A Fast Fourier Transform-Coupled Machine Learning-Based Ensemble Model for Disease Risk Prediction Using a Real-Life Dataset

Raid Lafta[1,2(✉)], Ji Zhang[1], Xiaohui Tao[1], Yan Li[1], Wessam Abbas[1],
Yonglong Luo[3], Fulong Chen[3], and Vincent S. Tseng[4]

[1] Faculty of Health, Engineering and Sciences, University of Southern Queensland,
Toowoomba, Australia
{RaidLuaibi.Lafta,ji.zhang,xtao,yan.li}@usq.edu.au
[2] Computer Center, University of Thi-Qar, Thi-Qar, Iraq
[3] School of Mathematics and Computer Science,
Anhui Normal University, Wuhu, China
ylluo@ustc.edu.cn, long005@mail.ahnu.edu.cn
[4] Department of Computer Science, National Chiao Tung University,
Hsinchu, Taiwan
vtseng@cs.nctu.edu.tw

Abstract. The use of intelligent technologies in clinical decision making have started playing a vital role in improving the quality of patients' life and helping in reduce cost and workload involved in their daily healthcare. In this paper, a novel fast Fourier transform-coupled machine learning based ensemble model is adopted for advising patients concerning whether they need to take the body test today or not based on the analysis of their medical data during the past a few days. The weighted-vote based ensemble attempts to predict the patients condition one day in advance by analyzing medical measurements of patient for the past k days. A combination of three algorithms namely neural networks, support vector machine and Naive Bayes are utilized to make an ensemble framework. A time series telehealth data recorded from patients is used for experimentations, evaluation and validation. The Tunstall dataset were collected from May to October 2012, from industry collaborator Tunstall. The experimental evaluation shows that the proposed model yields satisfactory recommendation accuracy, offers a promising way for reducing the risk of incorrect recommendations and also saving the workload for patients to conduct body tests every day. The proposed method is, therefore, a promising tool for analysis of time series data and providing appropriate recommendations to patients suffering chronic diseases with improved prediction accuracy.

Keywords: Fast Fourier transformation · Ensemble model · Recommender system · Heart failure · Time series prediction · Telehealth

J. Kim et al. (Eds.): PAKDD 2017, Part I, LNAI 10234, pp. 654–670, 2017.
DOI: 10.1007/978-3-319-57454-7_51

1 Introduction

The chronical diseases such as heart disease have become the main public health issue worldwide which accounting for 50% of global mortality burden [1]. Due to lack of chronical diseases prediction tools, most of the populations around the world may be suffering from chronical diseases [2]. Recently, the survival rates have been noticeably increased due to technological improvements in diseases prediction models.

One of the important problems in medical science is accurate prediction of disease based on analysing historical data of patients. The data mining techniques and statistical analysis have been extensively used to provide major assistance to experts in disease prediction [27].

Recommendation systems can be defined as computer applications that assist and support medical practitioners in improved decision-making recommendation [3,28,29]. Those systems can help in minimizing medical errors and providing more detailed data analysis in shorter time [4].

In the recent years, the ensemble methods have been very robust for the blend of various predictive models. The major purpose of ensemble model is to improve the overall accuracy of prediction model. An ensemble is a set of base learners that use to enhance the prediction performance of low-quality data [5]. Bagging is an ensemble algorithm that was proposed by Breiman in mid-1990's [6]. Empirical results showed that both regression and classification problem ensemble are often more accurate than individual classifier that make them up [5]. Therefore, much research efforts have been invested using machine learning ensemble for chronical diseases prediction.

Least square-support vector machine (LS-SVMs) are a relatively new kind of machine learning techniques that was proposed by [17]. They have been recently used in the field of disease prediction. There are several studies in disease prediction filed where LS-SVMs are used. LS-SVM has been used by [18] for heart disease prediction. Muscle fatigue prediction in electromyogram (sEMG) signal is implemented using LS-SVM that proposed by [19]. Finally, LS-SVM has also been successfully applied by [20] to predict breast cancer.

Due to the ensemble outperforms individual classifiers, several such ensemble approaches have been proposed recently. A combination of different data mining techniques have been applied on different datasets. An ensemble framework based on different classifiers has been used by Das et al. and Helmy et al. [7,8] to generate high prediction accuracy for heart disease patients. The results show that heterogeneous ensemble has better results as compared to individual classifiers. A novel ensemble has been proposed by Bashir et al. [9] to improve the classification and heart disease prediction. The proposed ensemble used a bagging algorithm with a multi-objective optimized weighted voting that contacted on heart disease datasets. Verma et al. [10] developed a novel hybrid model using data mining methods. In their model, the proposed ensemble was used to predict coronary artery disease cases using non-invasive clinical data of patients.

Fast Fourier transform, an efficient technique to compute the discrete and the inverse, is an emerging tool for prediction. It has recently been applied to: analyze and predict electricity consumption in buildings [11,12], to forecast water demand [13,14], to detect epileptic seizure in electroencephalography (EEG) [15,16].

Since the importance of the prediction in medical domain as well as the urgency of demanding more powerful analytic tools in this regard, further efforts are definitely needed to enhance evidence-based decisions quality. In this work, we propose a novel fast Fourier transform-coupled a machine learning ensemble to predict and assess the short-term risk of disease and provide patients with appropriate recommendations for necessity of taking a medical test in the incoming day.

The remainder of this paper is organised as follows. Section 2 explains the details of fast Fourier transformation and machine learning classifiers that constructing the proposed ensemble model. Section 3 briefly defines the proposed methodology including predictive model development and describes the used data set. Section 4 discussed in details the experimental evaluation results. Finally we conclude the paper and highlight the future work in Sect. 5.

2 Theoretical Background

2.1 Bootstrap Aggregation (Bagging)

An ensemble method is one of the combination approaches used to overcome the limited generalization performance of individual models and to generate more accurate predictions than single models. Bagging is a machine learning ensemble used to solve problems by combining the decisions of multiple classifiers [21]. During a bootstrap method, in a bagging method, classifiers are trained independently and then aggregated by an appropriate combination strategy. The proposed ensemble model can be divided into two phases. At the first phase, bagging uses bootstrap sampling to generate a number of training sets. At the second phase, training the base classifiers is performed using bootstrap training sets generated during the first step. The generic flowchart of bagging algorithm is shown in Fig. 1. In this study, the training set was divided into multiple datasets using bootstrap aggregation approach, and then individually classifiers are applied on these datasets to generate the final prediction. We argue that each individual classifier in the weighted-bagging approach should has a different individual performance level. We proposed, therefore, to assign a weight for each classifier's vote based on how well the classifier performed. The classifier's wight is calculated based on it's error rate so that the classifier that has the lower error rate is more accurate, and therefore, it should be assigned the higher weight for that classifier. The weight of classifier C_i's vote is calculated as follows [22]:

$$\log \frac{1 - error(C_i)}{error(C_i)} \tag{1}$$

The proposed weighted-bagging ensemble can be easily understood by the following example:

1. Suppose that the classifier training is performed for training data and error rate is calculated.
2. Neural Networks (NN), Support Vector Machine (SVM) and Naive Bayes (NB) are used as individual classifiers. Following error rate results are generated for each classifier: NN = 0.25, SVM = 0.14, NB = 0.30.
3. Now, according to the formula given in *Eq.* (6), the resultant weights are as follows: NN = 0.47, SVM = 0.78, NB = 0.36.
4. Suppose, the algorithms have predicted the following classes for a test day: NN = 0, SVM = 1, NB = 0. (0: no test required; 1: test needed).
5. Based ensemble classifier, the weighted vote will be generated the following prediction results:
   Class 0: $NN + NB \longrightarrow 0.47 + 0.36 \longrightarrow 0.83$,
   Class 1: $SVM \longrightarrow 0.78$.
6. Finally, according to weighted vote, the class 0 has higher value as compared to class 1. Therefore, the ensemble classifier for this test day will be classified as Class 0.

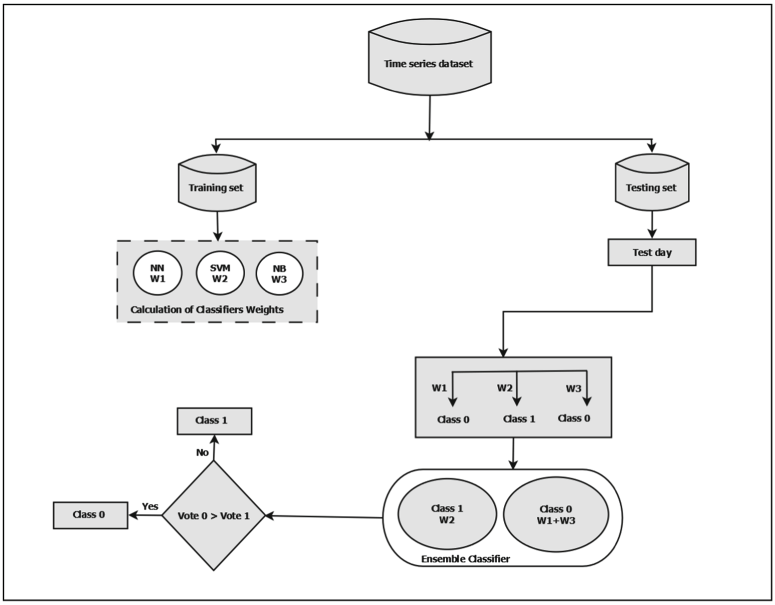

Fig. 1. The generic flowchart of bagging algorithm

2.2 Fast Fourier Transform-Coupled Machine Learning Based Ensemble (FFT-MLE) Model

The major purpose of this study was to demonstrate the effectiveness of fast Fourier transform-coupled machine learning based ensemble model for advising patients concerning whether they need to take the body test today or not based on analysis of their medical data during the past a few days. After developing a ensemble model that consists from three predictive models, Fast Fourier transformation algorithm, which is a data-preprocessing tool for non-stationary singles, is proposed. A fast Fourier transform (FFT) is an efficient algorithm to calculate the discrete Fourier transform (DFT) and the inverse. DFT decomposes the input sequence values to extract the frequency information in order to predict the next day. The discrete-time Fourier transform of a time series x(t) can be defined as:

$$X(c^{jw}) = \sum_{t=-\infty}^{\infty} x(t)c^{-jwt} \tag{2}$$

where t is the time index of discrete, and w refers to the frequency. There are T input time series x(t), so the transform pair of DFT can be defined as:

$$X(P) = \sum_{t=0}^{T-1} x(t)W_T^{tp} \Leftrightarrow x(t) = \frac{1}{T}\sum_{p=0}^{T-1} X(P)W_T^{-tp}, where \quad W = c^{-j2\Pi/T} \tag{3}$$

Furthermore, the DFT can be presented as discrete-time Fourier transform of a cyclic signal with period T.

$$x = \begin{bmatrix} x(0) \\ x(1) \\ \vdots \\ x(T-1) \end{bmatrix}, \quad X = \begin{bmatrix} X(0) \\ X(1) \\ \vdots \\ X(T-1) \end{bmatrix} \tag{4}$$

$$W = [W_T^{pt}] = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & W_T & \dots & W_T^{T-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & W_T^{T-1} & \dots & W_T^{(T-1)(T-1)} \end{bmatrix} \tag{5}$$

and the following equation presents the relationship between x and X as follows:

$$X = Wx \Leftrightarrow x = \frac{1}{T}W^H X \tag{6}$$

According to the above equations, the DFT matrix W requires T^2 complex multiplications for the implementation of a time series input signal x(t) with length T. Therefore, a required implementation cost for factorizing the fast Fourier transform W into a matrix is lower than the direct DFT. For each stage of fast Fourier transform requires $T/2$ multiplications and T additions [23].

In practical sense, the input time series data segments into set of slide windows with a length of k (the size of the sliding window used in time series data analysis). The input time series are decomposed using fast Fourier transform to extract the frequency information included in the input data in order to predict the following medical test day.

3 Methodology

3.1 Predictive Model Development

The predictive models are developed in MATALB environment on a desktop computer with the configurations of a 3.40 GHz Intel core i7 CPU processor with 8.00 GB RAM. The major purpose of this study was to investigate the performance of the fast Fourier transform-coupled machine learning ensemble to predict the short-term risk of disease and provide patients with appropriate recommendations for necessity of taking a medical test in the incoming day. The training data is used to train the classifiers that construct the ensemble and testing data to evaluate the performance of predictive model. In this study, the time series medical data were partitioned into about 75% as training data and 25% as testing data.

Figure 2 illustrates the different stages of the fast Fourier transform coupled-ensemble model. Basically, the input time series data was segmented into a set of sub-segments with overlapping of m based on a predefined value of k to identify the window size of sub-segment. Let $X = \{y_1, y_2, y_3, \ldots, y_n\}$ is a time series of n test measurements. The main idea is to separate X into a number of overlapping sub-segments. The overlapping value m is set to be a test measurement of one day. Then, each sub-segment is passed through fast Fourier transform in order to obtain the desired information. A resulting set of fast Fourier transform coefficients of 28 levels is tested to figure out the desired FFT level. Different combinations of statistical features from each level are tested and analysed the performance of the proposed model with different FFT characteristics. The purpose of using FFT in this paper is to study the properties of time series in frequency domain which could be difficult to obtain in time domain. The basic idea of frequency analysis is to re-express the original time series as a new sequence which determine the important of each frequency components. The fast Fourier transform was used to decompose each time series slide window to acquire five ($\alpha, \beta, \gamma, \delta$, and θ) frequency bands Fig. 2. Based on the literature, the high frequency band could be able to capture the desired information, therefore, the high frequency band was divided into 8 sub-frequency bands. In addition, the original time series slide window was also added, as a reference, to the feature extraction test. As a result, the total number of frequency bands sets is $(5+8+1=14)$ for each slide window. Furthermore, the power of the FFT coefficients was calculated for all the 14 sets of frequency bands. It allows to compute the square of the absolute value of the Fourier coefficients. These frequencies were grouped together and used as input to the proposed ensemble. As a result, the features were extracted from 28 set of frequency bands $(14+14=28)$. All the bands features were tested and analysed to figure out best combination of features. Figure 3 shows an example of decomposing a slide window of time series into 28 bands.

Two sets of statistical features were extracted from each band to find out the best combination of features to present the data. Two and four statistical features were extracted from each band. The mentioned 28 bands were tested and

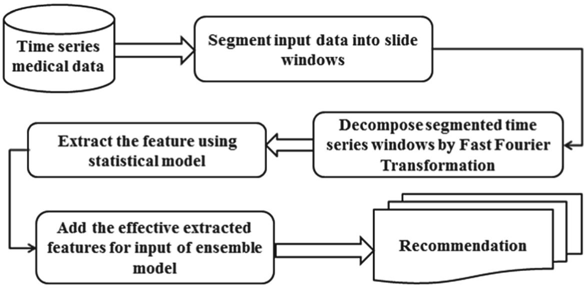
Time series medical data
Segment input data into slide windows
Decompose segmented time series windows by Fast Fourier Transformation
Extract the feature using statistical model
Add the effective extracted features for input of ensemble model
Recommendation

Fig. 2. The stages of developing a Fast Fourier Transform-Coupled Machine Learning based Ensemble (FFT-MLE) model

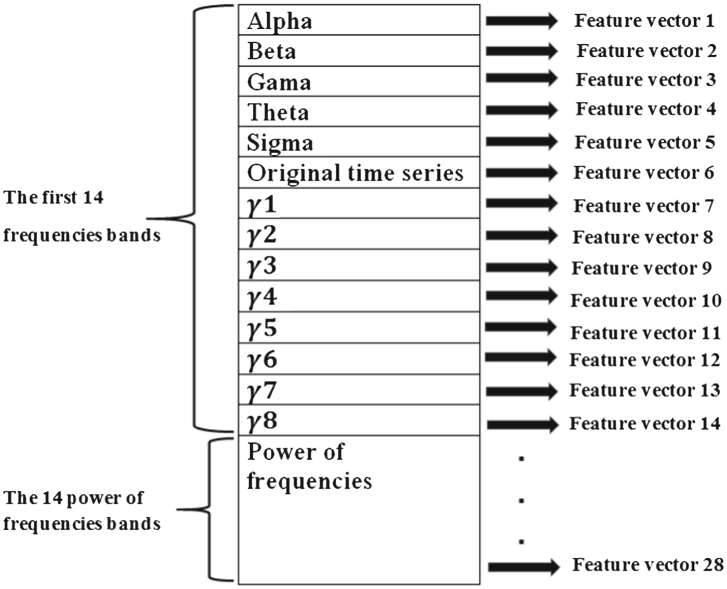

Fig. 3. The generic decomposition of slide window time series into 28 bands

analysed with those features sets at each stage, and the results were recorded. However, it was observed that using four statistical features yield better results, in term of accuracy, compared with using two features. The extracted features from all the 28 bands were used as key features to training the ensemble's classifiers in order to predict the following day.

The two features sets were included (max and min), while the four features sets were included (max, min, standard deviation, and median). The features were denoted $(X_{min}, X_{max}, X_{std}, X_{med})$. The short explanations of statistical features are provided in Table 1. To test the relationship between the features and the risk prediction, extensive experiments were conducted. As a result, two vectors of 2X28 and 4X28 were extracted and used to identify whether the patient requires to take a test today or not.

The features were tested separately by testing each band and by using all features of bands as a one vector. Our findings showed that the 4 features set gives a better results than those of two features set. The extracted features from each band, four feature set, were grouped in one vector and used as the features set to predict the next day. All the detailed are discussed in simulation and experiment results section.

In the experiments, we use a statistical approach to extract the statistical features from each band, and then put all the features from one segment in a vector to present the window. It was found that some of time series data are symmetric distribution and other skewed distribution. The min and max are considered appropriate measures for a time series with symmetric distribution, whilst, for a skewed distribution, mean and standard division are used to measure the center and spread of dataset [24,25].

Table 1. Short explanations of statistical features

Feature name	Formula	Description
Maximum value	$X_{max} = Max[x_n]$	Where $x_n = 1, 2, 3, \ldots, n$ is a time series, N is the size of slide window, AM is the mean of slide window
Minimum value	$X_{min} = Min[x_n]$	
Mean	$X_{mean} = \frac{1}{n}\sum_{1}^{n} x_i$	
Standard division	$X_{SD} = \sqrt{\sum_{n=1}^{N}(x_n - AM)}\frac{2}{n-1}$	

3.2 Evaluation Design

In this section, we offer details concerning the strategy of our experimental evaluation including datasets, performance metrics and the experimental platform.

As the predictive performance of the FFT-MLE model is quite important, assessment of potential predictions is critically dependent on the quality of the used dataset. For this reason, telehealth data from Tunstall dataset will be conducted in this work. We use a real-life dataset obtained from our industry collaborator Tunstall to test the practical applicability of the FFF-MLE model. A Tunstall dataset obtained from a pilot study has been conducted on a group of heart failure patients and the resulting data were collected for their day-to-day medical readings of different measurements in a tele-health care environment. The Tunstall database employed in the development of the algorithm consists of data from six patients with a total of 7,147 different time series records. Data were acquired between May and January 2012, using a remote telehealth collaborator. The dataset is by nature in a time series and contains a set of measurements taken from the patients on different days. Each record in the dataset consists of a few different meta-data attributes about the patients such as patient-id, visit-id, measurement type, measurement unit, measurement value, measurement question, date and date-received. The characteristics of the features of the dataset are shown in Table 2.

Table 2. Characteristic features of the dataset

Feature name	Feature type
id	Numeric
id-patient	Numeric
hcn	Numeric
visit-id	Numeric
measurement type	Nominal
measurement unit	Nominal
measurement value	Numeric
measurement question	Nominal
date	Numeric
date-received	Numeric

In addition, each record contains a few medical attributes including Ankles, Chest Pain, and Heart Rate, Diastolic Blood Pressure (DBP), Mean Arterial Pressure (MAP), Systolic Blood Pressure (SBP), Oxygen Saturation (SO2), Blood Glucose, and Weight. Ethical clearance was obtained from the University of Southern Queensland (USQ) Human Research Ethics Committee (HREC) prior to the onset of the study. This dataset is used as the ground truth result to test the performance of our recommendation system. The recommendations produced by our system will be compared with the actual readings of the measurement in question recorded in the dataset to see how accurate our recommendations are.

Due to the patient's historical medical data has often class-imbalanced problem (i.e. the number of normal data is much more than that of abnormal data), we are carefully dealt with the class-imbalanced problem for classifier building. The over-sampling and under-sampling have been proposed as a good means to address this problem. The predictive accuracy is usually used to evaluate the performance of machine learning algorithms. However, this measure is not appropriate when the used data is imbalanced [26].

The performance of individual classifiers as well as the proposed ensemble is evaluated by calculating the *accuracy*, *workload saving*, and *risk*. Accuracy refers to the percentage of correctly recommended days against the total number of days that recommendations are provided; workload saving refers to the percentage of the total number of days when recommendations are provided against the total number of days in the dataset, while risk refers to the percentage of incorrectly recommended days that recommendations are no test needed. Mathematically, Accuracy, workload saving and risk are defined as follows:

$$Accuracy = \frac{NN}{NN + NA} \times 100\% \tag{7}$$

$$Saving = \frac{NN + NA}{|\mathcal{D}|} \times 100\% \tag{8}$$

$$Risk = \frac{NR}{|\mathcal{D}|} \times 100\% \tag{9}$$

Where NN denotes the number of days with correct recommendations, NA denotes the number of days with incorrect recommendations, NR denotes the number of days with incorrect days that recommendations are no test needed, and $|\mathcal{D}|$ refers to the total number of days in the dataset. Here, a correct recommendation means that the model produces the recommendation of "no test required" for the following day and the actual reading for that day in the dataset is normal. If this is a case, the recommendation is considered accurate.

4 Result Analysis

The using of the fast Fourier transform-coupled machine learning based ensemble (FFT-MLE) model aims at short-term risk assessment in patients based on analytic of a patient's historical medical data using fast Fourier transform. As mentioned above, the time series slide windows were decomposed by using the FFT. Then, the suitable features were selected as input for the ensemble model. The new time series selected were employed as input of the ensemble's classifiers instead of the original time series data. Different sets of statistical features, as mentioned above, were used to determine the best number of features for each slide window. The detailed results are discussed in the following sub sections:

4.1 Prediction Accuracy with Different Number of Features

To evaluate the relationship between the number of the extracted features and the prediction accuracy, several experiments were conducted using different sets of features. Based on the experiment results, When the number of the statistical features is increased, the predictive accuracy of the proposed model is more significant. The three classifiers of the FFT-MLE model, neural networks (NN), least square-support vector machine (LS-SVM) and naive Bayes (NB), were trained with different sets of features.

Typically, an ensemble model is a supervised learning technique for combining multiple weak learners or models to produce a strong model [9]. Based on our findings, a group of classifiers is likely to make better decisions compared to individuals. The experimental results showed that the proposed method using an ensemble classifier gives a satisfactory recommendation accuracy compared to a single classifier.

Table 3. Prediction accuracy of Ensemble model based on the first 5 frequencies bands with two features.

Tunstall dataset (%)				
Measurement	Ensembles	Accuracy (%)	Saving (%)	Risk (%)
Heart rate	Neural network	70.64	54.33	09.92
	LS-SVM	75.37	62.34	07.21
	Naive Bayes	71.87	53.54	09.78
	Ensemble model	**85.38**	**60.60**	**5.90**
DBP	Neural network	71.30	58.18	09.76
	LS-SVM	78.45	64.75	07.44
	Naive Bayes	72.11	57.66	09.78
	Ensemble model	**86.27**	**65.57**	**5.88**
MAP	Neural network	69.70	50.33	09.92
	LS-SVM	77.98	64.34	07.21
	Naive Bayes	70.17	56.54	12.78
	Ensemble model	**83.33**	**62.60**	**06.45**
SO2	Neural network	70.30	54.33	09.98
	LS-SVM	74.39	65.34	07.95
	Naive Bayes	68.78	50.54	12.78
	Ensemble model	**84.38**	**66.47**	**5.90**

Two-Features Set. In this experiment, first, the first main frequencies bands ($\alpha, \beta, \gamma, \delta$, and θ) were selected from each slide window and then the two features of (X_{min} and X_{max}) for each band were utilized to evaluate the performance of the proposed model. Through this process, each slide window was converted to a vector of 10 extracted features. The extracted features were randomly divided into training and testing sets. Each classifier is trained on whole training set. Then, we individually apply the basic *bagging* algorithm on each day in test set by assigning a weighted vote for each classifier based on it's performance in the training stage. The classifier that has the lower error rate is more accurate, and therefore, it should be assigned the higher weight for that classifier. The final prediction of each day in test set is calculated based on the sum of weights for each class. As a result, the class that has the highest weight will be selected as class label for that day. Each day in test set is classified into "need test" or "no test needed" labels. Table 3 presents the comparison of accuracy, workload saving and risk results of ensemble model with individual classifier techniques for the four measurements. It is compared with different classifiers such as Neural network (NN), Least square- Support Vector Machine (LS-SVM) and Naive Bayes (NB).

From the obtained results in this table, although the proposed FFT-MLE model is achieved noticeably significant results using five frequencies bands.

Table 4. Prediction accuracy of Ensemble model based on 28 frequencies bands with two features

Tunstall dataset (%)				
Measurement	Ensembles	Accuracy (%)	Saving (%)	Risk (%)
Heart rate	Neural network	74.13	50.38	07.80
	LS-SVM	80.37	60.14	05.21
	Naive Bayes	75.20	50.55	06.10
	Ensemble model	**90.20**	**60.30**	**4.20**
DBP	Neural network	75.74	55.11	07.75
	LS-SVM	82.88	60.24	05.65
	Naive Bayes	77.74	55.74	06.95
	Ensemble model	**91.40**	**65.57**	**4.10**
MAP	Neural network	75.45	56.20	07.95
	LS-SVM	81.93	67.22	05.77
	Naive Bayes	76.28	53.13	07.30
	Ensemble model	**91.85**	**65.20**	**04.00**
SO2	Neural network	73.25	51.50	07.95
	LS-SVM	80.55	60.38	05.10
	Naive Bayes	74.11	54.36	07.82
	Ensemble model	**90.50**	**60.25**	**4.15**

However, the five bands were not enough to represent the slide windows because they did not appear the appropriate characteristics of slide windows. Therefore, all of 28 bands instead of the first five bands were considered to represent each slide window. In this case, $28 \times 2 = 56$ statistical features for each slide window were extracted and then used in the training of ensemble model.

From a observation the results in Table 4, it can be noticeably seen that the performance of the proposed FFT-MLE model, for all measurements, is improved compared with the previous results. This is because that the accuracies have significantly increased using all frequencies bands instead of the top five bands. According to the results in Table 4, the prediction accuracy of the proposed FFT-MLE model is increased by more than 5% compared to the obtained results in Table 3.

Four-Features Set. To improve the predictive performance of the FFT-MLE model, a four-features set of $(X_{min}, X_{max}, X_{std}, X_{med})$ was selected, tested to analyse the time series medical data. For each slide window, 5×4 statistical features were extracted and then used to evaluate the performance of the proposed FFT-MLE model, where 5 refers to the main bands of FFT that including: $\alpha, \beta, \gamma, \delta$, and θ, and the 4 indicates the number of selected features. The obtained results were showed that using four features set can be improved the prediction

Table 5. Prediction accuracy of Ensemble model based on the first 5 frequencies bands with a four-features set.

Tunstall dataset (%)				
Measurement	Ensembles	Accuracy (%)	Saving (%)	Risk (%)
Heart rate	Neural network	74.30	49.33	07.60
	LS-SVM	78.39	60.34	07.40
	Naive Bayes	74.73	52.42	07.50
	Ensemble model	**88.75**	**60.20**	**4.80**
DBP	Neural network	75.30	50.13	07.76
	LS-SVM	81.62	61.25	06.30
	Naive Bayes	74.45	57.30	07.82
	Ensemble model	**89.41**	**62.54**	**4.50**
MAP	Neural network	73.62	53.33	07.95
	LS-SVM	82.98	60.40	06.21
	Naive Bayes	75.33	51.40	07.50
	Ensemble model	**90.20**	**60.10**	**04.10**
SO2	Neural network	76.20	56.55	07.70
	LS-SVM	82.50	63.64	06.30
	Naive Bayes	74.60	55.50	07.5
	Ensemble model	**90.33**	**62.48**	**4.00**

accuracy of the proposed FFT-MLE model. As a result, an accuracy of 94%, for all measurements, was attained. Table 5 shows the prediction accuracies results and risk assessment using a four-features set with the five selected bands for all measurements. According to the obtained results in Table 5, the prediction accuracies for all measurements were noticeably improved using a four-features set. The obtained results proved that the four selected features have significantly improved the predictive performance of the proposed FFT-MLE model. The prediction accuracy of all measurements was increased by more than 6% compared with the two-features set results.

However, in order to further increase the prediction accuracy of the proposed model, all of 28 bands instead of the five bands were also selected, used to represent each slide window. The $28 \times 4 = 112$ statistical features for each slide window were extracted and then used in the training of ensemble model. From the obtained results in Table 6, it can be noticed that the performance of the proposed FFT-MLE model significantly improved for all measurements after having increased the number of selected bands with more features. Figure 4 shows the averaged accuracies using both of 5 and 28 bands with 2 and 4 features sets for all measurements.

Table 6. Prediction accuracy of Ensemble model based on 28 frequencies bands with a four-features set

Tunstall dataset (%)				
Measurement	Ensembles	Accuracy (%)	Saving (%)	Risk (%)
Heart rate	Neural network	79.57	53.88	06.90
	LS-SVM	86.50	64.75	04.95
	Naive Bayes	78.40	54.54	07.05
	Ensemble model	**94.15**	**63.25**	**3.30**
DBP	Neural network	78.35	56.14	07.10
	LS-SVM	90.50	64.50	04.10
	Naive Bayes	77.44	59.30	07.75
	Ensemble model	**95.30**	**64.13**	**2.85**
MAP	Neural network	80.30	55.50	06.65
	LS-SVM	89.50	63.40	05.50
	Naive Bayes	78.60	52.30	07.20
	Ensemble model	**94.50**	**64.30**	**03.00**
SO2	Neural network	79.25	54.60	06.95
	LS-SVM	89.40	64.50	05.25
	Naive Bayes	78.55	56.25	07.30
	Ensemble model	**95.20**	**63.13**	**02.95**

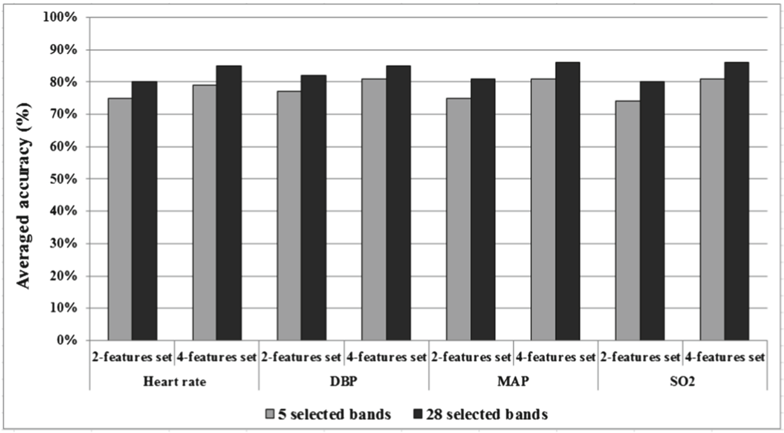

Fig. 4. The obtained results of two and four features sets after applying 5 and 28 FFT decomposition for all measurements

4.2 Prediction Time

In this experiment, the prediction time including training time and execution time of classifiers was proposed. Figure 5 shows the prediction time for each classifier and the proposed ensemble as well. From the results in Fig. 5, we observed that although the proposed model took more time compared with individual classifiers, it provided more accurate recommendation to patients suffering chronic diseases. On other hand, the least square-support vector machine (LS-SVM) was recorded the lowest prediction time compared with other individual classifiers.

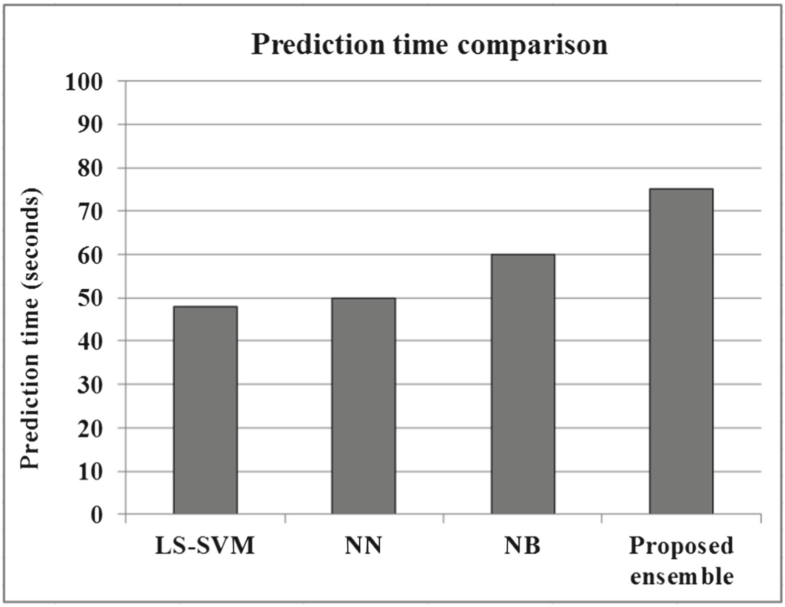

Fig. 5. Comparison of the prediction time between classifiers and the proposed ensemble

5 Conclusions and Future Work

In this work, a pilot study has been performed to evaluate the ability of a fast Fourier transform-coupled machine learning based ensemble model to predicts and assesses the short-term disease risk for patients suffering from chronical diseases such as heart disease. This study is considered one of the vital studies to use medical measurements of patient in the assessment and prediction of the short-term disease risk. This research presents a machine learning based ensemble model which incorporates fast Fourier transformation algorithm for pre-processing of input time series data. This ensemble is based on three heterogeneous learners named neural networks, least square support vector machine and naive Bayes in order to generate appropriate recommendations. The prediction model is developed aiming at improving the quality of clinical evidence-based decisions and helping reduce financial and timing cost taken by patients.

The experimental results showed that the four-features set yields a better predictive performance for all measurements compared to the two-features set. In addition, it is also mentioned that using the all of 28 bands give reasonable prediction accuracies under all measurements.

Future research directions include application of the proposed model on different datasets for more validations. We also plan to incorporate wavelet transformation with fast Fourier translation in order to pre-prossing the time series data.

Acknowledgement. The authors would like to thank the support from National Science Foundation of China through the research projects (Nos. 61572036, 61370050, and 61672039) and Guangxi Key Laboratory of Trusted Software (No. kx201615).

References

1. Kuh, D., Shlomo, Y.B.: A Life Course Approach to Chronic Disease Epidemiology. Inem Oxford University Press, London (2004)
2. Atlas, I.D.: International Diabetes Federation Diabetes Atlas, 6th edn. International Diabetes Federation, Basel (2013)
3. Thong, N.T.: HIFCF: an effective hybrid model between picture fuzzy clustering and intuitionistic fuzzy recommender systems for medical diagnosis. Expert Syst. Appl. **42**(7), 3682–3701 (2015)
4. Chen, D., Jin, D., Goh, T.-T., Li, N., Wei, L.: Context-awareness based personalized recommendation of anti-hypertension drugs. J. Med. Syst. **40**(9), 202 (2016)
5. Valentini, G., Masulli, F.: Ensembles of learning machines. In: Marinaro, M., Tagliaferri, R. (eds.) WIRN 2002. LNCS, vol. 2486, pp. 3–20. Springer, Heidelberg (2002). doi:10.1007/3-540-45808-5_1
6. Breiman, L.: Bagging predictors. Mach. Learn. **24**(2), 123–140 (1996)
7. Das, R., Turkoglu, I., Sengur, A.: Effective diagnosis of heart disease through neural networks ensembles. Expert Syst. Appl. **36**(4), 7675–7680 (2009)
8. Helmy, T., Rahman, S., Hossain, M.I., Abdelraheem, A.: Non-linear heterogeneous ensemble model for permeability prediction of oil reservoirs. Arab. J. Sci. Eng. **38**(6), 1379–1395 (2013)
9. Bashir, S., Qamar, U., Khan, F.H.: BagMOOV: a novel ensemble for heart disease prediction bootstrap aggregation with multi-objective optimized voting. Australas. Phys. Eng. Sci. Med. **38**(2), 305–323 (2015)
10. Verma, L., Srivastava, S., Negi, P.: A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data. J. Med. Syst. **40**(7), 1–7 (2016)
11. Tsai, C.-L., Chen, W.T., Chang, C.-S.: Polynomial-Fourier series model for analyzing and predicting electricity consumption in buildings. Energy Build. **127**, 301–312 (2016)
12. Ji, Y., Xu, P., Ye, Y.: HVAC terminal hourly end-use disaggregation in commercial buildings with Fourier series model. Energy Build. **97**, 33–46 (2015)
13. Brentan, B.M., Luvizotto Jr., E., Herrera, M., Izquierdo, J., Prez-Garca, R.: Hybrid regression model for near real-time urban water demand forecasting. J. Comput. Appl. Math. **309**, 532–541 (2016)
14. Odan, F.K., Reis, L.F.R.: Hybrid water demand forecasting model associating artificial neural network with Fourier series. J. Water Resour. Plan. Manag. **138**(3), 245–256 (2012)

15. Samiee, K., Kovcs, P., Gabbouj, M.: Epileptic seizure classification of EEG time-series using rational discrete short-time Fourier transform. IEEE Trans. Biomed. Eng. **62**(2), 541–552 (2015)
16. Kovacs, P., Samiee, K., Gabbouj, M.: On application of rational discrete short time Fourier transform in epileptic seizure classification. IEEE Trans. Biomed. Eng. 5839–5843 (2014)
17. Suykens, J.A., Vandewalle, J.: Least squares support vector machine classifiers. Neural Process. Lett. **9**(3), 293–300 (1999)
18. Bai, Y., Han, X., Chen, T., Yu, H.: Quadratic kernel-free least squares support vector machine for target diseases classification. J. Comb. Optim. **30**(4), 850–870 (2015)
19. Sharawardi, N.A., Choo, Y.-H., Chong, S.-H., Muda, A.K., Goh, O.S.: Single channel sEMG muscle fatigue prediction: an implementation using least square support vector machine. In: Information and Communication Technologies (WICT), pp. 320–325 (2014)
20. Li, S., Tang, B., He, H.: An imbalanced learning based MDR-TB early warning system. J. Med. Syst. **40**(7), 1–9 (2016)
21. Gao, H., Jian, S., Peng, Y., Liu, X.: A subspace ensemble framework for classification with high dimensional missing data. Multidimens. Syst. Sig. Process. 1–16 (2016)
22. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques. Elsevier, Amsterdam (2011)
23. Alfred, M.: Signal Analysis Wavelets, Filter Banks, Time-Frequency Transforms and Applications. Wiley, New York (1999)
24. Şen, B., Peker, M., Çavuşoğlu, A., Çelebi, F.V.: A comparative study on classification of sleep stage based on EEG signals using feature selection and classification algorithms. J. Med. Syst. **38**(3), 1–21 (2014)
25. Diykh, M., Li, Y.: Complex networks approach for EEG signal sleep stages classification. Expert Syst. Appl. **63**, 241–248 (2016)
26. Bach, M., Werner, A., Żywiec, J., Pluskiewicz, W.: The study of under-and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis. Inf. Sci. **384**, 174–190 (2016)
27. Weng, C.-H., Huang, T.C.-K., Han, R.-P.: Disease prediction with different types of neural network classifiers. Telemat. Inform. **33**(2), 277–292 (2016)
28. Zhang, J., Li, H., Gao, Q., Wang, H., Luo, Y.: Detecting anomalies from big network traffic data using an adaptive detection approach. Inf. Sci. **318**, 91–110 (2015). Elsevier Publisher
29. Zhang, J., Gao, Q., Wang, H.: SPOT: a system for detecting projected outliers from high-dimensional data streams. In: 24th IEEE International Conference on Data Engineering (ICDE 2008), pp. 1628–1631. IEEE Computer Society, Cancun, April 2008