Hindawi
Complexity
Volume 2022, Article ID 7816200, 11 pages
https://doi.org/10.1155/2022/7816200

WILEY | Hindawi

Research Article

Systematic Framework to Predict Early-Stage Liver Carcinoma Using Hybrid of Feature Selection Techniques and Regression Techniques

Marium Mehmood,[1] Nasser Alshammari ,[2] Saad Awadh Alanazi ,[2] and Fahad Ahmad [3]

[1]*Department of Computer Sciences, Kinnaird College for Women, Lahore 54700, Punjab, Pakistan*
[2]*Department of Computer Science, College of Computer and Information Sciences, Jouf University, Sakaka, Aljouf 72341, Saudi Arabia*
[3]*Department of Basic Sciences, Deanship of Common First Year, Jouf University, Sakaka, Aljouf 72341, Saudi Arabia*

Correspondence should be addressed to Fahad Ahmad; drfahadahmadmian@gmail.com

Received 10 October 2021; Accepted 7 December 2021; Published 6 January 2022

Academic Editor: Sampath Pradeep

The liver is the human body's mandatory organ, but detecting liver disease at an early stage is very difficult due to the hiddenness of symptoms. Liver diseases may cause loss of energy or weakness when some irregularities in the working of the liver get visible. Cancer is one of the most common diseases of the liver and also the most fatal of all. Uncontrolled growth of harmful cells is developed inside the liver. If diagnosed late, it may cause death. Treatment of liver diseases at an early stage is, therefore, an important issue as is designing a model to diagnose early disease. Firstly, an appropriate feature should be identified which plays a more significant part in the detection of liver cancer at an early stage. Therefore, it is essential to extract some essential features from thousands of unwanted features. So, these features will be mined using data mining and soft computing techniques. These techniques give optimized results that will be helpful in disease diagnosis at an early stage. In these techniques, we use feature selection methods to reduce the dataset's feature, which include Filter, Wrapper, and Embedded methods. Different Regression algorithms are then applied to these methods individually to evaluate the result. Regression algorithms include Linear Regression, Ridge Regression, LASSO Regression, Support Vector Regression, Decision Tree Regression, Multilayer Perceptron Regression, and Random Forest Regression. Based on the accuracy and error rates generated by these Regression algorithms, we have evaluated our results. The result shows that Random Forest Regression with the Wrapper Method from all the deployed Regression techniques is the best and gives the highest $R2$-Score of 0.8923 and lowest MSE of 0.0618.

1. Introduction

The liver is one of the essential organs in the human body [1]. Liver and gallbladder are used for absorption, digestion, and processing of food. The liver has multiple functions carried out through liver cells (hepatocytes) [2]. It creates half of the cholesterol of the body, and the rest of it comes from food, which will be helpful in making bile that supports digestion. Proteins and hormones production are the liver's primary task to control sugar level in blood and clotting in the blood. Therefore, the liver's location and functionality make it prone to diseases. Presently, many types of liver diseases lead to the highest death rate in human beings [3]. The most common liver diseases are hepatitis A, B, C, D, E fibrosis, cirrhosis, fatty liver, and liver cancer. The most severe of all is liver cancer; it can only be triggered if cancer cells are developed inside the liver.

Cancer can only be cured when spotted at an early stage [4]. Uncontrolled growth of abnormal cells in the liver can cause cancer in it. It is difficult to diagnose early cancer stage of the liver as its symptom's appearance is meager [5]. The liver may work abnormally when there is an excessive intake of alcohol or smoking, or a patient has hepatitis B or C and has type II diabetes. These abnormalities may lead to liver

cancer. Hepatocyte (liver cell), if affected, can cause cancer of the liver, which is hepatocellular carcinoma. Liver cancer symptoms may include white stools, hepatitis B or C, severe jaundice, severe weight loss, severe vomiting, abdominal pain, and many more [2]. Diagnosing liver diseases at an early stage will then be the next task with the help of soft computing techniques.

Soft computing provides methodologies to solve real-life problems [6]. Soft computing's main aim is to accomplish a higher accuracy rate, less ambiguity, estimated reasoning to robustness, cost-friendly solutions, and controllability. Filtering irrelevant attributes, data mining techniques of artificial intelligence are used, which have predictive model representation [7]. Data mining techniques include different Feature Selection Techniques that are used for filtering like Filter, Wrapper, Embedded, and so forth. Data mining is necessary to eliminate irrelevant data in the dataset [8]. The performance of many mining algorithms is reduced with extensive features or attributes. Hence, Feature Selection Techniques are applied to mine data. Feature selection's main objective is to improve model performance, reduce cost, and avoid overfitting for fast and accurate results [9]. Filter, Embedded, and Wrapper approaches are used for feature selection.

As the dataset and its proper refining are critical, the researcher focuses on feature extraction techniques. They provide the ability to construct features and select appropriate features, feature ranking, and assessment model for feature validation [10]. Data mining techniques play a significant role in mining the data as a suitable subset is selected for manipulations from the whole dataset. Data mining techniques include Feature Selection Techniques, which help in eliminating irrelevant features [11]. Feature Selection Techniques are grouped into three categories, that is, Filter Method, Wrapper Method, and the hybrid of these two (Embedded Method) [12].

In the Filter Method, data mining algorithms are not used and the significance of attributes is calculated by observing the fundamental properties of the data. Mostly, essential features are calculated, and low scoring attributes are removed [13]. The Wrapper approach's principal characteristic is to measure the feature subset quality by the performance of the data mining algorithm applied to that feature subset. The Embedded Method is another method of feature selection [14]. It combined feature selection algorithm with learning algorithms. The decision tree is most appropriate for all Embedded Methods. The performance of these selection techniques can be evaluated and the model which is comparatively most efficient can be chosen [15].

The Filter Method is one of the feature extraction methods, which involves statistical analysis of the dataset being extracted from the aggregate data without applying Machine Learning algorithms. Filter Methods are of univariate and multivariate types, which include Information gain, Pearson's Correlation, Chi-squared, Quasi-Constant Elimination, Odds Ratio, Duplicate Feature Elimination, Constant Feature Elimination, Correlated Feature Extraction, and many more [16]. Wrapper Method is another method in which the subset selection algorithm uses learning algorithms Forward Elimination, Backward Elimination, and Bidirectional Elimination to find the best subset from the entire feature subspace with higher prediction performance [17]. Researchers also give the hybrid techniques of Filter and Wrapper to evaluate more relevant results for feature selection, which includes Embedded Methods and provides a trade-off solution by embedding feature selection in learning algorithms and returns both the selected subset and learning algorithm, which will further be processed. Widely used Embedded Methods are Regularization Methods, which include L1 Regularization (LASSO) and L2 Regularization (Ridge) methods [18].

Finally, the above-mentioned Feature Selection Techniques will then be evaluated with the help of Regression techniques. These techniques help us in measuring accuracy and error rates for the selected dataset. Regression is a predictive analysis technique in data mining. It includes Linear Regression, Ridge Regression, LASSO Regression, Elastic Net Regression, Decision Tree Regression, Support Vector Regression, Multilayer Perceptron Regression (Neural Network Regression), Random Forest Regression, and many more [19–24]. These techniques will further be adopted by Feature Selection Techniques to evaluate some statistical results for data extraction.

The organization of the paper is as follows: in Section 2, the material and methods are described followed by the results of the identified models. Section 3 discusses the methods and their ultimate performance in carcinoma detection. Finally, the conclusion and future work are presented in the Sections 5 and 5, respectively.

2. Materials and Methods

Many artificial intelligence techniques solve medical-related issues. We will extract useful features that help in the detection of liver cancer with the help of feature extraction techniques in which Filter Methods, Embedded Methods, and Wrapper Methods are very helpful. These methods will be implemented in the regression model to train our data accordingly, and some useful features will be extracted from these models. These algorithms help us in the extraction of useful data that will be useful in further treatment while diagnosing early liver disease. The whole process is represented in Figure 1.

Hundreds of thousands of features are carried out in a dataset from which useful features are extracted from the whole dataset by using Feature Selection Techniques; these include Filter Methods, Wrapper Methods, and Embedded Methods for training the data. We will evaluate our Regression models on the Anaconda Python tool for the required results.

In the proposed model, data collection is the first step that we have collected from an online source National Institute of Health database for liver cancer [25]. Fifty features are extracted, and data of 240 patients are taken. The data include some demographic as well as medical-related features, that is, Age, City, Area, Education, Marital Status, Occupation, Hobbies, Siblings, Weight, Height, Gender, Arthritis, Family History, Hereditary Status, Blood Pressure, Diabetes, Smoking, Alcohol Consumption, Heart Disease,

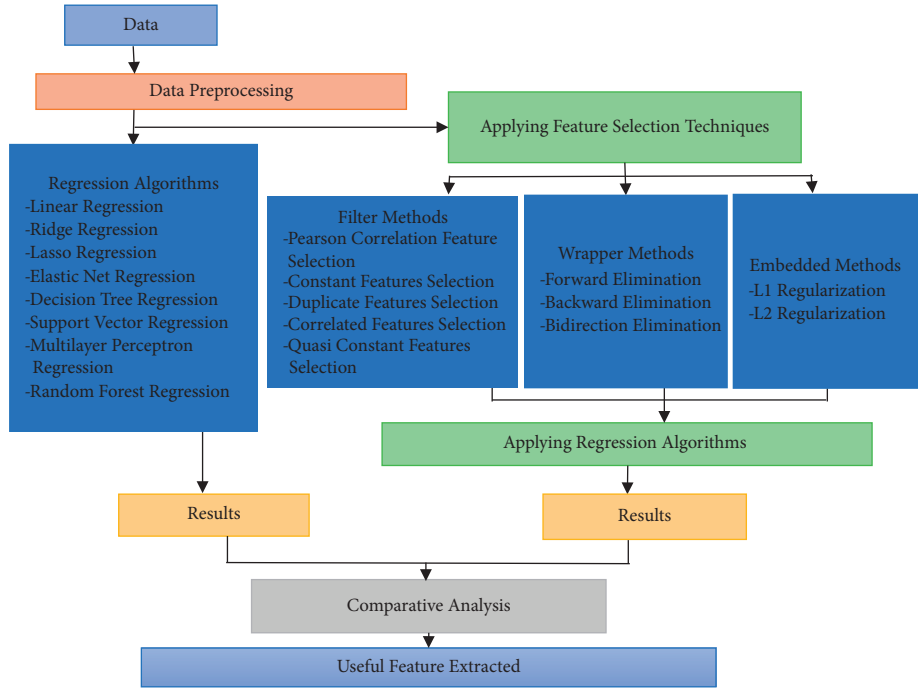

Figure 1: Liver cancer diagnosis using Feature Selection Techniques.

Osteoporosis, Medicine Intake, Jaundice, Gallbladder Inflammation, Kidney Stone, Vomiting, Nausea, Temperature, Liver Function Test, Asthma, White Stools, Eye Color, Last Blood Test, Cancer Patient, Pneumonia, Hepatitis Type, Chilling, Bronchitis, Cough, Weight Loss, Loss of Appetite, Back Pain, Enlarged Liver, Sputum Color, Calcium Level, Obesity, Fatigue/Weakness, Chest Pain, Hemoglobin Level, and Sputum Level Result. These data are converted into numerical values and are detailed in Table 1. We have used Anaconda Prompt (Jupyter Notebook) tool to solve our problem, and the language used in it is Python. The system specifications are as follows: 10th Generation Intel Core i7 processor, Windows 10 Pro 64-bit operating system, 32 GB DDR4 memory, 1 TB SSD hard drive, and NVIDIA RTX A3000 graphics card.

Initially, we will apply Regression techniques to compute the accuracy and error rates for the preprocessed dataset.

2.1. Regression. Regression is a statistical analysis method used to determine the relationship between one dependent variable and multiple independent variables [26]. There are different types of Regression models or algorithms by which one can easily estimate the criticality of the problem accordingly [27]; that is,

(i) Linear Regression identified the relationship existing between predicted values and targeted values [19].

(ii) Ridge Regression examines the labels based on a statistical-based fundamental relationship. This method gives lower values than the values of variance obtained from the least-square method and is more preferable [28].

(iii) LASSO Regression performs two main tasks, regularization and feature selection. Its goal is to minimize the prediction error [29].

(iv) Elastic Net Regression is the combination of Ridge Regression and LASSO Regression, which works by penalizing the model [30].

(v) Decision Tree Regression is like a model that decides with the help of tree structure. This model gives all possible results, costs for input and time complexity, and so forth and is a supervised learning algorithm [22].

(vi) Support Vector Regression method uses high-dimensional feature space to compute a linear function where the nonlinear function is the input data. It reduces error to increase Regression performance [31].

(vii) Multiple Layer Perceptron Regression is an algorithm to learn the potential nonlinear function approximator [32].

(viii) Random Forest Regression performs the Regression and classification with the use of multiple decision trees [33].

We apply these Regression algorithms with Feature Selection Techniques on our dataset, which is divided into training and testing datasets in the ratios of 80% and 20%, respectively.

2.2. Feature Selection Techniques. Feature Selection Techniques are involved in preprocessing of data in data mining. Reducing the size of the dataset and removing irrelevant features are the main tasks of data mining, which help in

Table 1: Illustration of the preprocessed dataset.

Identified parameters	Demographic and clinical values
Age	69
City	1
Area	1
Education	1
Marital Status	1
Occupation	3
Hobbies	3
Siblings	2
Weight	58
Height	158
Gender	0
Arthritis	0
Family History	0
Hereditary	0
Status	1
Blood Pressure	0
Diabetes	1
Smoking	0
Alcohol Consumption	0
Heart Disease/Attack	0
Osteoporosis	0
Medicine Intake	0
Jaundice	0
Vomiting	1
Nausea	1
Temperature	0
Liver Function Test	0
Asthma	0
White Stools	0
Eye Color	0
Cancer Patient	1
Pneumonia	0
Hepatitis Type	1
Chilling	0
Bronchitis	1
Cough	1
Weight Loss	1
Loss of Appetite	1
Back Pain	1
Sputum Color	0
Calcium Level	0
Obesity	1
Fatigue/Weakness	1
Chest Pain	1
Hemoglobin Level	1
Sputum Level Result	0.5

improving the efficiency and accuracy of Machine Learning algorithms. It also reduced overfitting [14, 34]. Feature Selection Techniques can be broadly divided into the Filter, Wrapper, and Embedded approaches.

2.2.1. Filter Method. The Filter Method's main criteria are based on the relationship among the variable to be predicted and set of features. In a Filter Method, a subset of features are selected independently from a set of all features and used as an input in any Machine Learning algorithm. It uses statistical techniques to find the relationship between the input variable and the predicted variable [35]. An example of the Filter Method is as follows:

(i) Pearson's Correlation is based on the correlation matrix. The coefficient can be calculated between input and output variables as a measure of a linear relationship between the two. It is calculated by dividing the covariance of two variables with the product of their standard deviations [36].

(ii) Quasi-Constant features show the same value for the vast majority of the observation of dataset, so we do not consider these features in predicting the results. No rule is set for what should be the variance threshold for Quasi-Constant features [35].

(iii) Constant Feature Elimination method eliminates the constant features, that is, having the same values and target values with zero variance [35]. It is better to eliminate these features to avoid the repetition of data.

(iv) Correlated Feature Extraction mined the features' subset that is good enough when it is highly correlated with the output and uncorrelated with each other. Two or more features are correlated if they are close to each other in the linear space.

(v) Duplicate Feature Elimination disregards identical values of features that make them duplicate. They harm results accuracy as it increases time delays and overheads and does not help in improving the algorithm's training [35].

2.2.2. Wrapper Method. A Wrapper Method is similar to the Filter Method, but it uses a predefined Machine Learning algorithm and uses its performance as evaluation criteria instead of an independent measure for the subset evaluation [37]. The following are different types of Wrapper Methods:

(i) Forward Selection is a recursive method having no feature in the model initially. After each iteration, we keep on adding features one by one until the addition of new variable does not improve the model's performance [38].

(ii) Backward Elimination works in an opposite direction compared to the Forward Selection; we start with a full set of features and then remove one by one the insignificant features with less significance level. First, choose a significance level and then fit a model using all features and then consider the feature which has value higher than the significance level and remove those features which have value less than the significance level and then repeat the procedure [39].

(iii) Bidirectional Elimination is a hybrid of both Forward Elimination and Backward Elimination. Firstly, choose the significance level for entering and exiting the model and then add features and check the value of feature which is less than the significance level value and then do not add that feature and then perform Backward Elimination steps and

check significance level value for exiting the model with feature value; if its value is less than the significance value of that particular feature, then remove it.

2.2.3. Embedded Method. Embedded Method provides a trade-off key between Filter and Wrapper Methods by embedding feature selection into the model learning. This method takes care of each iteration of the model training process and extracts features that contribute highest to the training for a particular iteration [40]. Regularization is the most common Embedded Method which penalizes a feature given a coefficient threshold. It includes LASSO, Ridge, and Elastic Net Regression, from which two are as follows:

(i) L1 Regularization (LASSO) penalizes a feature with coefficient to 0, if it is insignificant. Hence, features with coefficient = 0 will be removed, and the rest of the features will again pass through the LASSO Regularization technique [41].

(ii) L2 Regularization (Ridge) gives penalty if a feature is insignificant, equivalent to the square of the magnitude of the coefficients. It does not shrink coefficients to zero. So, Ridge regression puts limitations on the coefficients; if the value of the coefficients is considerable, then the optimization function is penalized [41].

Data mining and Machine Learning are twisted together with relatable properties in which feature purification or selection is considered as one of their core procedures. Our model's performance is highly dependent on them. We have used a dataset of 240 patients and trained our model on this dataset, and Regression techniques have also been applied to it. The following is Table 2 that will give initial results for R2-Score and Mean Square Error (MSE) for training and testing of data without applying any feature selection technique.

The R2-Score is a critical indicator for assessing the effectiveness of a Regression-based model. It is also known as the coefficient of determination and is abbreviated *R*-squared. It operates by calculating the amount of variation in the dataset-explained predictions. The MSE is the average of the squares of the inaccuracies. The bigger the number, the bigger the error. In this situation, error refers to the gap between observed and expected values. So negative and positive numbers do not even cancel each other out; we square each difference.

Now, the Wrapper Method of Feature Selection Technique is applied, which has the following types: Forward Selection, Backward Elimination, and Bidirectional Elimination. The results in Table 3 are for Feature Selection Technique using Wrapper Method with Forward Selection for all regression algorithms.

The results in Table 4 are for Feature Selection Technique using the Wrapper Method with Backward Elimination for all regression algorithms.

The results in Table 5 are for Feature Selection Technique using the Wrapper Method with Bidirectional Elimination for all regression algorithms.

Embedded Method has the two following methods that will be used to calculate accuracy and error rate for training and testing datasets. The following are results for L1 Regularization (LASSO) for all Regression algorithms represented in Table 6.

The following are results for L2 Regularization (Ridge) described in Table 7 for all regression algorithms.

The Filter Method has the following methods that will be used to calculate accuracy and error rate for training and testing datasets. The following are results for a Pearson Correlation for all Regression algorithms in Table 8.

Table 9 shows the results for Constant Feature Elimination for all Regression algorithms.

The following are results for Quasi-Constant Elimination described in Table 10 for all regression algorithms.

Table 11 shows the results of Correlated Feature Elimination for regression algorithms.

The following results in Table 12 are for Duplicate Feature Elimination for all regression algorithms.

The above tables show that Random Forest Regression has the best results from all Regression techniques. Table 13 gives collective results of Random Forest Regression for all Feature Selection Methods that have highest accuracy and lowest error rate in comparison to all Regression techniques.

2.2.4. Complete Data Training. As all Wrapper Methods give similar results, we need to calculate *R*2-Score and MSE using full dataset for all of them. Results are presented in Table 14.

2.2.5. Unseen Dataset. When all of the data are used to train the model using various algorithms, the problem of evaluating the models and choosing the best one remains. The main goal is to figure out which model has the lowest generalization error out of all the others. In other words, which model outperforms all others in predicting future or unseen datasets? This necessitates the use of a technique that allows the model to be trained on one dataset and tested on another. Now, we will execute these techniques using an unseen dataset. We have a history of 60 patients now. Then we will again extract features using Wrapper Methods for this dataset. Now the identified features from unseen dataset by using the trained model are shown in Figures 2–4 as.

2.2.6. Accuracy and Error Rates. We represent our statistical analysis in the form of the above and below given tables that show results for Feature Selection Techniques using Regression algorithms. The best given results during testing using Regression algorithms are from Random Forest Regression having highest *R*2-Score and lowest MSE as shown in Table 15.

3. Discussion

The last stage of cancer disease is the main cause of increasing mortality rate nowadays. In most cases, liver cancer at early stage is not detected, which is devastating for humans, and, due to late diagnosis, this cancer leads to

Table 2: Regression algorithms' with accuracy and error rate during training and testing.

Models	Accuracy ($R2$-Score) training (↑)	Accuracy ($R2$-Score) testing (↑)	MSE training (↓)	MSE testing (↓)
Linear Regression	0.5534	0.2594	0.126	0.1694
Ridge Regression	0.5534	0.2606	0.126	0.1692
LASSO Regression	0.3327	0.3528	0.154	0.1583
Elastic Net Regression	0.3954	0.4147	0.1466	0.1506
Decision Tree Regression	0.9993	0.0587	0.0051	0.1909
Support Vector Regression	0.5382	0.3475	0.1281	0.159
Multilayer Perceptron Regression	0.2222	0.0461	0.1662	0.1922
Random Forest Regression	**0.8921**	**0.4851**	**0.0619**	**0.1412**

Table 3: Feature Selection Technique (Wrapper Method with Forward Selection)-based Regression algorithms with accuracy and error rate during training and testing.

Models	Accuracy ($R2$-Score) training (↑)	Accuracy ($R2$-Score) testing (↑)	MSE training (↓)	MSE testing (↓)
Linear Regression	0.5534	0.2594	0.126	0.1694
Ridge Regression	0.5534	0.2606	0.126	0.1692
LASSO Regression	0.3327	0.3528	0.154	0.1583
Elastic Net Regression	0.3954	0.4147	0.1466	0.1506
Decision Tree	0.9993	0.0264	0.0051	0.1942
Support Vector Regression	0.5382	0.3475	0.1281	0.159
Multilayer Perceptron Regression	0.3268	0.3162	0.1547	0.1627
Random Forest Regression	**0.8923**	**0.5031**	**0.0618**	**0.1387**

Table 4: Feature Selection Technique (Wrapper Method with Backward Elimination)-based Regression algorithms with accuracy and error rate during training and testing.

Models	Accuracy ($R2$-Score) training (↑)	Accuracy ($R2$-Score) testing (↑)	MSE training (↓)	MSE testing (↓)
Linear Regression	0.5534	0.2594	0.126	0.1694
Ridge Regression	0.5534	0.2606	0.126	0.1692
LASSO Regression	0.3327	0.3528	0.154	0.1583
Elastic Net Regression	0.3954	0.4147	0.1466	0.1506
Decision Tree	0.9993	0.134	0.0051	0.1831
Support Vector Regression	0.5382	0.3475	0.1281	0.159
Multilayer Perceptron Regression	0.2775	0.0863	0.1602	0.1881
Random Forest Regression	**0.8923**	**0.5031**	**0.0618**	**0.1387**

Table 5: Feature Selection Technique (Wrapper Method with Bidirectional Elimination)-based Regression algorithms with accuracy and error rate during training and testing.

Models	Accuracy ($R2$-Score) training (↑)	Accuracy ($R2$-Score) testing (↑)	MSE training (↓)	MSE testing (↓)
Linear Regression	0.5534	0.2594	0.126	0.1694
Ridge Regression	0.5534	0.2606	0.126	0.1692
LASSO Regression	0.3327	0.3528	0.154	0.1583
Elastic Net Regression	0.3954	0.4147	0.1466	0.1506
Decision Tree	0.9993	0.1771	0.0051	0.1785
Support Vector Regression	0.5382	0.3475	0.1281	0.159
Multilayer Perceptron Regression	−0.5937	−0.751	0.238	0.2604
Random Forest Regression	0.8923	0.5031	0.0618	0.1387

death. For diagnosis, the initial step is to find significant features that will be best demonstration for someone's illness. Here, we have identified a scheme that will be helpful in extracting useful subset of features from a set of exhaustive features that could be helpful in further cancer or other diseases treatment. For this, data training is done using

Table 6: L1 Regularization-based Regression algorithms with accuracy and error rate during training and testing.

Models	Accuracy ($R2$-Score) training (↑)	Accuracy ($R2$-Score) testing (↑)	MSE training (↓)	MSE testing (↓)
Linear Regression	0.5534	0.2594	0.126	0.1694
Ridge Regression	0.5534	0.2606	0.126	0.1692
LASSO Regression	0.3327	0.3528	0.154	0.1583
Elastic Net Regression	0.3954	0.4147	0.1466	0.1506
Decision Tree	0.9993	0.2416	0.0051	0.1714
Support Vector Regression	0.5382	0.3475	0.1281	0.159
Multilayer Perceptron Regression	0.44	0.0821	0.1411	0.1885
Random Forest Regression	**0.8903**	**0.4844**	**0.0624**	**0.1413**

Table 7: L2 Regularization-based Regression algorithms with accuracy and error rate during training and testing.

Models	Accuracy ($R2$-Score) training (↑)	Accuracy ($R2$-Score) testing (↑)	MSE training (↓)	MSE testing (↓)
Linear Regression	0.5534	0.2594	0.126	0.1694
Ridge Regression	0.5534	0.2606	0.126	0.1692
LASSO Regression	0.3327	0.3528	0.154	0.1583
Elastic Net Regression	0.3954	0.4147	0.1466	0.1506
Decision Tree	0.9993	0.0103	0.0051	0.1958
Support Vector Regression	0.5382	0.3475	0.1281	0.159
Multilayer Perceptron Regression	0.4026	0.3289	0.1457	0.1612
Random Forest Regression	0.8917	0.4798	0.062	0.1419

Table 8: Pearson Correlation-based Regression algorithms with accuracy and error rate during training and testing.

Models	Accuracy ($R2$-Score) training (↑)	Accuracy ($R2$-Score) testing (↑)	MSE training (↓)	MSE testing (↓)
Linear Regression	0.5534	0.2594	0.126	0.1694
Ridge Regression	0.5534	0.2612	0.126	0.1692
LASSO Regression	0.3511	0.3364	0.1518	0.1603
Elastic Net Regression	0.4078	0.4019	0.1451	0.1522
Decision Tree	0.9993	0.0103	0.0051	0.1958
Support Vector Regression	0.7063	0.1806	0.1022	0.1782
Multilayer Perceptron Regression	−18.6064	−17.7664	0.8347	0.8526
Random Forest Regression	0.8912	0.4864	0.0622	0.141

Table 9: Constant Feature Elimination-based Regression algorithms with accuracy and error rate during training and testing.

Models	Accuracy ($R2$-Score) training (↑)	Accuracy ($R2$-Score) testing (↑)	MSE training (↓)	MSE testing (↓)
Linear Regression	0.5519	0.2923	0.1262	0.1656
Ridge Regression	0.5519	0.2933	0.1262	0.1654
LASSO Regression	0.3511	0.3364	0.1518	0.1603
Elastic Net Regression	0.4078	0.4019	0.1451	0.1522
Decision Tree	0.9993	−0.0004	0.0051	0.1969
Support Vector Regression	0.707	0.1797	0.102	0.1782
Multilayer Perceptron Regression	−4303.1142	−4577.799	12.3669	13.3172
Random Forest Regression	0.8907	0.4903	0.0623	0.1405

Table 10: Quasi-Constant Elimination-based Regression algorithms with accuracy and error rate during training and testing.

Models	Accuracy ($R2$-Score) training (↑)	Accuracy ($R2$-Score) testing (↑)	MSE training (↓)	MSE testing (↓)
Linear Regression	0.5534	0.2594	0.126	0.1694
Ridge Regression	0.5534	0.2612	0.126	0.1692
LASSO Regression	0.3511	0.3364	0.1518	0.1603
Elastic Net Regression	0.4078	0.4019	0.1451	0.1522
Decision Tree	0.9993	0.0318	0.0051	0.1936

Table 10: Continued.

Models	Accuracy ($R2$-Score) training (↑)	Accuracy ($R2$-Score) testing (↑)	MSE training (↓)	MSE testing (↓)
Support Vector Regression	0.7063	0.1806	0.1022	0.1782
Multilayer Perceptron Regression	−33.3501	−24.4743	1.1048	0.9933
Random Forest Regression	0.8912	0.4864	0.0622	0.141

Table 11: Correlated Feature Elimination-based Regression algorithms' with accuracy and error rate during training and testing.

Models	Accuracy ($R2$-Score) training (↑)	Accuracy ($R2$-Score) testing (↑)	MSE training (↓)	MSE testing (↓)
Linear Regression	0.5534	0.2594	0.126	0.1694
Ridge Regression	0.5534	0.2606	0.126	0.1692
LASSO Regression	0.3327	0.3528	0.154	0.1583
Elastic Net Regression	0.3954	0.4147	0.1466	0.1506
Decision Tree	0.9993	0.0103	0.0051	0.1958
Support Vector Regression	0.5382	0.3475	0.1281	0.159
Multilayer Perceptron Regression	0.4564	0.389	0.139	0.1538
Random Forest Regression	0.8903	0.4844	0.0624	0.1413

Table 12: Duplicate Feature Elimination-based Regression algorithms with accuracy and error rate during training and testing.

Models	Accuracy ($R2$-Score) training (↑)	Accuracy ($R2$-Score) testing (↑)	MSE training (↓)	MSE testing (↓)
Linear Regression	0.5534	0.2594	0.126	0.1694
Ridge Regression	0.5534	0.2612	0.126	0.1692
LASSO Regression	0.3511	0.3364	0.1518	0.1603
Elastic Net Regression	0.4078	0.4019	0.1451	0.1522
Decision Tree	0.9993	-0.0327	0.0051	0.2
Support Vector Regression	0.7063	0.1806	0.1022	0.1782
Multilayer Perceptron Regression	−238.646	−185.6239	2.9181	2.6886
Random Forest Regression	**0.8912**	**0.4864**	**0.0622**	**0.141**

Table 13: Random Forest-based Regression algorithms with accuracy and error rate during training and testing.

Models	Accuracy ($R2$-Score) training (↑)	Accuracy ($R2$-Score) testing (↑)	MSE training (↓)	MSE testing (↓)
Simple Random Forest Regression	0.8921	0.4851	0.0619	0.1412
Random Forest Regression with Forward Selection (Wrapper Method)	**0.8923**	**0.5031**	**0.0618**	**0.1387**
Random Forest Regression with Backward Elimination (Wrapper Method)	**0.8923**	**0.5031**	**0.0618**	**0.1387**
Random Forest Regression with Bidirectional Elimination (Wrapper Method)	**0.8923**	**0.5031**	**0.0618**	**0.1387**
Random Forest Regression with L1 Regularization (Embedded Method)	0.8903	0.4844	0.0624	0.1413
Random Forest Regression with L2 Regularization (Embedded Method)	0.8903	0.4844	0.0624	0.1413
Random Forest Regression with Pearson Correlation (Filter Method)	0.8917	0.4798	0.062	0.1419
Random Forest Regression with Constant Feature Elimination (Filter Method)	0.8912	0.4864	0.0622	0.141
Random Forest Regression with Quasi-Constant (Filter Method)	0.8907	0.4903	0.0623	0.1405
Random Forest Regression with Correlated Features (Filter Method)	0.8912	0.4864	0.0622	0.141
Random Forest Regression with Duplicate Features (Filter Method)	0.8912	0.4864	0.0622	0.141

Table 14: Wrapper Method-based accuracy and error rate for complete dataset during training.

Models	Accuracy ($R2$-Score) training (↑)	MSE training (↓)
Random Forest Regression (Forward Selection)	0.904	0.0591
Random Forest Regression (Backward Elimination)	0.9053	0.0587
Random Forest Regression (Bidirectional Elimination)	0.9053	0.0587

```
( ' Gender ',
  ' occupation ' ,
  ' nausea ' ,
  ' Area ' ,
  ' Marital Status ' ,
  ' Eye Color ' ,
  ' gall bladder inflamation ' ,
  ' Alcohal consumption ' ,
  ' Diabities ' ,
  ' Siblings ' ,
  ' back pain ')
```

Figure 2: Forward Selection process.

```
( ' Area ' ,
  ' Eye Color ' ,
  ' gall bladder inflamation ' ,
  ' Family history ' ,
  ' Heriditary Status ' ,
  ' LFT ' ,
  ' Jaundice ' ,
  ' Last Blood Test ' ,
  ' Age ' ,
  ' occupation ' ,
  ' back pain ' )
```

Figure 3: Backward Elimination process.

```
( ' Siblings ' ,
  ' occupation ' ,
  ' gall bladder inflamation ' ,
  ' Eye Color ' ,
  ' Alcohal consumption ' ,
  ' back pain ' ,
  ' Gender ' ,
  ' Diabities ' ,
  ' nausea ' ,
  ' Area ' ,
  ' Marital Status ' )
```

Figure 4: Bidirectional Elimination process.

Table 15: Wrapper Method-based accuracy and error rate for complete dataset during testing.

Models	Accuracy ($R2$-Score) testing (↑)	MSE testing (↓)
Random Forest Regression (Forward Selection)	0.9405	0.061
Random Forest Regression (Backward Elimination)	0.94	0.0613
Random Forest Regression (Bidirectional Elimination)	0.9416	0.0604

Feature Selection Techniques and Regression models to extract optimized features. Our framework will extract subset of features from a huge dataset and its results will be helpful to identify patient's health condition which further takes decisions either to take patient home or for further specialized diagnosis procedures. These techniques will also help the patients, medical facility providers, and governments to reduce the diagnosis expenses. Our results are

immensely gainful for early and efficient detection of patient's health that can facilitate a person with proper treatment for its health issues within time. Nowadays, Machine Learning is a vast field that can solve many of the medical-related issues; moreover, it is very important to diagnose a disease before it gets fatal and cancer is a slow poison that slowly gets through all the organs of the body, so it is very necessary to diagnose the disease at the right time [42]. Our focused work has been advantageous, and less time will be taken with superior $R2$-Score and MSE. The training and testing sets have individual results for $R2$-Score and MSE and hence it is easier to detect difference between the original and predicted values in the dataset using $R2$-Score and MSE for the identified problem.

4. Conclusion

Medical-related issues could be diagnosed at early stages mostly with the help of soft computing techniques, Machine Learning, and data mining. Data mining is the initial step to diagnose a disease as appropriate features' selection is of utmost importance. For this reason, we have used Feature Selection Techniques which provide appropriate selection of features which makes processing of disease detection easier. Those techniques prove them strongly helpful in data mining and Machine Learning techniques. Feature Selection Techniques have multiple methods to mine the dataset which include Filter Methods, Wrapper Methods, and Embedded Methods. Our work shows that Wrapper Method is most appropriate for the detection of features that are most important for the diagnosis of diseases as this method has highest $R2$-Score and lowest MSE for the extracted features. Then we have evaluated some Wrapper Methods which are Forward Elimination, Backward Elimination, and Bidirectional Elimination. The results of these methods were then tested by Regression algorithms. We calculated $R2$-Score and MSE with the help of these Regression techniques too. The higher $R2$-Score and the lower MSE show higher detection correctness of disease. Our research concludes that Wrapper Method-based Features Selection Techniques have better results and then Random Forest Regression is applied which gives best results for $R2$-Score and MSE on our dataset.

5. Future Work

After appropriate features' selection, our task could be to find whether patients' condition is critical or not. If patients' condition is critical, we can recommend them scanning: either Computed Tomography Scan or Positron Emission Tomography Scan. This procedure can be done using further soft computing techniques which include Neural Networks, Genetic Algorithms, and Adaptive Neurofuzzy System. Afterwards, image recognition could be done on these scans which will help to identify the location of the tumor in the specific organ. Further, the results of the image recognition techniques will identify cancer cells and their spreading in other organs.

Data Availability

The dataset generated for this study is publicaly available.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Authors' Contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

References

[1] 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.

[2] M. I. M. Obayya, N. F. F. Areed, and A. O. Abdulhadi, "Liver cancer identification using adaptive neuro-fuzzy inference system," *International Journal Of Computer Applications*, vol. 140, 2016.

[3] S. Karthik, A. Priyadarishini, J. Anuradha, and B. Tripathy, "Classification and rule extraction using rough set for diagnosis of liver disease and its types," *Advances in Applied Science Research*, vol. 2, no. 3, pp. 334–345, 2011.

[4] G. S. Stein and K. P. Luebbers, *Cancer: Prevention, Early Detection, Treatment and Recovery*, John Wiley & Sons, Hoboken, NJ, USA, 2019.

[5] J. Lafemina and G. F. Whalen, "Liver cancer: hepatocellular carcinoma," in *Cancer: Prevention, Early Detection, Treatment and Recovery*p. 17, John Wiley & Sons, Hoboken, NJ, USA, 2019.

[6] R. Janghel, A. Shukla, and K. Verma, "Soft computing based expert system for hepatitis and liver disorders," in *Proceedings of the 2016 IEEE International Conference on Engineering and Technology (Icetech)*, pp. 740–744, IEEE, Coimbatore, India, March 2016.

[7] N. Jothi, N. A. A. Rashid, and W. Husain, "Data mining in healthcare–a review," *Procedia Computer Science*, vol. 72, pp. 306–313, 2015.

[8] S. Beniwal and J. Arora, "Classification and feature selection techniques in data mining," *Research Gate*, vol. 7, 2012.

[9] T. V. Afanasieva and P. V. Platov, "Cardiovascular health analysis system using machine learning model," in *Proceedings of the 10th International Scientific and Practical Conference*, Aachen, Germany, September 2021.

[10] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning*, vol. 3, 2003.

[11] D. Oreskia, S. Oreski, and B. Klicek, "effects of dataset characteristics on the performance of feature selection techniques," *Applied Soft Computing*, Elsevier, vol. 52, , 2017.

[12] H. Shi, H. Li, D. Zhang, C. Cheng, and X. Cao, "An efficient feature generation approach based on deep learning and feature selection techniques for traffic classification," *Computer Networks*, vol. 132, 2018.

[13] N. Sanchez-Marono, A. Alonso-Betanzos, and M. Tombilla-Sanromán, "Filter methods for feature selection–a comparative study," in *Proceedings of the International Conference on Intelligent Data Engineering and Automated Learning*, Springer, Birmingham, UK, December 2007.

[14] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, Elsevier, vol. 40, 2014.

[15] F. Ros and S. Guillaume, "From supervised instance and feature selection algorithms to dual selection: a review," in *Sampling Techniques for Supervised and Unsupervised Taks*p. 45, Springer Nature Switzerland Ag, Heidelberg, Germany, 2020.

[16] M. Labani, P. Moradi, F. Ahmadizar, and M. Jalili, "A novel multivariate filter method for feature selection in text classification problems," *Engineering Applications of Artificial Intelligence*, vol. 70, pp. 25–37, 2018.

[17] S. S. M. Sumi and A. Narayanan, "Improving classification accuracy using combined Filter+Wrapper feature selection technique," in *Proceedings of the 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, Coimbatore, India, February 2019.

[18] M. Lu, "Embedded feature selection accounting for unknown data heterogeneity," *Expert Systems with Applications*, vol. 119, pp. 350–361, 2019.

[19] H. Kikuchi, C. Hamanaga, H. Yasunaga, H. Matsui, H. Hashimoto, and C.-I. Fan, "Privacy-preserving multiple linear regression of vertically pertitionaed real medical datasets," *Journal of Information Processing*, vol. 10, 2018.

[20] C.-D. Tseng, C.-S. Shieh, Y.-J. Huang, P.-J. Chao, and T.-F. Lee, "Using lasso regression based svm classification to improve the predictive performance of radiation-induced pneumonitis complication in breast cancer," *Journal of the Chinese Institute of Engineers*, vol. 41, no. 8, pp. 660–666, 2018.

[21] C. Giglio and S. D. Brown, "Using elastic net regression to perform spectrally relevant variable selection," *Journal of Chemometrics*, vol. 32, no. 8, p. E3034, 2018.

[22] S. Murugan, B. M. Kumar, and S. Amutha, "Classification and prediction of breast cancer using linear regression, decision tree and random forest," in *Proceedings of the 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC)*, Mysore, India, September 2017.

[23] I. K. Omurlu, F. Cantas, M. Ture, and H. Ozturk, "An empirical study on performances of multilayer Perceptron, logistic regression, ANFIS, KNN and bagging cart," *Journal of Statistics & Management Systems*, vol. 23, 2020.

[24] S. Sharma, T. Choudhury, and A. Aggarwal, "Breast cancer detection using machine learning algorithms," *IEEE*, vol. 5, 2018.

[25] "Liver Datasets [Online]," http://Biometry.Nci.Nih.Gov/Cdas/Datasets/Plco/13/.

[26] F. Hoffmann, T. Bertram, R. Mikut, M. Reischl, and O. Nelles, "Benchmarking in classification and regression," *Wires Data Mining And Knowledge Discovery*, vol. 9, 2019.

[27] R. V. Mccarthy, *Predictive Models Using Regression*, Springer Nature Switzerland Ag, Heidelberg, Germany, 2019.

[28] M. Toğaçar, Z. Cömert, and B. Ergen, "Application of breast cancer diagnosis based on a combination of convolutional neural networks, ridge regression and linear discriminant analysis using invasive breast cancer images processed with autoencoders," *Medical Hypotheses*, Elsevier, vol. 135, 2019.

[29] V. Fonti, "Feature selection using lasso," *Research Paper In Business Analytics*, vol. 26, 2017.

[30] S. D. Brown and C. Giglio, "Using elastic net regression to perform spectrally relevant variable selection," *Journal of Chemometrics*, vol. 32, 2018.

[31] H. Mahmoodian and L. Ebrahimian, "Using support vector regression in gene selection and fuzzy rule generation for relapse time prediction of breast cancer," *Biocybernetics and Biomedical Engineering*, Elsevier, vol. 36, 2016.

[32] T.-L. Lin, H.-W. Tseng, Y. Wen, F.-W. Lai, C.-H. Lin, and C.-J. Wang, "Reconstruction algorithm for lost frame of multiview videos in wireless multimedia sensor Network based on deep learning multilayer Perceptron regression," *IEEE Sensors Journal*, vol. 18, 2018.

[33] X. Zhou, X. Zhu, Z. Dong, and W. Guo, "Estimation of biomass in wheat using random forest regression algorithm and remote sensing data," *The Crop Journal*, vol. 4, no. 3, pp. 212–219, 2016.

[34] B. Ghotra, S. McIntosh, and A. E. Hassan, "A large-scale study of the impact of feature selection techniques on defect classification models," *IEEE*, vol. 12, 2017.

[35] A. Degeest, B. Frénay, and M. Verleysen, *About Filter Criteria For Feature Selection In Regression*, Springer Nature Swizerland Ag, Heidelberg, Germany, 2019.

[36] P. Schober, C. Boer, and L. A. Schwarte, "Correlation coefficients: appropriate use and interpretation," *Anesthesia & Analgesia*, vol. 126, no. No. 5, pp. 1763–1768, 2018.

[37] J. Suto, S. Oniga, and P. P. Sitar, "Comparison of wrapper and filter feature selection algorithms on human activity recognition," *IEEE-International Conference on Computers Communications and Control*, vol. 6, 2016.

[38] L. Puggini and S. Mcloone, "Forward selection component analysis: algorithms and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2395–2408, 2017.

[39] J. Park, K. Li, and H. Zhou, "K-fold subsampling based sequential backward feature elimination," in *Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods ICPRAM*, pp. 423–430, Rome, Italy, February 2016.

[40] H. Zhu, N. Bi, J. Tan, and D. Fan, "An embedded method for feature selection using kernel parameter descent support vector machine," in *Pattern Recognition And Computer Vision*, p. 11, Springer, Heidelberg, Germany, 2018.

[41] A. Farbahari, T. Dehesh, and M. H. Gozashti, "The usage of lasso, ridge, and linear regression to explore the most influential metabolic variables that affect fasting blood sugar in type 2 Diabetes patients," *Romanian Journal Of Diabetes Nutrition & Metabolic Diseases*, vol. 26, 2019.

[42] M. Mehmood, E. Ayub, F. Ahmad et al., "Machine learning enabled early detection of breast cancer by structural analysis of mammograms," *Computers, Materials & Continua*, vol. 67, no. 1, pp. 641–657, 2021.