the most complicated patterns in DM, thereby minimizing the risks and providing the most accurate predictions.^(8,9,19)

The risk level of that progressive disease can be detected via ML. It also supports in proving appropriate diagnosis for individuals at risk.^(10,22,36)

Then the following techniques are also employed for detecting diseases: LR, DT,RF, NB, SVM, and KNN (K-Nearest Neighbor). The above techniques are one of the ML techniques.⁽¹¹⁾ The issue will be addressed by selecting the accessible data type. Then unstructured data (images or text) are better suited for some techniques, and structured data will be better suited for a few techniques.^(23,27)

For predicting disease, the most commonly utilized technique is called LR. Then, for binary classification problems, this LR technique is effective, as it predicts the presence or absence of the disease. It can be employed in various disease predictions and provides superior outcomes, this LR can be easily trained.^(12,25)

To predict a disease, the common ML techniques named DT and RF are used. Utilizing a series of binary decisions for the sample classification, since these techniques are tree-like structures. It will be effective for addressing issues using input features and it is utilized in many clinical applications.^(13,23)

The effective classification technique is SVM that manages both the linear and non-linear boundaries. This technique is one of the ML techniques employed for predicting disease. It is also employed for cardiovascular disease and many medical datasets especially for Diabetes prediction.^(15,19)

KNN is employed in many datasets for predicting disease; it is a simple and effective ML. It depends on the KNN majority classes in the training data for classifying new sample.^(16,35)

For binary classification problems, NB is employed, it is a probabilistic technique. It also employs the Bayes theorem for estimating every class probability; it also assumes that data features are independent. Specially for predicting Breast cancer, and tuberculosis, NB is employed.^(15,16,18,24)

Thus, improved patient outcomes are provided by employing the ML technique. It also has the potential in predicting and diagnosing the disease in an effective manner thereby reducing risk factors of DM.^(32,33) Still, an improvement could be made in each algorithm's performance or simply by applying more advanced techniques for building a more accurate model by using ensemble or hybrid techniques.⁽³³⁾

Ensemble Machine learning models

Ensemble learning is a machine learning technique combining multiple models to produce improved predictions.⁽¹⁷⁾ The idea behind ensemble learning is that by combining the strengths of various models, the resulting ensemble can achieve higher accuracy and stability than any individual model. Some popular ensemble learning methods include Bagging, boosting, stacking, and voting.⁽²⁰⁾

- Bagging (Bootstrapped Aggregation) is an ensemble learning technique in machine learning that aims to improve the stability and accuracy of a model by combining the results of multiple instances of the same base algorithm, each trained on different randomly sampled subsets of the training data. In figure 1 A one can see how the bagging ensemble method works. By taking the average of predictions

from several instances, bagging helps in lowering the model's variance, which in turn cuts down the chance of it fitting too closely to the training data. People often use bagging when they work with decision trees though it's something you can use with any basic method. Machine learning models often use bagging to predict disease by training multiple instances of the same base algorithm on randomly sampled subsets of patient data. Each model can be combined to produce a more precise and robust outcome than any individual instance. The model's variance can be reduced, which can improve disease prediction. However, the success of Bagging for disease prediction will rely heavily on the quality and size of the patient data as well as the choice of the base algorithm.^(23,28)

- Boosting is another technique for machine learning ensemble that improves the accuracy of a model by combining the results of multiple weak learners into a single strong learner. Figure 1 B shows the boosting ensemble technique. Unlike Bagging, boosting focuses on training instances that previous models in the ensemble misclassified. This results in focusing on the most brutal examples, which can lead to improved accuracy. Boosting algorithms typically use a weighted combination of weak learners, where the weighting is based on the accuracy of each model. Some popular boosting algorithms include AdaBoost, Gradient Boosting, and XGBoost.^(23,26)

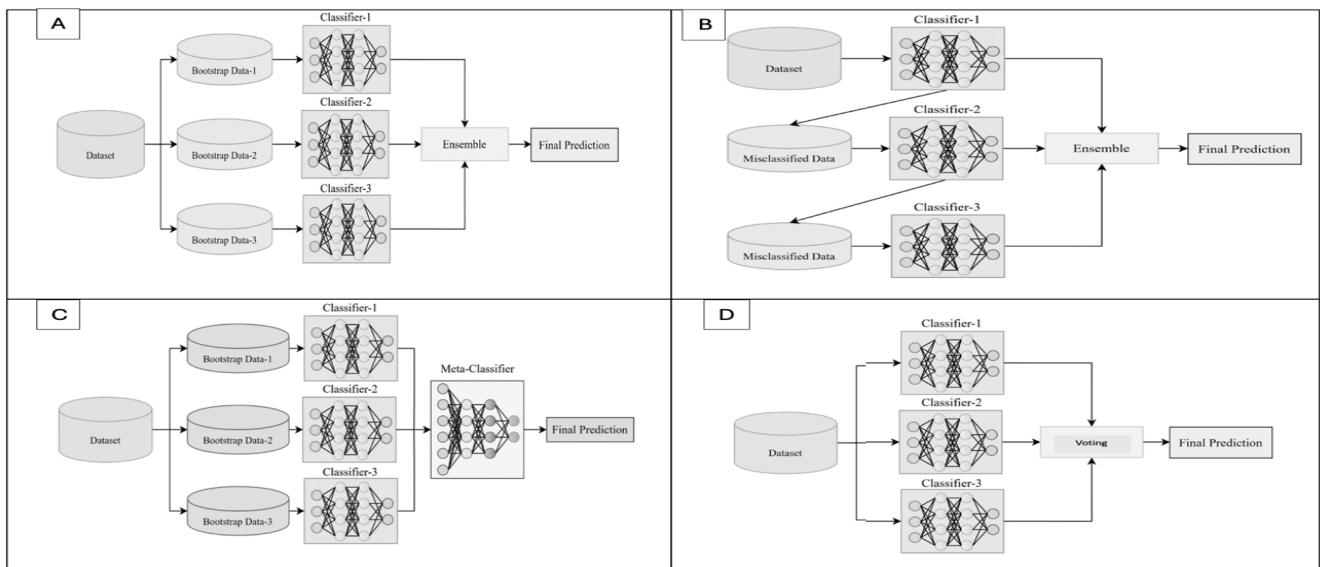


Figure 1. A. Bagging Ensemble Model B. Boosting Ensemble Model C. Stacking Ensemble Model D. Voting Ensemble Model

- Stacking is a machine learning ensemble technique that combines the prediction results of multiple models to improve the overall accuracy of a single prediction. Unlike bagging and boosting, stacking involves training a meta-model to make predictions using several base models' outputs (predictions) as input features. The idea behind stacking is to leverage the strengths of multiple base models to generate a more accurate forecast. Figure 1 C shows the stacking technique. The base models can be any machine learning algorithm trained on the same or different datasets. The key to stacking is carefully designing the meta-model to leverage the base models' predictions effectively.^(23,29)

- Voting: voting, on the other hand, is a broader concept in ensemble learning where multiple models (classifiers or regressors) are trained independently, and the final prediction is made based on a combination of the predictions of these individual models. Figure 1 D shows the stacking technique, which is similar to bagging, but the same data is used for each classifier and heterogeneous models are used. There are different types of voting, such as hard voting and soft voting. In the process known as hard voting, the choice that gets the most support from the models wins out in the end; on the other side, soft voting takes a different approach by looking at what all the models predict on average. Even though voting is often seen alongside bagging, like when they team up to form a voting classifier, it's not just limited to that; it also plays a role in other ways of combining models, such as boosting. Therefore, while voting falls under the umbrella of ensemble learning, it does not precisely fit into the category of a bagging ensemble model.^(23,31)

Recent studies have shown that using a mix of ML models together works well for guessing diseases. They show that putting different algorithms together can make these guesses more accurate.^(37,38)

For example, Wu et al.⁽³⁹⁾ developed an ensemble model for the prediction of breast cancer using gene expression data. They combined multiple ML algorithms to create a more accurate prediction model, including DT, RF, and SVM.

In their study, Salehi et al.⁽⁴⁰⁾ introduced a new mixed method that uses machine learning to guess if someone might get diabetes. Salehi et al.⁽⁴⁰⁾ applied a method that brings together extreme gradient boosting SVM and RF to see how likely it is for someone to develop Diabetes.

Huang et al.⁽²²⁾ applied a combined deep learning model to diagnose COVID-19 by looking at chest CT scans. They combined multiple DCNN to improve the accuracy of COVID-19 diagnosis.

METHOD

In this work, we utilized supervised learning to classify diabetes presence using various machine learning algorithms. It is a binary classification problem where we have to differentiate between diabetic and non-diabetic subjects with the set of input features. This retrospective study is based on previously collected data in hospital discharge files. The comprehensive dataset of people with and without diabetes obtained from the Healthcare Cost and Utilization Project (HCUP) was used. The dataset has many demographic, clinical, and lifestyle variables.

The main goal of this study is creating an ensemble model from other standard models for the purpose of predicting disease. Figure 2 shows the flowchart of the methodology. It includes 2 phases: one is various ML techniques training, and the other one is to construct an ensemble model based on phase 1 results.

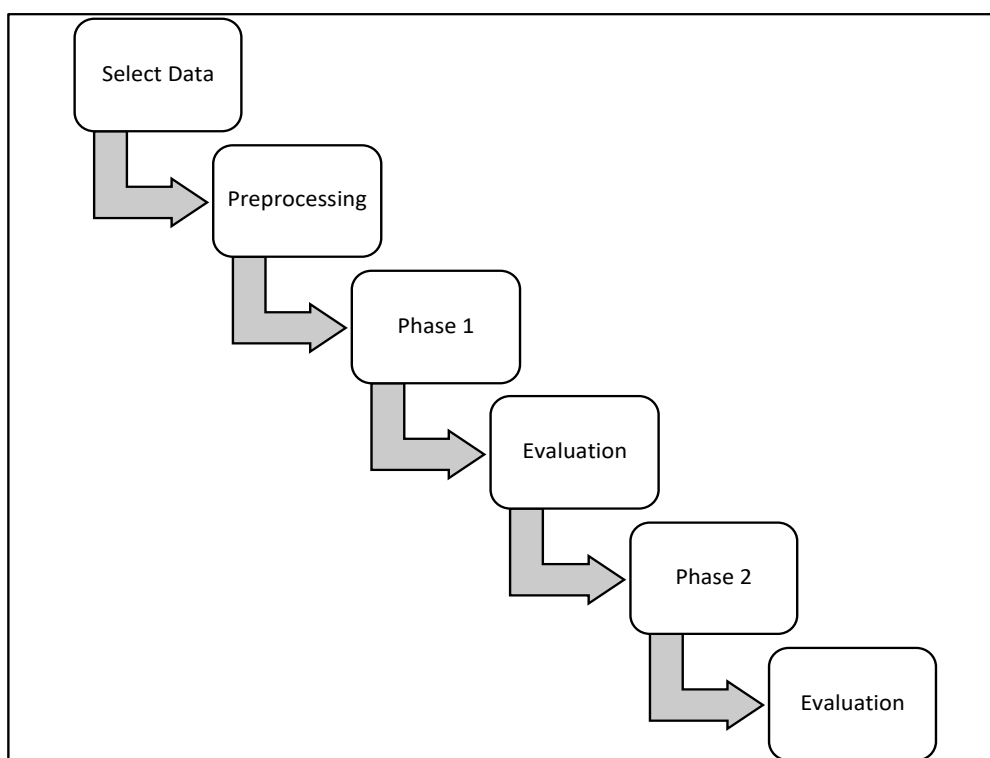


Figure 2. Methodology Flowchart

As shown in figure 2, the methodology we followed for applying Machine Learning includes the following steps:

Selecting the target data

We selected the data set relevant to the study that contained the necessary variables appropriate for applying machine-learning algorithms from the Healthcare Cost and Utilization Project (HCUP).⁽¹⁴⁾ The Healthcare Cost and Utilization Project (HCUP) is a family of databases and related tools and software developed through a federal-state-industry partnership and sponsored by the Agency for Healthcare Research and Quality (AHRQ). One of the key components of HCUP is the State Inpatient Databases (SID), which contains data on inpatient hospital stays from participating states. The SID database includes data on patient demographics, primary and secondary diagnoses, procedures performed, and hospital characteristics such as bed size, ownership, and teaching status. The database is designed to be used for a variety of purposes, including healthcare research, policy development, and healthcare quality improvement.^(14,15) We used code from the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM), and Clinical Classifications Software (CCS) to identify diabetic patients from HCUP records. The ICD-9-CM system maps each diagnosis to a numeric code. The following ICD-9-CM codes were used:

- diabetes mellitus without complications [49.]
- diabetes mellitus with complications [50.]

The types of variables in the HCUP SID dataset can be broadly categorized into the following:

- Demographic Variables: patient Age, Gender, and Race/Ethnicity.
- Clinical Variables: primary Diagnosis Codes, Secondary Diagnosis Codes, Procedure Codes, and Severity of Illness, The number of Chronic Conditions, and number of procedures.
- Hospital-Level Variables: hospital Identifier, Hospital Location, and Hospital Type.
- Admission and Discharge Information: admission Source, Admission Type, and Length of Stay.
- Payment and Cost Variables: payer Information, and Charges and Costs

The dataset used in this study contained several thousand records, nearly equally distributed between cases of diabetes and non-diabetes. The sample size may be considered large enough to train, validate, and apply machine learning models with the required statistical power to detect significant patterns and relations within the data.

Preprocessing

We performed a data cleaning procedure, which included removing missing values and noise outliers. To understand the data and know how to handle it, we applied a preliminary analysis after performing several data preparation steps to clean the data. We programmed and applied these steps in RapidMiner software.

The data is unbalanced because the majority of inpatients did not have Diabetes. To compensate for the unbalanced class and overcome this issue, we used several classifiers on a balanced sample. We trained and tested these classifiers on the chosen balanced sample. The sample size was 500 000 records, with 50 % of patients having Diabetes as one of their diagnoses.

Phase 1

The algorithms were trained and tested using the RapidMiner tool, and their performance was evaluated using different metrics such as accuracy, precision, recall, and F1 score.

The selected list of trained algorithms was based on previous research. These algorithms are:

- Naive Bayes (NB)
- Generalized Linear Model (GLM)
- Logistic Regression (LR)
- Fast Large Margin (FLM)
- Deep Learning (DL)
- Decision Tree (DT)
- Random Forest (RF)
- Gradient Boosted Trees (GBT)
- Support Vector Machine (SVM)

Evaluation

10- The k-fold cross-validation method is used to train and test the models. Also, for evaluation purposes, more standard measures were used: accuracy, sensitivity, specificity, precision, recall, and f1-score (F1).

For diabetic prediction, which is a binary classification task, confusion matrix was used for evaluation. Where:

- True Positive (TP) refers to a sample which is from the positive class, is correctly classified by the classification model.
- False Positive (FP) refers to a sample that is from a negative class, being incorrectly classified as belonging to the positive class.
- True Negative (TN) refers to a sample that is from the negative class, is correctly classified by the classification model.
- False Negative (FN) refers to a sample that is from the positive class, being incorrectly classified as a negative class by the model.

The evaluation metrics used based on the above confusion matrix definitions are:

- Sensitivity: ability to select what needs to be selected ($TP/(TP + FN)$)
- Specificity: ability to reject what needs to be rejected ($TN/(TN + FP)$)
- Precision: proportion of cases found that were relevant ($TP/(TP+FP)$)
- Recall: proportion of all relevant cases that were found ($TP/(TP+FN)$)
- Accuracy: aggregate measure of classifier performance ($(TP+TN)/(TP+TN+FP+FN)$)

Phase 2

In this stage, we used Stacking to build an ensemble meta-classifier to predict diabetics. Figure 3 shows how the decision tree was used as a meta-classifier after training Naive Bayes, Deep Learning, and Gradient Boosted Trees.

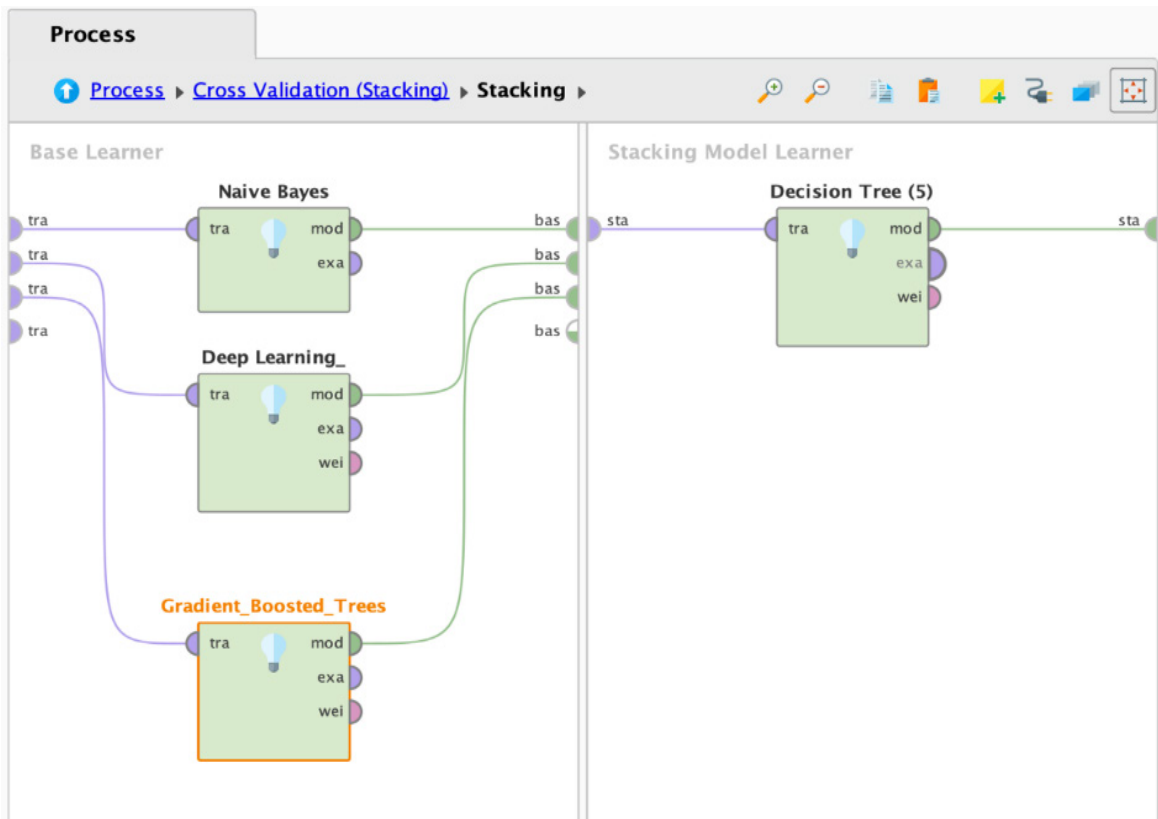


Figure 3. RapidMiner process for stacking

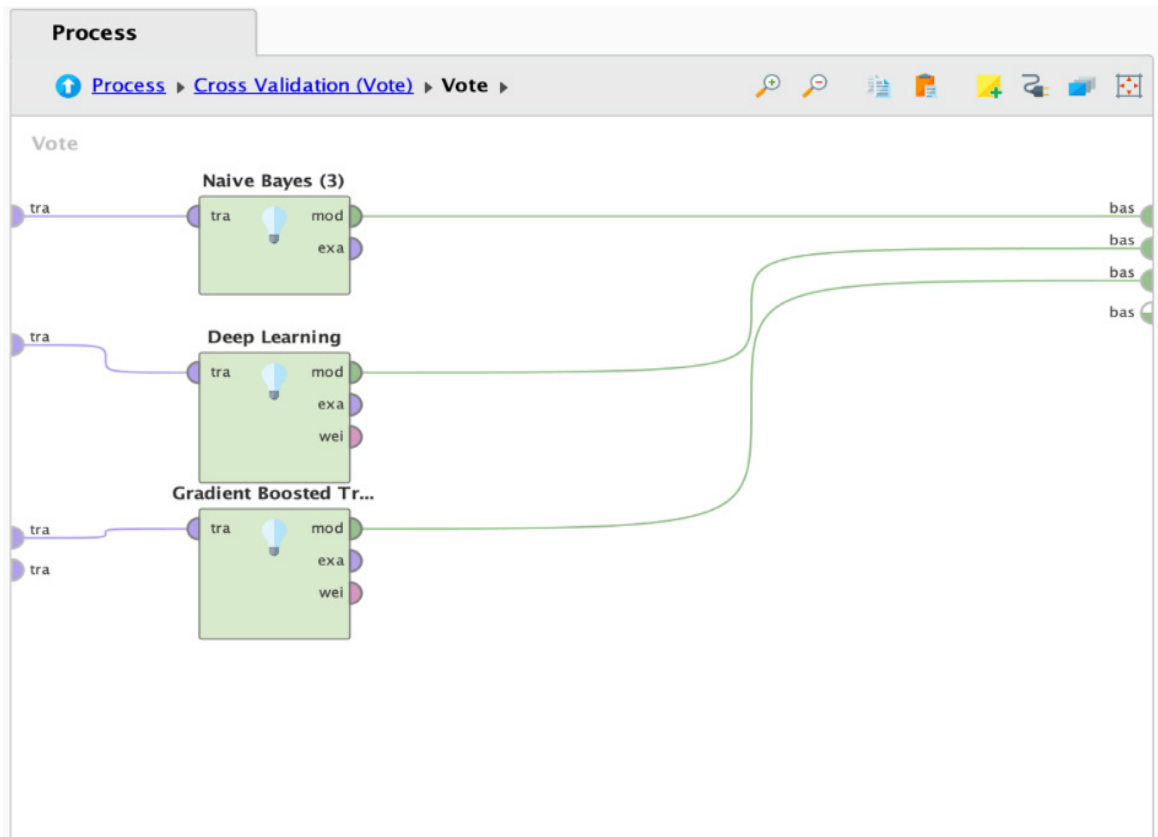


Figure 4. RapidMiner process for Voting ensemble model

We also used voting (shown in figure 4) to compare the effect of different ensemble techniques compared to each other on the same dataset. In voting, several algorithms can be combined together to create the ensemble model. For classification purposes, the majority vote of all classifiers is given as the prediction, and for regression, the average of all classifiers is given as the prediction as in the bagging algorithm. The RapidMiner is shown in figure 4. Inside the “Cross Validation” operator, we use the “Vote” operator. In the training phase, we used Naive Bayes, Deep Learning, and Gradient Boosted Trees.

Evaluation

Comparing obtained results from the stacking and voting model to other ensemble methods and individual algorithm performance.

The methodology of this study was done in a way that the developed classification models would not only be proficient but also generalizable. This study, therefore, adds great value to literature on machine learning applications in health informatics by ensuring that the sample used is representative, considering those variables relevant to it, and the rigorous evaluation in respect to the models.

RESULTS

The goal of this research is to identify the factors related to diabetes mellitus patients’ hospitalizations and complications using ensemble Machine Learning algorithms.

For this study, we define people with Diabetes by the existence of ICD-9-CM code 49 or 50 among patients’ diagnoses codes. We used a balanced sample of patients aged 18 or older. The sample size was 500 000, where 50 % were diabetic patients. Table 1 shows sample characteristics.

We used the RapidMiner tool to run KDD steps and apply the most common Machine Learning algorithms in literature (Naive Bayes, Generalized Linear Model, Logistic Regression, Fast Large Margin, Deep Learning, Decision Tree, Random Forest, Gradient Boosted Trees, and Support Vector Machine).

Table 1. Sample Characteristics

	Patients with Diabetes		Patients without Diabetes	
	Average	Std Dev	Average	Std Dev
Length of stay (in days)	5,05	6,34	4,25	5,96
Number of chronic conditions	7,27	3,22	3,88	3,28
Number of procedures	1,77	2,42	1,81	2,20

Table 2 shows the results of training and testing the applied Machine Learning algorithms to predict whether a patient has Diabetes and identify essential features in hospitalization records. The top three algorithms based on accuracy were Deep Learning, Naive Bayes, and gradient-boosted trees.

Table 2. Model Evaluation

Model	Accuracy	Classification Error	Precision	Recall	F Measure
Naive Bayes	0,74	0,26	0,73	0,75	0,74
Generalized Linear Model	0,72	0,28	0,71	0,73	0,72
Logistic Regression	0,72	0,28	0,71	0,73	0,72
Fast Large Margin	0,70	0,30	0,70	0,70	0,70
Deep Learning	0,77	0,23	0,76	0,77	0,76
Decision Tree	0,71	0,29	0,65	0,89	0,75
Random Forest	0,71	0,29	0,68	0,77	0,72
Gradient Boosted Trees	0,73	0,27	0,74	0,68	0,71
Support Vector Machine	0,49	0,51	0,49	1,00	0,66

Deep Learning was the best algorithm in accuracy, classification error, Area Under the ROC Curve (AUC), precision, and specificity. The decision tree was better at recall, f-measure, and sensitivity.

The results of the comparison showed that the performance of the algorithms varied depending on the dataset and the specific problem being solved. Individual algorithm performance was found to be less compared to combining them into one ensemble model using stacking.

For phase 2, we combined the top three algorithms (Deep Learning, Naive Bayes, and gradient-boosted trees) using two ensemble techniques:

- Ensemble model using Stacking.
- Ensemble model using Voting.

The performance of both built ensemble models outperformed the individual algorithms. As table 3 shows, the accuracy of the Ensemble model using stacking and the Ensemble model using voting was (99,94 % and 99,34 %) respectively. Stacking can give slightly better results due to the use of the meta-learner.

Model	Accuracy	Classification Error	Precision	Recall	Specificity	Sensitivity	F Measure
Stacking Model	99,94	0,06	100,00	99,88	99,88	100	99,94
Voting Model	99,34	0,66	100,00	98,68	100	98,68	99,33

The attributes that were highly correlated to the prediction of diabetics were:

- NCHRONIC: the number of Chronic Conditions in the record.
- NDX: number of diagnoses.
- AGE: patient Age in years.
- LOS: the length of Stay in Hospital.

These results were consistent with the literature since Diabetes comes with multiple chronic conditions related to it.^(24,26) Adding more individual data could help in personalizing the results, and integrating the prediction model with IoT collected data on patients could also provide a richer insight.

DISCUSSION

The proposed ensemble model for Diabetes Mellitus prediction using stacking and Voting methods showed significant improvement in predictive accuracy compared to individual machine learning algorithms. Specifically, the accuracy of the stacking ensemble in this research slightly surpassed the voting one and scored 99,94 %. Finally, these results clearly underline multiple classifier combinations in leveraging their strengths while mitigating the weaknesses of the individual ones. This has already been very well supported by the literature in recent years.^(23,32,39,40)

Adding to these, the ensemble models perform better, as present in the study samples. There is a finding in the literature where it was reported that the ensemble approaches are superior in their predictive ability compared to single models in the medical field. For example, Wu et al.⁽³⁹⁾ showed the same benefits of ensemble methods in predicting breast cancer patients and achieved higher accuracy than single classifiers. Furthermore, further to the prediction of DM, Salehi et al.⁽⁴⁰⁾ estimated the DL models using ensemble methods, respectively. There were improvements in the prediction accuracy and robustness, respectively, in line with this extensive body of evidence on the strength of ensemble methods and applications in health.

Furthermore, the setting within which this study uses the HCUP data will provide the model developed with real-world application. The identified predictors—number of chronic conditions, gender, and age—consist of established risk factors for DM, thereby augmenting the model with clinical relevance. This consistency with existing clinical knowledge is crucial, as it argues that the model's predictions are grounded in medically recognized risk factors, a point that is important, of course, for gaining acceptance in clinical settings.^(3,4,15,24)

Considerable clinical implications must be attributed to the findings in the current study. Such accurate prediction models, developed in the present work, would enable the health care providers to identify and intervene early in the developing subjects who would be at risk of progression of DM into lower levels of the disease.⁽³⁾ Early intervention in the case of DM will prevent the pace of the disease to the worsening stages, so better patient outcomes with lesser expenditure on health care can be achieved. The good predictive ability of the model regarding DM, based on easily obtainable clinical data points such as chronic conditions and demographic factors, suggests the potential for integration within the existing electronic health record systems.⁽⁵⁾

In addition, ensemble learning here could increase precision for personalized medicine. The model could consider more elements through multiple classifiers, hence having a risk assessment with greater nuances compared to classical methods, increasing the chance of more personalized and, therefore, effective treatment strategies.

Some ensemble models, particularly the iteration approach in stacking, become too complicated and computationally heavy regarding resources and interpretability. These factors may constrain the model's application in settings with constrained resources or in which model transparency is desired.

However, despite the application of cross-validation techniques, overfitting is another drawback of these

models. Specifically, ensemble models that combine more than one layer of classifiers—for example, when using stacking are more likely to overfit. This study took the necessary steps to mitigate the risk associated with this, but it's probably something future research will have to keep a close eye on.

In the future, more focus should be on the generalization of the model, whereby more diverse datasets from different regions of the globe and healthcare setups can be included. Similarly, specific approaches to integrating the other ensemble techniques or hybrid models would result in further significant improvements in prediction more robustly. Thirdly, further work can involve structure simplification whereby balance in accuracy vis-à-vis interpretability can be achieved, making it appropriate for universal clinical embracement.

Besides, the potential integration of this model into clinical workflows needs further exploration. The research should focus on developing user-friendly interfaces and decision-support systems to exploit this model and help clinicians make decisions about the choice of strategies for the prevention and management of DM.

This makes this research very recommendable for using the ensemble learning method to predict diabetes mellitus, meaning that when classifiers are combined by means of stacking and voting, they perform better in prediction. The findings contribute to the growing body of evidence on the use of machine learning in health and its application possibilities in improving clinical outcomes through better and timely predictions. Such a model will likely play an even more critical role in personalized medicine and preventive healthcare as machine learning evolves.

CONCLUSIONS

The results of this study provide insights into the relative strengths and limitations of different machine learning algorithms for diabetes prediction and the use of ensemble techniques. The choice of algorithm depended on the type of data available, and the specific problem being solved. Further combinations using ensemble techniques are needed to continue improving the accuracy of these algorithms and to expand their use to more complex medical problems related to Diabetes.

In this paper, performance analysis was made for the diabetes mellitus data set to improve the accuracy by using a classification algorithm. We compared nine prediction models using the important attributes. Then build two ensemble models with two different techniques: stacking and voting. The performance of the ensemble models exceeded that of the single classification model. More specifically, the ensemble model using stacking gave better results than voting.

Finally, we recommend encouraging stakeholders in the healthcare sector to enhance healthcare providers' decision-making abilities and enable more precise prediction of chronic diseases such as diabetes by training providers to use modern data analysis modalities such as Machine Learning and massive data sets.

Future research could include the application of SHAP explainable AI and the use of IoT data to personalize the ensemble model predictions.

BIBLIOGRAPHIC REFERENCES

1. World Health Organization. Diabetes [Internet]. 2021 [cited 2021 Jan 4]. Available from: <https://www.who.int/news-room/fact-sheets/detail/diabetes>
2. Runkler TA. Data Mining. Wiesbaden: Vieweg+Teubner; 2010.
3. Chaves L, Gonçalo M. Data mining techniques for early diagnosis of diabetes: A comparative study. *Appl Sci.* 2021;11(5):2218.
4. Guariguata L, Whiting DR, Hambleton I, Beagley J, Linnenkamp U, Shaw JE. Global estimates of diabetes prevalence for 2013 and projections for 2035. *Diabetes Res Clin Pract.* 2014;103(2):137-49.
5. Fiarni C, Sipayung EM, Maemunah S. Analysis and prediction of diabetes complication disease using data mining algorithm. *Procedia Comput Sci.* 2019;161:449-57.
6. Ajlouni K, et al. Time trends in diabetes mellitus in Jordan between 1994 and 2017. *Diabet Med.* 2019;36(9):1176-82.
7. Blair M. Diabetes mellitus review. *Urol Nurs.* 2016;36(1).
8. Yuvaraj D, Mavaluru D, Sivaram M, Nageswari S. An efficient data mining process on temporal data using relevance feedback method. *World Rev Sci Technol Sustain Dev.* 2022;18(1):20-30.
9. Munoz-Gama J, et al. Process mining for healthcare: Characteristics and challenges. *J Biomed Inform.* 2022;127:103994.

10. Kuatbayeva AA, Izteleuov NE, Kabdoldin A, Abdyzhalilova R. Data mining models for healthcare. *Adv Technol Comput Sci*. 2022;3:11-7.
11. Durugkar SR, Raja R, Nagwanshi KK, Kumar S. Introduction to data mining. In: *Data Mining and Machine Learning Applications*. 2022. p. 1-19.
12. Mavrogiorgou A, Kiourtis A, Manias G, Kyriazis D. An optimized KDD process for collecting and processing ingested and streaming healthcare data. In: *12th Int Conf on Information and Communication Systems (ICICS)*. IEEE; 2021. p. 49-56.
13. Traymbak S, Issar N. Data Mining Algorithms in Knowledge Management for Predicting Diabetes After Pregnancy by Using R. *Indian J Comput Sci Eng*. 2021;12(6).
14. Healthcare Cost and Utilization Project (HCUP). Agency for Healthcare Research and Quality. [Internet]. 2021 [cited 2021 Jan 4]. Available from: <https://www.ahrq.gov/data/hcup/index.html>
15. Al-Shanableh N. Using data mining to investigate hospitalization experiences of Parkinson's disease patients. ProQuest Dissertations Publishing; 2018.
16. Herle H, Padmaja KV. Relative merits of data mining algorithms of chronic kidney diseases. *Int J Adv Comput Sci Appl*. 2021;12(6):575-83.
17. Karrar AE. Investigate the ensemble model by intelligence analysis to improve the accuracy of the classification data in the diagnostic and treatment interventions for prostate cancer. *Int J Adv Comput Sci Appl*. 2022;13(1):181-8.
18. Tarawneh O, Otair M, Husni M, Abuaddous HY, Tarawneh M, Almomani MA. Comparative analysis of machine learning algorithms for heart disease predictions. *Int J Adv Comput Sci Appl*. 2022;13(4):1340-4.
19. Maliha SK, Mahmood MA. An efficient model for early prediction of diabetes utilizing classification algorithm. In: *6th Int Conf on Intelligent Computing and Control Systems (ICICCS)*. IEEE; 2022. p. 1607-11.
20. Anil KS, Jain R. Data mining techniques in diabetes prediction and diagnosis: A review. In: *6th Int Conf on Trends in Electronics and Informatics (ICOEI)*. IEEE; 2022. p. 1696-701.
21. The Middle East and North Africa. In: *IDF Diabetes Atlas*. 10th ed. 2022. p. 2000-45.
22. Huang K, Yang H, Zhu X, et al. Ensemble deep learning for COVID-19 diagnosis using chest CT scan images. *IEEE Trans Med Imaging*. 2020;39(8):2572-83.
23. Al Diabat M, Al-Shanableh N. Ensemble learning model for screening autism in children. *Int J Comput Sci Inf Technol*. 2019;11:45-62.
24. Alzyoud M, et al. Diagnosing diabetes mellitus using machine learning techniques. *Int J Data Netw Sci*. 2024;8(1):179-88.
25. Alsubihat D, Al-shanableh N. Predicting Student's Performance Using Combined Heterogeneous Classification Models. *Int J Eng Res Appl*. 2023;13(4):206-18.
26. Al-shanableh N, et al. Data Mining to Reveal Factors Associated with Quality of life among Jordanian Women with Breast Cancer. 2023;6:1-6.
27. Ababneh A, Al-shanableh N, Alzyoud M. A Review of Algorithms and Techniques for Analyzing Big Data. *Int J Emerg Trends Eng Res*. 2021;9(6):695-702.
28. Abu Salimeh A, Al-shanableh N, Alzyoud M. Natural Language Processing and Parallel Computing for Information Retrieval from Electronic Health Records. In: *ITM Web Conf*. 2022;42:01013.
29. Alghamdi A, Alshammari I. Diabetes Prediction Using Machine Learning Techniques. In: *2nd Int Conf on Computer Applications & Information Security (ICCAIS)*. IEEE; 2020. p. 1-6.

30. Yadav N, Tiwari A, Pal NR. Machine Learning Based Diabetes Prediction Using Clinical Data. In: 9th Int Conf on Cloud Computing Data Science & Engineering - Confluence. IEEE; 2019. p. 424-9.
31. Qureshi MA, Azad AKMA. Diabetes risk factor identification using machine learning techniques. In: Int Conf on Electrical Computer and Communication Engineering (ECCE). IEEE; 2019. p. 1-6.
32. Dheeraj K, Murugesan PR. Machine Learning based Risk Prediction for Type 2 Diabetes. In: Int Conf on Intelligent Techniques and Control (ITC). IEEE; 2020. p. 1-6.
33. Bano S, Siddiqui MH, Raza M, Raza MA. Diabetes Prediction and Risk Factors Identification using Machine Learning. In: Int Conf on Computer and Communication Technologies (IC3T). IEEE; 2020. p. 1-6.
34. Chen H, Li H, Huang G, Liu X, Xu J. A hybrid deep learning approach for accurate breast cancer diagnosis. IEEE Access. 2019;7:76314-23.
35. Surya DSK, Bhowmik SK, Kundu MK. Prediction of Heart Disease Using Machine Learning Algorithms: A Survey. IEEE Access. 2020;8:160504-18.
36. Rashid NS, Yahya SW, Razak RA, Hanafi FF. Deep Learning Techniques for Disease Detection and Classification: A Survey. IEEE Access. 2020;8:149937-65.
37. Qureshi MA, Islam MA, Ali MI. Machine Learning Techniques for Disease Diagnosis: A Review. In: 2nd Int Conf on Computing Mathematics and Engineering Technologies (iCoMET). IEEE; 2019. p. 1-6.
38. Chowdary SGS, Annapurna RGVJL. Machine Learning Algorithms for Disease Diagnosis: A Comprehensive Review. In: 5th Int Conf on Advanced Computing & Communication Systems (ICACCS). IEEE; 2019. p. 1009-14.
39. Wu Y, Liu X, Zhang C, et al. An ensemble model for the prediction of breast cancer using gene expression data. IEEE Access. 2018;6:16103-11.
40. Salehi M, Gandomi AH, Aghaei AH, Mirjalili SA. A novel ensemble machine learning approach for diagnosing and treating diseases. IEEE Access. 2019;7:55256-64.

FINANCING

The authors did not receive financing for the development of this research.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Najah Al-shanableh, Mazen Alzyoud, Raya Yousef Al-husban, Nail M. Alshanableh, Ashraf Al-Oun, Mohammad Subhi Al-Batah, Mowafaq Salem Alzboon.

Data curation: Najah Al-shanableh, Mazen Alzyoud, Raya Yousef Al-husban, Nail M. Alshanableh, Ashraf Al-Oun, Mohammad Subhi Al-Batah, Mowafaq Salem Alzboon.

Formal analysis: Najah Al-shanableh, Mazen Alzyoud, Raya Yousef Al-husban, Nail M. Alshanableh, Ashraf Al-Oun, Mohammad Subhi Al-Batah, Mowafaq Salem Alzboon.

Research: Najah Al-shanableh, Mazen Alzyoud, Raya Yousef Al-husban, Nail M. Alshanableh, Ashraf Al-Oun, Mohammad Subhi Al-Batah, Mowafaq Salem Alzboon.

Methodology: Najah Al-shanableh, Mazen Alzyoud, Raya Yousef Al-husban, Nail M. Alshanableh, Ashraf Al-Oun, Mohammad Subhi Al-Batah, Mowafaq Salem Alzboon.

Validation: Najah Al-shanableh, Mazen Alzyoud, Raya Yousef Al-husban, Nail M. Alshanableh, Ashraf Al-Oun, Mohammad Subhi Al-Batah, Mowafaq Salem Alzboon.

Drafting - original draft: Najah Al-shanableh, Mazen Alzyoud, Raya Yousef Al-husban, Nail M. Alshanableh, Ashraf Al-Oun, Mohammad Subhi Al-Batah, Mowafaq Salem Alzboon.

Writing - proofreading and editing: Najah Al-shanableh, Mazen Alzyoud, Raya Yousef Al-husban, Nail M. Alshanableh, Ashraf Al-Oun, Mohammad Subhi Al-Batah, Mowafaq Salem Alzboon.