International Journal of Advance Research in Science and Engineering
Volume No. 11, Issue No. 07, July 2022
www.ijarse.com
IJARSE
ISSN 2319 - 8354

A Machine Learning Approach for the Prediction of Liver Disease with feature selection

K. Venkateswara Rao[1] , L. Mary Gladence[2]

[1]*Research Scholar, School of Computing, Sathyabama Institute of Science & Technology, Chennai, TN.*

[2]*Associate Professor, School of Computing, Sathyabama Institute of Science & Technology, Chennai, TN.*

Corresponding Author : venkat545@gmail.com

Abstract:

The number of people suffering from liver illness has been rapidly increasing in recent years. This is due to an unhealthy lifestyle and excessive alcohol consumption. The patients suffering from Liver disease has grown rapidly in the recent times, so in order to be cautions, we have to come up with a prediction model for predicting whether a patient is suffering from Liver related diseases or not. As a result, early detection of liver illness can save a person's life. The dataset used in this paper consists of 10predictive attributes and 1 class. The main aim of this paper is to predict the liver disease using various classification algorithms with and without feature reduction and without feature reduction datasets. The performance measures such as precision, recall, f-measure, ROC area, MAE, RMSE, accuracy are considered and compared with and without feature selection.

Keywords-: Liver disease, alcohol, precision, recall, f-measure

1. Introduction:

The largest solid organ in the human body is the liver. It removes impurities from the body's blood supply, controls blood coagulation, and performs hundreds of additional tasks. It is located beneath the rib cage in the right upper abdomen. The liver filters all of the blood in the body and breaks down harmful substances such as alcohol and drugs. Bile is a bile-like fluid produced by the liver that aids in fat breakdown and waste disposal. Each lobe of the liver has eight sections and thousands of lobules (or small lobes). It is possible to pass on liver disease from one generation to the next (genetic). Viruses, alcohol consumption, and obesity are just a few factors that can affect the liver.The Human Liver is pictorially shown below:

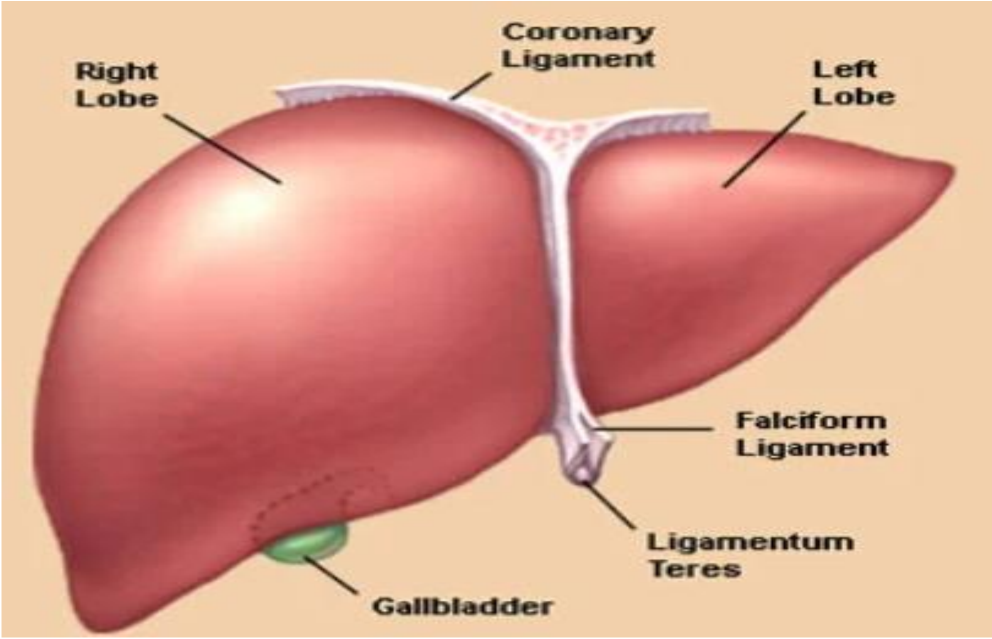

Fig :1 Liver with Right Lobe & Left Lobe.

The liver is divided into four lobes: the right and left lobes, as well as the caudate and quadrate lobes, which are smaller. The left and right lobes are separated by the falciform (Latin for "sickle-shaped") ligament, which connects the liver to the abdominal wall. The liver lobes are further subdivided into eight segments, each with thousands of lobules (small lobes). Each of these lobules has a duct that connects to the common hepatic duct, which drains bile from the liver.

Damage to the liver over time can result in scarring (cirrhosis), which can progress to liver failure, which can be fatal. Early therapy, on the other hand, may allow the liver to heal.

The Following are some of the symptoms that cause due to Liver disease. They are:

- ✓ jaundice,
- ✓ abdominal pain and swelling,
- ✓ confusion,
- ✓ bleeding,
- ✓ fatigue, and
- ✓ weight loss.

2. Related Study:

Research on machine learning has been extensive, and it has been used in a wide variety of fields around the world. Machine learning has proven its worth in medicine, where it has been used to handle a variety of urgent issues such as cancer therapy, heart disease diagnostics, and dengue fever diagnosis, among others. Many studies have used Decision Tree algorithms, which are one of numerous exceptional methodologies.

International Journal of Advance Research in Science and Engineering
Volume No. 11, Issue No. 07, July 2022
www.ijarse.com
IJARSE
ISSN 2319 - 8354

3. Implementation

3.1. Dataset Description:

The Sample dataset used in this paper is shown below:

Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumin_and_Globulin_Ratio	outcome
65	Female	0.7	0.1	187	16	18	6.8	3.3	0.9	1
62	Male	10.9	5.5	699	64	100	7.5	3.2	0.74	1
62	Male	7.3	4.1	490	60	68	7	3.3	0.89	1
58	Male	1	0.4	182	14	20	6.8	3.4	1	1
72	Male	3.9	2	195	27	59	7.3	2.4	0.4	1
46	Male	1.8	0.7	208	19	14	7.6	4.4	1.3	1
26	Female	0.9	0.2	154	16	12	7	3.5	1	1
29	Female	0.9	0.3	202	14	11	6.7	3.6	1.1	1
17	Male	0.9	0.3	202	22	19	7.4	4.1	1.2	0
55	Male	0.7	0.2	290	53	58	6.8	3.4	1	1
57	Male	0.6	0.1	210	51	59	5.9	2.7	0.8	1
72	Male	2.7	1.3	260	31	56	7.4	3	0.6	1
64	Male	0.9	0.3	310	61	58	7	3.4	0.9	0
74	Female	1.1	0.4	214	22	30	8.1	4.1	1	1
61	Male	0.7	0.2	145	53	41	5.8	2.7	0.87	1
25	Male	0.6	0.1	183	91	53	5.5	2.3	0.7	0
38	Male	1.8	0.8	342	168	441	7.6	4.4	1.3	1
33	Male	1.6	0.5	165	15	23	7.3	3.5	0.92	0

Figure:1 Snapshot of Sample dataset

The dataset used in this paper consists of 10 attributes and 1 outcome. The number of instances taken are around 600 samples. The attributes considered are: Age, Gender, Total Bilirubin, Direct Bilirubin, Alkaline Phosphotase, Alamine Aminotransferase, Total Protiens, Albumin, Aspartate_Aminotransferase, Albumin and Globulin Ratio and 1 outcome. Here we are taking 80% of the samples as training data and 20% of the samples as testing data.

3.2 Logistic Regression:

Logistic regression is a Machine learning technique which is very simple and yet very effective classification algorithm. It is commonly used for many binary classification tasks.When the value of the target variable is categorical in nature, then we go for logistic regression. When the outcome is either 1 or 0 then we prefer this classification technique.

3.3 J48 Algorithm:

For data classification, we've been using the most common method J48. In order to identify distinct applications, the J48 algorithm is utilized. In terms of categorical and continuous data analysis, the J48 algorithm is one of the most effective machine learning algorithms. However, when it is used to identify medical data, it consumes more memory and reduces efficiency.

3.4 Random Forest Algorithm:

Random Forest comes under the category of supervised learning algorithms. And this method can be used to tackle classification and regression issues, however it is most commonly used to address classification problems. There are numerous classifiers that work together to tackle a complex problem and improve the model's accuracy. With a number of decision-making trees on distinct subsets of the given dataset, Random Forest chooses the mean to enhance the prediction accuracy, as its name implies.

3.5 K-Nearest Neighbor (KNN) Algorithm:

KNN algorithm is one of the most basic machine-learning approaches under supervised learning algorithms. In KNN Algorithm, all training data is stored and a new data point is designated on the same basis. This suggests that the KNN Algorithm may be quickly grouped into a well-suited group when fresh data is introduced. If you want to do regression or classification, the KNN algorithm can be employed.

3.6 Rep Tree Algorithm:

The full form of REP is "Reduced Error Pruning" tree.Rep Tree algorithm is a fast decision tree learner it is also based on C4. 5 algorithm and can produce classification (discrete outcome) or regression trees (continuous outcome). It builds a regression/decision tree using information gain/variance and prunes it using reduced-error pruning (with back-fitting).

4. Results Comparison

In this work, liver disease data set taken form UCI machine learning repository having 10 predictive attribute and 1class. Data set subjected to various machine learning algorithms such as logistic regression, J48,Random forest, K nearest neighbor(K=7) and REP Tree with 80 percent of training data. Some attributes which are showing less impact on prediction accuracy were identified and required features were selected using Info gain and classical attribute evaluation methods both are giving the same and better features in view of prediction accuracy. This work depicted in two cases showing performance of algorithms without and with feature selection.

Case 1: prediction of liver disease without feature selection

In this case various machine learning classification algorithms were applied on original data set and performance evaluation parameters are compared. Table.1 shows numerical results of performance measures with all features both logistic regression and REP tree are showing best prediction accuracy when compared to other methods. Figure 3 and 4 depicts graphical analysis of prediction accuracy with various algorithms. Figure 5 gives graphical variation of precision, recall and f-measure for different machine learning algorithms. Figure 9 shows graphical variation of prediction accuracy of liver disease data set and linear regression gives the better prediction accuracy with feature selection rather than without reducing the feature.

Table.1 performance measures for liver data set with different classification algorithms

Algorithm	Precision	Recall	f-measure	ROC Area	MAE	RMSE	Accuracy (%)
Logistic Regression	0.678	0.724	0.681	0.766	0.3413	0.4061	72.4138
J48	0.718	0.716	0.717	0.655	0.3335	0.4508	71.5517
Random Forest	0.672	0.707	0.681	0.724	0.3442	0.4232	70.6897
KNN	0.614	0.638	0.625	0.636	0.3782	0.4554	63.7931
REP Tree	0.659	0.724	0.654	0.654	0.3275	0.4077	72.4138

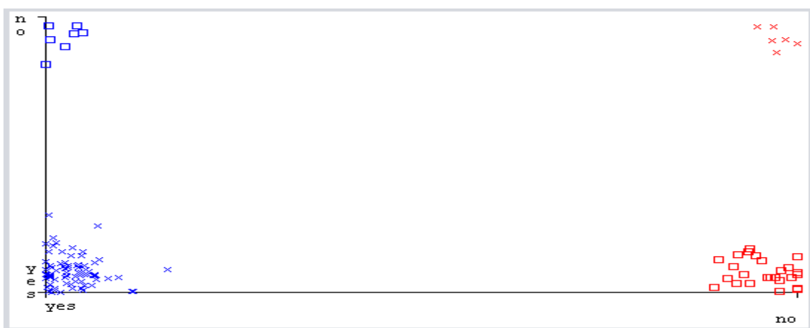

Fig.3. prediction analysis of liver disease

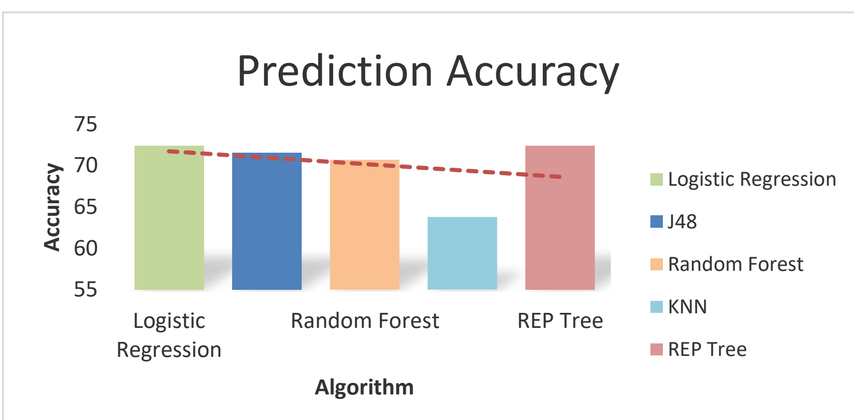

Fig.4 Liver disease Prediction accuracy without feature selection with different algorithms

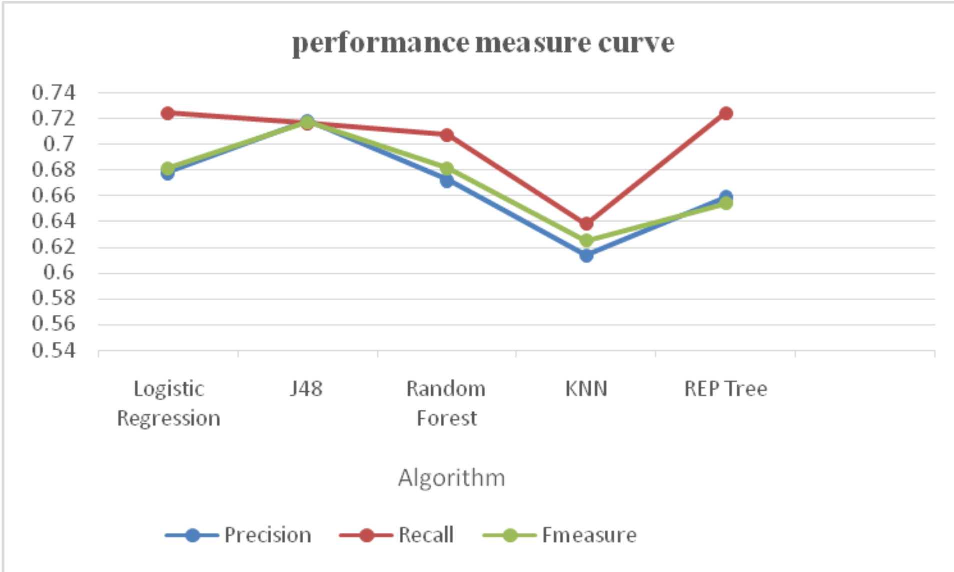

Fig.5 Comparison of performance evaluation parameters without feature selection

Case 2: prediction of liver disease with Info gain and Classical attribute evaluation

In this case various machine learning classification algorithms were applied on data set with reduced features and performance evaluation parameters are compared. Info gain and classical attribute evaluation methods are showing similar attributes to consider (1 to 7 out of 10). Table.2 shows numerical results of performance measures with reduced features logistic regression is showing best prediction accuracy when compared to other methods. Figure 6 and 7 depicts graphical analysis of prediction accuracy with various algorithms. Figure 8 gives graphical variation of precision, recall and f-measure for different machine learning algorithms.

Algorithm	**Precision**	**Recall**	**f-measure**	**ROC Area**	**MAE**	**RMSE**	**Accuracy (%)**
Logistic Regression	0.723	0.750	0.691	0.760	0.3507	0.407	75
J48	0.733	0.724	0.733	0.500	0.3502	0.4431	73.2759
Random Forest	0.672	0.707	0.681	0.714	0.34	0.4224	70.6897
KNN	0.644	0.672	0.656	0.635	0.3703	0.4615	67.2414
REP Tree	0.684	0.733	0.711	0.500	0.402	0.4431	73.2759

Table.2 performance measures for liver data set with different classification algorithms

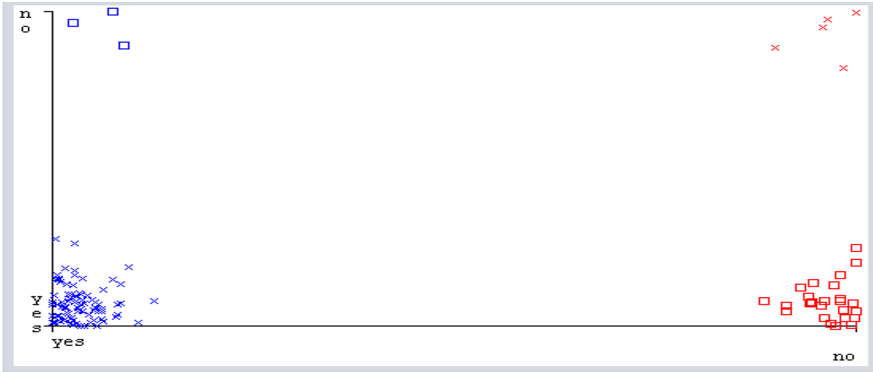

Fig.6. prediction analysis of liver disease with info gain and Classical attribute evaluation

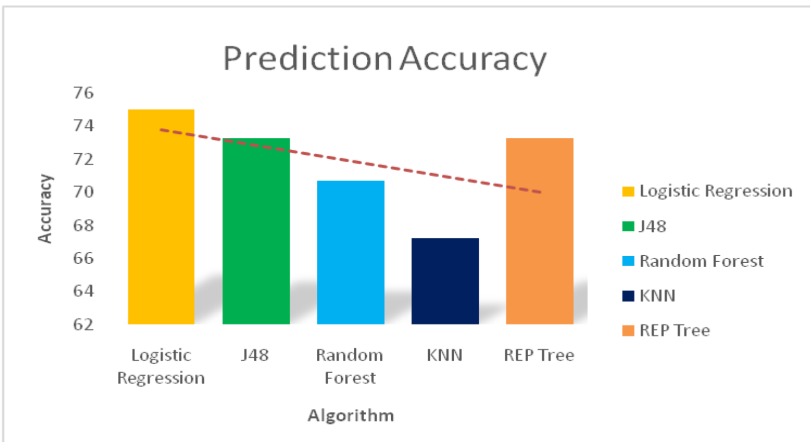

Fig.7 Liver disease Prediction accuracy with feature selection and different algorithms

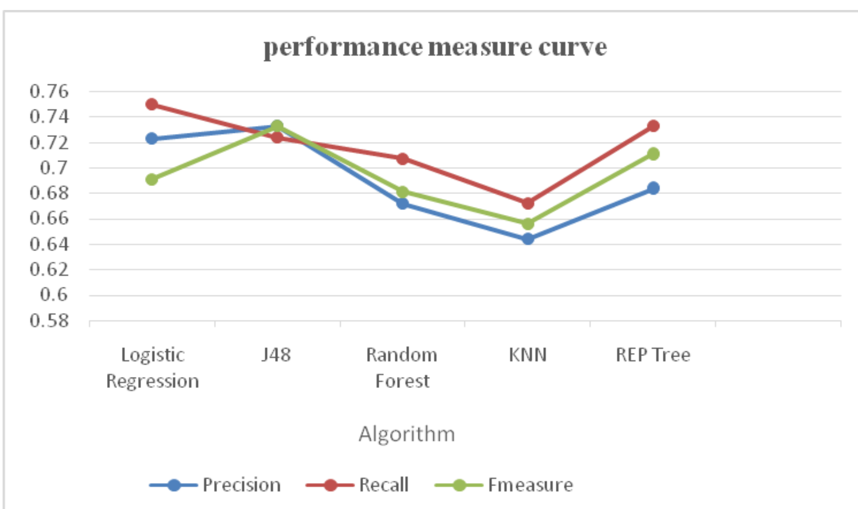

Fig.8 Comparison of performance evaluation parameters with feature selection

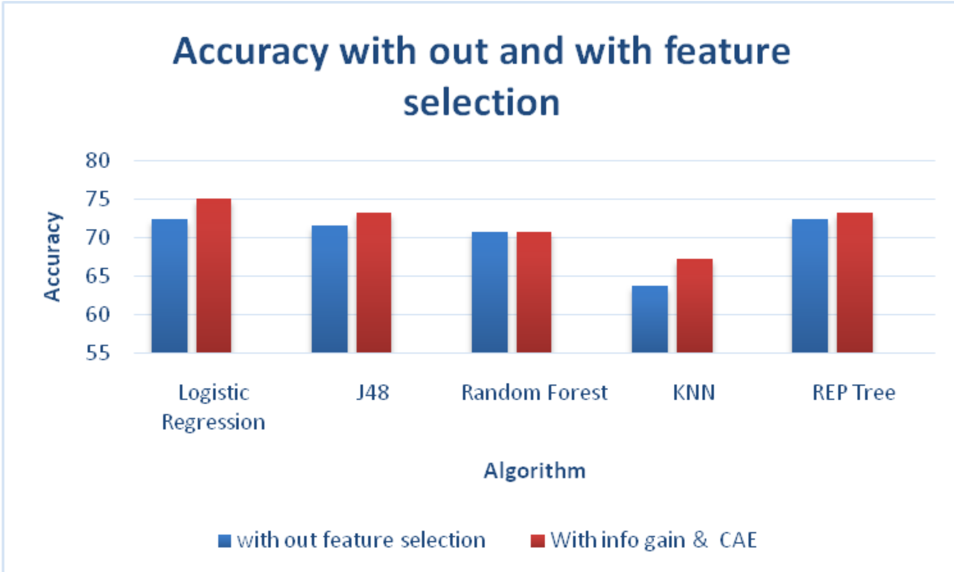

Fig.9 Comparison of prediction accuracy with and without feature selection

5. Conclusion

Many researchers are yet trying to apply machine learning techniques for various data analysis and prediction issues related to may engineering applications. This work aims to predict the liver disease prior to avoid death cases. Prediction analysis carried out without and with seven essential attributes in two different cases with five different classification algorithms such as linear regression, J48, Random forest, K-nearest neighbor and REP tree. After the implementation of various algorithms Linear regression showing good results for prediction of liver disease by considering essential features with infogain and classical attribute evaluation methods. The study can also be expanded to include other data mining methods, such as time series, clustering and association rules, vector support systems and genetic algorithms.

References

[1] Jagdeep Singh, Sachin Bagga, Ranjodh Kaur " Software-based Prediction of Liver Disease with Feature Selection and Classification Techniques", ICCIDS 2019, PP:1970-1980.

[2] M. Banu Priya, P. Laura Juliet, P.R. Tamilselvi "Performance Analysis of Liver Disease Prediction Using Machine Learning Algorithms", IRJET Vvol-5, Issue:1, 2018

[3]Nazmun Nahar , Ferdous Ara "LIVER DISEASE PREDICTION BY USING DIFFERENT DECISION TREE TECHNIQUES",(IJDKP) Vol.8, No.2, March 2018, PP: 01-09

[5]. K.Venkateswara Rao, D.Srilatha Disease prediction and diagnosis implementing fuzzy neural classifier based on IOT and Cloud, International Journal Of Advanced Science and Technology, 2020 , Vol. 29 Issue .5, PP. 737-745.

[6]. D.Srilatha, Dr.S.Sivanagaraju Analyzing Power Flow Solution With Optimal Unified Power Flow Controller, International Journal Of Engineering And Technology (IJET), 2017, Vol 9, No 3,PP: 2278-2289.

[7] K.Venkateswara Rao, Research of feature selection methods to predict breast cancer, International Journal of Recent Technolgy and Engineering (IJRTE), ISSN-2277-3878, Vol-8,Issue-2S11,Sep 2019.

[8] R. Armañanzas, C. Bielza, K. R. Chaudhuri, P. Martinez-Martin, and P. Larrañaga, “Unveiling relevant non-motor Parkinson's disease severity symptoms using a machine learning approach,” Artificial Intelligence in Medicine, vol. 58, no. 3, pp. 195–202, 2013.

[9]R. Armañanzas, C. Bielza, K. R. Chaudhuri, P. Martinez-Martin, and P. Larrañaga, “Unveiling relevant non-motor Parkinson's disease severity symptoms using a machine learning approach,” Artificial Intelligence in Medicine, vol. 58, no. 3, pp. 195–202, 2013.

[10]H.-L. Chen, G. Wang, C. Ma, Z.-N. Cai, W.-B. Liu, and S.-J. Wang, “An efficient hybrid kernel extreme learning machine approach for early diagnosis of Parkinson's disease,” Neurocomputing, vol. 184, no. 4745, pp. 131–144, 2016.

[11] N. P. Pérez, M. A. Guevara López, A. Silva, and I. Ramos, “Improving the Mann-Whitney statistical test for feature selection: an approach in breast cancer diagnosis on mammography,” Artificial Intelligence in Medicine, vol. 63, no. 1, pp. 19–31, 2015.

[12] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, “Extreme learning machine for regression and multiclass classification,” IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 42, no. 2, pp. 513–529, 2012.

[13] G.-B. Huang, D. H. Wang, and Y. Lan, “Extreme learning machines: a survey,” International Journal of Machine Learning & Cybernetics, vol. 2, no. 2, pp. 107–122, 2011.

[14] L. Duan, S. Dong, S. Cui et al., “Extreme learning machine with gaussian kernel based relevance feedback scheme for image retrieval,” in Proceedings of ELM-2015 Volume 1: Theory, Algorithms and Applications (I), vol. 6 of Proceedings in Adaptation, Learning and Optimization, pp. 397–408, Springer, Berlin, Germany, 2016.

[15] J. H. Holland, “Genetic algorithms,” Scientific American, vol. 267, no. 1, pp. 66–72, 1992.

[16] D.-S. Huang and H.-J. Yu, “Normalized feature vectors: a novel alignment-free sequence comparison method based on the numbers of adjacent amino acids,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 10, no. 2, pp. 457–467, 2013.