A Machine Learning Model to Predict the Onset of Alzheimer Disease using Potential Cerebrospinal Fluid (CSF) Biomarkers

Syed Asif Hassan
Department of Computer Science, Faculty of Computing and Information Technology Rabigh (FCITR)
King Abdulaziz University,
Jeddah, Saudi Arabia

Tabrej Khan
Department of Information Sciences, Faculty of Computing and Information Technology Rabigh (FCITR)
King Abdulaziz University,
Jeddah, Saudi Arabia

***Abstract*—Clinical studies in the past have shown that the pathology of Alzheimer's disease (AD) initiates, 10 to 15 years before the visible clinical symptoms of cognitive impairment starts to appear in AD diagnosed patients. Therefore, early diagnosis of the AD using potential early stage cerebrospinal fluid (CSF) biomarkers will be valuable in designing a clinical trial and proper care of AD patients. Therefore, the goal of our study was to generate a classification model to predict earlier stages of the AD using specific early-stage CSF biomarkers obtained from a clinical Alzheimer dataset. The dataset was segmented into variable sizes and classification models based on three machine learning (ML) algorithms, such as Sequential Minimal Optimization (SMO), Naïve Bayes (NB), and J48 were generated. The efficacy of the models to accurately predict the cognitive impairment status was evaluated and compared using various model performance parameters available in Weka software tool. The current findings show that J48 based classification model can be effectively employed for classifying cognitive impaired Alzheimer patient from normal healthy individuals with an accuracy of 98.82%, area under curve (AUC) value of 0.992 and sensitivity & specificity of 99.19% and 97.87%, respectively. The sample size (60% training and 40% independent test data) showed significant improvement in T-test with J48 algorithm when compared with other classifiers tested on Alzheimer dataset.**

***Keywords*—Alzheimer disease; early-stage biomarker; machine learning algorithm; classification model; accuracy; sensitivity**

I. Introduction

Alzheimer's disease (AD) is a type of dementia that usually affects elderly persons leading to progressive cognitive impairment disorder such as memory loss and a decline in functional abilities of the brain [1], [2]. As per world Alzheimer report, 2016 around 46.8 million people are affected by Alzheimer and related dementia. It is estimated the incidence of Alzheimer will double in every 20 years and by 2050 the prevalence of Alzheimer will be around 131.5 million across the globe [3]. With the current diagnostic technology, only one out of four individuals with the AD is diagnosed [3]. Currently, no permanent cures for AD exist, but there are many treatments which can delay the advancing trait of this disorder. In this regard, it is important to early identify an individual with mild cognitive impairment (MCI) who are most likely of progressing to late stages of the AD. The severity of the AD increases if the Alzheimer is not diagnosed in the earlier stages. Diagnosis of the AD is primarily focused on genetic (Apolipoprotein E genotype) and demographic (gender and age) data, CSF biomarker, neuropsychological test and medical imaging data. Multidimensionality of aforementioned clinical diagnostic factors makes it difficult for us to analyze and infer the information from the same. In this regard, a review describing a computer-based diagnostic method using Random Forest (RF) algorithms on medical imaging data have demonstrated high reliability in classifying early stage MCI patients which later progresses to advanced stages of AD [4]. Similarly, multi-kernel Support Vector Machine (SVM) was employed to predict future clinical symptoms of MCI patients using both baseline and longitudinal multimodal biomarkers data [5]. Various studies related to the implementation of multivariate and ML analysis for the prediction of early stages of the AD from the data obtained from Magnetic Resonance Imaging (MRI), Positron Emission Tomography and CSF biomarkers Data were discussed [6]-[9]. Even though ML methods using neuroimaging data are widely employed for predicting the early stages of AD still the method is inadequately applied to potential low-cost CSF biomarker to detect AD in its initial stages. The biochemical change in the brain due to progressive nature of AD provides a reasonable pool of diagnostic CSF fluid biomarkers. In this regard, a clinical study on the subject with MCI and healthy control was conducted to screen appropriate CSF biomarker required to classify subjects under study as impaired or healthy control [10]. Therefore, the goal of the present study is to generate a classification model using the dataset comprising of early stage CSF biomarker and demographic data generated by Craig-Schapiro et al. 2011 to predict subject with early stages of the AD. The above-mentioned clinical dataset was obtained from a recent Kaggle competition on classifying early stage (AD patient) from healthy subjects.

The proposed classification model will have a remarkable impact on the application of ML-based methods to screen MCI patients before the onset of clinical indicators of the AD. The present research paper is divided into three sections: Section II describes the materials and methods to build a classification model. Section III explains the obtained results of various

classifier based model tested on Alzheimer clinical dataset and the required discussion for the same and Section IV deliver the closing remarks about the current research work and future scope. A summary of the approach involved in building a classification model for screening MCI from healthy control is represented in Fig. 1.

II. Materials and Methods

This section defines the dataset, data preprocessing methodology as well as describes the ML algorithm involved in building a classification model. Furthermore, this section also presents the statistical model performance evaluator of Weka for assessing and comparing the robustness and reliability of the built models.

A. Data Source

The Alzheimer clinical dataset was acquired from Kaggle dataset (https://www.kaggle.com). The clinical dataset describes a clinical study of 333 subjects comprising of MCI patients (n=91) and healthy control subjects (n=242). Data collected from each subject consisted of a set of non-imaging biomarkers namely protein level of amyloid $-\beta 42$ or $A\beta 42$, native Tau protein, phosphorylated form of Tau (pTau), and Apolipoprotein E genotype (E2, E3, and E4) [10].

The most significant variant of Apolipoprotein E genotype is allele E4 which is mostly associated with AD [11]. The data collected on each subject also includes 124 probable CSF biomarker, and other demographic parameters namely gender and age. The goal of the clinical study was to differentiate healthy control from patients with mild cognitive impairment.

B. Processing of Clinical Dataset

1) Preparation of data: To store and process the data in Weka, the clinical dataset obtained from Kaggle was converted into ARFF format [12]. A nominal value namely unhealthy or control for each subject was amended in the last column of the dataset which represented an extra feature labeled as "class".

2) Pre-processing the dataset: Normally not all the parameters in a dataset contribute towards an efficient model building process [13]. The key idea behind screening the best-fit features is to reduce the computation time of the model and decrease the dimensionality of the dataset. In this regard, the feature selection algorithm search across the dataset to present a subset of the attribute that contributes most towards the model building [14], [15]. Basically, the feature selection in Weka is performed by a combination of methods namely an attribute evaluator and search method. In the current study, we have applied InfoGainAttributeEval in combination with Ranker search method. InfoGainAttributeEval assesses a feature based on the information gained with respect to a given class. While Ranker search method gives rank to an attribute based on its evaluation. The list of the feature selected based on above-mentioned technique are listed below:

(1)AXL; (2) Creatine_Kinase_MB; (3) Eotaxin_3; (4) FAS; (5) GRO _alpha; (6) IGF_BP_2; (7) IL_7; (8) MIF; (9) MIP_1alpha; (10) MMP10; (11) MMP7; (12) PAI_1; (13) Pancreatic polypeptide; (14) TRAIL_R3; (15) Thrombopoitein; (16) VEGF; (17) Age; (18) Tau; (19) p_tau; (20) Aβ_42; (21) Male; (22) E4.

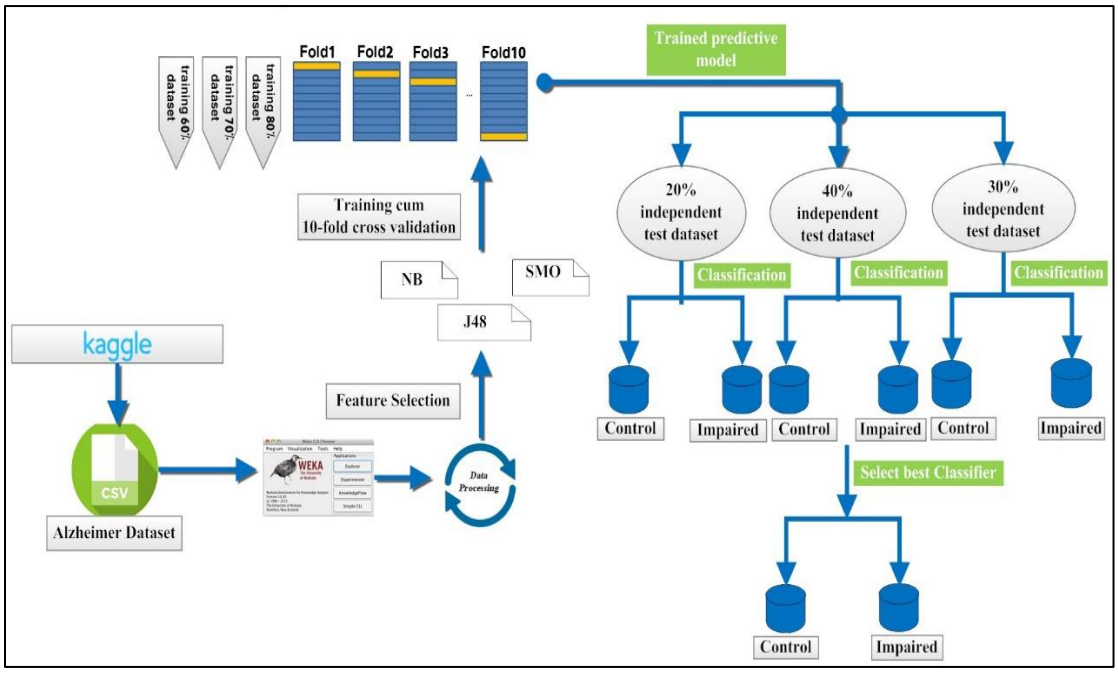

Fig. 1. An illustration representing the activities involved in building a classification model for predicting the early stages of the AD using appropriate CSF Biomarkers.

3) Data segmentation and cross-validation studies: Finally, the subjects of the clinical dataset were systematically arranged based on their respective class. Further using the resample procedure of Weka the clinical dataset was split into varying independent test set i.e., 20%, 30%, and 40%, respectively and training-cum validation set i.e., 60%, 70%, and 80%, respectively across 10 folds, used for the current study. The invertSelection was set to false and noReplacement was set to true for creating subsample of training data of various sizes. While for preparing independent test data of varying sizes both invertSelection and noReplacement was set to true. Independent datasets were generated to evaluate the performance of the trained classification models [16]. The training-cum-cross validation randomly divides each training data of different sizes (i.e., 80%, 70% and 60%) into 10 equal group of data and during each iteration, one group of data is used for testing and remaining n-1 groups are used for training the model with a specific classifier. This process is repeated until each fold have been used as a test fold at least once during the 10 folds cross-validation protocol. The average accuracy of each test fold for a given training data sizes was calculated. Eventually, the each trained model was tested on its respective independent test data. The values obtained for each statistic evaluator for each model tested on individual test dataset provides efficacy of each model to differentiate between the impaired subject from control subjects from any clinical dataset involving AD patient and healthy controls.

C. Machine Learning Algorithms for Model Building

Classification based on ML algorithm assigns subjects based on similar attributes to a specific class. In this paper three best-known ML algorithm namely Naïve Bayes (NB), Sequential Minimal Optimization (SMO) and J48 were used for classifying subjects based on selected attributes into impaired and healthy control. The predictive capability of each model based on various statistical measures was calculated and compared. A short description of the ML algorithms used in the current study to build classification model to differentiate MCI patients from healthy controls are stated as follows:

1) Naive bayes: The NB algorithm relies on the assumptions that each predictive attributes (X_1, X_2… X_n) in the training dataset are conditionally independent. The NB algorithm classifies attributes in the test dataset based on Bayes theorem which calculate the prior possibility and likelihood of an attribute to be classified in any one of the given classes. As per Bayes rule, the prior possibility of an attribute is based on previous experience i.e., in this case, the subjects in test case are classified based on the conditional probability of attributes for a given class. Secondly, the likelihood of a subject to be classified in either of the classes is based on the percentage of subjects in any one of the classes with similar attributes. In NB analysis, the final classification of the subject in a dataset is determined by multiplying both prior and likelihood information regarding an attribute, to form a posterior possibility. The subject with the maximum posterior possibility for attributes for a given class is classified in the same [17], [18]. The NB algorithm can be explained as follows:

Let us assume, that the probability of a subject "X" with attributes $Z = < z_1,...,z_n>$ belongs to class impaired denoted by "I" is represented as follows:

$$P(\frac{I_1}{z_i}) = \frac{P(\frac{z_i}{I_1})P(I_1)}{P(\frac{z_i}{I_1})P(I_1)+P(\frac{z_i}{I_2})P(I_2)} \quad (1)$$

In the present case, $P(I_1)$ is the previous probability linked with class I_1, while $P(I_1|z_i)$ is a posterior probability.

Therefore, for "n" different hypotheses, we have

$$P(z) = \sum_{j=1}^{n} P(\frac{z_i}{I_j})P(I_j) \quad (2)$$

Therefore, we have

$$P(\frac{I_1}{z_i}) = \frac{P(\frac{z_i}{I_1})P(I_1)}{P(n_i)} \quad (3)$$

2) J48: The principle of the C4.5 algorithm is implemented in Java-based decision tree J48 developed by Weka team. The C4.5 algorithm based classification creates decision tree on the basis of information gain i.e., the attribute with maximum information gain is identified as the starting point for splitting. Now, for a given instance if there is no ambiguity regarding the appropriateness of the attribute value for a given dependent variable i.e., class value, then that point is considered as the leaf node. A leaf node in a decision tree specifies the dependent variable i.e., class. If otherwise, then we look for other attributes which provide the next highest information gain. Likewise, we continue from top to bottom along the tree to identify correct combination of attributes for which the data instances have values falling within a particular range of value specific for a given dependent variable [19]. The execution of J48 classification algorithm is shown as follows:

a) Find the normalized information gain for each attribute in a given dataset for a given instance.

b) Let us suppose we found x as the attribute with maximum normalized information gain.

c) Make x as the root node and split the node based on splitting parameter into branches with independent variable values suppose x1 and x2.

d) If the value x1 of the attribute x is considered as a decision point then generate a leaf node and tag it with a specific dependent variable i.e., class.

e) If otherwise, at the branch if some unambiguity exist then find the next attribute with highest information gain by splitting on attribute x, and add those nodes as node for next cycle of selection of children node until a decision point is reached for this path and a particular instance is classified to a specific dependent variable i.e., class label.

3) Sequential Minimal Optimization (SMO): The algorithm Sequential minimal optimization (SMO) is used for solving the problem associated with optimizing linearly constrained quadratic function that appears during the training of support vector machines (SVM) [20]. The quadratic

problem associated with the training of SVM is solved using SMO as follow:

For any binary classification problem for a given dataset such as $(a_1, b_1), \ldots,(a_n, b_n)$, where a_i is the input variable and $b_i \in \{-1, +1\}$denotes the binary tag associated with b_i. The quadratic programming problem in dual form is expressed as follows:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - 1/2 \sum_{i=1}^{n} \sum_{j=1}^{n} b_i b_j K(a_i, a_j)\alpha_i \alpha_j \quad (4)$$

If:

$0 \leq \alpha_i \leq c$, for i=1, 2… n.

$$\sum_{i=1}^{n} b_i \alpha_i = 0 \quad (5)$$

Where K (a_i, a_j) is the kernel function and C is a hyperplane parameter of SVM, and the variables α are Lagrange multipliers. The large QP optimization problem is divided into a sequence of small subproblem by using SMO algorithm. The subproblem is then solved analytically using SMO. Since the Lagrange multiplier (LG) α_i is associated with linear equality constraint. Therefore, the smallest QP subproblem involves two LG i.e., α_1 and α_2. Then the constraints related to each LG (α_1 and α_2), are reduced as follows:

$$0 \leq \alpha_1, \alpha_2 \leq C \quad (6)$$

$$y_1\alpha_1 + y_2\alpha_2 = k \quad (7)$$

Subsequently, the reduced QP quadratic equations are solved logically using the one-dimensional quadratic function. K is fixed in each iteration for solving QP problem in SMO.

D. Classification Model Performances Evaluation

Various classification models trained based on three base classifiers (namely, NB, J48, and SMO) with varying training data sizes (60%, 70%, and 80%) were evaluated using respective independent test data using various statistical measure available in Weka data mining tool. True positive rate (TPR) determines the proportion of predicted True Positives (TP) (i.e., number of correctly classified impaired subjects) from the total number of impaired subjects (i.e., True Positive (TP) + False Negative (FN)) and is calculated as TP/TP + FP. False Positive Rate (FPR) determines the proportion of False Positive (FP) i.e., incorrectly classified as a healthy instance when compared to the total number of predicted impaired instances (TN + FP) and is calculated as FP/FP+TN.

Specificity is defined as the competence of the classification model to predict the negative instances such as TN and FP (i.e., impaired instances in the current study) and is calculated as TN/TN+FP, whereas sensitivity represents the capability of the classification model to identify healthy controls predicted as TP and FN. The classification model which shows higher sensitivity and specificity will always have lower error value. Another, model performance evaluator is accuracy which determines the overall nearness of the predicted accuracy of the model to its ideal value i.e., 1. In this study, accuracy calculates the proportion of accurately classified healthy (TP) and impaired subjects (TN) when compared to the total number predicted instances i.e., TP + TN + FP +FN from a given independent test dataset. The accuracy of the model in general is calculated as ([TP +TN]/[TP +TN +FP +FN]). The Area under the Curve (AUC) value is used to plot the Receiver Operating Characteristic (ROC) curve which determines the reliability of the classifier to predict accurately positive instances in a given dataset. The FPR and TPR of each instance in the dataset are plotted in x and y-axis, respectively to determine the AUC value for each classifier employed in building a classification model. The AUC closer to 1 is considered as the most reliable predictive model.

1) t-test for model evaluation: A two-sample paired t-test was performed to evaluate the significant difference between the classification model of different data sizes built using three important base classifier, namely, NB, J48, and SMO. Studies in the past have used paired t-test to evaluate the significance of a model over other models [21]. The dataset was segmented using Weka resample tool into 20%, 30% and 40% independent test data. The accuracy values obtained for each classifier based model when tested on each independent test data (20%, 30% and 40% data sizes) were grouped and compared for significance using paired sample t-test.

2) Gain and lift chart analysis: Gain and lift is a measure of the effectiveness of a classification model calculated as the ratio between the results of TP obtained with and without the model. The greater the area between the lift curve and the baseline, the better the model. Moreover, lift chart shows how much more likely we are able to predict impaired instances accurately than if we do the screening of impaired instances without the use of any classification model [22]. A comparative gain and lift chart analysis was performed between the classification models for a specific independent testing data size that showed better results to screen impaired instances from the given clinical dataset.

III. Results and Discussion

The AD clinical dataset obtained from Kaggle dataset consisted of 91 patients with MIC and rest 242 were healthy subjects. The dataset obtained from Kaggle was converted into an ARFF format using Weka. The training and testing of different classifier were based on independent variables i.e., the non-imaging biomarkers obtained from each subject (instance) involved in the clinical study. The class label (i.e., Healthy control or Impaired) was assigned as the dependent variable of the clinical dataset. The subset of a feature or independent variable which contributes most towards model building was selected using InfoGainAttributeEval in combination with Ranker search method. The dataset was modified accordingly, that is, constituting of a subset of the independent variable for all instances (subjects) involved in the clinical study. The modified dataset with selected features was segmented into 80%, 70% and 60% training data and 20% and 30% and 40% independent test data. The model based on each base classifier algorithm namely NB, J48 and SMO were trained using 80%, 70%, and 60% training data. Subsequently, the trained classification models were tested on independent test data i.e., 20%, 30% and 40%, respectively.

TABLE I. THE PERFORMANCE STATISTICS OF NB, SMO AND J48 CLASSIFIERS BASED MODEL TESTED ON 20 %, 30 AND 40 % INDEPENDENT TEST DATA

Classifier	Test data size	Ac	ROC	Sn	Sp	TPR	FPR
J48	20 %	95.35	0.972	83.33	100.00	0.833	0.000
	30 %	96.50	0.979	93.75	97.56	0.938	0.024
	40 %	96.92	0.985	100.00	95.74	1.000	0.043
NB	20 %	83.72	0.868	66.67	90.32	0.667	0.097
	30 %	82.46	0.834	68.75	87.80	0.688	0.122
	40 %	76.92	0.872	66.67	80.85	0.667	0.191
SMO	20 %	90.70	0.807	66.67	100.00	0.667	0.000
	30 %	87.72	0.800	62.50	97.56	0.625	0.024
	40 %	87.70	0.812	66.67	95.74	0.667	0.043

Here Ac = Accuracy; ROC = Receiver Operating Characteristic; Sn = Sensitivity; Sp = Specificity; TPR = True Positive Rate; FPR = False Positive Rate

The results shown in Table I provide a comparative performance evaluation of different classification model based on three important base classifiers, namely, NB, SMO, and J48. The results of the various statistical evaluators for the model performance of each classifier based model are based on independent test data. The results obtained by using base classifiers (NB, J48, and SMO) on 60% training data and 40% independent testing data showed better results as compared to other learning and testing data sizes as shown in Table I. The J48 based classification model showed an accuracy value of 98.96%, which is far better when compared to the results obtained for both SMO and NB. A measure of sensitivity assesses the ability of the classification model to accurately screen TP instances from a dataset (i.e., the impaired sample in the present study), while specificity evaluates the ability of the classification model to accurately screen TN (i.e., healthy control in the present study) instances from a given dataset. A classification model which achieves 100% sensitivity and specificity in predicting TP and TN instances from a dataset is considered as an ideal model. As shown in Table I, each base classifier based model showed better sensitivity and specificity with 40% independent testing data. A comparative sensitivity and specificity analysis of NB, J48, and SMO based classification model on 40% independent test data are illustrated in Fig. 2. The J48 based classification model with a sensitivity of 100% was found to be an ideal system to screen impaired instances (TP) from a given clinical dataset. Moreover, the same model was found be highly specific in screening healthy control instances (TN) from the clinical dataset with a specificity percentage of 97.56%.

While, the NB based classification model showed lower sensitivity and specificity to predict TP and TN instances in a given dataset as compared to a model built using J48 and SMO, respectively. The efficacy of the classification models to distinguish between TP and TN was determined by plotting a ROC curve using the AUC values generated by plotting the TPR and FPR of each instance in the dataset. In the present study, the ROC curve analysis demonstrates the accuracy of classification models to accurately discriminate impaired from healthy control instances present in 40% independent test data as shown in Fig. 3. It can be observed from the comparative ROC plot of different classifier based model, that the classifier J48 showed a maximum AUC value (i.e., 0.985) as compared to SMO and NB based classification model. In the present study, the classification accuracy in terms of AUC value was minimum for SMO (i.e., 0.862) and maximum for J48 (i.e., 0.985) classifier based model. Statistically, the ideal value of AUC is 1, therefore, the model with AUC value closer to 1 is considered significant in discriminating binary two class dataset.

The basic concept of applying classification model is to enhance the accuracy of screening TP from a given set of instances in a dataset when compared to random screening. The gain or lift in the screening of TP i.e., accuracy by any given classification model can be determined by plotting a gain or a lift chart between the cumulative percentage of instances on the X-axis and cumulative percentage of TP on the Y-axis. A comparative evaluation of enrichment potential of NB, SMO and J48 classification model to screen TP as compared to random screening was performed and is illustrated in Fig. 4 and 5, respectively.

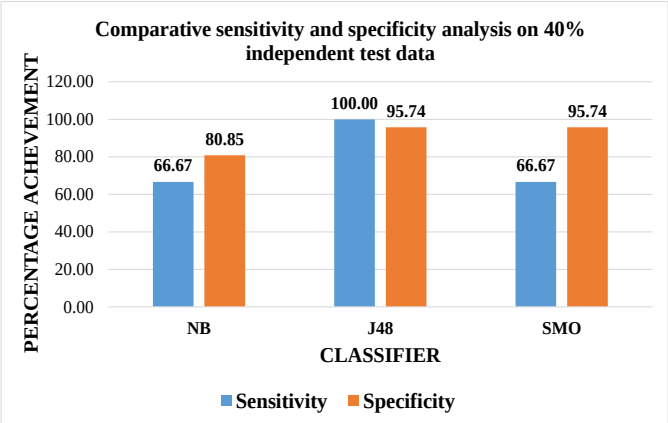

Fig. 2. Comparative study of Sensitivity (Sn) and Specificity (Sp) of NB, SMO and J48machine learning algorithm based classification model.

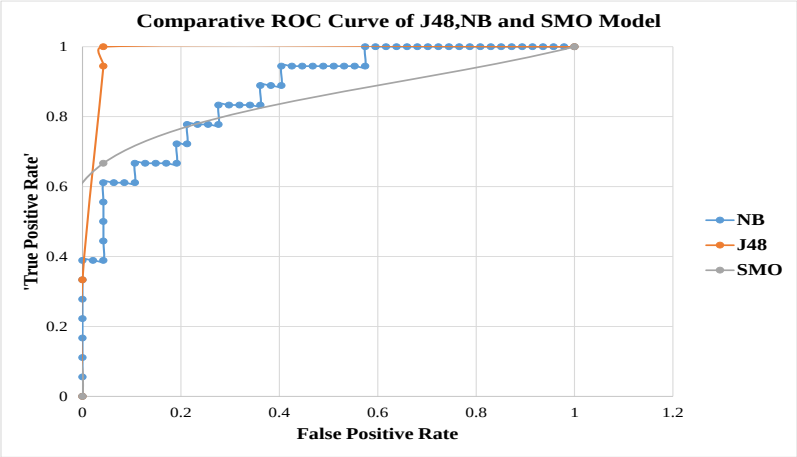

Fig. 3. Comparative plot of ROC representing the AUC values of J48, Sequential Minimal Optimization (SMO) and Naïve Bayes (NB).

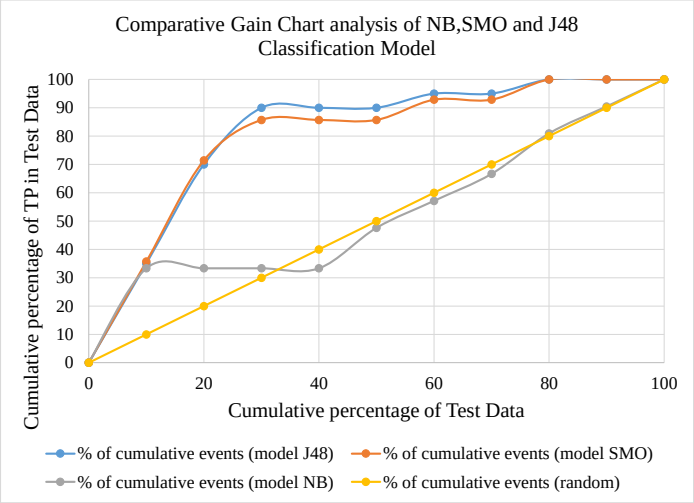

Fig. 4. Comparative gain chart analysis of NB, SMO and J48 classification model over random screening method.

As shown in Fig. 4 and 5 the gain or lift for J48 based model tested on top 10%, 20% and 30% of the 40% test data show a lift or gain of 3.5, 3.5 and 3.0, respectively. Similar gains were obtained for SMO for top 10 and 20% of the test instances. However, for SMO based model the lift values for the remaining instances (i.e., 30% to 100%) of data were comparatively lower than J48 based model. Moreover, the gain or lift values for NB based model were far inferior when compared to SMO and J48 base predictive model. These gain or lift values obtained from J48 classification model show that an enrichment of more than a fold of TP's can be attained using the J48 model as compared to any other random screening protocols. Since in the present study, the J48 model was found have better gain or lift (i.e., the ratio of TP obtained with and without the model) values as compared to NB and SMO based classification model. Therefore, the J48 classifier based model is recommended as a reliable model to discriminate and screen cognitively impaired individuals from a given Alzheimer dataset.

The statistical significance of the J48 classifier based classification model over SMO and NB classifier based model was evaluated using paired sample t-test. The accuracy obtained by J48 based classifier when tested on 20, 30 and 40 % independent test data was compared with SMO and NB based model. The mean, standard deviation, standard error and significance value obtained by comparing the accuracy results of J48 & NB and J48 & SMO when tested on various test data are tabulated in Tables II and III, respectively. The significance value of 0.026 and 0.035 was obtained when the results of 20%, 30% and 40% independent test data of J48 was compared with NB and SMO, respectively. The significance value obtained show that the accuracy results obtained by the J48 based classification model over SMO and NB are statistically significant as the generated significance values are lower than 0.05.

Even though, neuroimaging data are widely used to classify subjects with early stages of the AD, the novelty in our approach is to adequately apply low-cost CSF biomarker to detect AD in its initial stages. Therefore the present study provides a novel CSF biomarker-based classification tool to efficiently classify a subject with an early stage of cognitive impairment from healthy subjects with higher accuracy and sensitivity.

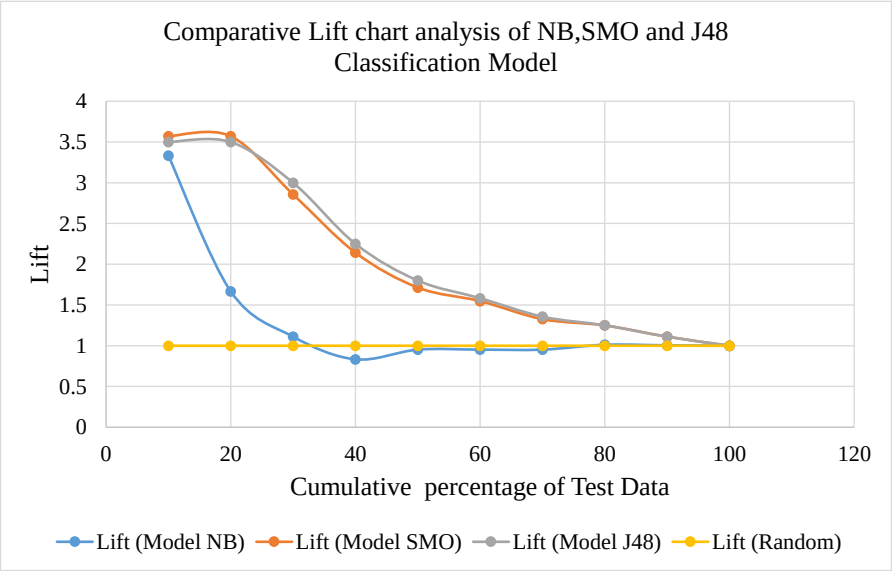

Fig. 5. Comparative lift chart analysis of NB, SMO and J48 classification model over random screening method.

TABLE II. PAIRED SAMPLES TEST BETWEEN J48 AND NB CLASSIFICATION MODEL

Algorithms	**Paired Differences**							
	Mean	***Std. Deviation***	***Std. Error Mean***	***95% Confidence Interval of the Difference***		***t***	***df***	***Sig. (2-tailed)***
				Lower	***Upper***			
J48 and NB	15.22333	4.30865	2.48760	4.52006	25.92660	6.120	2	.026

TABLE III. PAIRED SAMPLES TEST BETWEEN J48 AND SMO CLASSIFICATION MODEL

Algorithms	**Paired Differences**							
	Mean	***Std. Deviation***	***Std. Error Mean***	***95% Confidence Interval of the Difference***		***t***	***df***	***Sig. (2-tailed)***
				Lower	***Upper***			
J48 and SMO	7.55000	2.52109	1.45555	1.28726	13.81274	5.187	2	.035

IV. CONCLUSION AND FUTURE SCOPE

In the present study, we have proposed a supervised classification model based on a J48 algorithm that can efficiently discriminate between patients with MCI and healthy subjects using clinical CSF biomarkers. The ability of the model to predict patients with early stages of the AD was based on appropriate training attributes selected using feature selection method. The efficiency of the model built using NB, J48 and SMO were evaluated using various statistical performance evaluators and compared. Based on the performance J48 based classification model was selected as the best model to discriminate between the given dependent variable (MCI patients and healthy controls) with high accuracy, sensitivity, and specificity.

The significance of the accuracy obtained by J48 on various independent test data sizes was compared with models based on NB and SMO, respectively and was found to be significant by paired two-tailed t-test at 0.1 significance level. The comparative lift and gain chart analysis of the models on independent test data showed that J48 based model can enhance the prediction of the MCI subjects by three folds. Therefore, the present study is a step forward in predicting the early stages of Alzheimer disease using the ML-based classification model based on early stage CSF biomarkers. In future, the authors have planned to build an online prediction system to screen subjects with initial stages of cognitive impairment using the early stage biomarker attributes of the clinical Alzheimer dataset.

ACKNOWLEDGEMENT

We are thankful to the Faculty of Computing and Information Technology Rabigh (FCITR) of King Abdulaziz University, Jeddah for providing the state of art facility to perform our experiments.

REFERENCES

[1] N. C. Berchtold and C. W. Cotman, “Evolution in the conceptualization of dementia and Alzheimer’s disease: Greco-Roman period to the 1960s.,” Neurobiol. Aging, vol. 19, no. 3, pp. 173–189, 1998.

[2] A. Collie and P. Maruff, “The neuropsychology of preclinical Alzheimer’s disease and mild cognitive impairment.,” Neurosci. Biobehav. Rev., vol. 24, no. 3, pp. 365–374, May 2000.

[3] M. Prince, A. Comas-Herrera, M. Knapp, M. Guerchet, and M. Karagiannidou, “World Alzheimer Report 2016 Improving healthcare for people living with dementia. Coverage, Quality and costs now and in the future,” Alzheimer’s Dis. Int., pp. 1–140, 2016.

[4] A. Sarica, A. Cerasa, and A. Quattrone, "Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer’s Disease: A Systematic Review," Front Aging Neurosci., vol. 9, 2017.

[5] D. Zhang and D. Shen, “Predicting Future Clinical Changes of MCI Patients Using Longitudinal and Multimodal Biomarkers,” PLoS One, vol. 7, no. 3, p. e33182, Mar. 2012.

[6] F. Falahati, E. Westman, and A. Simmons, “Multivariate data analysis and machine learning in Alzheimer’s disease with a focus on structural magnetic resonance imaging.,” J. Alzheimers. Dis., vol. 41, no. 3, pp. 685–708, 2014.

[7] P. T. Trzepacz, P. Yu, J. Sun, K. Schuh, M. Case, M. M. Witte, H. Hochstetler, and A. Hake, “Comparison of neuroimaging modalities for the prediction of conversion from mild cognitive impairment to Alzheimer’s dementia.,” Neurobiol. Aging, vol. 35, no. 1, pp. 143–151, Jan. 2014.

[8] A. Khan and M. Usman, “Early diagnosis of Alzheimer’s disease using machine learning techniques: A review paper,” in 2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 2015, vol. 01, pp. 380–387.

[9] X. Long, L. Chen, C. Jiang, and L. Zhang, “Prediction and classification of Alzheimer disease based on quantification of MRI deformation,” PLoS One, vol. 12, no. 3, p. e0173372, Mar. 2017.

[10] R. Craig-Schapiro, M. Kuhn, C. Xiong, E. H. Pickering, J. Liu, T. P. Misko, R. J. Perrin, K. R. Bales, H. Soares, A. M. Fagan, and D. M. Holtzman, “Multiplexed Immunoassay Panel Identifies Novel CSF Biomarkers for Alzheimer’s Disease Diagnosis and Prognosis,” PLoS One, vol. 6, no. 4, p. e18850, Apr. 2011.

[11] J. Kim, J. M. Basak, and D. M. Holtzman, “The role of apolipoprotein E in Alzheimer’s disease.,” Neuron, vol. 63, no. 3, pp. 287–303, Aug. 2009.

[12] K. P. Soman, S. Diwakar, and V. Ajay, “Insight into data mining : theory and practice,” p. 403, 2006.

[13] A. Z. Dudek, T. Arodz, and J. Galvez, “Computational methods in developing quantitative structure-activity relationships (QSAR): a review.,” Comb. Chem. High Throughput Screen., vol. 9, no. 3, pp. 213–228, Mar. 2006.

[14] M. Yuwono, Y. Guo, J. Wall, J. Li, S. West, G. Platt, and S. W. Su, “Unsupervised feature selection using swarm intelligence and consensus clustering for automatic fault detection and diagnosis in Heating Ventilation and Air Conditioning systems,” Appl. Soft Comput., vol. 34, no. Supplement C, pp. 402–425, 2015.

[15] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, “Inductive learning algorithms and representations for text categorization,” in Proceedings of the seventh international conference on Information and knowledge management - CIKM ’98, 1998, pp. 148–155.

[16] R. H. Fagard, “Exercise characteristics and the blood pressure response to dynamic physical training.,” Med. Sci. Sports Exerc., vol. 33, no. 6 Suppl, pp. S484–92; discussion S493–4, Jun. 2001.

[17] N. Friedman, D. Geiger, and M. Goldszmidt, “Bayesian Network Classifiers,” Mach. Learn., vol. 29, no. 2, pp. 131–163, 1997.

[18] T. R. Patil, “Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification,” Int. J. Comput. Sci. Appl. ISSN 0974-1011, vol. 6, no. 2, pp. 256–261, 2013.

[19] J. R. (John R. Quinlan and J. Ross, C4.5 : programs for machine learning. Morgan Kaufmann Publishers, 1993.

[20] J. Platt, “Sequential minimal optimization: A fast algorithm for training support vector machines,” 1998.

[21] A. H. Osman, N. Salim, M. S. Binwahlan, R. Alteeb, and A. Abuobieda, “An improved plagiarism detection scheme based on semantic role labeling,” Appl. Soft Comput., vol. 12, no. 5, pp. 1493–1502, 2012.

[22] M. Bichler and C. Kiss, “A Comparison of Logistic Regression, k-Nearest Neighbor, and Decision Tree Induction for Campaign Management,” AMCIS 2004 Proc., Dec. 2004