Hindawi
Mobile Information Systems
Volume 2020, Article ID 8870240, 9 pages
https://doi.org/10.1155/2020/8870240

Research Article

Development of Hepatitis Disease Detection System by Exploiting Sparsity in Linear Support Vector Machine to Improve Strength of AdaBoost Ensemble Model

Wasif Akbar,[1] **Wei-ping Wu**,[1] **Sehrish Saleem**,[2] **Muhammad Farhan**,[3] **Muhammad Asim Saleem**,[4] **Ashir Javeed**,[4] **and Liaqat Ali**[5,6]

[1]*School of Computer Science and Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China*
[2]*Department of Computer Science, MNS University of Engineering and Technology Multan, Multan, Pakistan*
[3]*Department of Computer Science, COMSATS University Islamabad,Lahore Campus, Lahore, Pakistan*
[4]*School of Information and Software Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China*
[5]*School of Information and Communication Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China*
[6]*Department of Electrical Engineering, University of Science and Technology, Bannu, Pakistan*

Correspondence should be addressed to Wasif Akbar; sewasif@hotmail.com

Received 19 March 2020; Revised 22 June 2020; Accepted 9 October 2020; Published 3 November 2020

Academic Editor: Ali Kashif Bashir

Hepatitis disease is a deadliest disease. The management and diagnosis of hepatitis disease is expensive and requires high level of human expertise which poses challenges for the health care system in underdeveloped and developing countries. Hence, development of automated methods for accurate prediction of hepatitis disease is inevitable. In this paper, we develop a diagnostic system which hybridizes a linear support vector machine (SVM) model with adaptive boosting (AdaBoost) model. We exploit sparsity in linear SVM that is caused by L_1 regularization. The sparse L_1-regularized SVM is capable of eliminating redundant or irrelevant features from feature space. After filtering features through the sparse linear SVM, the output of the SVM is applied to the AdaBoost ensemble model which is used for classification purposes. Two types of numerical experiments are performed on the clinical features of hepatitis disease collected from UCI machine learning repository. In the first experiment, only conventional AdaBoost model is used, while in the second experiment, a feature vector is applied to the sparse linear SVM before its application to the AdaBoost model. Simulation results demonstrate that the strength of a conventional AdaBoost model is enhanced by 6.39% by the proposed method, and its time complexity is also reduced. In addition, the proposed method shows better performance than many previously developed methods for hepatitis disease prediction.

1. Introduction

Hepatitis is considered a major chronic liver disease worldwide. The liver is considered to be the heaviest and one of the largest organs of the human body [1]. The liver is one of the key organs of a human body responsible for different functions. These functions include bile secretion, protein formation, and elimination of toxins from body. Hence, inflammation of liver (caused by hepatitis) results in dysfunction of the liver, and consequently, the health of the subject is deteriorated. The symptoms of hepatitis are different in different patients, with some subjects showing no signs. Well-known symptoms include yellowish eyes and skin, abdominal pain, poor appetite, and tiredness [2, 3]. Hepatitis can be acute or chronic depending on duration. If it lasts for less than six months, it is acute; however, if it lasts for more than six months, it is chronic [4]. It has been reported that hepatitis results in more than a million deaths

each year. Diagnosis of hepatitis through conventional methods is a difficult job and requires expensive medical tests [5]. Additionally, the diagnosis of such disease through intelligent system reduces the cost and also examines the patient in shorter time. Hence, development of intelligent diagnostic systems for such type of disease prediction is very important.

In the past, numerous hybrid models for disease detection have been developed by different researchers. These include automated systems for Parkinson's disease prediction [6–8], mortality prediction [9, 10], cancer detection [11, 12], and heart disease [6, 13, 14]. These models are developed by hybridizing data mining models (for feature preprocessing) such as principal component analysis (PCA) and Fisher discriminant analysis (FDA) with machine learning models such as decision trees, logistic regression, support vector machine (SVM), Naive Bayes, neural network models, ensembles of neural networks, K-nearest neighbors, deep neural networks, and optimized and stacked SVMs [15–24]. For example, Adamczak developed different automated models for hepatitis prediction. These models include MLP + BP, RBF (Tooldiag), and FSM without rotation and achieved a prediction accuracy of 77.4%, 79%, and 88.5%, respectively [25]. In another study conducted by Passi, MLO was developed for hepatitis which resulted in hepatitis prediction of 79.70% [26, 27]. Stern and Dobnikar developed AIS, LDA, and FDA models which achieved the hepatitis prediction accuracy of 82%, 84.5%, and 86.40%, respectively [27]. Nilashi et al. developed KNN, ANFIS, NN, and SVM and achieved hepatitis prediction accuracy of 71.41%, 79.67%, 78.31%, and 81.17%, respectively [28]. Recently, Polat and Gunes discussed the hybridization of the feature extraction through the principal component analysis model with classification through artificial immune recognition system for the prediction of hepatitis disease [1, 29].

In this paper, we develop a hybrid intelligent diagnostic system. To improve the strength of AdaBoost predictive model, we propose to use L_1-penalized linear SVM. The L_1 penalty makes the linear SVM sparse, thus making it capable of eliminating redundant features by making their coefficients zero through sparse solutions. After elimination of redundant features through the sparse linear SVM, the remaining features are supplied to the AdaBoost model for classification. In order to analyze the impact of the sparse linear SVM on the AdaBoost model, we performed two types of numerical experiments. In the first experiment, we developed the conventional AdaBoost model, while in the second experiment, we constructed a learning system by stacking the sparse SVM with the AdaBoost model. The performance of both the models, developed in the two experiments, was evaluated using an online hepatitis disease data. Experimental results demonstrated that the sparse linear SVM enhances the accuracy of conventional AdaBoost (for the hepatitis disease prediction based on the collected clinical features). Additionally, the sparse linear SVM also reduces AdaBoost model's complexity as the optimal subset of features contains less number of features.

The rest of the manuscript is organized as follows. Datasets, the proposed sparse linear SVM, and AdaBoost-based learning system are elaborated in Section 2. Section 3 discusses various schemes for validation as well as multiple metrics for evaluation used in the manuscript. Section 4 discusses experimental setup and obtained results, whereas the last section concludes the paper.

2. Materials and Methods

2.1. Dataset Description. The hepatitis dataset consists of 155 samples, and each sample contains 19 features. Details about the 19 commonly used features for the hepatitis dataset are given in Table 1. The label of the dataset is binary, i.e., it can have a value of 1 or 2, where 1 means the sample belongs to a patient who died, while 2 means the sample is that of a subject who survived. There are 32 samples having label 1 and 123 samples having the label value of 2, i.e., the dataset contains 123 samples belonging to healthy class and 32 samples belonging to patient class. In machine learning, we split the data into two parts, namely, training and testing. The training part is used to train the model, and its performance is checked by testing the trained model on the testing data. In this study, the dataset is divided into training and testing datasets using 70–30 data portioning. Hence, out of the 155 samples, 108 samples are used for training purposes, and the remaining 47 samples are used for testing purposes. Out of the 108 training samples, 23 samples belong to the patient class, and 85 patients belong to healthy class. On the other hand, out of the 47 testing samples, 7 samples belong to the patient group, and 38 samples belong to the healthy group. It can be noticed that lower class distribution of the patient class is a limitation of the dataset.

2.2. Proposed Method. As discussed above, in this paper, we exploit the sparsity in linear SVM to improve the strength of machine learning models, namely, k-nearest neighbours (KNN), Gaussian Naive Bayes (GNB), linear discriminant analysis (LDA), and AdaBoost ensemble model. Initially, L_1-penalized linear SVM is used to generate sparse features, i.e., to process the full set of features, null the redundant features, and yield a subset of features containing relevant features only. The generated subset of features by sparse linear SVM is supplied to machine learning models for classification purposes. The sparsity of the linear SVM is controlled by its hyperparameter λ. Hence, for distinct values of λ, various distinct features will be nullified resulting in different subsets of features. Thus, for achieving better hepatitis prediction accuracies, it is necessary to develop a sparse linear SVM that would nullify the most redundant or irrelevant features and generate a subset of the most relevant features. This can be accomplished by tuning the hyperparameter λ. In order to better comprehend the functioning of the proposed learning system, it is pertinent to briefly discuss the L_1-penalized linear SVM model and its formulation. The formulation is as follows.

Support vector machines (SVMs) are considered powerful learning methods and have been widely used in different biomedical- and health informatics-related problems [30]. During the training process, SVM tries to construct an

Table 1: Details of the 19 hepatitis features.

Feature no	Feature code	Feature description	Values
1	D_1	Age	10, 20, 30, . . . , 70, 80
2	D_2	Sex	Male, female
3	D_3	Steroid	1, 2
4	D_4	Antivirals	1, 2
5	D_5	Fatigue	1, 2
6	D_6	Malaise	1, 2
7	D_7	Anorexia	1, 2
8	D_8	Liver big	1, 2
9	D_9	Liver firm	1, 2
10	D_{10}	Spleen palpable	1, 2
11	D_{11}	Spiders	1, 2
12	D_{12}	Ascites	1, 2
13	D_{13}	Varices	1, 2
14	D_{14}	Bilirubin	0.39, 0.8, 1.2, 2.0, 3.0, 4.0
15	D_{15}	Alkaline phosphatase	33, 80, 120, 160, 200, 250
16	D_{16}	SGOT	100, 200, 300, 400, 500
17	D_{17}	Albumin	2.1, 3.0, 3.8, 4.5, 5.0, 6.0
18	D_{18}	Protime	10, 20, 30, 40, . . . , 80, 90
19	D_{19}	Histology	1, 2

optimal hyperplane that can better differentiate the data points of the two classes (in case of binary classification) [31]. The major reason that motivates machine learning researchers to use SVM for their problems is that SVMs have powerful generalization capabilities to unseen data and they depend on very small number of hyperparameters [32].

Considering a dataset D_S with S instances $D = \{(p_i, q_i) | p_i \in R^Q, q_i \in \{-1, 1\}\}_{i=1}^{S}$, where p_i stands for i^{th} instance, Q represents the dimension of the original feature space of hepatitis data, and q_i denotes the class labels, i.e., presence or absence of hepatitis disease. The value is 19 for the hepatitis dataset considered in this paper. The SVM model determines a hyperplane calculated by $g(x) = \beta^T * x + \delta$, where δ represents the bias and β denotes the weight vector. Based on the training data, the hyperplane $g(x)$ of SVM augments the margin, whereas it curtails the classification error [33]. The sum of the distances between the closest negative and closest positive instances is called margin. In other words, the hyperplane augments the margin distance $2/\|\beta\|_2^2$.

SVM uses a set of slack variables denoted by θ_i, $i = 1, \ldots, S$ and a penalty parameter, i.e., λ, and attempts to maximize $\|\beta\|_2^2$ and minimize the errors of misclassification [34]. This fact is formulated as follows:

$$\min_{\beta,\delta,\theta} \underbrace{\frac{1}{2}\|\beta\|_2^2}_{\text{Regularizer}} + \lambda \underbrace{\sum_{i=1}^{S} \theta_i}_{\text{Error loss}}, \quad \text{subject to} \begin{cases} y_i(\beta x_i + \delta) \geq 1 - \xi_i \\ \xi_i \geq 0, i = 1, \ldots, S \end{cases}, \tag{1}$$

where θ is the slack variable that calibrates the degree of misclassification and Euclidean norm or L_2-norm is the penalty term. A varied version of SVM was introduced by Bradley and Mangasarian which replaces the Euclidean norm, i.e., L_2-norm with L_1-penalty function [35]. The L_1-penalized SVM produces sparse solutions and has the feature selection property due to its competence of overthrowing irrelevant or noisy features automatically and hence can be used for feature selection. The formulation of L_1-penalized SVM is given as follows:

$$\min_{\beta,\beta,\xi} \underbrace{\|\beta\|_1}_{\text{Regularizer}} + \lambda \underbrace{\sum_{i=1}^{S} \theta_i}_{\text{Error loss}}, \quad \text{subject to} \begin{cases} y_i(\beta x_i + \delta) \geq 1 - \theta_i \\ \theta_i \geq 0, i = 1, \ldots, S. \end{cases} \tag{2}$$

From the above formulas, it can be seen that, for different settings of the hyperparameter of the L_1 SVM, i.e., λ, different features will be nulled; consequently, a different subset of features will be produced [36]. The goal is to tune the value of λ in such a way to produce a subset of features which will show best performance in terms of hepatitis disease prediction accuracies. This is done by using exhaustive search methodology. After production of the features' subset, its application to AdaBoost machine learning models is carried out. The AdaBoost model is used for classification task.

AdaBoost (also known as adaptive boosting classifier) is an ensemble learning model. It utilizes boosting approach to construct a metaclassifier by combining the strengths of base classifiers, i.e., weak estimators. The boosting operation helps convert the weak estimators into a stronger or boosted model. During the process of boosting, weighted sum of the base learners or estimators is evaluated to produce the final output of the boosted model. This fact is reflected in the following formulation:

$$G(x) = \text{sign}\left(\sum_{m=1}^{M} \alpha_m B_m(x)\right), \tag{3}$$

where the m^{th} base classifier is denoted by B_m and α_m denotes the weight of the m^{th} classifier or estimator. To implement the AdaBoost model, we used scikit-learn python API [37]. In the following discussion, E denotes the total

number of classifiers or estimators used for constructing the eventual AdaBoost model.

The primary objective of this paper is to investigate and exploit the sparsity in the linear L_1-regularized SVM to further improve the strength of the AdaBoost model. To meet this objective, we develop a cascade of the L_1 linear sparse SVM and AdaBoost model. The full feature set is supplied at the input of L_1 SVM which produces different subset of features based on the value of its hyperparameter λ. Performance of the subset of features is evaluated by their application to AdaBoost model. Thus, in the initial stages, we need to discretize the λ hyperparameter. After discretization of λ, we will have to search the optimal value of λ that will produce optimal subset of features which will show best classification performance. The whole process of the proposed method is shown in the Figure 1. From the figure, it can be seen that initially, a subset of features is generated by utilizing a specific value of λ. The subset of features is given to the AdaBoost model which is trained using one value of E. For the subset of features, performance is evaluated under optimal E. Furthermore, another subset of features is generated by utilizing another discrete value of λ, and again the AdaBoost model is trained and evaluated under optimal value of E. The process is repeated until all the subset of features are evaluated and tested. At the end, the optimal subset of features is selected based on the performance.

3. Evaluation of the Proposed Method

In literature, different researchers have utilized various metrics for performance evaluation of their proposed methods. However, for a more realistic evaluation of the performance of our proposed method, we utilized the following five evaluation metrics known as accuracy (ACC), specificity (Spec.), sensitivity (Sen.), and Matthews correlation coefficient (MCC). Accuracy gives information about the total number of correctly classified subjects (whether healthy or patients). Specificity conveys information about the number of healthy subjects which are classified correctly. Similarly, sensitivity represents the percentage of subjects which are classified correctly. MCC is used to measure the quality of binary classification. The basic formulas for these metrics are given as follows:

$$\begin{aligned}
\text{ACC} &= \frac{tp + tn}{tp + fp + tn + fn}, \\
\text{Sen} &= \frac{tp}{tp + fn}, \\
\text{Spec} &= \frac{tn}{tn + fp}, \\
\text{MCC} &= \frac{tp \times tn - fp \times fn}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}}.
\end{aligned} \tag{4}$$

4. Results and Discussion

In this section, the experimental setting and the obtained results are analyzed and discussed. All the experiments (including conventional machine learning-based experiments and the proposed method-based experiments) are performed using Python software (scikit-learn). The experiments were simulated using Intel Core i5 processor with 8 GB RAM and 64-bit operating systems. For the purpose of comparison, we performed two types of experiments. First, the conventional AdaBoost model is developed for the prediction of hepatitis disease. Second, the proposed hybrid model is developed to predict hepatitis disease based on the filtered set of features.

4.1. Simulation of Conventional AdaBoost Model on Hepatitis Data. In this experiment, we develop the conventional AdaBoost model for the hepatitis disease data. The model is trained using 70% of the dataset and tested on the remaining 30% of the data. An exhaustive grid search algorithm is used to search the optimized version of the AdaBoost model. The results on both optimal hyperparameters and nonoptimal hyperparameters are given in Table 2. It is evident from the table that best performance of 82.97% accuracy, 11.11% sensitivity, 100% specificity, and MCC of 0.302 is obtained at optimal hyperparameter, i.e., $E = 3$.

4.2. Simulation of the Proposed Method Using the Sparse Linear SVM and AdaBoost Model on Hepatitis Data. In this experiment, the proposed learning system is developed by using both the models, i.e., sparse linear SVM and AdaBoost model. The simulation results are reported in Table 3. As can be seen in the table, different values of λ for the sparse SVM generate different subsets of features with different sizes. For subset of features with sizes from $N = 1$–10, no improvement in the performance is observed. However, from $N = 10$ onwards, we see changes in performance of the system. It is evident from the table that best performance of 89.36% is obtained at $N = 16$, i.e., with subset of features having only 16 features. However, the best performance on full feature set, i.e., on conventional AdaBoost is 82.97% which is shown in the last row of the table. Hence, it can be observed that coupling the conventional AdaBoost model with sparse linear SVM model improves the performance by 6.39%.

To statistically analyze the results on the testing data, we utilize confusion matrix. As discussed above, the dataset is divided into training and testing datasets using 70–30 data portioning. Hence, out of the 155 samples, 108 samples are used for training purposes, and the remaining 47 samples are used for testing purposes. Out of the 108 training samples, 23 samples belong to the patient class, and 85 patients belong to healthy class. On the other hand, out of the 47 testing samples, 7 samples belong to the patient group, and 38 samples belong to the healthy group. The predicted results of the proposed L_1 SVM-AdaBoost model are depicted statistically in the confusion matrix in Figure 2.

To further show that the coupling of the sparse linear SVM with conventional AdaBoost model enhances the

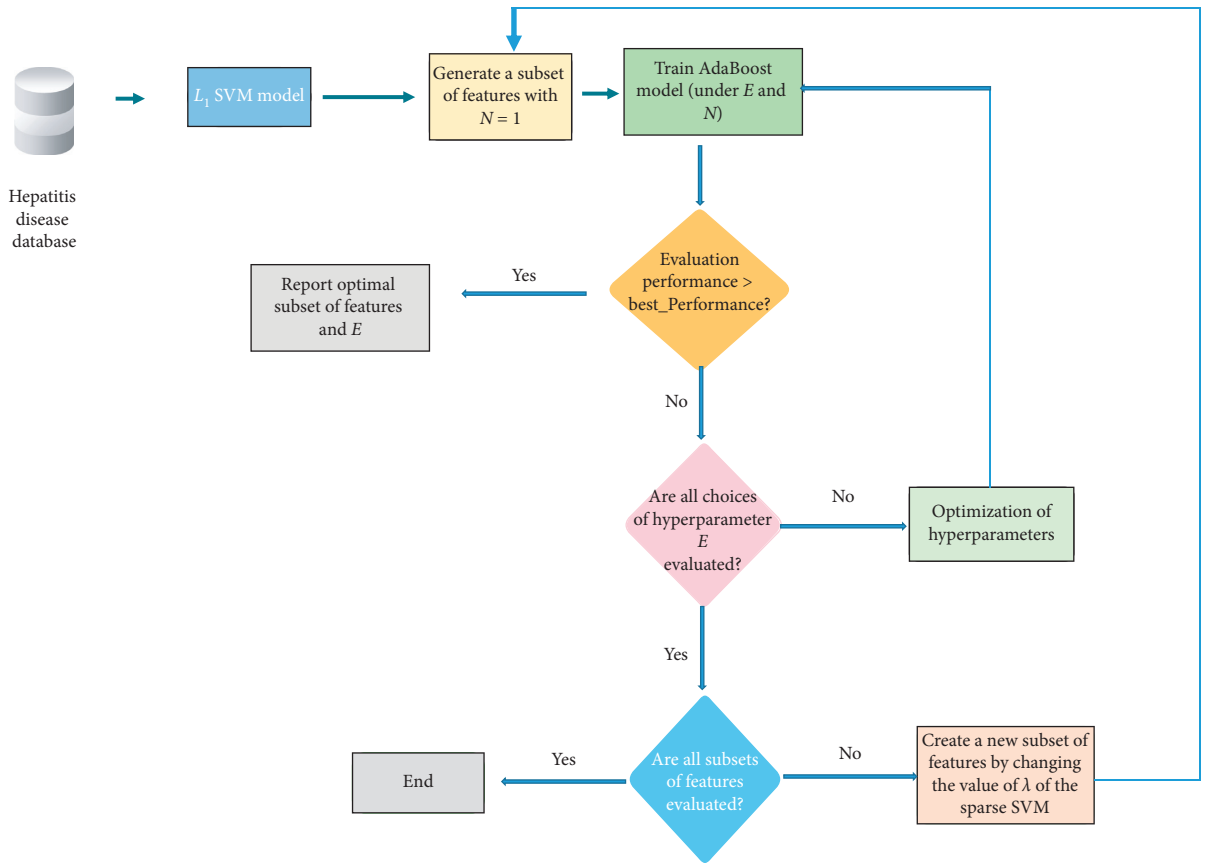

Figure 1: Block diagram of the proposed diagnostic system. E: number of estimators used by the AdaBoost model, N: size of subset of features, and λ: hyperparameter of the L_1 SVM model that controls the sparsity.

Table 2: Performance of the conventional AdaBoost model on HF data.

E	Acc_{test}	$\text{Acc}_{\text{train}}$ (%)	Sens. (%)	Spec. (%)	MCC
3	**82.97**	**85.18**	**11.11**	**100.0**	**0.302**
10	76.59	93.51	11.11	92.10	0.045
12	74.46	96.29	11.11	89.47	0.007
14	74.46	97.22	11.11	89.47	0.007

Bold values indicate optimal performance.

Table 3: Performance of the proposed sparse SVM and AdaBoost-based learning system at optimal hyperparameters of the two models on hepatitis disease data.

N	λ	E	Acc_{test}	$\text{Acc}_{\text{train}}$ (%)	Sens. (%)	Spec. (%)	MCC
1	0.01	1	80.85	86.11	22.22	94.73	0.239
2	0.015	1	80.85	86.11	22.22	94.73	0.239
3	0.02	1	80.85	86.11	22.22	94.73	0.239
4	0.04	1	80.85	86.11	22.22	94.73	0.239
5	0.06	1	80.85	86.11	22.22	94.73	0.239
6	0.065	1	80.85	86.11	22.22	94.73	0.239
7	0.07	1	80.85	86.11	22.22	94.73	0.239
8	0.085	1	80.85	86.11	22.22	94.73	0.239
9	0.088	1	80.85	86.11	22.22	94.73	0.239
10	0.09	1	80.85	86.11	22.22	94.73	0.239
11	0.1	8	82.97	90.74	22.22	97.36	0.315
16	**0.3**	**75**	**89.36**	**100.0**	**44.44**	**100**	**0.626**
17	0.9	36	87.23	100.0	33.33	100.0	0.536
18	3	3	82.97	85.18	11.11	100.0	0.302
19	—	3	82.97	85.18	11.11	100.0	0.302

Bold values indicate optimal performance.

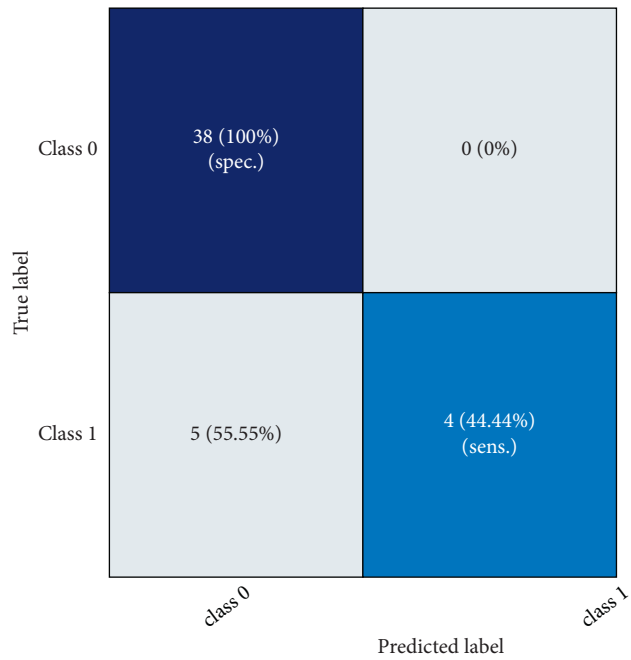

Figure 2: Graphical depiction of statistics of the obtained results on the testing dataset in terms of confusion matrix. Spec: specificity; Sens: sensitivity.

performance of conventional AdaBoost model, we use AUC. The AUC in case of conventional AdaBoost model is 0.587, while AUC in case of the proposed method is 0.649. Hence, the ROC charts further validate the fact that the coupling of the sparse linear SVM enhances the performance of AdaBoost for hepatitis disease data.

4.3. Comparison of the Proposed Method with Some Other Proposed Methods Applied to Hepatitis Data. The above discussion validates that the learning system proposed in this paper significantly augments the strength of the conventional AdaBoost model. In this section, the effectiveness of the learning system thus developed is further validated by carrying out a comparison of its performance with some of the well-known models presented in previous studies. The prediction accuracies and brief details about the models are given in Table 4. It is evident that our proposed method promises better performance upon 23 other machine learning models.

By analyzing Table 4, it can be seen that previous methods have exploited various machine learning-based methods to improve the hepatitis disease prediction accuracy. For example, Stern and Dobnikar developed methods based on discriminant analysis (including linear discriminant analysis and quadratic discriminant analysis) and could achieve a classification accuracy of 85.8% with quadratic discriminant analysis. Similarly, Ozyildirim and Yildirim developed a number of models for searching out optimum model with better classification accuracy. They obtained the highest classification accuracy of 83.75% using radial basis function (RBF). Moreover, if we analyze the results tabulated in Table 4, the previous methods have carried out analysis of their proposed method by only considering classification accuracy. In this paper, we analyzed the results of the proposed hybrid method with a number of metrics and proved the robustness of the proposed method from two key metrics, i.e., classification accuracy and area under the curve (AUC).

4.4. Limitations of the Study. Although this paper demonstrated the effectiveness of exploitation of sparsity in feature space to improve the performance of the machine learning models, the main limitation is lower sensitivity rate. This is due to the low representation of the patient class in the dataset. The main limitation of the hepatitis disease dataset is its imbalanced nature. The dataset has uneven class distribution, i.e., out of 155 samples, 123 samples belong to the healthy class, and 32 samples belong to the patient class. Recent research pointed out that machine learning models trained under such imbalanced classes show biased performance against the minority class (i.e., the models show very poor performance on the minority class) [40]. On the other hand, the models are biased towards the majority class, i.e., the models will show very good performance on the majority class. In case of the hepatitis disease dataset, the minority class is the patient class, and the majority class is the healthy class. From the results, it can be seen that the majority class has 100% detection accuracy (i.e. 100% specificity) while the minority class has poor detection accuracy, i.e., 44%. In future studies, we need to collect balance datasets, i.e., having the same representation for both the classes. Machine learning models trained under such balanced scenario are supposed to show better sensitivity. Moreover, the exhaustive search method for hyperparameters optimization is time-consuming. In future, application of metaheuristic algorithms [41, 42] should be explored.

Table 4: Comparison of the proposed method with some well-known methods proposed for hepatitis disease in terms of prediction accuracy [25, 28, 38, 39].

Model or method number	Model or method	Study or authors	Acc. (%)
1	K-nearest neighbours (KNN)	Nilashi et al.	71.41
2	Neural network	Nilashi et al.	78.31
3	ANaFIS	Nilashi et al.	79.67
4	SVM	Nilashi et al.	81.17
5	ASI	Stern and Dobnikar	82.0
6	Multilayer perceptron + backpropagation	Adamczak	77.4
7	Linear discriminant analysis (LDA)	Stern and Dobnikar	86.4
8	Multilayer perceptron (MLP)	Ozyildirim, yildirim	74.37
9	Radial basis function (Tooldiag)	Adamczak	79.0
10	1NN	Stern and Dobnikar	85.3
11	Radial basis function (RBF)	Ozyildirim, yildirim	83.75
12	15NN, stand. Euclidean	Grudzinski	89.0
13	FSM with rotations	Adamczak	89.7
14	FSM without rotations	Adamczak	88.5
15	Multilayer perceptron with backpropagation	Stern and Dobnikar	82.1
16	Quadratic discriminant analysis	Stern and Dobnikar	85.8
17	(NB and semi-NB), i.e., Naive Bayes and semi-NB	Stern and Dobnikar	86.3
18	Fisher discriminant analysis (FDA)	Stern and Dobnikar	84.5
19	LVQ	Stern and Dobnikar	83.2
20	GRNN	Ozyildirim, yildirim	80.0
21	ASR	Stern and Dobnikar	85.0
22	IncNet	Norbert jankowski	86.0
23	Classification and regression tree (decision tree)	Stern and Dobnikar	82.7
24	LFC	Stern and Dobnikar	81.9
25	L_1-SVM-AdaBoost	The proposed method	89.36

5. Conclusion and Future Work

This work developed an automatic hepatitis disease detection system by using machine learning methods. The AdaBoost model was developed for the hepatitis disease prediction. To improve the classification strength of the AdaBoost model, sparsity in the linear SVM model was exploited. The SVM model eliminated redundant or irrelevant features and thus improved the prediction accuracy of the AdaBoost model. It was also shown that the proposed sparse linear SVM also proves helpful in decreasing the time complexity of the AdaBoost model. Moreover, as evident by the simulation results, our proposed method surpassed many previously published methods in terms of hepatitis disease prediction accuracy. Given the experimental quantitative figures and results, it can thus be safely concluded that the proposed methodology can also be exploited to improve performance of other machine learning models and thus can help to make quality decisions in various other disease detection problems as well.

As discussed above, although the proposed method can be used as a tool to improve the performance of machine learning models, the obtained accuracy still needs considerable amount of improvement. Thus, in future studies, more robust cascaded models should be developed by using deep learning approaches for classification. Additionally, the low rate of sensitivity that is caused by lower class representation of the patient class in the dataset is also a limitation of the study that should be considered as an open challenge for the future work. In future studies, extended hepatitis disease datasets should be collected that will have balanced class distribution.

Data Availability

All the data used in this study are available at the UCI Machine Learning Repository.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] K. Polat and S. Güneş, "Prediction of hepatitis disease based on principal component analysis and artificial immune recognition system," *Applied Mathematics and Computation*, vol. 189, no. 2, pp. 1282–1291, 2007.

[2] E. Dogantekin, A. Dogantekin, and D. Avci, "Automatic hepatitis diagnosis system based on linear discriminant analysis and adaptive network based on fuzzy inference system," *Expert Systems with Applications*, vol. 36, no. 8, pp. 11282–11286, 2009.

[3] Y. F. Liaw and C. M. Chu, "Hepatitis b virus infection," *The Lancet*, vol. 373, no. 9663, pp. 582–592, 2009.

[4] B. Rehermann and M. Nascimbeni, "Immunology of hepatitis b virus and hepatitis c virus infection," *Nature Reviews Immunology*, vol. 5, no. 3, pp. 215–229, 2005.

[5] K. Polat and S. Güneş, "Hepatitis disease diagnosis using a new hybrid system based on feature selection (FS) and artificial immune recognition system with fuzzy resource allocation," *Digital Signal Processing*, vol. 16, no. 6, pp. 889–901, 2006.

[6] L. Ali, C. Zhu, M. Zhou, and Y. Liu, "Early diagnosis of Parkinson's disease from multiple voice recordings by

simultaneous sample and feature selection," *Expert Systems with Applications*, vol. 137, pp. 22–28, 2019.

[7] L. Ali, C. Zhu, Z. Zhang, and Y. Liu, "Automated detection of Parkinson's disease based on multiple types of sustained phonations using linear discriminant analysis and genetically optimized neural network," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 7, pp. 1–10, 2019.

[8] L. Ali, C. Zhu, N. A. Golilarz, A. Javeed, M. Zhou, and Y. Liu, "Reliable Parkinson's disease detection by analyzing handwritten drawings: construction of an unbiased cascaded learning system based on feature selection and adaptive boosting model," *IEEE Access*, vol. 7, pp. 116480–116489, 2019.

[9] F. S. Ahmad, L. Ali, H. A. Khattak et al., "A hybrid machine learning framework to predict mortality in paralytic ileus patients using electronic health records (EHRS)," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–11, 2020.

[10] F. S. Ahmed, L. Ali, B. A. Joseph, A. Ikram, R. Ul Mustafa, and S. A. C. Bukhari, "A statistically rigorous deep neural network approach to predict mortality in trauma patients admitted to the intensive care unit," *The Journal of Trauma and Acute Care Surgery*, vol. 89, no. 4, pp. 736–742, 2020.

[11] T. Meraj, A. Hassan, S. Zahoor et al., "Lungs nodule detection using semantic segmentation and classification with optimal features," *Neural Computing and Applicationsdoi*, vol. 1, 2019.

[12] L. Ali, I. Wajahat, N. A. Golilarz, F. Keshtkar, and S. A. Chan Bukhari, "LDA-GA-SVM: improved hepatocellular carcinoma prediction through dimensionality reduction and genetically optimized support vector machine," *Neural Computing and Applications*, 2020.

[13] L. Ali and S. Bukhari, "An approach based on mutually informed neural networks to optimize the generalization capabilities of decision support systems developed for heart failure prediction," IRBM, 2020.

[14] L. Ali, S. U. Khan, N. A. Golilarz et al., "A feature-driven decision support system for heart failure prediction based on statistical model and Gaussian naive bayes," *Computational and Mathematical Methods in Medicine*, vol. 2019, Article ID 6314328, 2019.

[15] X. Tian, Y. Chong, Y. Huang et al., "Using machine learning algorithms to predict hepatitis b surface antigen seroclearance," *Computational and Mathematical Methods in Medicine*, vol. 2019, Article ID 6915850, 2019.

[16] N. K. Kumar and D. Vigneswari, "Hepatitis-infectious disease prediction using classification algorithms," *Research Journal of Pharmacy and Technology*, vol. 12, no. 8, pp. 3720–3725, 2019.

[17] V. K. Yarasuri, G. K. Indukuri, and A. K. Nair, "Prediction of hepatitis disease using machine learning technique," in *Proceedings of the Third International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pp. 265–269, IEEE, Palladam, India, 2019.

[18] G. Ahmad, M. A. Khan, S. Abbas, A. Athar, B. S. Khan, and M. S. Aslam, "Automated diagnosis of hepatitis b using multilayer mamdani fuzzy inference system," *Journal of Healthcare Engineering*, vol. 2019, Article ID 6361318, 2019.

[19] M. P. McRae, B. Bozkurt, C. M. Ballantyne et al., "Cardiac scorecard: a diagnostic multivariate index assay system for predicting a spectrum of cardiovascular disease," *Expert Systems with Applications*, vol. 54, pp. 136–147, 2016.

[20] O. Altay, T. Gurgenc, M. Ulas, and C. Özel, "Prediction of wear loss quantities of ferro-alloy coating using different machine learning algorithms," *Friction*, vol. 8, no. 1, pp. 107–114, 2020.

[21] G. Manogaran, R. Varatharajan, and M. K. Priyan, "Hybrid recommendation system for heart disease diagnosis based on multiple kernel learning with adaptive neuro-fuzzy inference system," *Multimedia Tools and Applications*, vol. 77, no. 4, pp. 4379–4399, 2018.

[22] U. R. Acharya, H. Fujita, O. S. Lih, M. Adam, J. H. Tan, and C. K. Chua, "Automated detection of coronary artery disease using different durations of ECG segments with convolutional neural network," *Knowledge-Based Systems*, vol. 132, pp. 62–71, 2017.

[23] A. D. Dolatabadi, S. E. Z. Khadem, and B. M. Asl, "Automated diagnosis of coronary artery disease (cad) patients using optimized SVM," *Computer Methods and Programs in Biomedicine*, vol. 138, pp. 117–126, 2017.

[24] M. Abdar, M. Zomorodi-Moghadam, R. Das, and I.-H. Ting, "Performance analysis of classification algorithms on early detection of liver disease," *Expert Systems with Applications*, vol. 67, pp. 239–251, 2017.

[25] D. Çalişir and E. Dogantekin, "A new intelligent hepatitis diagnosis system: PCA–LSSVM," *Expert Systems with Applications*, vol. 38, no. 8, pp. 10705–10708, 2011.

[26] P. Luukka, "Similarity classifier using similarities based on modified probabilistic equivalence relations," *Knowledge-Based Systems*, vol. 22, no. 1, pp. 57–62, 2009.

[27] Y. Kaya and M. Uyar, "A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease," *Applied Soft Computing*, vol. 13, no. 8, pp. 3429–3438, 2013.

[28] M. Nilashi, H. Ahmadi, L. Shahmoradi, O. Ibrahim, and E. Akbari, "A predictive method for hepatitis disease diagnosis using ensembles of neuro-fuzzy technique," *Journal of Infection and Public Health*, vol. 12, no. 1, pp. 13–20, 2019.

[29] K. Polat and S. Güneş, "An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease," *Digital Signal Processing*, vol. 17, no. 4, pp. 702–710, 2007.

[30] L. Ali, A. Niamat, J. A. Khan et al., "An optimized stacked support vector machines based expert system for the effective prediction of heart failure," *IEEE Access*, vol. 7, pp. 54007–54014, 2019.

[31] S. A. Naghibi, K. Ahmadi, and A. Daneshi, "Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping," *Water Resources Management*, vol. 31, no. 9, pp. 2761–2775, 2017.

[32] S. Maldonado, J. Pérez, R. Weber, and M. Labbé, "Feature selection for support vector machines via mixed integer linear programming," *Information Sciences*, vol. 279, pp. 163–175, 2014.

[33] X. Yuan, Q. Tan, X. Lei, Y. Yuan, and X. Wu, "Wind power prediction using hybrid autoregressive fractionally integrated moving average and least square support vector machine," *Energy*, vol. 129, pp. 122–137, 2017.

[34] H. S. Jang, K. Y. Bae, H. S. Park, and D. K. Sung, "Solar power prediction based on satellite images and support vector machine," *IEEE Transactions on Sustainable Energy*, vol. 7, no. 3, pp. 1255–1263, 2016.

[35] P. S. Bradley and O. L. Mangasarian, "Feature selection via concave minimization and support vector machines," in *Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998)*, Madison, WI, USA, 2020.

[36] J. Zhu and H. Zou, "Variable selection for the linear support vector machine," *Studies in Computational Intelligence Book Series*, vol. 35, pp. 35–39, Springer, Berlin, Germany, 2017.
[37] F. Pedregosa, G. Varoquaux, A. Gramfort et al., "SCIKIT-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[38] H. L. Chen, D. Y. Liu, B. Yang, J. Liu, and G. Wang, "A new hybrid method based on local fisher discriminant analysis and support vector machines for hepatitis disease diagnosis," *Expert Systems with Applications*, vol. 38, no. 9, pp. 11796–11803, 2011.
[39] J. S. Sartakhti, M. H. Zangooei, and K. Mozafari, "Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA)," *Computer Methods and Programs in Biomedicine*, vol. 108, no. 2, pp. 570–579, 2012.
[40] Y. Sun, A. K. C. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, no. 4, pp. 687–719, 2009.
[41] S. U. Khan, M. Rahim, and L. Ali, "Correction of array failure using grey wolf optimizer hybridized with an interior point algorithm," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 9, pp. 1191–1202, 2018.
[42] N. A. Golilarz, H. Gao, R. Kumar, L. Ali, Y. Fu, and C. Li, "Adaptive wavelet based MRI brain image de-noising," *Frontiers in Neuroscience*, vol. 14, 2020.