Multimedia Tools and Applications (2024) 83:59021–59042
https://doi.org/10.1007/s11042-023-17825-1

RG-SVM: Recursive gaussian support vector machine based feature selection algorithm for liver disease classification

Prasannavenkatesan Theerthagiri[1] · Sahana Devarayapattana Siddalingaiah[1]

Received: 18 May 2023 / Revised: 1 December 2023 / Accepted: 4 December 2023 /
Published online: 22 December 2023

Abstract

Health is an essential concern for everyone, so it is necessary to facilitate medical services that are easily accessible to everyone. The primary goal of this work is to predict liver diseases using a machine-learning strategy that makes use of feature selection and classification techniques. This work proposes the recursive Gaussian support vector machine-based feature selection (RG-SVM) algorithm. It uses the Gaussian kernel of support vector machine and recursive feature selection algorithm for the prediction of liver disease. The proposed RG-SVM algorithm has been evaluated on the Indian liver patient records dataset. Various classification algorithms such as logistic regression, decision tree, k-nearest neighbour, and Naive Bayes are implemented and compared in order to assess the accuracy, confusion matrix and area under curve. The proposed RG-SVM has been compared with other existing algorithms such as logistic regression (LR), decision tree (DT), k-nearest neighbour (KNN), Naïve Bayes (NB), and proposed RG-SVM algorithms. The algorithms LR, DT, KNN, NB, and proposed RG-SVM have accuracy values of 73, 80, 81, 54, and 93%, respectively. It clearly shows that the proposed RG-SVM with the support of a recursive feature selection algorithm, outperformed other existing algorithms with an improved accuracy of 14 – 39% 12- 20% of reduced MSE error over other compared algorithms. Similarly, the sensitivity and specificity of RG-SVM algorithm produced 5–26% and 34–72% improved results over the existing algorithms. The results of the proposed algorithm will be useful for physicians to make better decisions for liver disease patients.

Keywords Recursive Feature selection · Feature ranking, RG-SVM, decision tree · Liver disease Prediction

✉ Prasannavenkatesan Theerthagiri
prasannait91@gmail.com

[1] Department of Computer Science and Engineering, GITAM School of Technology, GITAM University Bengaluru, Bengaluru, India

1 Introduction

The liver, the largest functional organ in the body, performs numerous crucial functions that are necessary for survival, including the metabolism of carbohydrates, production and elimination of bile, and the synthesis of proteins, lipids, and fatty acids. It is the second-most important organ in humans [1]. According to the WHO, chronic diseases cause nearly 35 million deaths worldwide, accounting for roughly 46% of all illnesses and 59% of all fatalities. Liver illness has become more common over time. According to government statistics, liver illnesses are the sixth most common cause of death in the UK. Across all digestive conditions, liver diseases are the second largest cause of death in the US [2]. Over the past few decades, liver cancer has been more common overall, especially in wealthy nations [3]. The survival rate for liver cancer could be increased by early detection, but this is challenging given the absence of recognisable symptoms and incomplete understanding of the aetiology of oncogenesis [4].

The growth of artificial intelligence technology with advanced algorithms facilitates effective solutions for disease diagnosis and prognosis. The prediction and classification of diseases can be accomplished by machine learning techniques. Numerous approaches like Random Forest (RF), Decision tree, Naive Bayes algorithm, and Logistic regression have been used to provide support by analysis and prediction about the disease or illness [5]. In any kind of prediction model, the selection of features is more crucial to provide a better outcome. The filter and wrapper approaches are often employed for the feature selection techniques. The filter approach, which is independent of any classification algorithm, takes into account how each characteristic is related to the class label [6]. Feature selection enhances machine learning and improves the predictive power of machine learning algorithms in the medical field [7].

Additionally, it improves prediction performance, reduces processing requirements, alleviates the curse of dimensionality, and facilitates data comprehension. The aim of selecting a feature is to consider a part of input variables that can characterise the input data accurately while minimising the influence of noise or auxiliary components and still producing accurate prediction results. The design of the classification model will be influenced by irrelevant attributes if a dataset contains multiple features and multi-class datasets, which could reduce classification accuracy. Therefore, feature selection may be a crucial pre-processing method for solving classification issues. Improving the quality of categorization models can also aid in the reduction of redundant and unused attributes [8–11].

Section 2 elaborates on the current literature on liver disease prediction using machine learning and feature selection techniques employed in various healthcare fields. The proposed RG-SVM algorithm and its schematic view are given in Section 3. The results and discussions are detailed in Section 4. Section 5 concludes the paper with future enhancements.

2 Literature Survey

Hassan et al. [12] suggested ensemble filter techniques that utilised symmetrical uncertainty, gain ratio, Information Gain (IG), and Support Vector Machine- Recursive Feature Elimination (SVM-RFE) to rank all genes and choose the best genes. This approach is used to assess three binary-labelled gene expression datasets from leukaemia, lung cancer, and breast cancer. Decision trees, support vector machines (SVM), Random Forest (RF), K-nearest neighbours (KNN), and naive Bayes were the five classifiers utilised to evaluate the chosen attributes (NB). They claimed that across all datasets, the novel approach outperformed the earlier methods in terms of classification accuracy.

In the study by Dong et al. [13], RFEs (Recursive Feature Eliminations) and SVMs were suggested. The methodology used methylation chip data from The Cancer Genome Atlas (TCGA) database to examine 377 Hepatocellular carcinoma (HCC) patients and 50 normal samples. A total of 47,099 samples were examined for 134 methylation locations using the SVM-RFE, Cox regression, and Frank-Wolfe (FW)-SVM algorithms. This technique predicted patient survival rates based on the assessment of the model's high, moderate, and low-risk categories.

Hepatitis C was recognised as a virus by Azam Orooji et al.[14] The liver-attacking virus hepatitis C has a significant death rate. By tackling the imbalanced datasets issue, concentrate on this study. The proposed strategy makes use of random over- and under-sampling techniques. When the oversampling method was combined with the RF method, the accuracy of the results was at its highest. The authors, G. Shobana et al. [15], investigate the recursive feature removal feature reduction technique to improve prediction accuracy. Simple machine learning models were applied to the dataset, and the findings revealed that multi-layer perceptrons and logistic regression provided higher prediction accuracy with fewer data.

In order to choose the right amount of features from SVM-RFE, the authors Xiaohui Lin et al. [16] suggested a method known as SVM-RFE-OA(Overlapping Ratio)that integrates the performance of the classifier and the collection period of the data. A modified SVM-RFE-OA technique is suggested to temporally filter off the information occurring in heavily corresponding pixels in each iteration to estimate the feature weights between iterations more precisely. The condition of the test set's liver was examined by Tsehay Admassu Assegie et al. [17]. The ideal feature set was used to train the SVM during the pre-processing phase, and the RF model was used to reduce repetitive features. The experimental findings demonstrate improved accuracy for the proposed SVM model.

Assegie et al. [18] developed a predictive training model of liver disease using SVM and KNN learning techniques, and the performance of the approach was assessed using an Indian liver disease data source. The outcome analysis reveals improved SVM accuracy. It has been shown that SVM outperforms the KNN algorithm for predicting liver disease based on the accuracy ratings of SVM and KNN on the analysis results. The analysis of patient characteristics and genome expression by S. Sontakke et al. [19] aims to enhance the detection of liver illnesses. The molecular biology approach is influenced by factors such as age, ethnicity, and diet. The chemical method of forecasting is more reliable. Most likely, molecular biology research can save lives while illuminating the mysteries of the human anatomy.

Compared to earlier studies on liver disease, the cutting-edge decision tree-based system used by Moloud Abdar et al. [20] displayed good accuracy forecasts while considering more factors. The collection includes 167 data for a healthy liver and 416 records for liver diseases. It is examined by the two algorithms, Chi-square automatic interaction detection (CHAID) and Boosted C5.0, which are frequently used to pinpoint risk factors for liver disease. The findings demonstrate that both algorithms significantly affect the prediction of liver illness based on the rules they produce. According to Marwa I. M. et al. [21], liver tumours can be classified as benign or malignant by looking at CT liver pictures and using the adaptive neuro-fuzzy inference system (ANFIS) model. The decision-making approach involved four steps: liver extraction using thresholding, picture augmentation to boost image quality and boundary extraction methods. Then, using the Discrete Wavelet Transformation characteristics that were recovered using the Fuzzy C-mean (FCM) clustering technique, the interior of the tumour object is segmented. Finally, using the least squares strategy and the backpropagation gradient descent method, the ANFIS classifier is trained using these extracted features. A series of patient CT pictures were used to assess the effectiveness of the suggested technique.

By gathering crucial laboratory values, the Farokhzad et al. approach [22] for performing the diagnosis of liver illness using fuzzy logic was proposed. Fuzzy heuristic systems are built using two different variants of Triangle membership functions and Gussy membership functions. By carefully selecting their input parameters, the quantity of membership functions, and the kind of membership functions, they were able to achieve an accuracy of 83 percent. A suitable selection of features is provided by Marium Mehmood et al. of the feature selection techniques in [23], which makes it easier to identify illnesses. These methods demonstrate their value for data mining and machine learning. According to the study, the Wrapper Method has the highest R2-Score, making it the best at identifying traits that are crucial for disease detection. A higher R2-Score and a lower MSE signify more accurate sickness detection.

Sampling-Continuous Re-RX was introduced as a revolutionary technique by the authors Y Hayashi et al. [24] for developing highly precise and understandable rules for the British United Provident Association (BUPA) and Hepatitis datasets. They demonstrated an extracted rule set from the BUPA dataset and offered a healthcare information explanation of the found rules. Since the suggested approach was close to the trade-off curve, it was more precise and easy to understand, making it more suitable for use in medical decision-making. Padmakala et al. [25] suggested using a group SVM-based sample weighted RF with a brand-new improved colliding body optimization (NICBO) method to identify liver illnesses. The patient data are pre-processed using the ELTA technique for collection, packing, modification, and evaluation. It combined the appropriate model and the filter-based procedures, resulting in the relevant feature.

Admassu et al. [17] automated method for diagnosing liver disease makes use of SVM and RF detection methods. The proposed technique SVM and RF-based hybrid model successfully diagnoses liver illness in the test set. The SVM is trained using the ideal feature set during the well-before phase, and the RF model is used for recurrent feature reduction. The results of the experiment show that the proposed SVM model has an accuracy rate of 78.3%. Abdalrada et al. [35] dealt with the issue of predicting the progression of liver disease. The logistic regression based predictive model was utilised to estimate the likelihood of developing liver disease.

The patient's liver condition is examined using machine learning techniques by Tokala et al. [36]. This work used the proportion of people who get the condition as both positive and negative data. Various ML classifiers and confusion matrix were used to process the percentages of liver disease. According to Madhusudan et al. [37], the main driving force behind the effort was to put a machine learning (ML) based real-time framework for classifying liver illnesses onto the cloud in order to lighten the workload of clinicians. Convolutional neural networks (CNN) were used, and their output from the flatten layer was then delivered to classifiers. The performance of the model was assessed using the stratified K-fold approach.

Various artificial intelligence algorithms are studied by Khan et al. [38], in order to identify the presence of liver disease in a patient at an early stage. Sensitivity analysis was conducted on the dataset to look at how each attribute affects how well the model performs. It was shown that the Alanine Aminotransferase characteristic has the greatest influence on the prognosis of liver illness and is employed as a support system for the early detection of liver disease. A deep-learning approach was presented by Sun et al. [39] for the classification of histological images of liver cancer. Patch features are extracted and completely utilised to compensate for the lack of comprehensive cancer region annotations in those images. To obtain the image-level features for classification, transfer learning is paired with multiple-instance learning to provide the patch-level features.

2.1 Feature Selection

The redundant and irrelevant attributes from the dataset may be removed without affecting accuracy and the classification performance of learning models can be enhanced by feature selection algorithms. Feature selection algorithm that distinguishes crucial characteristics from less significant ones. Also, the dimensionality of training and testing data points is decreased via feature selection. The benefits of feature selection include decreased lifting, shorter training sessions, more accuracy, and more. These techniques can aid in identifying key features that can be utilized to classify various liver diseases [55–58].

Ruhul et al. [40] presented a system that generates a feature space by considering the covariance between observed variables, maximum class separation, and a linear combination of observed variables. Various statistical techniques were also applied to handle missing values, outliers, and data balancing to prevent bias and overfitting. Kumar et al. used the neighbourhood-weighted K-NN (NWKNN), fuzzy neighbourhood-weighted K-NN, and variable neighbourhood-weighted fuzzy K-NN classifiers to categorise liver patients [41, 43]. Tomek link and redundancy-based under-sampling technology (TR-RUS) is employed to avoid the unbalanced nature of the dataset and claims an improved of accuracy of 87.71% for NWFKNN classifier. A feature extraction approach has been employed to increase prediction performance by Salau et al. [42]. The author claims that the novel feature union prediction algorithm outperforms the existing classification algorithms in terms of accuracy and F1 score.

According to Admassu et al., the classification performance of machine learning models is enhanced by feature selection. This work makes use of a multivariate sample similarity metric for feature selection and chooses features that significantly contribute to the model [55]. By applying dimensionality reduction strategies, authors Ruhul Amin et al. investigated enhanced feature extraction systems for liver patient classification using statistical machine learning techniques. The system retrieved an improved feature space that takes into consideration the covariance between the observed variables, the linear combination of observed variables that maximizes class separation, and the maximum variation in the data. To deal with missing values, outliers, and data balance to prevent bias and overfitting, various robust statistical methods were applied [56].

Filter technique, wrapper approach, and embedding method were three different feature selection algorithms that authors Shruthi Jain et al. explored. The filter method is a preprocessing method for obtaining the greatest qualities. Highly ranking qualities are prioritized and used as predictors in this method. Predictors from the Wrapper Method are combined with a search algorithm that selects a subset and assigns the best possible predictor to that subset [57]. Finding the most pertinent and instructive subset of features in a given dataset is the aim of feature selection, according to the paper in [58]. This strategy helps to lessen dimensionality, improve model performance, and decrease the curse of dimensionality. The goal of this study is to create a brand-new framework that is snake-optimized. Five machine learning algorithms are used, along with the snake optimization (SO) method, to choose and categorize the best medical data, resulting in a highly accurate prediction of kidney and liver disease [58]. State of the art on liver disease prediction techniques has been summarized in Table 1.

Numerous researchers have demonstrated their work on predicting and classifying liver diseases using machine learning and deep learning algorithms. Few works are concentrated on feature selection and grid search algorithms for the better selection of features from the dataset. However, many research studies have not concentrated on the various combination of features and their importance, feature ranking, and statistical analysis on the features for the prediction and classification of liver disease. This work analyses the features in different factors, combinations, and priority of features investigated, and they ranked with statistical analysis support. Thus, this research work intends to predict and classify liver diseases by employing a recursive feature selection algorithm, feature ranking and Gaussian kernel-based support vector machine learning algorithms. The main contributions of this work are given as follows.

- This work proposes the recursive Gaussian support vector machine-based feature selection (RG-SVM) algorithm.
- The feature importance and feature ranking has been evaluated for the disease classification with the support of Gaussian kernel-based SVM algorithm.
- The results of this approach have been compared with the other existing algorithms, and performances and error metrics were evaluated.
- The results of the proposed algorithm will be useful for physicians to make better decisions for liver disease patients.

Early prediction and diagnosis of liver disease more accurately is the main challenge where many research works are going on. The main contribution of this paper is a novel recursive feature selection algorithm developed with a Gaussian-based SVM algorithm for liver disease prediction.

3 Proposed Methodology

The dataset for the classification of liver diseases has been taken from the Indian Liver Patient Records collected from Northeast of Andhra Pradesh, India [54]. It has the liver disease details about 583 patients with features, age, gender, total bilirubin, direct bilirubin, alkaline phosphatase, alamine aminotransferase, aspartate aminotransferase, proteins, albumin, albumin & globulin ratio, and class. Table 2 gives the sample patient details from the Indian Liver Patient dataset. Figure 1 illustrates the swarm plot of gender with respect to age.

The imbalanced classification involves developing a predictive model on the liver disease classification datasets with a severe class imbalance and, in turn, poor performance in the minority class. So, in this proposed work, an oversampling approach called the Synthetic.

Minority Oversampling Technique (SMOTE) has been used to address the issue of imbalanced datasets. SMOTE is the simplest approach that involves duplicating examples in the minority class, although it does not add any new information to the model. Instead, new examples can be synthesized from the existing examples. Figure 2 depicts the imbalanced dataset before applying the SMOTE algorithm to the dataset. It can be

Table 1 State of the art on liver disease prediction techniques

Authors	Algorithm	Objectives	Inference	Limitations	Performance Metrics
Zaheer, and Nirmala [44]	K-nearest neighbour (KNN)	Various machine learning and deep learning techniques have been applied to solve real-world problems in predicting liver disease	Multiple patient liver imaging datasets have been examined, and novel liver condition identification has been achieved	Data inequality is not considered, and the true positive and false positive rates are not analysed	The mean Accuracy rate of KNN is94.30%
Kumar et al. [45]	Neighbour Weighted Fuzzy K-Nearest Neighbour	It aims to enhance the classification of data for liver illness and imbalance	Prediction performance was increased with the usage of normalization, Tomek-linked, and redundancy-based undersampling	Higher computing complexity and processing time	Accuracy rate for BUPA dataset is 78.46%, ILPD dataset is 78.46%, and MPRLPD dataset is 95.79%
Padmakala et al. [46]	Ensemble SVM-based sample weighted random forests (eSVM-swRF)	To forecast liver disorders, using a novel improved colliding body optimization algorithm	The dataset has been preprocessed utilizing the extraction, loading, transformation, and analysis (ELTA) method, which generates significant features using a filter-based approach	Feature selection and analysis are not considered in this work	Accuracy rate of 98% is achieved using eSVM-swRF
Dritsas et al. [47]	Voting classifier	To forecast the occurrence of liver disease using various ML models and ensemble approaches	Class balancing and features ranking has been performed to produce better results	This work does not compare the results with existing works	F-measure is 80.1%, precision is 80.4%, AUC is 88.4%
Ghazal et al. [48]	Artificial Neural Network (ANN)	Various machine learning methods have been examined to reduce the expensive diagnostic cost of liver disease through prediction	The result of the work has been compared with various machine learning techniques with lower errors in liver disease prediction	The feature selection methods are not incorporated in this work	Accuracy rate is 88.4%, miss rate is 11.6%

Table 1 (continued)

Authors	Algorithm	Objectives	Inference	Limitations	Performance Metrics
Marium et al. [49]	Feature extraction Technique	Feature selection techniques are presented for disease diagnosis and simplify the process of disease detection	Filter, Embedded, and Wrapper Methods produce the R2-Score of 0.89 and the lowest MSE of 0.06 for the extracted features	True positive, false positive rates and f1 score values are not computed and analysed	R2 Score is 0.8923 lowest MSE score is 0.0618
Jayaram et al. [50]	KNN and AdaBoost models	Intelligent Liver Disease Prediction system developed to predict liver diseases	Empirical statistical analysis performed at an early stage, low-cost prediction of liver disorders	The dataset is imbalanced; thus, the result of the predictive classification model is questionable	KNN and AdaBoost models have an accuracy rate of 100%
Sateesh et al. [51]	Random Forest	To predict liver diseases using univariate and bivariate analysis and hyperparameter tuning with grid search and feature selection	The feature selection and hyperparameter adjustment improve the performance. Class balances, oversampling and undersampling were analysed	The results are having the over-fitting issue	Random Forest with Data Balancing Technique Accuracy rate is 100%
Rifky et al. [52]	Random forest algorithm, SVM-RFE (Recursive Feature Elimination)	SVM-RFE with optimization algorithm has been developed with grid-based hyper-parameters search to predict liver disease	The synthetic minority oversampling technique (SMOTE) is used to handle imbalanced datasets and for classification	The comparative analysis has not been performed with the existing works for classification analysis and error analysis	The accuracy rate is 0.879, precision is 0.902, and ROC is 0.966
Marwa et al. [53]	Adaptive neuro-fuzzy inference system (ANFIS)	To classify the malignant and benign tumours using the ANFIS classifier	The ANFIS system classifies the liver tumour using texture features and DWT features	Fails to identify the kind of infection in the liver and extract advanced tumour characteristics	Texture Feature extraction's accuracy rate is 90%

Table 2 Indian Liver Patient dataset

Features/Patients	Patient1	Patient2	Patient3	Patient4	Patient5	Patient6
Age	65	62	62	58	72	46
Gender	Female	Male	Male	Male	Male	Male
Total Bilirubin	0.7	10.9	7.3	1	3.9	1.8
Direct Bilirubin	0.1	5.5	4.1	0.4	2	0.7
Alkaline Phosphotase	187	699	490	182	195	208
Alamine Aminotransferase	16	64	60	14	27	19
Aspartate Aminotransferase	18	100	68	20	59	14
Protiens	6.8	7.5	7	6.8	7.3	7.6
Albumin	3.3	3.2	3.3	3.4	2.4	4.4
Albumin Globulin Ratio	0.9	0.74	0.89	1	0.4	1.3
Outcome	1	1	1	1	1	1

Fig. 1 Gender over Age

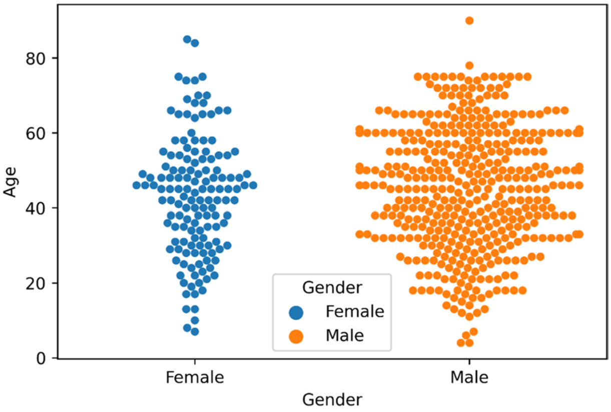
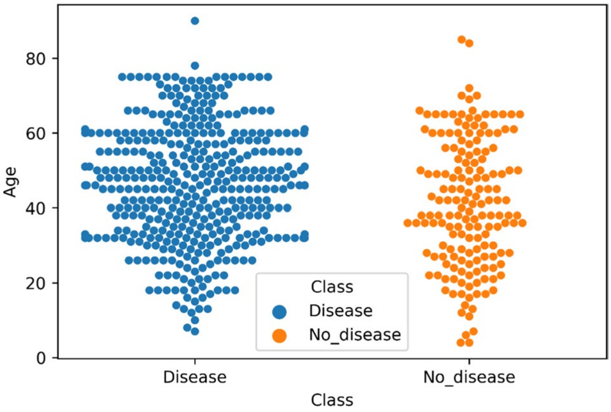


Fig. 2 Imbalanced Class



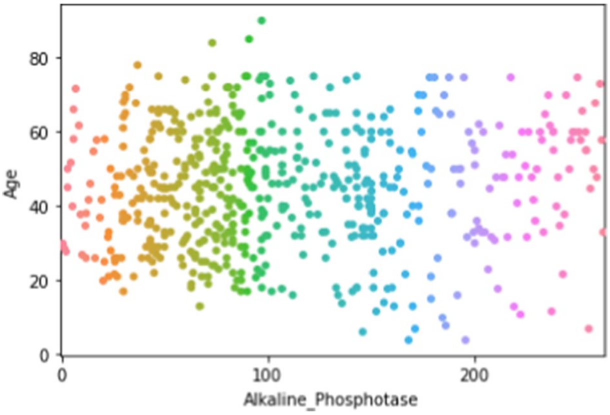


Fig. 3 Alkaline Phosphotase over Age

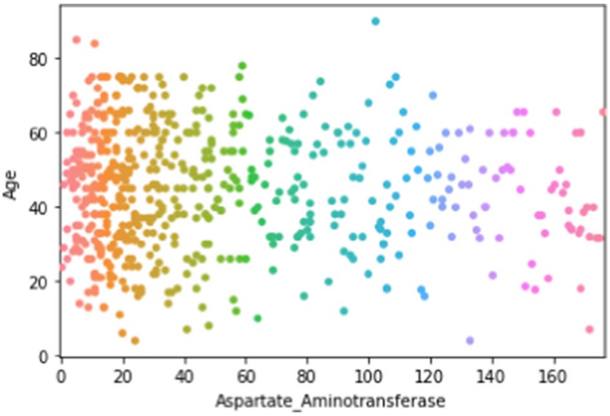


Fig. 4 Alamine Aminotransferase over age

clearly seen that patients without liver disease are the minority class. Figures 3 and 4 show the alkaline phosphatase distribution and alamine aminotransferase over age.

3.1 Recursive Feature Selection Algorithm and Ranking

The recursive Gaussian SVM-based feature selection algorithm involves selecting the predictors in reverse. The first step in this technique is to create a model using all of the predictors and determine the relevance of each predictor. The model is rebuilt, significance scores are calculated once more, and the least significant predictor(s) are discarded. In reality, the analyst defines each subset's size and the number of predictor subgroups to be evaluated. As a result, the subset size is the recursive feature elimination tuning parameter. Based on the importance rankings, the predictors are chosen using the subset size that optimises the performance requirements. The final model is then trained using the best subset [26].

Feature ranking involves ordering the characteristics in accordance with the outcome of a scoring function, which typically determines the relevance of the features. The score S(fi) determines S(fi) criteria for all features, which is computed using the training data. The S(fi) criteria determine the high score for ten liver features, which are denoted as beneficial features. The k highest ranked features according to S(fi) are chosen using the feature

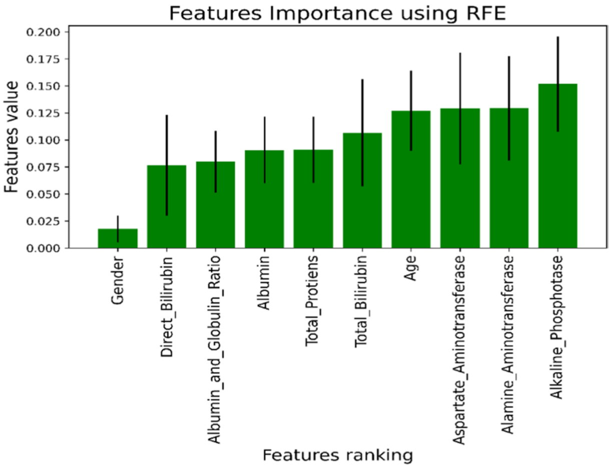


Fig. 5 Liver Feature Ranking

selection method that makes use of variable ranking. It simply requires the calculation and sorting of n scores, which is computationally efficient [27]. Figure 5 illustrates the feature ranking for all ten liver features. The selected features based on the recursive feature selection are three, which are clearly depicted in Fig. 6. The optimal features are Alkaline_Phosphotase, Alamine_Aminotransferase, and Aspartate_Aminotransferase. Figure 7 gives the schematic diagram of the proposed RG-SVM methodology. The liver dataset has been preprocessed with the support of the AMOTE algorithm, and then a recursive feature selection, extraction and ranking algorithm is employed. As a result, the data is transformed into a feature vector. The statistical test on the features is performed using the chi-square test, and finally, the optimal features are selected for the liver disease classification.

3.2 Gaussian-based Support Vector Machine Algorithm

A promising classification method for identifying a complex condition like liver disease using widespread, straightforward data is support vector machine modelling. In circumstances where sample sizes are limited and there are many variables present, the SVM

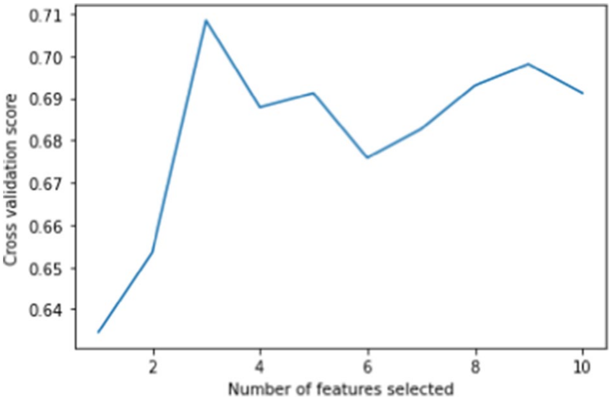


Fig. 6 Features selected for liver disease prediction

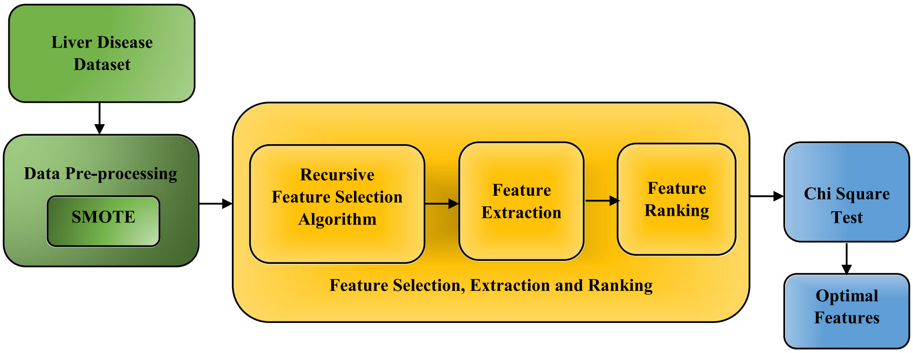


Fig. 7 Schematic diagram of Proposed RG-SVM methodology

technique, which is data-driven and model-free, may offer significant discriminative potential for classification. Recent advances in disease detection techniques and automated disease classification have both benefited from the usage of this technology [28–31]. SVM with Gaussian kernel function has been developed for the nonlinearly separable liver dataset. The Gaussian kernel function enables the separation of nonlinearly separable liver dataset by converting the input vector to a Hilbert space representation. The Gaussian kernel is an exponential function that includes the real constant and norm, as given in Eq. (1).

$$K_G(u,v) = exp\left(-\frac{\|u-v\|^2}{2\rho^2}\right) \tag{1}$$

where u and v are input vectors, the Euclidean norm in the exponential expression's numerator part is determined using input vectors. It is a real constant, a freely chosen value, in the denominator part. The Gaussian kernel function exhibits hyper-spherical outlines because of its exponential decay in the input feature space and uniform distribution around the support vector. Iterative, time- and energy-intensive, the experimental search is a process. Therefore, one of the key solutions to SVM classification issues would be the creation of an effective technique for adjusting to an ideal width for the data.

4 Results and Discussion

The proposed recursive Gaussian SVM-based feature selection algorithm has been developed and evaluated with a system configuring 6 GB RAM, an Intel i3 processor, and Python libraries. The liver dataset was divided into training and testing with 70% and 30%, respectively, with a k-fold cross-validation of 10. The results of the proposed work are analysed using accuracy, mean square error (MSE), precision, recall, sensitivity, specificity, confusion matrix, and area under receiver operating characteristic curve (AUROC).

Precision is the quality of positive instances produced by the proposed model, whereas recall is the proportion of correctly classified samples (also known as sensitivity). As a result, relevance serves as the foundation for accuracy and f score. A common statistic evaluation model is a mean squared error. The average of the squared prediction errors overall test set occurrences is used to calculate the mean squared error of a model

Fig. 8 Accuracy of algorithms

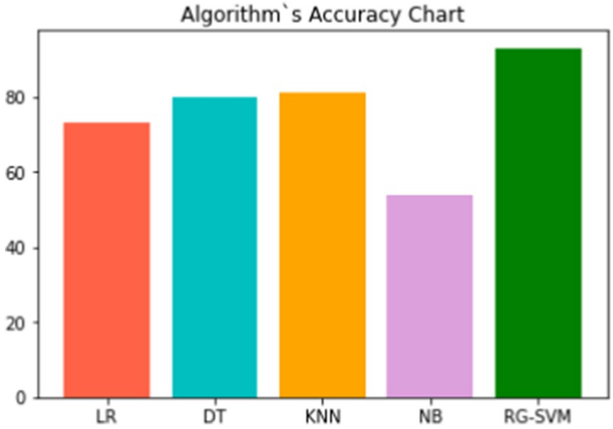
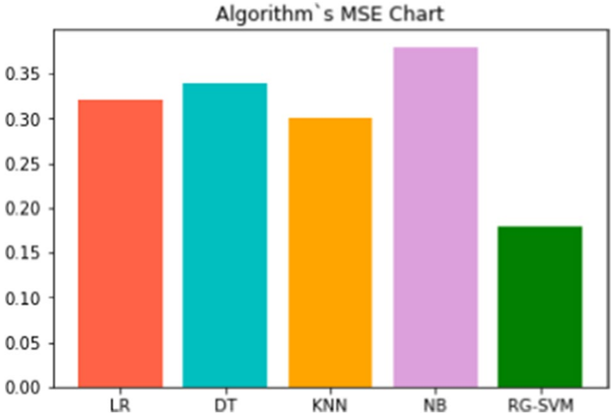


Fig. 9 MSE values of algorithms



with respect to the test set. The variance between a model's actual value and anticipated value is known as prediction error, as given in Eq. (2).

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \lambda(x_i))^2}{n} \tag{2}$$

where n indicates instances, 'y_i' is the real-world goal value for the test case 'x_i' and λ (x_i) is the target value anticipated for the test instance x_i [32, 33].

A predictive analytics tool, a confusion matrix list, contrasts expected and actual predictions. A confusion matrix is a statistic used to examine a machine-learning classifier's efficacy in a machine-learning situation. The confusion matrix is applied when the classifier's result includes two or more categories. Confusion matrices are used to display the key predictive parameters, including recall, specificity, accuracy, and precision [34]. Figure 8 gives the accuracy of algorithms, where the logistic regression (LR), decision tree (DT), k-nearest neighbour (KNN), Naïve Bayes (NB), and proposed RG-SVM algorithms are compared. The algorithms LR, DT, KNN, NB, and proposed RG-SVM have accuracy values of 73, 80, 81, 54, and 93%, respectively. It clearly shows

Fig. 10 Specificity Score of Algorithms

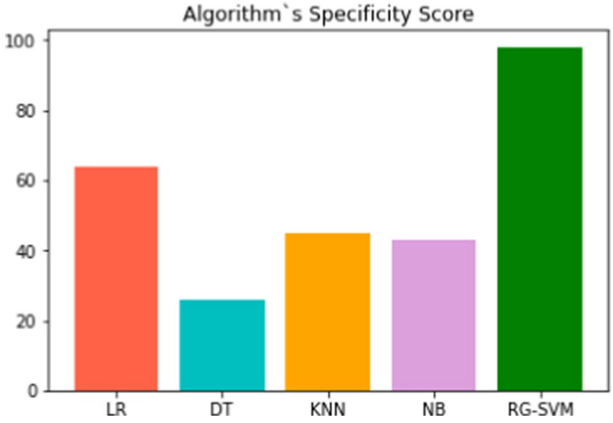
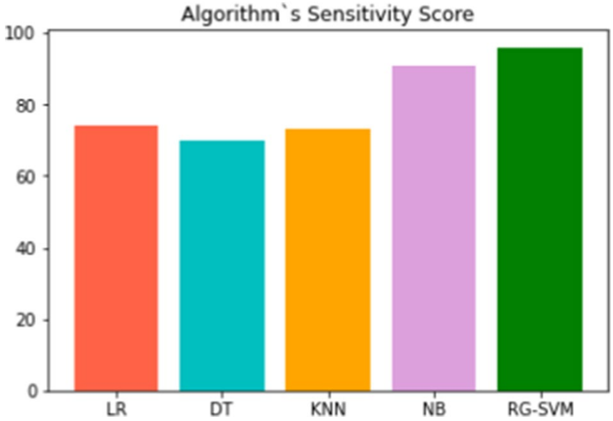


Fig. 11 Sensitivity Score of Algorithms



that the proposed RG-SVM, with the support of a recursive feature selection algorithm, outperformed other existing algorithms with an improved accuracy of 14 – 39%.

The values 0.32, 0.34, 0.3, 0.38, and 0.18 are the mean square errors for the LR, DT, KNN, NB, and RG-SVM algorithms, respectively, shown in Fig. 9. The proposed RG-SVM algorithm has 12- 20% of reduced error over the compared LR, DT, KNN, and NB algorithms. The sensitivity and specificity scores for the LR, DT, KNN, NB, and RG-SVM algorithms are depicted in Figs. 10 and 11, respectively. As shown in Fig. 10, the sensitivity scores are 74, 70, 73, 91, and 96 for the LR, DT, KNN, NB, and RG-SVM algorithms. Similarly, the specificity scores are 64, 26, 45, 43, and 98 for the LR, DT, KNN, NB, and RG-SVM algorithms. Comparatively, Figs. 8, 9, 10, and 11 indicate that the proposed RG-SVM algorithm outperforms all performance metrics. The foremost significance of improvement for the RG-SVM algorithm is the features selected recursively for liver disease prediction.

Figures 12, 13, 14, and 15 give the confusion matrix for the LR, DT, KNN, and proposed RG-SVM algorithms. The RG-SVM's functions change degree and alter width simultaneously, which improves the classification performance. Moreover, the influences of these functions produce significant results for the confusion matrix of the proposed RG-SVM algorithm compared to other algorithms. The RG-SVM works in

Fig. 12 Confusion matrix of LR

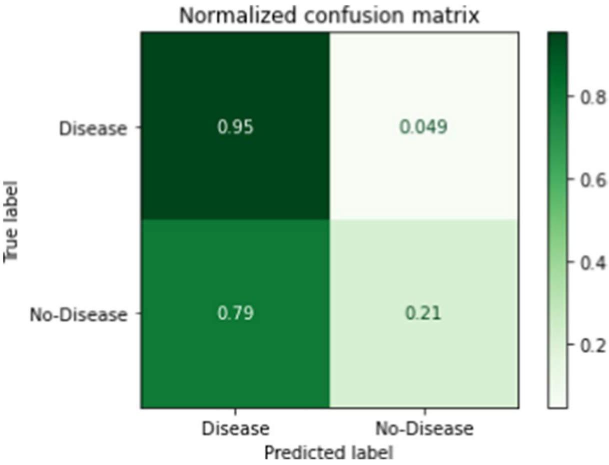


Fig. 13 Confusion matrix of DT

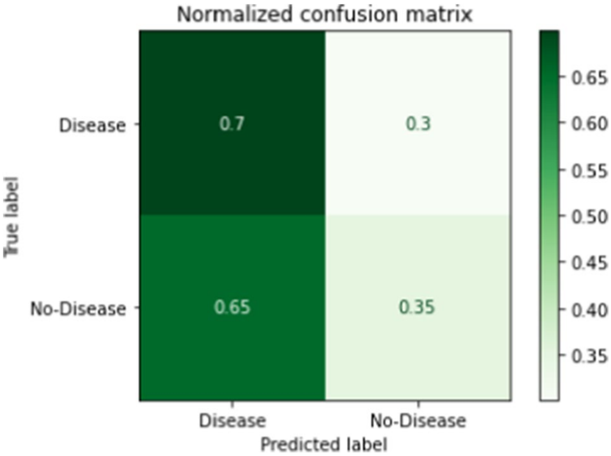


Fig. 14 Confusion matrix of KNN

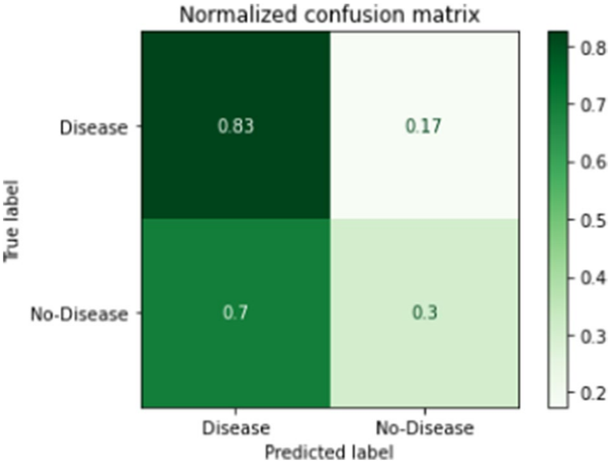
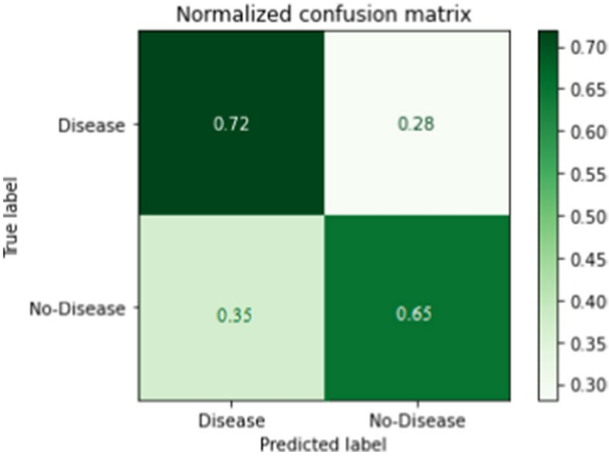


Fig. 15 Confusion matrix of RG-SVM



parallel with determining the feature weights in building the learning model. It gradually removes the attributes with the least values. The order in which the features are deleted provides an approximation of the ordering of feature importance. In essence, SVM-Recursive ranks the features based on the sequence in which features were eliminated during iterations. The most important components are those at the top of the list that were eliminated in the most recent iteration, as opposed to the least informative features at the bottom of the list that were eliminated in the first iteration. Figure 16 presents the AUROC graph for the LR, DT, KNN, and proposed RG-SVM algorithms. Table 3 gives the comparative analysis of performance metrics of various existing algorithms with the proposed RG-SVM algorithm. The metrics such as accuracy, precision, recall, F1-score, MSE, sensitivity, specificity and false positive rates are compared for the LR, DT, KNN, NB, MLP, Ensemble Classifier, Random Forest, and proposed RG-SVM algorithm. In Table 3, boldface indicates proposed RG-SVM algorithm, which outperforms over other methods. According to our evaluation, the proposed Gaussian kernel-based SVM with a recursive feature selection algorithm (RG-SVM) appears to be more effective than the other models.

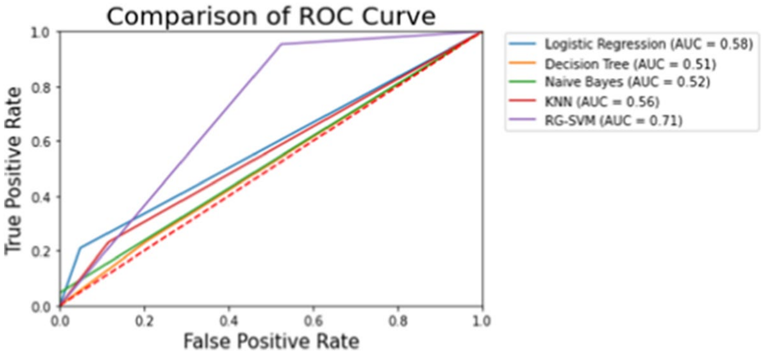


Fig. 16 AUROC Chart

Table 3 Comparative Analysis of Existing Algorithm's Performance Metrics

Performance Metrics	Learning Algorithms							
	LR	DT	KNN	NB	MLP	Ensemble Classifier	Random Forest	RG-SVM
Accuracy	73	80	81	54	84	82	82	**93**
Precision	69	49	59	71	79	75	78	**86**
Recall	58	50	56	62	97	97	59	**82**
F1-Score	57	48	55	54	83	84	55	**89**
Mean Square Error (MSE)	32	34	30	38	33	31	28	**18**
Sensitivity	74	70	73	91	82	86	83	**96**
Specificity	64	26	45	43	75	88	85	**98**
False Positive Rate	58	51	52	52	68	66	55	**71**

4.1 SHAP Result Analysis

The SHAP (SHapley Additive exPlanations) method measures the significance of each input variable to a model's ability to make predictions. The force plot is an approach to assess the impact of each characteristic on liver disease prediction. The model's score for liver disease prediction is indicated by the force plot in Fig. 17, as boldface 0.65. Lower scores cause the model to predict 0, whereas higher values cause it to anticipate 1. Red represents features (Age, Aspartate_Aminotransferase, Albumin, Alamine_Aminotransferase) that pushed the model score higher, and blue represents features that pushed the score lower. These features were crucial to predicting liver disease

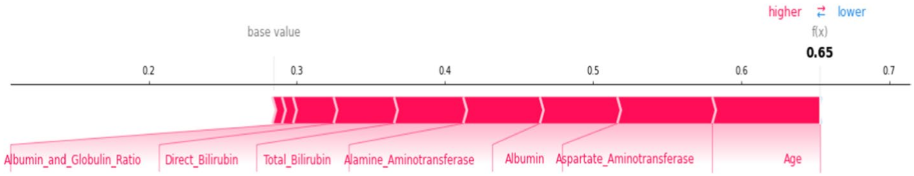
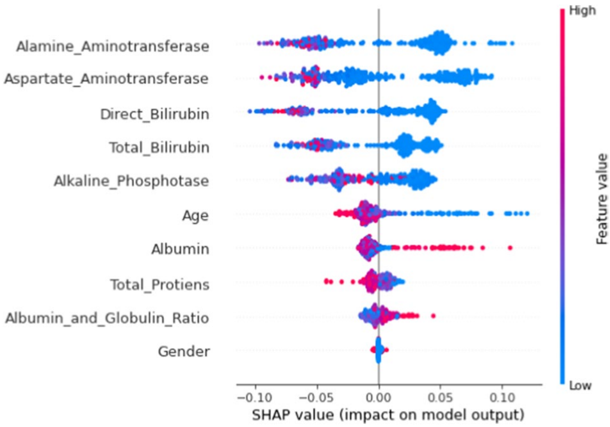


Fig. 17 SHAP Force Plot

Fig. 18 SHAP Summary Plot



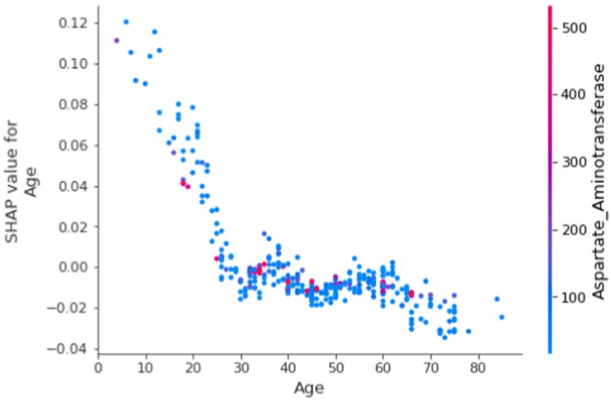


Fig. 19 SHAP Dependence Plot

prediction. The features closest to the line separating red from blue indicate how much of an influence it had on the prediction of liver disease, and the size of the bar reflects how much of an impact it had.

The summary plot in Fig. 18 combines the importance and the impacts of the features. The link between a feature's value and its influence on the prediction may be seen in the summary plot. However, we need to look at SHAP-dependent graphs to see the precise shape of the relationship. A Shapley value for a feature and an instance may be found at each point on the summary plot. The feature determines the location on the y-axis, while the Shapley value determines the position on the x-axis. From low to high, the colour denotes the value of the characteristic. As overlapping points are jittered in the direction of the y-axis, we can see how the Shapley values are distributed over each feature. The relevance of the features determines their ranking. Figure 19 shows the dependence plot on Age and Aspartate_Aminotransferase.

5 Conclusion and Recommendation for Future Work

This section elaborates on the conclusion and recommendations for future work.

5.1 Conclusion

In this work, a recursive Gaussian support vector machine-based feature selection (RG-SVM) algorithm has been proposed to predict liver disease. The RG-SVM works in parallel with determining the feature weights in building the learning model. It gradually removes the attributes with the least values. The order in which the features are deleted provides an approximation of the ordering of feature importance. The proposed RG-SVM with the support of the recursive feature selection algorithm has outperformed other existing algorithms. The improved accuracy of 14 – 39% and 12- 20% of reduced MSE error over the LR, DT, KNN, and NB algorithms. The proposed RG-SVM algorithm produces 96% sensitivity score, which is 5–26% higher than the scores of the LR, DT, KNN, and NB algorithms such as 74, 70, 73, and 91 respectively. Similarly, the specificity scores are 64, 26, 45, 43, and 98 for the LR, DT, KNN, NB, and RG-SVM algorithms. The specificity of RG-SVM algorithm produced 34–72% improved results over the existing algorithms.

Our study clearly showed that a simple model consisting of RG-SVM could identify liver disease patients with a clinically significant prediction with a high degree of accuracy. As a result, the physicians can use these features that have been chosen in this work to diagnose the disease phenotype and disease process.

5.2 Recommendation for Future Work

Other parameters which are affecting liver such as, smoking, alcohol intake, etc. will be considered for better prediction results. The application of this model will be further enhanced to predict liver biopsy or decrease the need for it among liver disease patients.

Authors contribution Prasannavenkatesan Theerthagiri: Conceptualization, Methodology, Software, Software, Validation, Writing- Reviewing and Editing. Sahana Devarayapattana Siddalingaiah: Data curation, Writing- Original draft preparation.

Data availability Data available on request from the authors.

Declarations

Competing Interests Nil.

Ethical and informed consent for data used: Nil.

Conflict of Interests The authors declare no conflict of interest.

References

1. Karthik S, Priyadarishini A, Anuradha J, Tripathy BK (2011) Classification and rule extraction using rough set for diagnosis of liver disease and its types. AdvApplSci Res 2(3):334–345
2. Sepanlou SG, Safiri S, Bisignano C et al (2020) The global, regional, and national burden of cirrhosis by cause in 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet Gastroenterol Hepatol 5:245–266
3. Stewart WB, Wild CP (2014) World cancer report 2014. WHO Press, Geneva, Switzerland 978-92-832-0432-9
4. Theerthagiri P (2022) Predictive analysis of cardiovascular disease using gradient boosting based learning and recursive feature elimination technique. Intell Syst Appl 16:200121
5. Lazar C, Taminau J, Meganck S, Steenhoff D, Coletta A, Molter C et al (2012) A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Trans Comput Biol Bioinform 9(4):1106–1119. https://doi.org/10.1109/TCBB.2012.33
6. Murugesan S, Bhuvaneswaran RS, Khanna Nehemiah H, KeerthanaSankari S, Nancy JY (2021) Feature Selection and Classification of Clinical Datasets Using Bioinspired Algorithms and Super Learner. Comput Math Methods Med 17(2021):6662420. https://doi.org/10.1155/2021/6662420
7. Wang XD, Chen RC, Yan F et al (2019) Fast adaptive K-means subspace clustering for high-dimensional data. IEEE Access 7:42639–42651
8. Chandrashekar G, Sahin F (2014) A survey on feature selection methods. Comput Electr Eng 40:16–28
9. Suthaharan, S (2016) Support vector machine. In Machine Learning Models and Algorithms for Big Data Classification. Integrated Series in Information Systems; Springer: Boston, MA, USA; pp. 207–235, ISBN 978–1–4899–7640–6. https://doi.org/10.1007/978-1-4899-7641-3_9
10. Butkiewicz M, Lowe E, Mueller R, Mendenhall J, Teixeira P, Weaver C, Meiler J (2013) Benchmarking ligand-based virtual high-throughput screening with the pubchem database. Molecules 18:735–756
11. Sanz H, Valim C, Vegas E et al (2018) SVM-RFE: selection and visualization of the most relevant features through non-linear kernels. BMC Bioinf 19:432. https://doi.org/10.1186/s12859-018-2451-4

12. Hassan A, Abou-Taleb AS, Mohamed OA, Hassan AA (2013) Hybrid Feature Selection approach of ensemble multiple Filter methods and wrapper method for Improving the Classification Accuracy of Microarray Data Set. Int J Comput Sci Inf Technol Secur 3:185–190
13. Dong RZ, Yang X, Zhang XY, Gao PT, Ke AW, Sun HC, Zhou J, Fan J, Cai JB, Shi GM (2019) Predicting overall survival of patients with hepatocellular carcinoma using a three-category method based on DNA methylation and machine learning. J Cell Mol Med 23(5):3369–3374
14. Orooji Azam, Kermani Farzaneh (2021) Machine Learning Based Methods for Handling Imbalanced Data in Hepatitis Diagnosis. Front Health Inf 10:57. https://doi.org/10.30699/fhi.v10i1.259
15. G. Shobana and K. Umamaheswari (2021) Prediction of Liver Disease using Gradient Boost Machine Learning Techniques with Feature Scaling, 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), pp. 1223–1229, https://doi.org/10.1109/ICCMC51019.2021.9418333
16. Lin X, Li C, Zhang Y, Su B, Fan M, Wei H (2017) Selecting Feature Subsets Based on SVM-RFE and the Overlapping Ratio with Applications in Bioinformatics. Molecules 23(1):52. https://doi.org/10.3390/molecules23010052
17. Admassu Tsehay, Subhashni Rajkumar, Napa Komal Kumar, Manivannan Jijendira, Duraisamy Pradeep, Engidaye Minychil (2022) Random forest and support vector machine based hybrid liver disease detection. Bull Electric Eng Inf 11:1650–1656. https://doi.org/10.11591/eei.v11i3.3787
18. Assegie TA (2021) Support Vector Machine And K-Nearest Neighbor Based Liver Disease Classification Model. https://doi.org/10.35882/ijeeemi.v3i1.2
19. Sontakke S, Lohokare J, Dani R (2017) Diagnosis of liver diseases using machine learning. Int Conf Emerg Trends Innov ICT (ICEI) 2017:129–133. https://doi.org/10.1109/ETIICT.2017.7977023
20. Abdar M, Zomorodi-Moghadam M, Das R (2017) I-Hsien Ting Corrigendum to “Performance Analysis of Classification Algorithms on early detection of Liver disease.” Expert Syst Appl 67:239–251
21 Obayya Marwa I M, Areed Nihal F F, Abdulhadi Abdulhadi Omar (2016) Article: Liver Cancer Identification using Adaptive Neuro-Fuzzy Inference System. Int J Comput Appl 140(8):1–7
22. Farokhzad MR, Ebrahimi L (2016) A novel adaptive neuro fuzzy inference system for the diagnosis of liver disease. International Journal of Academic Research in Computer Engineering 1(1):61–66
23. Mehmood M, Alshammari N, Alanazi SA, Ahmad F (2022) Systematic Framework to Predict Early-Stage Liver Carcinoma Using Hybrid of Feature Selection Techniques and Regression Techniques, Complexity, vol. 2022, Article ID 7816200, 11 pages. https://doi.org/10.1155/2022/7816200
24. Hayashi Y, Fukunaga K (2016) Accuracy of rule extraction using a recursive-rule extraction algorithm with continuous attributes combined with a sampling selection technique for the diagnosis of liver disease. Inform Med Unlocked 5:26–38
25. Padmakala S, Subasini CA, Karuppiah SP, Sheeba A (2021) ESVM-SWRF: Ensemble SVM-based sample weighted random forests for liver disease classification. Int J Numer Meth Biomed Engng 37(12):e3525. https://doi.org/10.1002/cnm.3525
26. Deshmukh S, Kawale P, Khopade M, Sawant A, Palan Y (2022) Liver disease diagnosis using machine learning algorithm. International Journal of Scientific Research in Engineering & Management 06(05)
27. Liu Y-X, Liu X, Cen C, Li X, Liu J-M, Ming Z-Y, Yu S-F, Tang X-F, Zhou L, Yu J, Huang K-J, Zheng S-S (2021) Comparison and development of advanced machine learning tools to predict nonalcoholic fatty liver disease: An extended study. Hepatobil Pancreatic Dis Int, Volume 20, Issue 5, Pages 409–415, ISSN 1499–3872, https://doi.org/10.1016/j.hbpd.2021.08.004
28. Sharma N, Dev J, Mangla M et al (2021) A Heterogeneous Ensemble Forecasting Model for Disease Prediction. New Gener Comput 39:701–715. https://doi.org/10.1007/s00354-020-00119-7
29. Bukhari SNH, Webber J, Mehbodniya A (2022) Decision tree based ensemble machine learning model for the prediction of Zika virus T-cell epitopes as potential vaccine candidates. Sci Rep 12:7810. https://doi.org/10.1038/s41598-022-11731-6
30. Abdullah DM, Abdulazeez AM (2021) Machine learning applications based on SVM classification a review. Qubahan Academic Journal 1(2):81–90
31. Krause, J, Gulshan, V, Rahimy, E, Karth, P, Widner, K, Corrado, GS, ... , Webster, DR (2018) Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. Ophthalmol, 125(8), 1264–1272. https://doi.org/10.1016/j.ophtha.2018.01.034
32. Theerthagiri P, Ruby AU (2022) RFFS: Recursive random forest feature selection based ensemble algorithm for chronic kidney disease prediction. Expert Syst 39(9):e13048
33. Botchkarev A (2018) Performance metrics (error measures) in machine learning regression, forecasting and prognostics: properties and typology. arXiv preprint arXiv:1809.03006

34. Theerthagiri P, Vidya J (2022) Cardiovascular disease prediction using recursive feature elimination and gradient boosting classification techniques. Expert Syst 39(9):e13064
35. Chicco D, Tötsch N, Jurman G (2021) The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Mining 14(1):1–22
36. Abdalrada, AS, Yahya, O, Alaidi, AH, Alaidi, M, Hussein, A, Alrikabi, H, Alquraishi, T (2019) A Predictive model for liver disease progression based on logistic regression algorithm. Periodic Eng Nat Sci (PEN). 7 1255–1264. https://doi.org/10.21533/pen.v7i3.667
37. Tokala, S, Hajarathaiah K, Sai G, Srinivasrao B, Lakshmikanth N, Pathipati N, Satish A, Murali KE (2023) Liver Disease Prediction and Classification using Machine Learning Techniques. Int J Adv Comput Sci Appl 14. https://doi.org/10.14569/IJACSA.2023.0140299
38. Lanjewar M, Parab J, Shaikh A, Sequeira M (2022) CNN with machine learning approaches using ExtraTreesClassifer and MRMR feature selection techniques to detect liver diseases on cloud. Clust Comput. https://doi.org/10.1007/s10586-022-03752-7
39. Khan MAR, Afrin F, Prity FS et al (2023) An effective approach for early liver disease prediction and sensitivity analysis. Iran J Comput Sci. https://doi.org/10.1007/s42044-023-00138-9
40. Sun C, Xu A, Liu D, Xiong Z, Zhao F, Ding W (2020) Deep Learning-Based Classification of Liver Cancer Histopathology Images Using Only Global Labels. IEEE J Biomed Health Inform 24(6):1643–1651. https://doi.org/10.1109/JBHI.2019.2949837
41. Amin R, Yasmin R, Ruhi S, Rahman MH, Reza MS (2023) Prediction of chronic liver disease patients using integrated projection based statistical feature extraction with machine learning algorithms. Inf Med Unlocked 36:101155. https://doi.org/10.1016/j.imu.2022.101155
42. Salau AO, Jain S (2019) Feature extraction: a survey of the types, techniques, applications. In: 2019 International Conference on Signal Processing and Communication (ICSC). IEEE, pp 158–164
43. Rubia Y, Ruhul A (2023) Design of Novel Feature Union for Prediction of Liver Disease Patients: A Machine Learning Approach. https://doi.org/10.1007/978-981-19-8032-9_36
44. Zaheer MM, Nirmala P (2022) An Effective Approach to Detect Liver Disorder using KNN Algorithm in Comparison with Decision Tree Algorithm to Measure Accuracy. Cardiometry; Issue 25; December 2022; p.1038–1046; https://doi.org/10.18137/cardiometry.2022.25.10381046
45. Kumar, P, Thakur, R (2021) Liver disorder detection using variable- neighbor weighted fuzzy K nearest neighbor approach. Multimed Tools Appl 80. https://doi.org/10.1007/s11042-019-07978-3
46 Padmakala S, Subasini CA, Karuppiah SP, Sheeba A (2021) ESVM-SWRF: Ensemble SVM-based sample weighted random forests for liver disease classification. Int J Numer Method Biomed Eng Dec;37(12):e3525. https://doi.org/10.1002/cnm.3525. (**Epub 2021 Sep 21. PMID: 34431606**)
47. Dritsas E, Trigka M (2023) Supervised Machine Learning Models for Liver Disease Risk Prediction. Computers 12(1):19. https://doi.org/10.3390/computers12010019
48. Ghazal TM, Rehman AU, Saleem M, Ahmad M, Ahmad S, Mehmood F (2022) Intelligent Model to Predict Early Liver Disease using Machine Learning Technique, 2022 International Conference on Business Analytics for Technology and Security (ICBATS), Dubai, United Arab Emirates, pp. 1–5, https://doi.org/10.1109/ICBATS54253.2022.9758929
49. Mehmood M, Alshammari N, Alanazi SA, Ahmad F (2022) Systematic Framework to Predict Early-Stage Liver Carcinoma Using Hybrid of Feature Selection Techniques and Regression Techniques, Complexity, vol. 2022, Article ID 7816200, 11 pages https://doi.org/10.1155/2022/7816200.
50. Praveen AD, Vital TP, Jayaram D, Satyanarayana LV (2021) Intelligent liver disease prediction (ILDP) system using machine learning models. In: Intelligent Computing in Control and Communication: Proceeding of the First International Conference on Intelligent Computing in Control and Communication (ICCC 2020). Springer, Singapore, pp 609–625
51. Ambesange S, Uppin VAR, Patil S, Patil V (2020) Optimizing Liver disease prediction with Random Forest by various Data balancing Techniques, 2020 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM), Bengaluru, India, pp. 98–102, https://doi.org/10.1109/CCEM50674.2020.00030.
52. Krisnabayu, RY, Ridok, A, Budi, AS (2021) Hepatitis detection using random forest based on SVM-RFE (recursive feature elimination) feature selection and SMOTE. In: ACM International Conference on Proceeding Series, pp. 151–156. https://doi.org/10.1145/3479645.3479668
53. Marwa IM, Nihal FF, Abdulhadi O (2016) Liver Cancer Identification using Adaptive Neuro-Fuzzy Inference System. Int J Comput Appl 140:1–7. https://doi.org/10.5120/ijca2016909402

54. Kaggle repository: https://www.kaggle.com/datasets/uciml/indian-liver-patient-records. Accessed 15 May 2023
55. Admassu Tsehay, Salau Ayodeji, Omeje Crescent, Braide Sepiribo (2023) Multivariate sample similarity measure for feature selection with a resemblance model. Int J Electric Comput Eng 13:3359–3366. https://doi.org/10.11591/ijece.v13i3.pp3359-3366
56. Amin Ruhul, Yasmin Rubia, SabbaRuhi Md, Habibur Rahman Md, Reza Shamim (2023) Prediction of chronic liver disease patients using integrated projection based statistical feature extraction with machine learning algorithms. Inf Med Unlocked 36:101155. https://doi.org/10.1016/j.imu.2022.101155
57. Jain S, Salau AO, Meng W (Reviewing editor) (2019) An image feature selection approach for dimensionality reduction based on kNN and SVM for AkT proteins, Cogent Engineering, 6:1https://doi.org/10.1080/23311916.2019.1599537
58. Ismail WN (2023) Snake-Efficient Feature Selection-Based Framework for Precise Early Detection of Chronic Kidney Disease. Diagnostics 13(15):2501. https://doi.org/10.3390/diagnostics13152501

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.