Vol: 01 Issue: 01

International Journal of Data Science and Advanced Analytics

Machine Learning Approaches for Liver Disease Diagnosing

Bilal Khan[1], Rashid Naseem[2], Mumtaz Ali[3], Muhammad Arshad[4], Nazir Jan[5]
[1-5]City University of Science and Information Technology,
Peshawar, Pakistan
[1]bilalsoft63@gmail.com, [2]rashid@cusit.edu.pk, [3]mumtazali@cusit.edu.pk, [4]arshad12@aup.edu.pk, [5]engr.nazirjan@gmail.com

***Abstract*- In ongoing time liver disease that is any damage in the liver capacity, are exceptionally normal everywhere throughout the world. It has been found that liver disease is discovered more in youthful people as a contrast with other age people. At the point when liver capacity becomes end up, life endures just can be up to 1 or 2 days scarcely. Analysts or moving towards the arrangement of early forecasting of liver disease utilizing various data mining and machine learning approaches. However, this study proposes a new model based on CHIRP methods for the early finding of liver disease. This examination center around MAE, RAE, and Accuracy assessment measurements for the benchmarking of the proposed model with other existing models. The exploratory outcomes show a better consequence of applying CHIRP assessing on MAE and RAE while utilizing the Accuracy of the exhibition of RF and MLP is seldom productive than CHIRP. The outcomes acquired utilizing the proposed model are; MAE 0.2870, RAE 58.8765%, and Accuracy is 71.30%, which demonstrates that this method performs well as opposed to other people.**

***Keywords*- Liver Disease, Composite Hypercube on Iterated Random Projection (CHIRP)**

1. Introduction

In the human body, liver is considered as the main organ, which plays a central role in several bodily functions [1]. In the human body the production of glucose, processing waste products, producing protein, removing worn-out tissue or cell, blood clotting to cholesterol, and iron metabolism are the core functions of the liver [2]. Somehow, Liver disease cased due to the failure of any of these functions. According to the World Gastroenterology Organization (WGO)and World Health Organization (WHO), 35 million death cases causes occur due to liver failure [3]. Liver infection frequently cause due to hepatotropic viruses executes a broad channel on health care resources. Liver diseases are basically classified into two classes that are acute and chronic. The acute liver disorder is an uncommon failure where fast debilitating of liver capacity results in coagulopathy, habitually with an International Normalized Ratio (INR) of greater than 1.5, and variation in the intellectual status (encephalopathy) of an earlier healthy person. For the most part, the youngsters are influenced because of acute liver disorder which conveys a high proportion of death cases [4]. The chronic liver disorder is a disease process of the liver which includes a procedure of dynamic devastation and recovery of the liver parenchyma prompting fibrosis and cirrhosis [5]. We can endure just a couple of days on the off-chance that the liver closes down. At the point when the liver ends up unhealthy, it can do genuine harm our health. There can be several effect and wellbeing conditions that can unconsciously cause liver harm [6].

Alcohol: Substantial Alcohol Consumption is the most widely recognized reason for liver disease. While drinking alcohol, the liver continues its role from the normal to focusing mostly on renovating alcohol into fewer toxic forms.

Obesity: People with the substantial fat on their muscles mostly accrue around the liver, cause fatty liver disease.

Diabetes: Diabetes patients have the risk of 50 percent liver disease, due to the high level of insulin that results in fatty liver disease.

Researchers countenance moving tasks in the Healthcare associations to foresee any sort of disease for the early forecast and treatment from the enormous number of medical data. Nowadays data mining and machine learning become basic in healthcare due to its strategies e.g. for example classification, clustering, association rule mining for discovering frequent patterns pragmatic for disease prediction on medical data [7]. The purpose of this study is to proposed a new solution for liver disease diagnoses based on Composite Hypercube on Iterated Random Projection (CHIRP). This study also includes the comparison of previously used models that are based on MLP, KNN, SVM, J48, RF, DS, RT and LR. The performance of each technique on the dataset is taken from UCI Machine Learning Repository is evaluated using MAE, RAE and Accuracy metrics. The rest of the paper is organized as follow: Section 2 and 3 describe the datasets and evaluation metrics employed in this exploration. Section 4 and 5 contain the overview of the existing models along with the proposed solution, while Section 6 includes the results obtained from experiments and discussion on the results.

2. Dataset Description

The dataset used in this study is taken from the UCI Machine Learning Repository available in (*https://archive.ics.uci.edu/ml/datasets/liver+disorders*). This dataset contains seven attributes in which the first 5 are all blood tests which are thought to be sensitive to liver disorders that might arise from excessive alcohol consumption that are mentioned in Table 1 and 345 instances.

Table 1. List of Dataset Attributes and Descriptions

S No	Attribute	Value Type	Description
1	mcv	Integer	Mean Corpuscular Volume
2	alkphos	Integer	Alkaline Phosphatase
3	sqpt	Integer	Alamine Aminotransferase
4	sqot	Integer	Aspartate Aminotransferase

5	gammagt	Integer	Gamma-Glutamyl Transpeptidase
6	drinks	Real	Number of Half-Pint Equivalents of Alcoholic Beverages % Drunk Per Day
7	selector	Selector {1, 2}	Field Used to Split Data into Two Sets

3. Evaluation Metrics

Evaluating your model is an important part of any research study. Your model may give you satisfactory results when you evaluate it with some standard evaluation metrics. In this study, the following measurements are used for the evaluation of the proposed model as compare to other models.

3.1 Mean Absolute Error (MAE)

MAE is used for model valuation based on regression models. The MAE of a technique or model regarding an assessment set is the mean of the absolute estimation of the discrete desire error on inclusive events in the assessment set, that is the distinction between the predicted error and the true error for overall events [8]. MAE is calculated as follow:

$$MAE = \frac{\sum_{i=1}^{n} abs\,(y_i - \lambda(x_i)}{n} \quad (1)$$

3.2 Relative Absolute Error (RAE)

RAE is essentially the same as the relative squared error as likewise, it is comparative with a simple predictor, which is only the average of the literal values. For this situation, however, the error is only the total absolute error rather than the absolute squared error.

The RAE of a single instance i can be calculated using the following equation:

$$RAE = \frac{\sum_{j=1}^{n} |P_{(ij)} - T_j|}{\sum_{j=1}^{n} |T_j - \bar{T}|} \quad (2)$$

3.3 Accuracy

This evaluation metric is used for the measurement of classification models, that predict it got right [9].

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \quad (3)$$

For binary classification, accuracy can likewise be determined as far as positives and negatives as pursues:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

here, TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

4. Overview Of Employed Techniques

4.1 Artificial Neural Network

Artificial neural networks (ANN) [10] focus happening on suppositious problems in the opinion of information processing. It processes a huge number of consistent processing features known as neurons, which are further connected with each other through connection link. These connection links are connected with weights containing information regarding input signals to solve particular problems. Any elementary neural network involves nomuries input applied by some weights, mutual together to give an output. To train the network, the response of the given output o'er put into inputs to regulate the applied weights, which is helpful to solve the practical, decision making, and non-linear problems easily. In this paper, the Multi-Layer Perceptron (MLP) [1] network is used. MLP combines several layers of knobs in a directed graph with each associated layer to the subsequent one excluding for the input knob, and every knob is a neuron with a non-linear simulation function.

4.2 K-Nearest Neighbor

K-Nearest Neighbor (KNN) [1], is the supervised learning technique being used for pattern recognition and statistical assessments. There are basically two types of techniques Euclidean Distance and Hamming Distance. Euclidean is used for continuous data while Hamming is used for categorical data. The targeted attributed is classified by the mainstream of its neighbors. For the number of neighbors, K value is used which always has to be a positive integer. The value of the K cannot be exceeded than the number of the population in the dataset.

4.3 Support Vector Machine

Support Vector Machine (SVM) [11], is created for gainfully plan straight learning machines in kernel induced component chairs by smearing the concept speculation of Vapnik and associates. It makes a small twofold depiction of the created hypothesis which prompts capable learning techniques, that can be taken care of by an enhancement system because of the Karush–Kuhn–Tucker conditions. Likewise, on account of Mercer's conditions on the pieces, the optimization issues are bent and the course of action combines to an overall perfect point. These highlights make SVM stand isolated among other models of recognition strategies, for instance, neural structures [12]. The objective of the support vector learning machine is to discover f(x, α) with α comparable to the weights and prejudices to delineate an essential relationship in the input data and their outcomes. The SVM system trains machines utilizing the instrument of reducing an upper bound on the disentanglement error while different procedures, for example, neural systems diminish training error on training data.

4.4 Decision Tree (J48)

This is a basic C4.5 decision tree utilized for classification problems [13] which make a doubletree. This methodology is recorded significant in classification problems. Utilizing this strategy, a tree is worked to consummate the classification procedure, that is additionally pragmatic to an individual record in the dataset and item in classification for that record. During this procedure, J48 algorithm mocks the lost values e.g. the value for that element can be forecasted grounded on what is perceived close by the classification value for different records. The basic idea is to parcel the data into run reliant on the quality regards for that thing that is found in the working out test. It permits classification through in addition decision trees or rubrics created from scratch [14].

4.5 Random Forest

Random Forests (RF)s [12] are a collective learning method for regression and classification that work by building lots of decision trees at learning time and outputting the class, which is the technique for the class outcome by every tree. RF does, as an outfit strategy for numerous trees, better to deal with categorical data in the wake of getting the last arrangement in the larger part casting a voting system for the outcomes of each tree is mediated [15].

4.6 Logistic Regression

Logistic Regression (LR) [15][16], measures the connection between a categorical reliant on variable and at least one autonomous variables, which are generally continuous, by utilizing likelihood scores as the anticipated estimations of the reliant on a variable. Chances are the proportion between the likelihood of accomplishment over the likelihood of disappointment, that is, pi / (1-pi), where p is the probability of a record belongs to Class 0. When $p > 0.5$, a record would be classified as Class 0. Else it would be critic as Class 1.

4.7 Random Tree

Random Tree (RT) [2], is a collective learning technique that makes numerous individual learners. It is a technique for assembling a tree that treats K arbitrary highlights at every node. It includes a snaring thought to make an arbitrary arrangement of data for structuring a decision tree. To structure a standard tree, every node is part of utilizing the best part among all variables.

4.8 Decision Stump

A Decision Stump (DS) is a machine learning technique consist of a one-level decision tree [17]. DS is fundamentally decision trees with a solitary label [2]. A stump is against a tree which has various layers. It fundamentally stops after the primary split. Decision stumps are normally utilized in a huge amount of data. Barely, they likewise help to make straightforward yes/no decision model for a little dataset.

5. Proposed Solution

Researchers are attempting to discover models for early diagnosis of illness utilizing biomedical information. Since the most recent couple of decades, they have utilized a parcel of models for early finding, each with their very own advantages and disadvantages. In this research, a CHIRP based model is proposed for the early forecast of liver disease.

5.1 Composite Hypercube on Iterated Random Projection

Composite Hypercube on Iterated Random Projection (CHIRP), is an iterative mechanism of three phases; anticipating, covering, and binning, which expected to a pact with the scourge of dimensionality, computational unconventionality, and nonlinear noticeability [18]. This technique is not the hybridization of other models, also not the enhancement or alteration of prevailing models; it utilizes new casing techniques. The exactness of CHIRP on generally utilized target datasets surpasses the precision of contenders. The CHIRP algorithm was created to process information gathered by the long-pattern Event Horizon Telescope, the global cooperation that in 2019 caught the black-hole picture of M87* for the first time. This algorithm was not used to produce images [19] but was a mathematical solution for the extraction of data from radio signals fabricating data by a variety of radio telescopes spread far and wide [20]. This technique uses the computationally effective approaches to build 2D forecasts and sets of quadrangular sections on those forecasts, that comprises opinions from an individual group of data. CHIRP classifies these groups of predictions and sections them into a conclusion incline for counting new data opinions.

Supervised classifiers facet around authentic difficulties. At first, here is the blight of dimensionality: adjacent neighbors in large-dimensional areas veer aggressively with estimation. The resulting issue is computational multifaceted nature: multinomial-time procedures are impracticable in large-dimensional spaces. A third issue is a uniqueness: for some theme sets in a direction space, for example, a 2D hover incorporated by a sphere, there is no equivalency that can let the division of areas by means of hyperplanes in the target space. There are various methods that can overcome these issues, such as k-mean clustering or random projection, reduction of dimensionality through principal component or hybridization of classifier [21][22].

6. Results and Discussion

In this study liver patient, the dataset is used taken from the UCI machine learning repository containing 7 attributes and 345 instances. Firstly, MLP, KNN, SVM, J48, RF, DS, RT and LR were employed on the dataset, and then the proposed model CHIRP was employed. The results obtained from the experiments show that CHIRP outperforms well in lessening the error rate of evaluation metrics that is 0.2870 for MAE and 58.8765% for RAE, as shown in Table 2.

Table 2. MAE and RAE Comparison Results

S No	Model	MAE	RAE %
1	CHIRP	0.2870	58.8765
2	RT	0.2928	60.0659
3	MLP	0.3543	72.6840
4	J48	0.3673	75.3511
5	KNN	0.3718	76.2906
6	RF	0.3803	78.0322
7	LR	0.4151	85.1648
8	SVM	0.4174	85.6386
9	DS	0.4751	97.4695

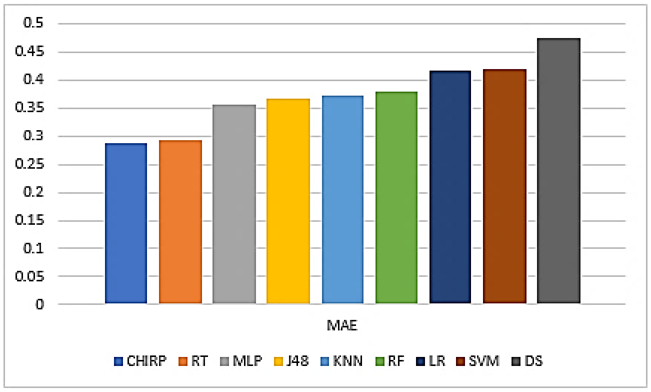

Figure 1. MAE Comparison Results

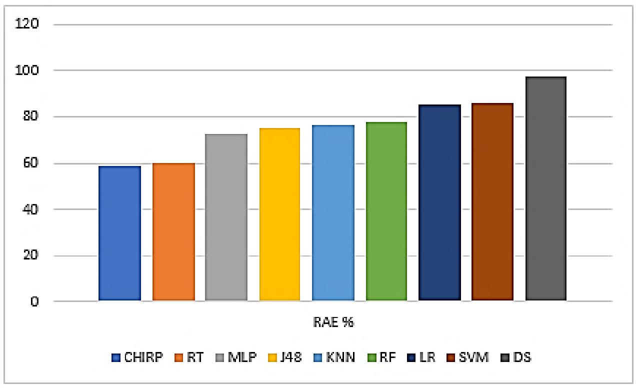

Figure 2. RAE Comparison Results

While comparing the performance in term of accuracy, RF performance is decent instead of MPL and CHIRP, that are listed in Table 3.

Table 3. Accuracy Comparison Results

S No	Model	Accuracy
1	RF	72.17%
2	MLP	71.60%
3	CHIRP	71.30%
4	RT	70.72%
5	J48	68.70%
6	LR	68.11%
7	KNN	62.90%
8	SVM	58.30%
9	DS	57.70%

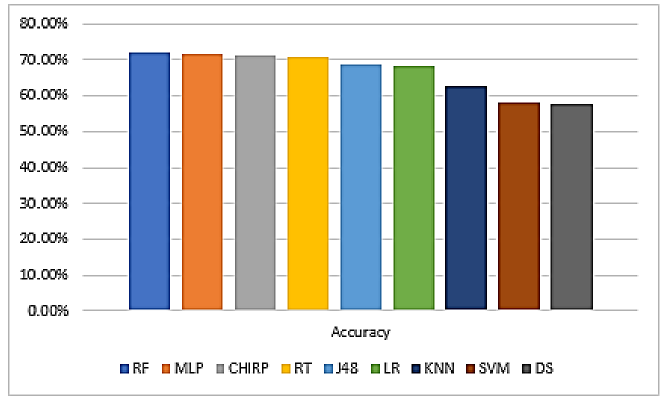

Figure 3. Accuracy Comparison Results

To measure accuracy, RF and MLP performance is decent than CHIRP, that's why the difference between each of them is compared. The percentage difference of accuracy between RF and CHIRP is 1.21% and between MLP and CHIRP is 0.4% shown in Table 4. But if we look forward towards the difference of error rate among these models; using MAE the difference between RF and CHIRP is 28.09%, for MLP and CHIRP the difference is 20.98%. In-case of using RAE the difference between RF and CHIRP is 27.98%, for MLP and CHIRP the difference is 20.99%.

Table 4. Percentage Difference in Results of RF, MLP, and CHIRP

S No	Metrics	RF and CHIRP	MLP and CHIRP
1	Accuracy	1.21%	0.45
2	MAE	28.09%	20.98%
3	RAE	27.98%	20.99%

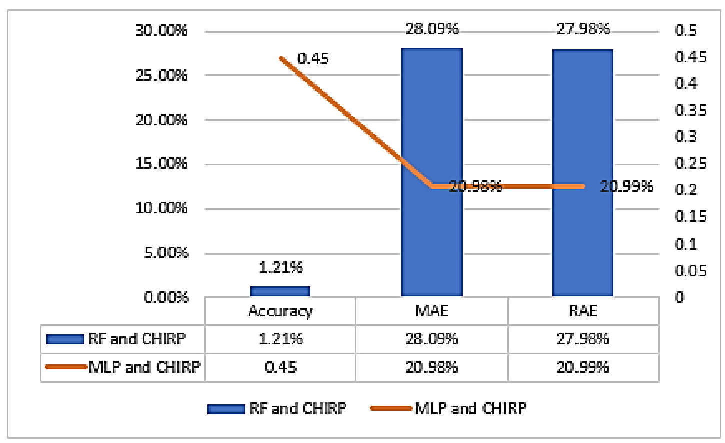

Figure 4. Percentage Difference in Results of RF, MLP and CHIRP

7. Conclusion

Liver diseases are expanding on a regular schedule, and it's hard to foresee these ailments in the early premise. Researchers have utilized a huge number of data mining models and machine learning strategies to foresee such sicknesses in the beginning period. Notwithstanding, in this research, CHIRP based model is presented for early analysis of liver disease. Based on experimental outcomes, it is seen that CHIRP performs well in lessening the error rate in assessment measurements rather than another utilized model. While looking at Accuracy RF and MLP perform well as opposed to CHIRP. Yet, there isn't more distinction between the Accuracy benchmark of RF, MLP, and CHIRP as a contrast with the benchmark in error rate that is high as discussed in Section 6 shown in Table 4.

References

[1] A. N. Arbain and B. Y. P. Balakrishnan, "A Comparison of Data Mining Algorithms for Liver Disease Prediction on Imbalanced Data," *Int. J. Data Sci. Adv. Anal.*, 2019.

[2] N. Nahar and F. Ara, "Liver Disease Prediction by Using Different Decision Tree Techniques," *Int. J. Data Min. Knowl. Manag. Process*, vol. 8, no. 2, pp. 01–09, 2018.

[3] U. Student, U. Sriwijaya, U. Sriwijaya, and U. Sriwijaya, "Issn 2598 0580," vol. 3, no. 1, pp. 51–63, 2019.

[4] J. E. Hay, "Acute liver failure," *Curr. Treat. Options Gastroenterol.*, vol. 7, no. 6, pp. 459–468, 2004.

[5] R. A. Hauser, T. A. Zesiewicz, A. S. Rosemurgy, C. Martinez, and C. W. Olanow, "Manganese intoxication and chronic liver failure," *Ann. Neurol.*, vol. 36, no. 6, pp. 871–875, 1994.

[6] S.Dhamodharan, "Liver Disease Prediction Using Bayesian Classfication," *4th Natl. Conf. Adv. Comput. Appl. Technol.*, no. May, pp. 1–3, 2014.

[7] M. S. D. Dr. S. Vijayarani1, "Liver Disease Prediction using SVM and Naïve Bayes Algorithms," *Int. J. Sci. Eng. Technol. Res.*, vol. 4, no. 4, pp. 816–820, 2015.

[8] K. Morik, "Medicine: Applications of Machine Learning," *Encycl. Mach. Learn. Data Min.*, pp. 809–817, 2017.

[9] A. Gunawardana and G. Shani, "A survey of accuracy evaluation metrics of recommendation tasks," *J. Mach. Learn. Res.*, vol. 10, pp. 2935–2962, 2009.

[10] A. Kaushik, A. Chauhan, D. Mittal, and S. Gupta, "COCOMO Estimates Using Neural Networks," *Int. J. Intell. Syst. Appl.*, vol. 4, no. 9, pp. 22–28, 2012.

[11] A. Al-Anazi and I. D. Gates, "A support vector machine algorithm to classify lithofacies and model permeability in heterogeneous reservoirs," *Eng. Geol.*, vol. 114, no. 3–4, pp. 267–277, 2010.

[12] A. Gulia, R. Vohra, and P. Rani, "Liver Patient Classification Using Intelligent Techniques," vol. 5, no. 4, pp. 5110–5115, 2014.

[13] M. M. Saritas, "Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification," *Int. J. Intell. Syst. Appl. Eng.*, vol. 7, no. 2, pp. 88–91, 2019.

[14] T. R. Baitharu and S. K. Pani, "Analysis of Data Mining Techniques for Healthcare Decision Support System Using Liver Disorder Dataset," *Procedia Comput. Sci.*, vol. 85, no. Cms, pp. 862–870, 2016.

[15] H. Jin, S. Kim, and J. Kim, "Decision factors on effective liver patient data prediction," *Int. J. Bio-Science Bio-Technology*, vol. 6, no. 4, pp. 167–178, 2014.

[16] H. H. Rau *et al.*, "Development of a web-based liver cancer prediction model for type II diabetes patients by using an artificial neural network," *Comput. Methods Programs Biomed.*, vol. 125, pp. 58–65, 2016.

[17] A. Naik and L. Samant, "Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime," *Procedia Comput. Sci.*, vol. 85, no. Cms, pp. 662–668, 2016.

[18] L. Wilkinson, A. Anand, and D. N. Tuan, "CHIRP: A new classifier based on composite hypercubes on iterated random projections," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 6–14, 2011.

[19] *et al.*, "First M87 Event Horizon Telescope Results. IV. Imaging the Central Supermassive Black Hole," *Astrophys. J.*, vol. 875, no. 1, p. L4, 2019.

[20] K. L. Bouman, M. D. Johnson, D. Zoran, V. L. Fish, S. S. Doeleman, and W. T. Freeman, "Computational imaging for VLBI image reconstruction," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-December, pp. 913–922, 2016.

[21] R. Tishirani and T. Hastie, "Margin trees for high-dimensional classification," *J. Mach. Learn. Res.*, vol. 8, pp. 637–652, 2007.

[22] C. Ding and X. He, "K-means clustering via principal component analysis," *Proceedings, Twenty-First Int. Conf. Mach. Learn. ICML 2004*, pp. 225–232, 2004.