ASEAN Journal of Science and Engineering

Journal homepage: http://ejournal.upi.edu/index.php/AJSE/

A Hybrid Classification Algorithm for Abdomen Disease Prediction

S. Vijayarani[1], C. Sivamathi[2,], P. Tamilarasi[3]*

[1]Assistant Professor, Department of Computer Science, Bharathiar University, Coimbatore, Tamilnadu, India
[2]Assistant Professor, Department of Computer Science, PSG College of Arts & Science, Coimbatore, Tamilnadu, India
[3]M.Phil Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore, Tamilnadu, India.
Correspondence: E-mail: sivamathi@psgcas.ac.in

ABSTRACT

Data mining is the process of sorting through large data sets to identify patterns and establish relationships to solve problems through data analysis. Data mining holds great potential for the healthcare industry to enable health systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and reduce costs. Data mining techniques consist of detection of Anomaly, learning the Association rules, Classification, Clustering, Regression, Time series analysis, and Summarization. In data mining, classification techniques are much popular in medical diagnosis and predicting diseases. Classification techniques are used to predict various diseases such as heart disease, lung cancer, breast cancer, liver diseases, and kidney diseases. The main objective of this work is to predict abdomen diseases like kidney and liver diseases. The work aims to predict liver diseases such as Cirrhosis, Bile Duct, Chronic Hepatitis, Liver Cancer, and Acute Hepatitis using Classification algorithms. The work also aims to predict kidney diseases such as Acute Nephritic Syndrome, Chronic Kidney disease, Acute Renal Failure, and Chronic Glomerulonephritis using Classification algorithms. This work proposes a novel hybrid classification algorithm called WRFSVM (Weighted Random Forest Support Vector Machine) for the prediction of liver diseases and kidney diseases.

ARTICLE INFO

Article History:
Submitted/Received 23 Des 2021
First revised 19 Feb 2022
Accepted 27 Apr 2022
First available online 28 Apr 2022
Publication date 01 Dec 2023

Keyword:
Acute nephritic syndrome,
Acute renal failure,
Chronic glomerulonephritis,
Chronic kidney disease,
Classification algorithms,
Liver diseases,
Liver function test,
Random forest,
Ripper,
SVM.

1. INTRODUCTION

Data mining can be referred to as the detection of relationships in large databases mechanically and in some cases, it is used for predicting relationships based on the results discovered (Vijayarani & Dhayanand, 2015). Large data items can be accessed by using data mining concepts (Karthik *et al*., 2011). In the current generation, lots of data is being collected and warehoused in profitable viewpoints, such as web data, e-commerce, purchases at department/ grocery stores, bank/credit card transactions, and in the medical field. In the health industry, data mining provides numerous benefits such as detection of fraud in health insurance, accessibility of medical solutions to the patients at lower cost, detection of causes of diseases, and identification of medical treatment methods. The medical field contains a large amount of data that are required to be processed. Data mining in the medical field improves the quality of patient care and the prediction of health care patterns. Data mining tools facilitate us to discover unknown patterns, group the related items, and decisions making of healthcare-oriented problems.

The liver plays an important role in many bodily functions from protein production and blood clotting to cholesterol, glucose, and iron metabolism. The liver is liable for many critical functions within the body and should it happen to the diseased, the loss of those functions can cause significant damage to the body. Liver disease is also referred to as hepatic disease (Kumari & Godara, 2011). Symptoms of liver diseases include weakness and fatigue, weight loss, nausea, vomiting, and yellow discoloration of the skin (jaundice) (Kumari & Godara, 2011). Liver diseases are typically caused by inflammation or damaged hepatocytes; register a continual presence on the list of top ten fatal diseases in the world (Vijayarani & Divya, 2011). Risk factors of liver disease are alcoholism, autoimmune diseases, exposure to toxins, hereditary conditions, and viruses. The World Health Organization (WHO) gave reports that approximately 3% of the world ‘s population is infected with hepatitis C. 170 million people are chronically polluted and 3-4 million are newly infected every year.

Kidney disease is the result of the gradual loss of kidney function. The function of the kidneys is to filter wastes and excess fluids from the blood. When kidney disease reaches an advanced stage, dangerous levels of fluid, electrolytes, and wastes can build up inside the body. In the early stages of kidney disease, there may be few symptoms. Hence it is necessary to predict kidney diseases at an early stage. Chronic kidney disease occurs when a disease or condition impairs kidney function, causing kidney damage to worsen over several months or years. Diabetes, High blood pressure, Glomerulonephritis, Interstitial nephritis, Polycystic kidney disease, Prolonged obstruction of the urinary tract, kidney stones, and Vesicoureteral are the reasons of kidney diseases.

The paper is organized as follows. Section 2 gives the literature review, Section 3 discusses existing classification algorithms, Section 4 describes the proposed work, Section 5 shows the experimental results, and the conclusion is given in Section 6.

2. LITERATURE REVIEW

Vijayarani & Dhayanand, (2015) have classified four types of kidney diseases. They compared Support Vector Machine (SVM) and Naïve Bayes classification algorithms based on the performance factors of classification accuracy and execution time. From the results, it was found that SVM has produced accurate results. Hence it is considered as best classifier when compared with a Naïve Bayes classifier algorithm.

DOI: http://dx.doi.org/10. 17509/xxxx.xxxx
p- ISSN 2776-6098 e- ISSN 2776-5938

Vijayarani *et al*. (2015) have compared two neural network techniques, Back Propagation Algorithm (BPA) and Radial Basis Function (RBF) with one non-linear classifier Support Vector Machine (SVM). They compared the algorithms using WEKA 3.6.5 tool. The main objective of their work was to predict kidney diseases using classification algorithms. From the experimental results, they concluded, that the backpropagation (BPA) was the best and has high classification accuracy.

Karthik *et al*., (2011) has predicted the types of liver diseases using the Naivebayes, MLP (Multilayer Perceptron), and RBF (Radial Basis Function) algorithms. In the first phase, Artificial Neural Networks are applied for classifying liver disease. In a second phase, rough set rule induction using LEM (Learn by Example) algorithm is applied to generate classification rules. In the third phase, fuzzy rules are applied to recognize the type of liver disease.

Kara *et al*. (2006) had concentrated on the diagnosis of optic nerve disease through the analysis of pattern electroretinography (PERG) signals with the help of an artificial neural network (ANN). They implemented Multilayer feed-forward ANN trained with a Levenberg Marquart (LM) backpropagation algorithm. The classified patient details as healthy and diseased. The stated results show that the proposed method PERG could make an effective interpretation.

Omar *et al*. (2021) proposed a classification system for predicting HCV (Hepatitis C Virus). This work integrates PCA (Principle Component Analysis), Modified-Particle Swarm Optimization, and LS-SVM (Least Squares Support Vector Machine algorithms. The proposed system is composed of 4 main phases, they are Data Pre-Processing, Features Extraction, Parameter Optimization, and Classification. The PCA algorithm extracts the most efficient HCV patient features that support diagnoses and treatment. The input parameters for LS-SVM were optimized using a modified version of the PSO algorithm. LS-SVM algorithm is used to classify HCV patients into one of two classes – Live or Die.

Rajeswari and Reena (2010) has predicted different types of liver disorders like hepatitis, cirrhosis, liver tumors, and liver abscess. The algorithms used in this work are Naive Bayes, FT Tree, and K-Star algorithm. Finally, the author concluded that the FT Tree Algorithm is the best than the other ones. FT Tree algorithm gives 97.10% of accuracy.

Learning vector quantization (LVQ), two layers feed-forward perceptron trained with a backpropagation training algorithm, and Radial basis function (RBF) networks for the diagnosis of kidney stone disease (Emeto & Ugwu, 2016). They have compared the performance of all three neural networks based on their accuracy, time is taken to build the model, and training data set size. They used Waikato Environment for Knowledge Analysis (WEKA) tool for execution. Finally, from the experimental results, the authors concluded that multilayer layer perceptron trained with backpropagation is the best algorithm for kidney stone diagnosis.

Gulia *et al*. (2014) proposed computational intelligence techniques for Liver Patient Classification. The classification algorithms considered in this work are Multiple Linear Regression, Support Vector Machine, Multilayer FeedForward Neural Network, J-48, Random Forest, and Genetic Programming. The authors used ILPD (INDIA Liver Data Set) Data Set. Authors employed under sampling and oversampling for balancing it. The results obtained from experiments indicated that Random Forest over sampling outperformed all the other techniques.

Modified Rotation Forest algorithm for accurate liver classification. Modified Rotation Forest algorithm for UCI liver data set was based on the Multilayer Perceptron (MP) classification algorithm and Random Subset feature selection technique (Gulia *et al*., 2014).

DOI: http://dx.doi.org/10.17509/xxxx.xxxx
p- ISSN 2775-6793 e- ISSN 2775-6815

The dataset used in this work was INDIA Liver Data Set (ILPD). The highest obtained classification accuracy was produced by the Multilayer Perceptron algorithm whose accuracy was stated to be 75% for the BUPA data and the KStar classification algorithm was reported to classify with 73% accuracy, the records of ILPD

3. EXISTING CLASSIFICATION ALGORITHMS

Classification techniques are widely used in data mining to classify data among various classes. Classification approaches usually use a training set where all objects are previously associated with known class labels. Some of the significant classification algorithms are Rule-based classifier, Decision tree induction, K-Nearest neighbor classifier, Bayesian classifier, Artificial neural network, Support vector machine, Ensemble classifier and Regression trees.

3.1. Support Vector Machine

SVMs were first suggested by Vapnik in the 1960s for classification and have recently become an area of powerful research owing to developments in the techniques and theory coupled with extensions to regression and density estimation (Alves *et al*., 2015). Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for both classification and regression challenges. In SVM, each data item is plotted as a point in n-dimensional space, (where n is several features) with the value of each feature being the value of a particular coordinate. Then, classification is used for finding the hyper-plane that differentiates the two classes. SVM has a technique called the kernel trick. These are functions that take low-dimensional input space and transform it into a higher-dimensional space. It converts not separable problem to separable problem; these functions are called kernels (Banu & Gomanthy, 2013).

3.2. Ripper

RIPPER stands for Repeated Incremental Pruning to Produce Error Reduction. This classification algorithm was proposed by Cohen. This algorithm is based on association rules with reduced error pruning (REP), a very common and effective technique found in decision tree algorithms. To generate association rules with the REP algorithm, the training data are separated into a growing set and a pruning set. The growing set is the primary association rule which can be generated purely from the dataset using some heuristic methods. The growing set contains a huge set of rules that should be frequently simplified to form the pruning set. Thus, the simplification is done using typical pruning operators which may allow erasing a phrase from any single rule or different association rules.

3.3. Random Forest

The random forest algorithm is a supervised classification algorithm. This algorithm creates a forest with many trees. The term came from Random decision forests that were first proposed by Tin Kam Ho by Bell Labs in 1995. The method combines Bremen's "bagging" idea and the random selection of features. The random forest algorithm starts by randomly selecting "k" features out of the total "m" features. In the next stage, using the randomly selected "k" features the root node is found by using the best split approach. In the next stage, the daughter nodes are calculated using the same best split approach. The steps are repeated until the trees are formed, with a root node and having the target as the leaf node. Thus, Random Forests are a grouping of tree predictors where each tree depends on the values of a random vector sampled independently with the same distribution for all trees in the forest.

DOI: http://dx.doi.org/10. 17509/xxxx.xxxx
p- ISSN 2776-6098 e- ISSN 2776-5938

4. PROPOSED WORK: HYBRID WEIGHTED RANDOM FOREST SUPPORT VECTOR MACHINE (WRFSVM)

The proposed system takes a hybrid approach of SVM and Weighted Random Forest for classification. Initially, SVM is applied to the dataset, and class labels along with instances are acquired. In each classified group, Weighted Random Forests are applied to get a highly accurate classification. Since the proposed methodology is applied to a medical data field, accuracy is highly important. The proposed methodology uses 0.1 as the weight value in a weighted random forest. This weight value gives more accurate classifiers. Support Vector Machines provide a method for creating classification functions from a set of labelled training data, from which predictions can be made for subsequent data sets. The algorithm works as follows: First, the SVM parameters are initialized and then training of data is done for L parts. Now support vector measures are calculated. Then the dataset is classified based on SVM classification. Then move the classified SVM classes to the random forest classifier.

In each class best split on each code by using an individual decision tree is selected. It will grow to the largest extent possible without pruning. Then each decision tree predicts test data according to the values of n features and the final classification for a given test sample is selected by a majority of the T trees. Now retrieve the result for the random forest classifier. Finally, the hybrid weighted random forest classifier is calculated using formulae:

Algorithm: Hybrid WRFSVM

Input: Dataset X=(x1, y1)..... (x_n, y_L),

// x -attributes, y class, n data instance, and L number of class labels –contain both //training and testing data

Output: Liver disease prediction

1. Initializing parameter of SVM C, ε

   //Generate support vector model by Train SVM with train data
2. divides train data into L parts, each part containing vectors of class 1 to L
3. Calculate support vector $\vec{w}.\vec{x} + b = y$
4. Minimize in $(\vec{w}, b)\{\frac{||\vec{w}||^2}{2}$ subject to $y_i\,(\vec{w}.\vec{x} + b) \geq 1 \text{ for any } i = 1, \ldots, n\}$
5. For each $x_i \in X$ from test data
6. Classify xi using ($\vec{w}$,b)
7. Store sub classes in R_{svm}

   // Random forest classifier
8. For each sub class in R_{svm} do
9. From the training set with M samples, generate T bootstrap subsets
10. For each of the T subsets, randomly grow on independent decision trees by selecting the best split with n features on each code
11. Each decision tree is grown to the largest extent possible without pruning
12. Each decision tree predicts test data according to the values of n features, the final predicted classification for a given test sample is selected by the majority of the T trees.

DOI: http://dx.doi.org/10.17509/xxxx.xxxx
p- ISSN 2775-6793 e- ISSN 2775-6815

13. Store result in R_{rfc}
    // Hybrid weighted WRFSVM
14. For each test sample
15. $R_{final} = w_1 . R_{svm} + w_2 . R_{rfc}$ where $\sum w1 + w2 = 1$
16. End for
17. End for

Pseudo Code of Hybrid WRFSVM

4. RESULTS AND DISCUSSION

4.1. Results of Liver Diseases

To evaluate the performance of the proposed algorithm, experiments are conducted on the Matlab tool. The liver dataset was used in this experiment. It contains 583 instances. Here 169 instances are taken for testing and 414 for training. Here the proposed algorithm is compared with existing SVM, Random Forest, and Ripper algorithms. In this section, performance results are reported and discussed. The performance factors used are Accuracy, Precision, Recall, and F-Measures of the algorithms. These measures are shown in **Table 1**. **Figure 1** shows the chart representation of the accuracy of the algorithms.

Table 1. Accuracy measure for liver disease dataset.

Algorithms	Accuracy	Precision	Recall	F-measure
SVM	74.80	0.59	0.76	0.67
RIPPER	84.73	0.67	0.82	0.74
RANDOM FOREST	88.93	0.75	0.89	0.81
HYBRID WRFSVM	91.22	0.79	0.93	0.86

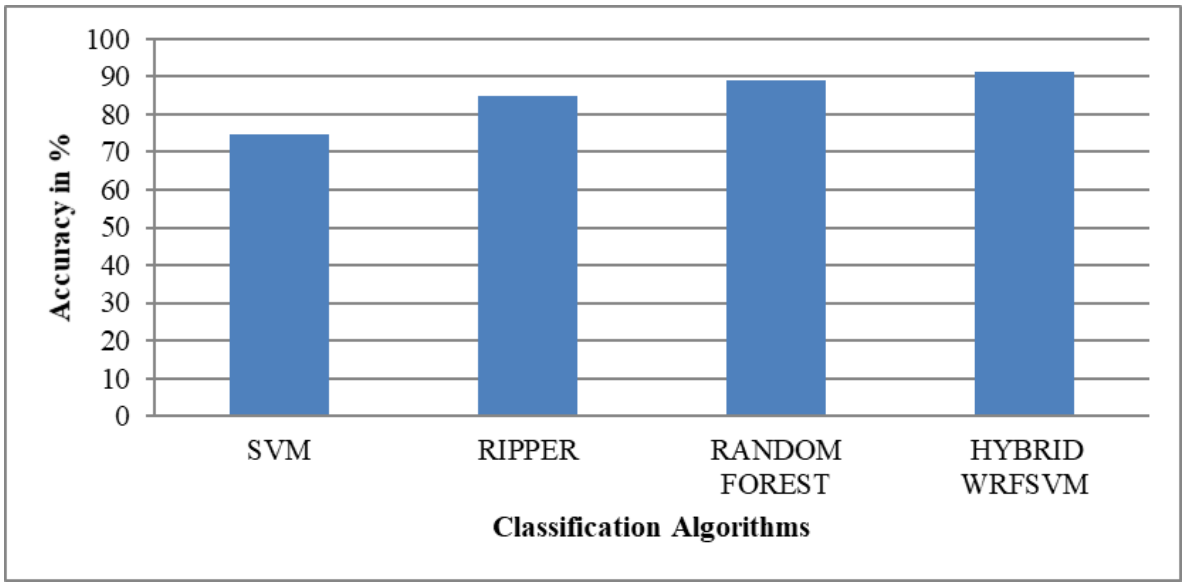

Figure 1. Accuracy measure for liver disease dataset.

DOI: http://dx.doi.org/10. 17509/xxxx.xxxx
p- ISSN 2776-6098 e- ISSN 2776-5938

Figure 2 shows the precision, recall, and f-measures of the algorithms. From the experimental result, it was found that the proposed hybrid WRFSVM has produced better results than the existing algorithms.

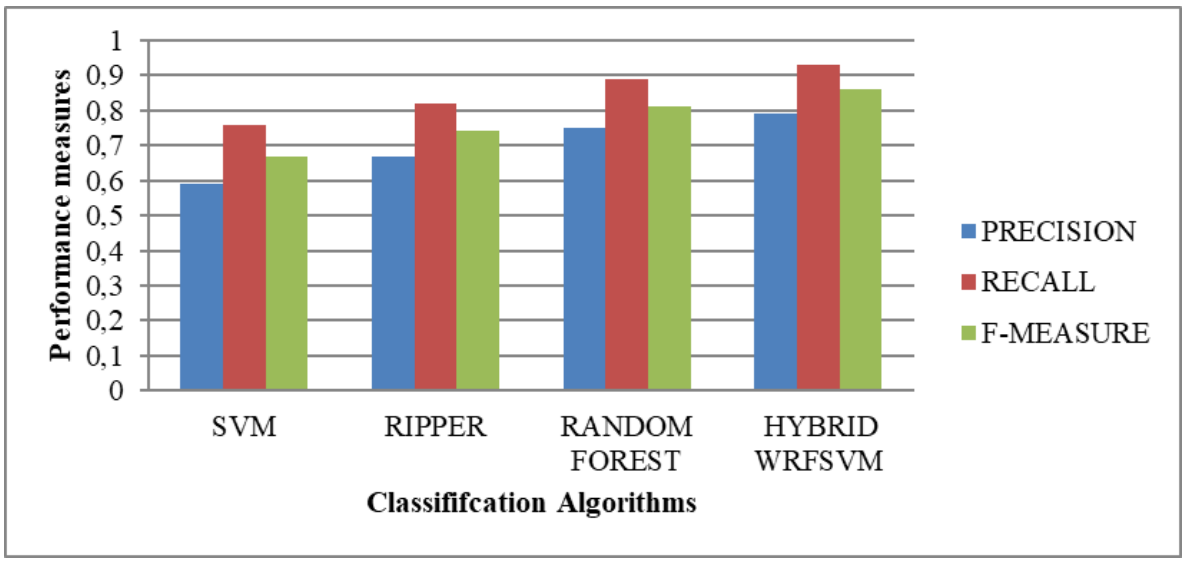

Figure 2. Precision, Recall, F-measure for liver dataset.

Table 2 shows the execution time of algorithms. **Figure 3** represents the execution time taken by the algorithms. It was found that hybrid WRFSVM executes with a minimum period of execution time than the other algorithms.

Table 2. Execution time analysis for liver disease prediction.

Algorithms	Execution Time in milli Seconds
Ripper	6300
Random Forest	5400
SVM	3290
WRFSVM	1450

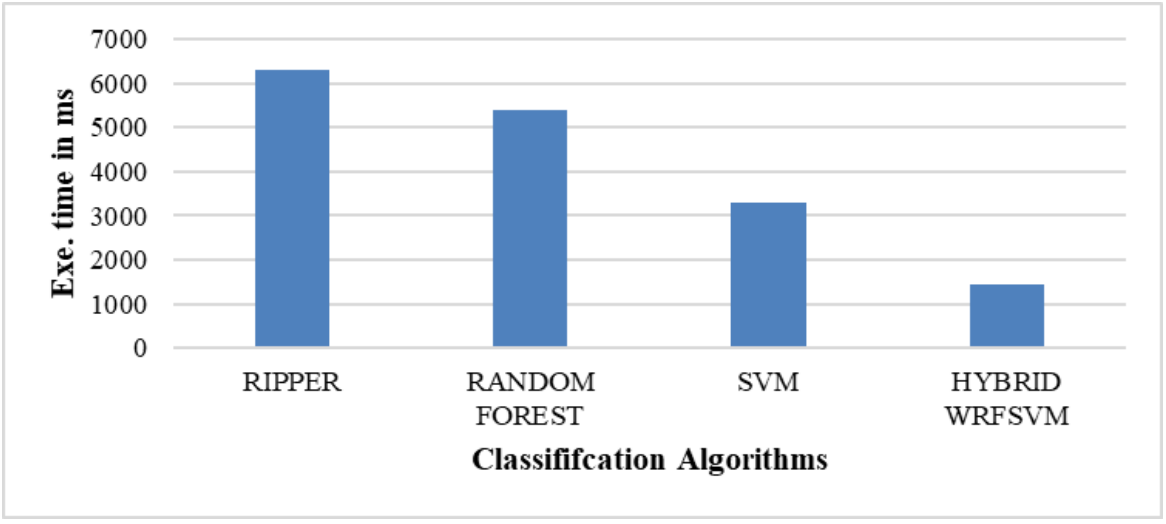

Figure 3. Execution time for liver disease dataset.

DOI: http://dx.doi.org/10.17509/xxxx.xxxx
p- ISSN 2775-6793 e- ISSN 2775-6815

Table 3 represents the number of patients who suffered from the different classifications of liver diseases, i.e. cirrhosis, bile duct, chronic hepatitis, liver cancer, and acute hepatitis. **Figure 4** shows the classification of liver diseases.

Table 3. Classification of liver disease.

Liver disease	Svm	Ripper	Random forest	Hybrid wrfsvm
Normal	126	145	148	152
Cirrhosis	15	18	15	18
Bileduct	11	10	13	13
Chronic hepatitis	8	8	9	10
Liver cancer	36	41	48	46

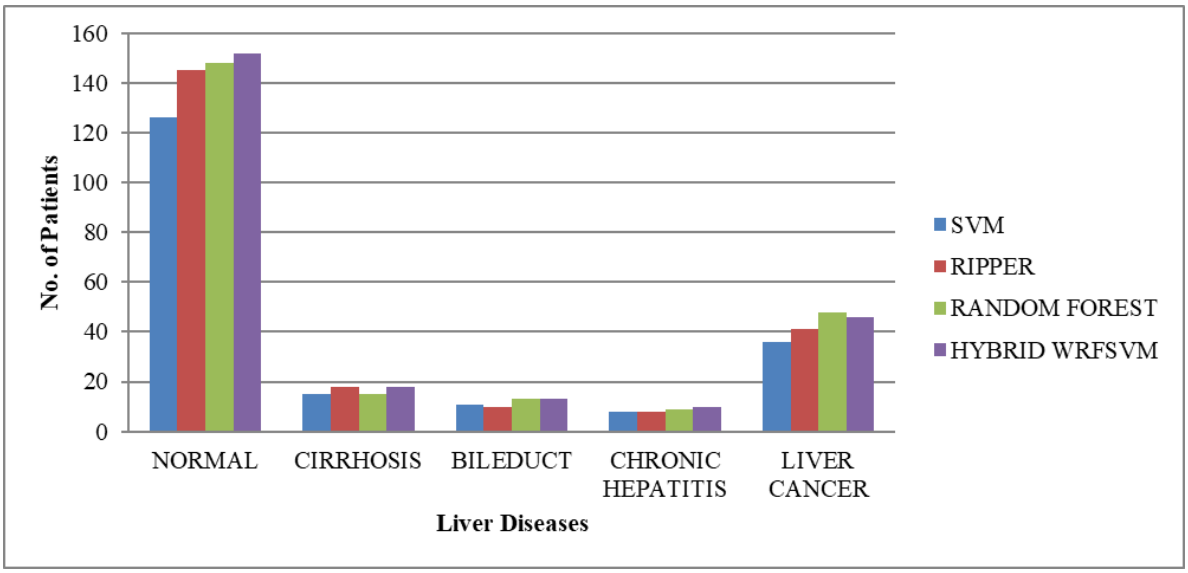

Figure 4. Liver disease classifications.

4.2. Results of Kidney Diseases

Kidney function test data were collected from several laboratories, medical centers, and hospitals. From these patient data, the synthetic kidney function test (KFT) dataset has been created for analysis of kidney disease. This dataset contains five hundred and eighty-four instances and five attributes. The renal affected diseases information's given in this dataset. **Table 4** shows the performance measures of existing and proposed algorithms like accuracy, precision, recall, and F-measure.

Table 4. Accuracy measure for kidney disease dataset.

Algorithm	Accuracy	Precision	Recall	F measure
Svm	75.1908	0.6839	0.7240	0.7034
Ripper	83.2061	0.7792	0.8262	0.8020
Random forest	88.8462	0.8292	0.8563	0.8425
Wrfsvm	93.4866	0.9039	0.9280	0.9158

DOI: http://dx.doi.org/10. 17509/xxxx.xxxx
p- ISSN 2776-6098 e- ISSN 2776-5938

Figure 5 shows a graphical representation of the table. **Table 5** shows the execution Time of existing and proposed algorithms for Kidney Disease Prediction.

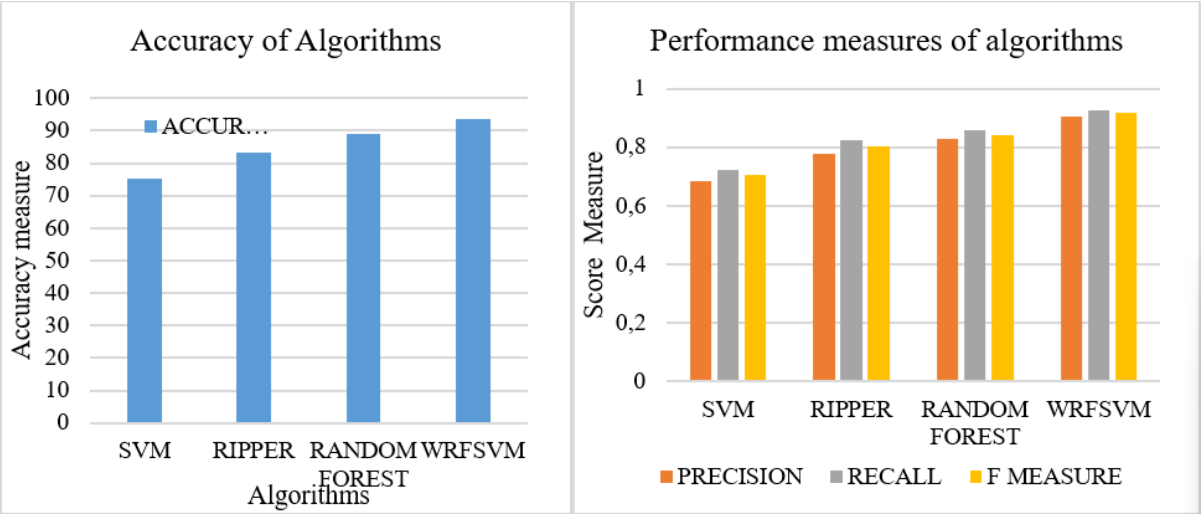

Figure 5. Performance measure for kidney disease dataset.

Table 5. Execution time for kidney disease prediction.

Algorithms	Execution Time in Seconds
Random forest	7.85
RIPPER	4.67
SVM	3.41
WRFSVM	1.67

Table 6 shows the result of classification of kidney diseases using proposed algorithm. **Figure 6** shows the chart representation. From the figure, it was found that the proposed algorithm executes better than the existing algorithms. **Figure 7** shows the chart for the same.

Figure 7 represents the kidney diseases classified by different types of classification algorithms: SVM, RIPPER, and RANDOM FOREST. A newly proposed hybrid WEIGHTED RANDOM FOREST SUPPORT VECTOR MACHINE is also used for kidney disease prediction. Based on chart analysis, the proposed WRFSVM gives the overall best classification of kidney diseases than other algorithms.

Table 6. Classification of kidney diseases.

Kidney diseases	SVM	Ripper	Random forest	WRFSVM
Acute nephritic syndrome	12	14	13	15
Chronic kidney disease	22	27	24	28
Acute renal failure	39	43	47	47
Chronic glomerulonephritis	89	98	105	108
Damaged	27	25	33	35

DOI: http://dx.doi.org/10.17509/xxxx.xxxx
p- ISSN 2775-6793 e- ISSN 2775-6815

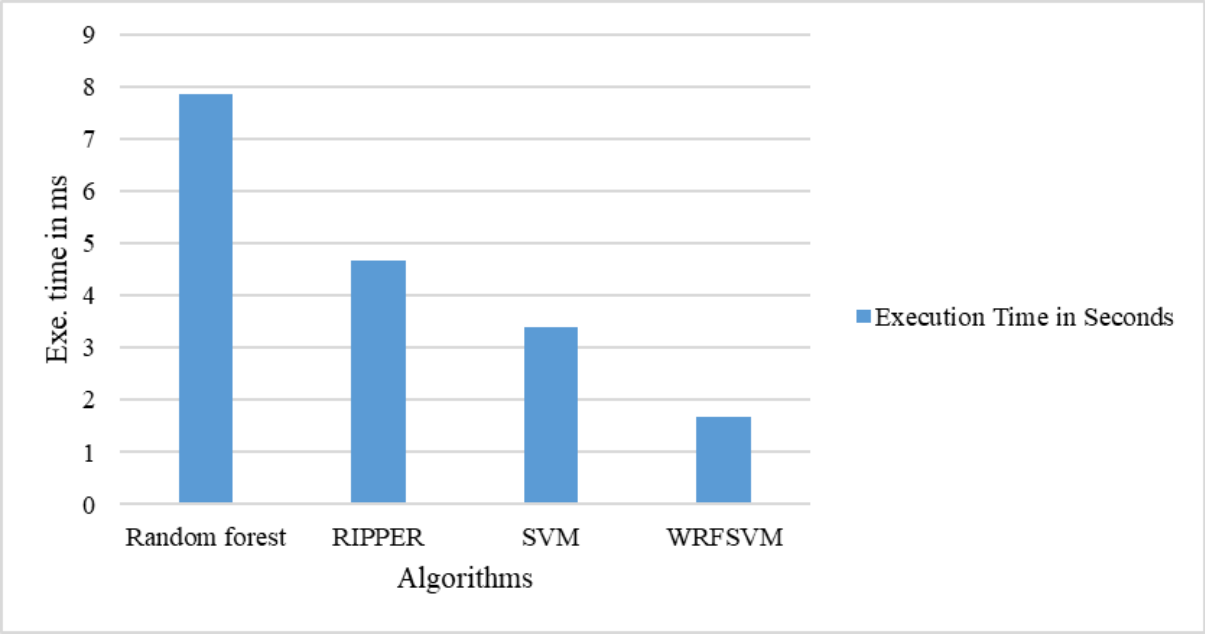

Figure 6. Execution time analysis for kidney disease dataset.

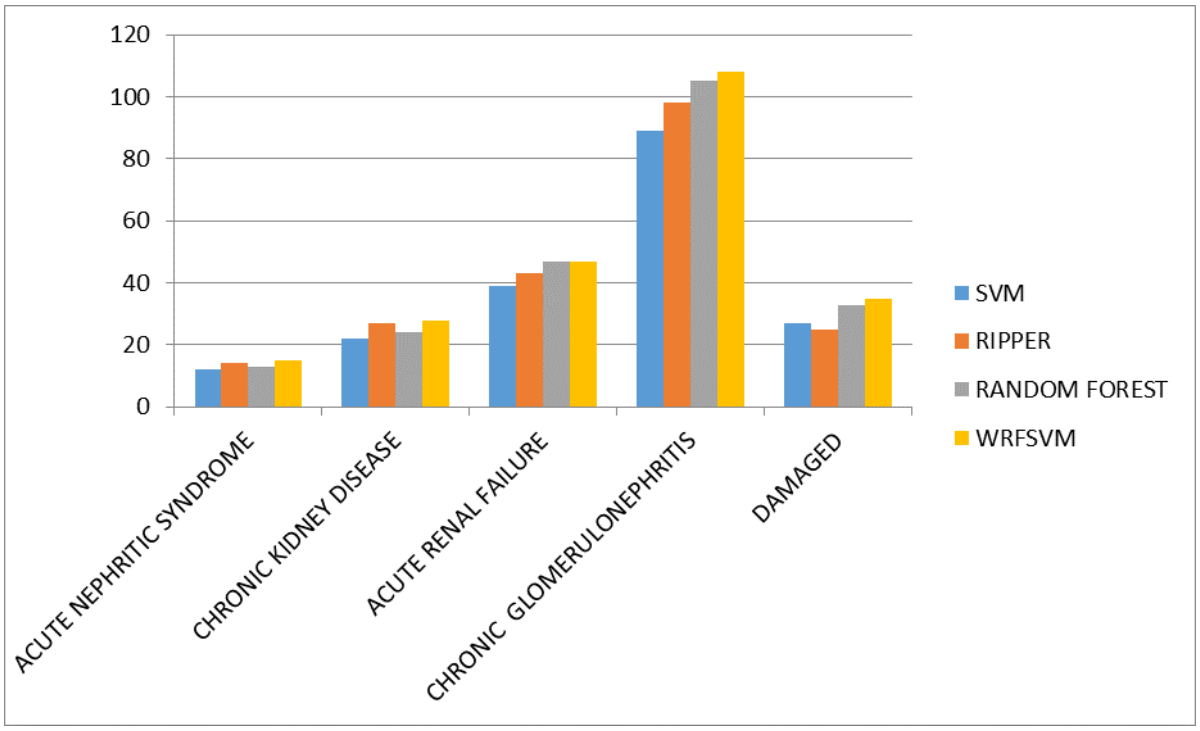

Figure 7. Classification of kidney disease.

5. CONCLUSION

Data mining has great importance in the area of medicine, and it represents a comprehensive process that demands a thorough understanding of the needs of healthcare organizations. Classification is the most important data mining technique which is primarily used in healthcare sectors for medical diagnosis and predicting diseases. This research work

DOI: http://dx.doi.org/10. 17509/xxxx.xxxx
p- ISSN 2776-6098 e- ISSN 2776-5938

proposed a hybrid classification algorithm WRFSVM (Weighted Random Forest Support Vector Machine) for kidney and liver disease prediction. From the experimental results, this work concludes that the WRFSVM classifier is considered the best algorithm because of its highest classification accuracy.

6. AUTHORS' NOTE

The authors declare that there is no conflict of interest regarding the publication of this article. The authors confirmed that the paper was free of plagiarism.

7. REFERENCES

Alves, V., Cury, A., Roitman, N., Magluta, C., and Cremona, C. (2015). Structural modification assessment using supervised learning methods applied to vibration data. *Engineering Structures*, *99*, 439-448.

Banu, M. N., and Gomathy, B. (2013). Disease predicting system using data mining techniques. *International Journal of Technical Research and Applications*, *1*(5), 41-45.

Emeto, I. C., and Ugwu, C. (2016). A Hybrid-based Medical Decision Support System. *International Journal of Computer Applications*, *975*, 8887.

Gulia, A., Vohra, R., and Rani, P. (2014). Liver patient classification using intelligent techniques. *International Journal of Computer Science and Information Technologies*, *5*(4), 5110-5115.

Gulia, A., Vohra, R., and Rani, P. (2014). Liver patient classification using intelligent techniques. *International Journal of Computer Science and Information Technologies*, *5*(4), 5110-5115.

Kara, S., Güven, A., and Öner, A. Ö. (2006). Utilization of artificial neural networks in the diagnosis of optic nerve diseases. *Computers in Biology and Medicine*, *36*(4), 428-437.

Karthik, S., Priyadarishini, A., Anuradha, J., and Tripathy, B. K. (2011). Classification and rule extraction using rough set for diagnosis of liver disease and its types. *Advances in Applied Science Research*, *2*(3), 334-345.

Kumari M. and Godara S. (2011). Comparative study of data mining classification methods in cardiovascular disease prediction. *International Journal of Computer Science and Technology (IJCST), 2*(2), 304-308.

Omar, M., Farid, K., Emran, T., El-Taweel, F., Tabll, A., and Omran, M. (2021). HCC-Mark: a simple non-invasive model based on routine parameters for predicting hepatitis C virus related hepatocellular carcinoma. *British Journal of Biomedical Science*, *78*(2), 72-77.

Rajeswari, P., and Reena, G. S. (2010). Analysis of liver disorder using data mining algorithm. *Global Journal of Computer Science and Technology*, *10*(4), 48-52.

DOI: http://dx.doi.org/10.17509/xxxx.xxxx
p- ISSN 2775-6793 e- ISSN 2775-6815

Vijayarani, S., and Dhayanand, S. (2015). Data mining classification algorithms for kidney disease prediction. *International Journal of Cybernetics and Informatics*, *4*(4), 13-25.

Vijayarani, S., and Divya, M. (2011). An efficient algorithm for generating classification rules. *International Journal of Computer Science and Technology*, *2*(4), 512-515.

Vijayarani, S., Dhayanand, S., and Phil, M. (2015). Kidney disease prediction using SVM and ANN algorithms. *International Journal of Computing and Business Research (IJCBR)*, *6*(2), 1-12.

DOI: http://dx.doi.org/10. 17509/xxxx.xxxx
p- ISSN 2776-6098 e- ISSN 2776-5938