Informatics in Medicine Unlocked 36 (2023) 101155

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Prediction of chronic liver disease patients using integrated projection based statistical feature extraction with machine learning algorithms

Ruhul Amin [a], Rubia Yasmin [a], Sabba Ruhi [a], Md Habibur Rahman [b], Md Shamim Reza [a,*]

[a] Department of Statistics, Pabna University of Science & Technology, Pabna, 6600, Bangladesh
[b] Department of Computer Science and Engineering, Islamic University, Kustia, Bangladesh

ARTICLE INFO

Keywords:
Machine learning
Outliers
Z-Score
Feature extraction
Feature integration
Liver disease

ABSTRACT

The healthy liver plays more than 500 organic roles in the human body, while a malfunction may be dangerous or even deadly. Early diagnosis and treatment of liver disease can improve the likelihood of survival. Machine learning (ML) is a powerful tool that can assist healthcare professionals during the diagnostic process for a hepatic patient. The standard ML system includes the methods of data pre-processing, feature extraction, and classification. In the feature extraction stage, ML researchers frequently use projection-based feature extraction approaches to remove data redundancy, but this does not produce the desired results. In addition, most statistical projection methods have different purposes when projecting original features. The Indian liver patient dataset (ILPD) from the University of California, Irvin (UCI) repository is used in this study to classify chronic liver disease. The data set has 583 patient disease records; 416 patients have liver disease, and 167 do not. Using several projection methods, we proposed an integrated feature extraction approach to categorize liver patients. In the pipeline, the proposed method first imputes the missing values and outliers for pre-treatment. Then, integrated feature extraction applies the pre-processed data to extract the significant features for classification. A simulation study is also being conducted to strengthen the suggested methodology. The proposed approach incorporates several ML algorithms, including logistic regression (LR), random forest (RF), K-nearest neighbor (KNN), support vector machine (SVM), multilayer perceptron (MLP), and the ensemble voting classifier. The offered system has an accuracy of 88.10%, a precision of 85.33%, a recall of 92.30%, an F1 score of 88.68%, and an AUC score of 88.20% in predicting liver diseases. Our proposed technique yielded 0.10–18.5% better results than the latest existing studies. The findings suggest that the recommended system could be used to supplement a physician's diagnosis of liver disease.

1. Introduction

In today's world, more than a million people are diagnosed with liver disease each year [1]. Liver cirrhosis, hepatitis (A, B, C), and liver cancer are common liver diseases. Globally, 1.32 million more people died of liver cirrhosis in 2017 than in 1990, of which 66.7% were men and 33.3% were women. Although overall death is declining because of improvements in advanced treatment and maintaining a sound lifestyle [2]. However, liver-related fatalities accounted for 3.5% of all deaths this century [3]. Excessive uses of drugs, alcohol, obesity, and diabetes are the main causes of liver disease [4].

These life-threatening diseases are manageable if they are diagnosed in their early stages. Machine learning techniques are widely used in the healthcare sector, in particular for the diagnosis and classification of certain diseases based on characteristic information [5]. These systems will help clinicians make accurate decisions about patients [6]. The input raw feature space is typically saturated with a significant amount of irrelevant feature information and frequently exhibits high dimensionality when the data is acquired using feature generation techniques in conventional ML systems [7]. Projection-based statistical methods such as principal component analysis (PCA), factor analysis (FA), and linear discriminant analysis (LDA) work well to reduce dimensionality. PCA reduces the dimensionality of the dataset without losing significant feature information. Factor analysis is an extension of PCA that describes the covariance relationships between variables in terms of some underlying factors [8]. LDA uses the class label to compute the matrix

* Corresponding author.
E-mail addresses: ruhulstat6@gmail.com (R. Amin), rubiayasmin0179@yahoo.com (R. Yasmin), sabba.ruhi@gmail.com (S. Ruhi), habib@iu.ac.bd (M.H. Rahman), shamim.reza@pust.ac.bd (M.S. Reza).

https://doi.org/10.1016/j.imu.2022.101155
Received 18 June 2022; Received in revised form 27 December 2022; Accepted 28 December 2022
Available online 31 December 2022

between and within the class and seeks the directions along which the classes are best separated. However, projection-based feature extraction, how many components should be retained is still an unsolved issue. The majority of the authors in existing research on ILPD data employ a single-feature extraction approach [9–11]. Different tactics are used by the PCA, FA, and LDA algorithms to transform the original features [12]. However, the main contribution of the research is an attempt to provide a single feature space that integrates PCA, FA, and LDA projection. Additionally, medical data frequently reveals an imbalance problem among classes. The projection-based dimension reduction approach, as well as the ML algorithm, do not operate well when the dataset is imbalanced and frequently suffers from overfitting [13]. Our working ILPD data reveals missing values, outliers, and higher-class imbalances where the positive class is more than twice as large as the negative class. To achieve better liver patient prediction using a computer-assisted diagnosis process we are accounting for all the stated data preprocessing techniques. The proposed integrated method enhances liver disease classification accuracy, prevents misdiagnosis of liver disease, and increases patient survival. In this paper, we propose a statistical feature integration approach that aims to improve AUC and accuracy.

The complete literature review is covered in section 2. Section 3 contains a description of the ILPD and artificial data set. Section 4, discussed different dimension reduction techniques. The methodology section is included in Section 5. The evaluation protocols, results, and discussion are presented in section 6. The final section introduces the conclusion.

2. Literature review

Human diseases are becoming more prevalent today than in past decades. If we compared liver diseases to other serious diseases, the number of people affected by liver diseases is growing all the time [14]. The majority of liver diseases however, do not present any significant symptoms in their early stages. In the age of modern databases, it is simple to extract data and get insights to assist in the treatment of any disease [15]. There are several strategies that the researcher tries to extract insight from the dataset. Some of them are used with ML classifiers in feature selection or extraction and some are not. Presently, data are generated and stored unremarkably. The available data gave the easiest path to the researcher for solving any object in such fields as medical imagining, finance, genomics, transaction, encroachment, and, etc. If we have a large volume of necessary and unnecessary data that could affect the ML algorithm. There are various methods to select and extract the most correlated feature space to predict any disease as well as any objects.

To predict heart diseases Pasha and Mohamed et al. [16] worked on the Cleveland, Hungarian, Statlog, and Switzerland heart disease datasets. They employed a novel feature reduction (NFR) model in their working methodology to predict cardiac disease. Their methods involved initially processing the dataset, followed by the identification of the features that contributed significantly using a variety of statistical techniques, including weighted least squares (WLS), correlation matrices, etc. The ML and Data Mining (DM) algorithms on the reduced set of features and then individual reduced features measure the AUC and accuracy. In their proposed NRF model, the boosted regression trees (BRT) achieved the highest AUC of 96.68%, and an accuracy of 93.53% by LR on the Cleveland dataset. The LR achieved the highest AUC of 92.51%, and an accuracy of 85.06% by BRT, SGB, and SVM on the Hungarian dataset. BRT achieved the highest AUC of 91.79%, and an accuracy of 87.65% by SVM, and RF in the Statlog dataset. Lastly, the BRT achieved the highest AUC and an accuracy of 99.20%, and 95.52% in the Switzerland dataset. The author demonstrates that the proposed NFR model has greater AUC, and accuracy in predicting heart disease, according to the comparison between the NFR and without NFR model. To develop heart disease risk prediction, the same author [17] worked on the same datasets using an advanced hybrid ensemble gain ratio (AHEGR) feature selection technique. The four feature selection methods, including ensemble feature selection, gain ratio feature selection, backward feature removal, and area under the curve (AUC), have all been applied in their proposed ensemble system. Their suggested feature selection method aids in improving the prognosis of heart diseases. In the Cleveland dataset, with the AHEGR-FS technique among the nine classifications, the AdaBoost, and KNN classifiers achieved the highest AUC, and accuracy of 93.20%, and 87.38%, the RF, and NB classifiers achieved the highest AUC, and accuracy of 95.00%, and 92.00% in the Hungarian dataset, the BRT and NB classifier achieved an AUC, and accuracy of 93.77%, and 89.13% in the Statlog dataset, and the RF, KNN classifier achieved an AUC, and accuracy of 99.00%, and 97.53% in the Switzerland dataset.

To classify liver patients early and precisely, numerous researchers have tried to utilize ML algorithms in various ways. Sreejith et al. [10] evaluated the classification performance by using ILPD, Thoracic Surgery (TSD), and Pima Indian Diabetes (PID) datasets in their paper. The main goal of their work is to compare performance before and after feature selection, as well as to use the class balancing Synthetic Minority Over-sampling Technique (SMOTE) technique and the Chaotic Multi-Verse Optimization (CMVO) evolutionary feature selection approach. They obtained an accuracy of 69.43% on the ILPD dataset using a random forest classifier without OSMOTE and CMVO-based feature selection, 82.620% with OSMOTE and without CMVO-based feature selection, and 82.46% with OSMOTE and without CMVO-based feature selection. Kuzhippallil et al. [11] compare and improve the performance of chronic liver disease classification employed by a variety of data preprocessing strategies (missing value imputation, outlier detection and elimination using isolation forest, duplicate value removal, and so on) and feature selection methods. They used MLP, KNN, LR, DT, RF, Gradient Boosting, AdaBoost, XGBoost, Light GBN, and Stacking Estimator to classify liver patients and achieved an accuracy after the proposed method is 82%, 79%, 76%, 84%, 88%, 84%, 83%,86%, 86%, and 85%.

Gan et al. [18] implemented four classification techniques, including AdaC-TANBN and TANBN, BN, and SVM. After experimenting with his proposed method, the integrated TANBN using a cost-sensitive method (AdaC-TANBN) provided an accuracy of 69.03%, which outperformed the results compared to the others. Abdar et al. [19] use various classification algorithms such as the decision tree (C5.0), the classification and regression tree (CART), and the automatic chi-square interaction detector (CHAID) with boosting technique. Based on the author's research protocol, they achieved 93.75% accuracy at the first stage using the boosted decision tree (B-C5.0) algorithm, and then their proposed method was a combination of multilayer perceptron neural network (MLPNN), namely MLPNNB-C5.0, offers the highest accuracy of 94.13%. Anagaw et al. [20] proposed a method called the compliment naive Bayesian (CNB) classification method and compared it to the naive Bayes classifier and a few other classifiers. The result of the proposed method is 71.36%, which scores better than the others. The author Babu et al. [21] suggested a K-means clustering strategy for detecting liver patients using the various classification model in their work. Following the implementation of the classification model, the accuracy of NBC, KNN, and C 4.5 was 56%, 64%, and 69%, respectively. P. Kumar et al. [22] worked on the ILPD dataset to classify liver patients more accurately. To classify the liver patients faultlessly, they used a 10-fold cross-validation technique. The authors used neighbor-weighted K-NN (NWKNN), fuzzy-neighbor-weighted K-NN, and variable-neighbor-weighted fuzzy KNN classifiers to diagnose liver patients. Since the data set is imbalanced, they work with this data using Tomek link and redundancy-based under-sampling technology (TR-RUS) to balance the data set and achieve an accuracy of 72.31% for the NWKNN classifier, 76.61% accuracy on the fuzzy NWKNN classifier, and finally their proposed method Variable-NWFKNN achieved an accuracy of 87.71%, which is above the other two classifiers. I. Straw et al. [23] diagnosed ILPD liver disease using various ML algorithms based on gender (male and female) stratification. They incorporated the SMOTE data

Table 1
ILPD dataset features description.

Features Name	Features Description	Features Type	Missing Values	Mean ± SD	Shapiro-Wilk test value
Age	Age of the patient	Integer	No	44.75 ± 16.19	0.0036
Gender	Gender of the patient	Categorical	No	–	–
TB	Total Bilirubin	Real number	No	3.3 ± 6.21	0.000
DB	Direct Bilirubin	Real number	No	1.41 ± 2.81	0.000
Alkphos	Alkaline Phosphatase	Integer	No	290.58 ± 242.94	0.000
Sgpt	Alamine Aminotransferase	Integer	No	80.71 ± 182.62	0.000
Sgot	Aspartate Aminotransferase	Integer	No	109.91 ± 288.92	0.000
TP	Total Proteins	Real number	No	6.48 ± 1.09	0.0037
ALB	Albumin	Real number	No	3.14 ± 0.8	0.006
A/G	Albumin and Globulin Ratio	Real number	4	0.95 ± 0.32	0.000
Target	Disease/non-disease	Binary integer	No	1.29 ± 0.45	–

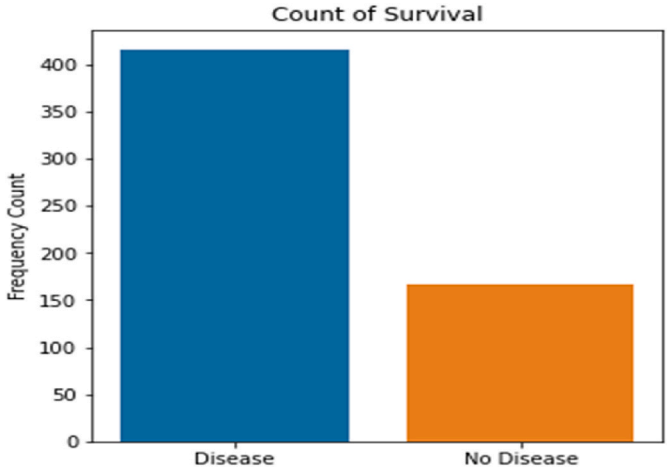

Fig. 1. Patients class distribution bar diagram.

balancing technique in their proposed method, and feature selection is performed using Recursive Feature Elimination (RFE) to improve performances and reduce biases. They used RF, LR, SVM, and Gaussian Nave Bayes (GNB) classifiers in their study, both with and without applying data balancing techniques, and feature selection based on sex disparities. In the all-classifier RF, LR gives higher performance than others. The results of the experiment show that the greatest false negative rate disparity of RF is −21.02% and LR is −24.04%.

3. Datasets description

3.1. ILPD dataset

The North East of Andhra Pradesh, India, is where the ILPD dataset was gathered. This dataset contains 583 observations with ten features and one target output. Table 1 provides information on the ILPD dataset in more depth. Then we retrieved the dataset from the UCI ML repository to evaluate our research [24]. A vast collection of databases, domain theories, and data generators are hosted at UCI, which serves as a hub for machine learning and information systems. This resource is used by the machine learning community such as students, experts, researchers, instructors, and others as the primary and key source to assess ML problems. Fig. 1 shows that 416 are positive/disease cases, and 167 are controlled cases. In general, ML algorithms presumptively assume that the data will be evenly distributed between classes; otherwise, the results would be heavily biased in favor of a few classes [25]. To detract from the ML algorithm biases, we employed the random oversampling data balancing technique. After that, we achieved 832 cases, of which 416 were disease/positive cases and 416 were non-disease/negative cases. Total bilirubin, direct bilirubin, alkaline phosphatase, alamine aminotransferase, and aspartate aminotransferase are shown as outliers in Fig. 2. We employed the Q-Q plot and Shapiro-Wilk test to check the normality of the data set. Fig. 3 and the Shapiro-Wilk test value show that none of the ILPD dataset's features follow the exact normal distribution. Although, based on Fig. 3, we may argue that the age, total protein, albumin, and globulin ratio features follows approximate normality.

3.2. Simulation data

To validate our strategic plan on ILPD's data, we generated simulation data from the Python scikit-learn library using the make classification command. The simulated database comprises 1000 sample observations with 10 features, five of which are informative, three of which are redundant, and two of which are number repeated features. The number of target classes is two.

4. Dimension reduction methods

A data set can contain thousands or millions of features. When we want to analyze this massive type of dataset the computational costs and time, as well as the analytical complexity will increase the machine-learning algorithm. The feature extraction or selection method can eliminate these problems. In the feature extraction methods, we extracted the influential features using PCA, FA, and LDA from the ILPD dataset to predict liver patients.

4.1. Principle components analysis (PCA)

Principal component analysis (PCA) is one of the most important techniques for feature reduction. It extracts the feature from the higher dimension to the lower dimension without losing significant information. To obtain an optimal number of PCAs from the given data set, we erected a 95% variation explanation tactic that incorporates 95% of the entire dataset information.

Suppose we have a random vector X.

$$X = [X_1, X_2, \ldots, X_r]^T$$

Calculate the variance-covariance matrix.

$$Cov(X) = \begin{bmatrix} \sigma_1^2 & \sigma_{12}\ldots & \sigma_{1r} \\ \sigma_{21} & \sigma_2^2\ldots & \sigma_{2r} \\ \sigma_{r1} & \sigma_{r2}\ldots & \sigma_r^2 \end{bmatrix}$$

Calculate the eigenvalues $(\omega_1, \omega_2, \omega_3, \ldots\ldots, \omega_r)$ and eigenvectors $(v_1, v_2, v_3\ldots\ldots, v_r)$ from the variance-covariance matrix. After sorting the eigenvalues, we choose the number *(p)* of principal components, i.e.

$$Y_1 = v_{11}X_1 + v_{12}X_2 + \ldots + v_{1r}X_r$$

$$Y_2 = v_{21}X_1 + v_{22}X_2 + \ldots + v_{2r}X_r$$

⋮⋮⋮⋮

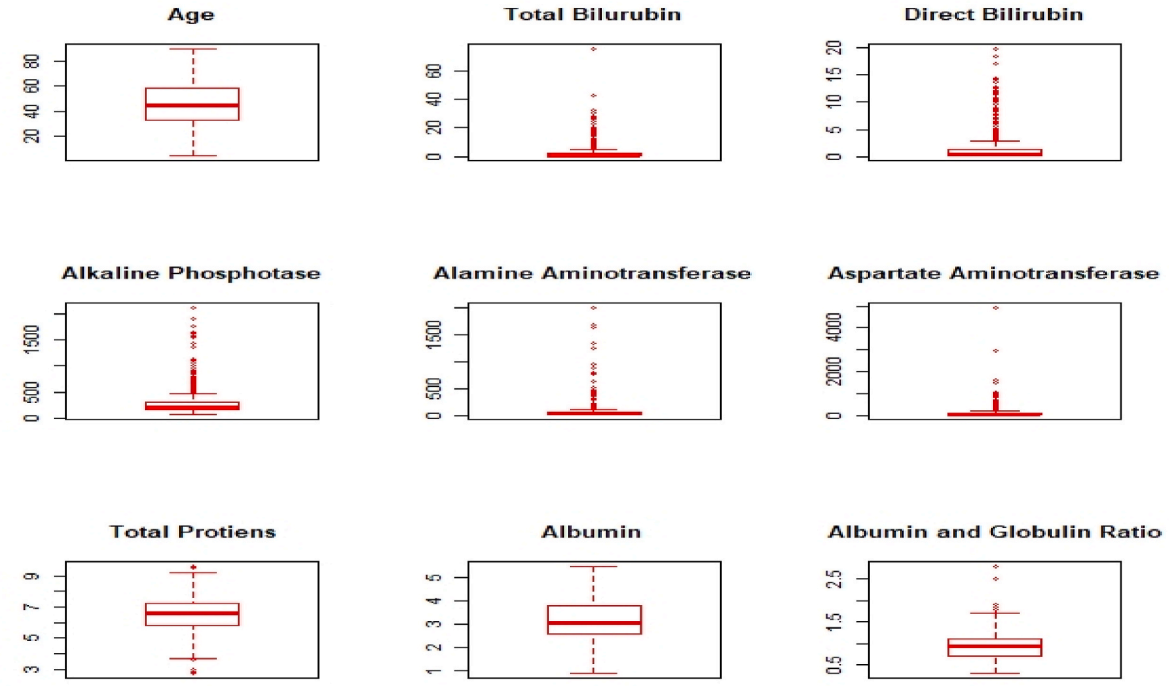

Fig. 2. Boxplot of the variables in the ILPD data set.

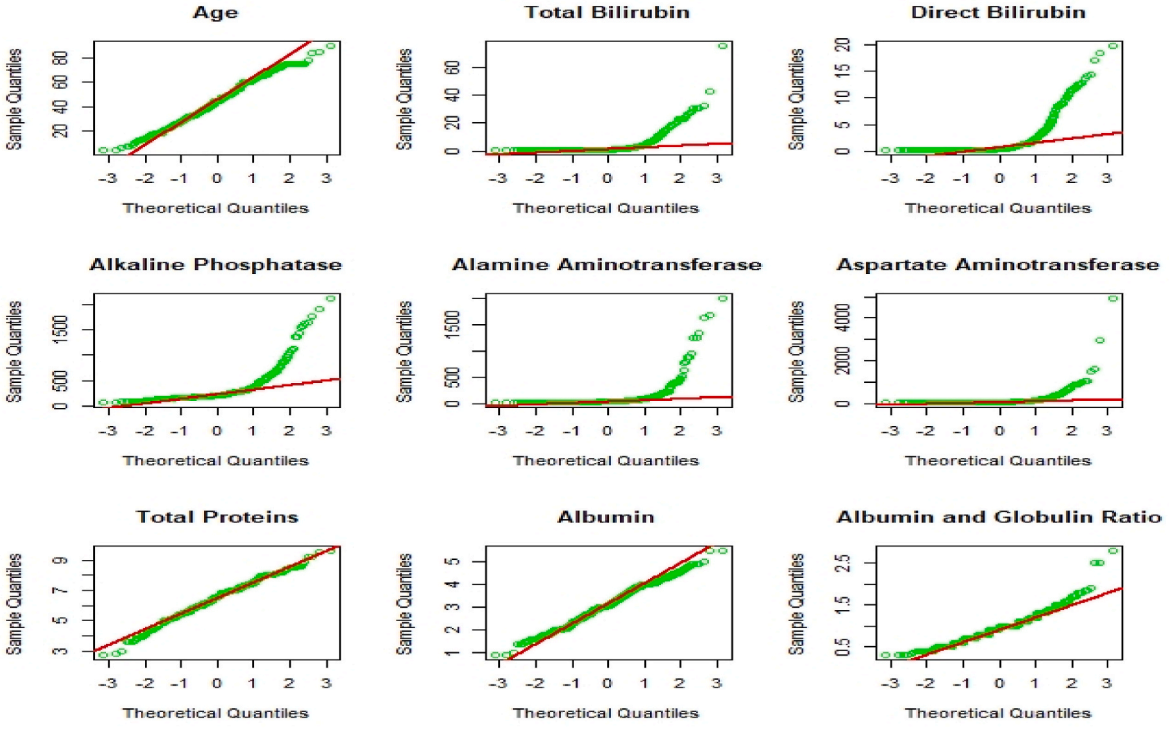

Fig. 3. Q-Q plot of the ILPD features.

$$Y_r = v_{r1}X_1 + v_{r2}X_2 + \ldots + v_{rr}X_r$$

Here the first PCA represents the maximum variance among the total number of the linear combination [26,27]. The proportion of the total population variance due to the r^{th} principal component is $\frac{\lambda r}{\sum_{i=1}^{r} \lambda i}$.

4.2. Factor analysis (FA)

Like PCA, factor analysis is a feature reduction method in which an unobserved or latent variable is sought out from the observed variable or the manifest variables. This method extracts all maximum common variance from the observed variable and inserts it into a common score so that we can use this for further analysis [28]. There are several methods for extracting the factor from a dataset, including principal

component analysis, common factor analysis, image factoring, maximum likelihood method, and so on. In general, the extraction of too many factors produces undesirable results, whereas the extraction of a few factors reduces the common variance without undesirable results. Consequently, it is important to carefully select the number of components in an analysis. The most commonly used techniques for determining the optimal number of factors are the eigenvalue, scree plot, Kaiser's criterion, and Jolliffe's criterion. We used the 'scree plot' approach, which is based on eigenvalues, to determine the number of expected factors.

Suppose we have a random vector X.

$$X = [X_1, X_2, \ldots, X_r]^T$$

For this random vector calculate the mean vector γ.

$$\gamma = [\gamma_1, \gamma_2, \ldots, \gamma_r]^T$$

The q common factor collected from the observed variable is:

$$f = [f_1, f_2, \ldots, f_q]^T \; ; \; [\text{Here q} < \text{r}]$$

Finally, our factor model will be a multiple regression model predicting all of the q-observed variables [29].

$$X_1 = \gamma_1 + k_{11}f_1 + k_{12}f_2 + \ldots + k_{1n}f_n + \varepsilon_1$$

$$X_2 = \gamma_2 + k_{21}f_1 + k_{22}f_2 + \ldots + k_{2n}f_n + \varepsilon_2$$

$$\vdots\vdots\vdots$$

$$X_r = \gamma_r + k_{q1}f_1 + k_{q2}f_2 + \ldots + k_{qn}f_n + \varepsilon_q$$

The general matrix form is $X = \gamma + KF + \varepsilon$ [30].

4.3. Linear discriminant analysis (LDA)

Linear Discriminant Analysis (LDA) is a supervised dimensionality reduction technique used to classify two or more classes. The main goal of this technique is to reduce the n-dimensional spaces to m-dimensional spaces. Typically, the total number of projected axes extracted by LDA is less than one of the number of classes in a dataset. In this approach, the projected new data matrix consists of a lower dimension which minimizes the within-class variance and maximizes the between variances. Each class has a single dimension that distinguishes it. Suppose we have a sample group and its class mean $\overline{X_i}$.

$$\text{i.e.,} \overline{X_i} = \frac{1}{N_i}\sum_{i=1}^{M_i} X_{ij}$$

Where M_i denotes data point in a class group.

The sample variance-covariance matrix is defined as:

$$S_i = \frac{1}{M_i - 1}\sum_{i=1}^{n}\left(X_{ij} - \overline{X_i}\right)\left(X_{ij} - \overline{X_i}\right)^T$$

Compute the within-class variance that measures the distance mean and sample of each class i.e.,

$$R_w = \sum_{i=1}^{n}(M_i - 1)S_i$$

Compute the between-class variance, which measures the distance between the mean of different classes i.e.,

$$R_b = \sum_{i=1}^{n} M_i\left(\overline{X_i} - \overline{X}\right)\left(\overline{X_i} - \overline{X}\right)^T$$

Where $\overline{X} = \frac{1}{M}\sum_{i=1}^{M_i} N_i\overline{X_i}$ is the grand mean.

Now finally created the lower-dimensional projection space using the above-mentioned variances within and between the classes. Let Q as a lower-dimensional space i.e., $\arg\max_Q \left|\frac{Q^t R_b Q}{Q^t R_w Q}\right|$. The characteristic (C-1) is extracted using LDA for any data set to find a projection matrix that maximizes the between-class (R_b) and minimizing the within the class (R_w) [31].

5. Methodology

In this big data era, having a large number of data points while having a low number of features as well as if exist meaningless feature space the ML algorithm faces challenges known as the curse of dimensionality [32]. To deal with this problem, we proposed a statistical projection-based (i.e., PCA, FA, and LDA) feature integration strategy that extracts useful features and makes use of all of the suggested approaches. The mathematical execution of the proposed integration method is as follows:

Firstly, using PCA, let $x_1, x_2, \ldots, x_r$ be an r-dimensional data matrix i.e., $x \in \mathbb{R}^r$. Searching set of a basis vector $v_1, v_2, \ldots, v_{\xi_1}$, when $v_i^t v = 0$, $\|v_j\| = 1$. Summarize an r-dimensional vector X with ξ_1 ($\xi_1 < r$) dimensional feature vector $h(X)$

$$\therefore Y_j = V_j \,.\, X$$

$$\therefore h(X) = \left(y_1, y_2, \ldots, y_{\xi_1}\right)^T = V^T X = V^T(X - \mu_0)$$

The projected new data representation is:

$$X \in \mathbb{R}^r \rightarrow V^T x \in \mathbb{R}^{\xi_1} \tag{1}$$

Where ξ_1 chosen $\frac{\sum_{i=1}^{\xi_1}\sigma_i^2}{\sum_{i=1}^{r}\sigma_i^2} > p, p = 0.95$.

In the next step, FA to consider the error term, and the proposed integration feature space can assume a factor model. Let ξ_2 input factor generating r-observables ($\xi_2 < r$).

$$X_i - \gamma_i = K_{i1}f_1 + K_{i2}f_2 + \ldots + K_{\xi_2}f_{\xi_2} + \in_i$$

$$X - \gamma = KF + \in$$

Searching K such that $S = KK^T + \psi$, where S is the estimation of the covariance matrix and K is the loading, $E(\varepsilon_i) = \psi_i$, $K \in \mathbb{R}^{r*\xi_2}$ ($\xi_2 < r$). Solutions using eigen values and eigen vectors are:

$$F = XW = XS^{-1}K$$

Then the reduced dimensionality using the FA model is:

$$X \in \mathbb{R}^r \rightarrow F = XW = XS^{-1}K \in \mathbb{R}^{\xi_2} \tag{2}$$

After applying the FA model, the suggested system employs LDA to make use of the class information. Let's assume C-classes constructing M_i sample means, $i = 1, 2, \ldots, C$.

$$M = \sum_{i=1}^{c} m_i \; ; \mu = \frac{1}{c}\sum_{i=1}^{c}\mu_i$$

LDA seems the projection transformation, $Y = U^T x$. That maximizes $Max = \frac{|R_b|}{|R_w|}$, where $R_w = \Sigma_{i=1}^{c}\Sigma_{j=1}^{m_i}(X_{li} - \mu_i)(X_{lj} - \mu_i)^T$, $R_b = \sum_{i=1}^{c}(u_i - \mu)(u_i - \mu)^T$. The column of the matrix U are eigenvalues and correspond to the largest eigenvectors.

$$\therefore R_b u_{\xi_3} = \lambda_{\xi_3} R_w u_{\xi_3} \tag{3}$$

The Max dimensionality sub-space is $(C - 1)$, since R_b has the most rank $(C - 1)$. Finally, the proposed 'augmenting' features obtain from (1), (2), and (3) are as follows:

$$S = \left[\, Z \in \mathbb{R}^{\xi_1} \,\vdots F \in \mathbb{R}^{\xi_2} \vdots Y = U^T x \in \mathbb{R}^{(c-1)} \,\right]$$

The ML algorithm is fed by this integrated feature space S.

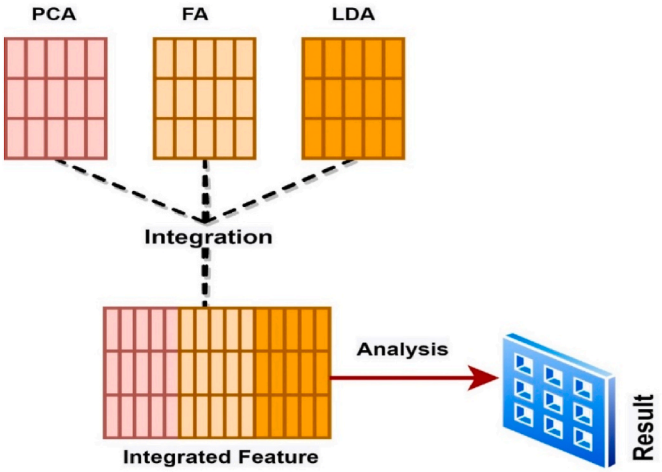

Fig. 4. Features integration strategy diagram.

Thereafter, to detect, and replace outliers the standard normal Z-Score statistic $\frac{x-\mu}{\sigma}$ (μ and σ are the mean and standard deviation, respectively) the rule is used for outlier detection, whereby values of the second quartile are selected for the corresponding outlier implementation. In the working ILPD dataset, some features consist of missing values and outliers. Outliers and missing values affect the classification results in the ML system [33]. To reduce the effect of the outliers and missing values, we replaced them with the value of the second quartile. Mathematically,

$$|Z| = \left|\frac{Xi - \overline{X}}{\sigma}\right| \tag{4}$$

Outliers $= \left|\frac{Xi-\overline{X}}{\sigma}\right| > 3$, where 3 is a commonly used cut-off number for detecting outliers [34]. To make the programming analysis easier, we use an absolute function to convert the z-score value unidirectional. After replacing missing values and outliers, standard scaling transformation removes the different sizes, units, and ranges of the features. A dataset y can be standardized scaler as follows:

$$y = \frac{x - mean}{standard\ deviation}$$

Where $mean = \frac{\sum_{i=1}^{N} x_i}{N}$ and the standard deviation $= \sqrt{\frac{\sum_{i=1}^{N}(x_i - mean)^2}{N-1}}$.

Furthermore, to avoid bias and overfitting in our experimental results, we endeavor a random oversampling approach. Fig. 4 depicts the feature integration notion, and details of the proposed method are shown in Fig. 5.

5.1. Pseudo-code of the proposed feature extraction method

Input: Pre-processed Dataset $X \in \mathbb{R}^r$.
Output: Extracted Feature Matrix.

Algorithm. Step 1: Extract feature vectors $v_1, v_2, \ldots, v_{\xi_1}$ from the processed dataset using principal component analysis (PCA) with 95% variation, then store the resultant reduced feature vectors, that is $X \in \mathbb{R}^r \rightarrow V^T x \in \mathbb{R}^{\xi_1}$
Step 2: Using Factor Analysis select reduced features, $X \in \mathbb{R}^r \rightarrow F = XW = XS^{-1}K\mathbb{R}^{\xi_2}$ and store the features.
Step 3: Utilizing LDA, separate the optimal (C-1) discriminant features from the input dataset, then store the feature vectors.
Step 4: Integrate all of the stored features into a new matrix space as:

$$S = \left[Z \in \mathbb{R}^{\xi_1} \vdots F \in \mathbb{R}^{\xi_2} \vdots Y = U^T x \in \mathbb{R}^{(c-1)} \right]$$

Step 5: Update the matrix space until the desired data variation is required.
Step 6: Return the extracted feature matrix S.

6. Evaluation protocols, results, and discussion

To first and accurately diagnose liver disease and enable a fair

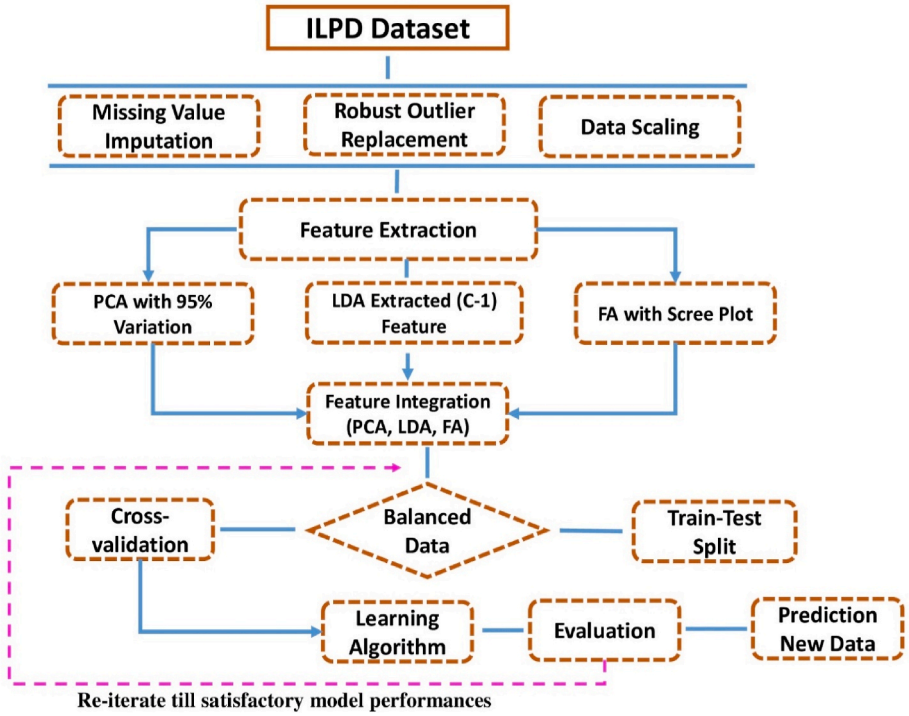

Fig. 5. Proposed integrated feature extraction and methodology diagram.

Table 2
Classifier's performance of the proposed method using10-fold cross-validation on simulated data.

Features Extraction methods	Classifiers Name	Accuracy (%)	Precision (%)	Recall (%)	F-1 Score (%)	AUC (%)	Time (sec)
PCA	LR	76.90	76.64	77.10	76.87	76.90	0.015
	RF	88.70	88.57	88.75	88.66	88.70	0.015
	K-NN	89.10	90.70	86.94	88.82	89.09	0.015
	SVM	90.70	89.78	91.76	90.76	90.70	0.014
	MLP	88.70	88.88	88.35	88.62	88.69	0.015
	Ensemble	**91.00**	**90.80**	**91.16**	**90.98**	**91.00**	**0.016**
FA	LR	76.30	75.84	76.90	76.37	76.30	0.015
	RF	87.70	86.54	89.15	87.83	87.70	0.015
	KNN	91.30	90.05	92.77	91.39	91.30	0.022
	SVM	90.90	89.21	92.97	91.05	90.90	0.015
	MLP	**91.70**	**92.43**	**90.76**	**71.59**	**91.69**	**0.015**
	Ensemble	90.70	89.94	91.56	70.74	90.70	0.019
LDA	LR	77.20	76.67	77.91	77.29	77.20	0.015
	RF	68.60	69.16	66.66	67.89	68.59	0.015
	KNN	72.20	76.19	64.25	96.71	72.16	0.016
	SVM	**77.90**	**75.60**	**82.12**	**78.72**	**77.91**	**0.015**
	MLP	75.90	72.98	81.92	77.19	75.92	0.015
	Ensemble	77.77	75.79	81.12	78.37	77.71	0.015
Proposed	LR	76.30	75.84	76.90	76.37	76.30	0.016
	RF	87.90	87.17	88.75	87.96	87.90	0.015
	K-NN	**92.00**	**93.36**	**90.36**	**91.83**	**91.99**	**0.016**
	SVM	90.90	89.21	92.97	91.05	90.90	0.017
	MLP	91.10	90.65	91.56	91.10	91.10	0.015
	Ensemble	91.40	90.87	91.96	91.41	41.40	0.015

comparison of existing approaches, the performance of the proposed method will be assessed by both the train-test split and cross-validation method. In an experimental protocol for evaluating the results, the data set was randomly divided into 75% training and 25% testing set using stratified random samples. We also perform 10-fold cross-validation to compare other existing works. In addition, the receiver-operating curve, known simply as the ROC uses for evaluating and comparing the performance of various classifiers. The ROC curve is created by the true positive rate versus the false positive rate. The total area under the curve is in the range of 0.5–1 [35]. Python programming on the Google cloud computing GPU hardware accelerator platform and R (desktop version x64 3.5.1) programming for graphical visualization is utilized to implement all machine learning algorithms. We carried out our research using a workstation with an Intel Core i7 with 8 GB RAM and an 11-GEN CPU processor.

6.1. Performance evaluation metrics

Classification or prediction is one of the contentious subjects that has received the greatest attention in scientific circles globally. To ascertain whether the specified classification algorithm performs well or unsatisfactorily, we must measure classification performance. The performance evaluation matrices such as accuracy, precision, recall, f-1, and AUC score have been taken into consideration while conducting the proposed research. However, we methodically discuss the following metrics to evaluate the Classification algorithm:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 - Score = 2 * \frac{precsion \cdot recall}{precision + recall}$$

The terms TP, TN, FP, and FN have been used to denote true positive, false positive, and false negative cases, respectively.

Table 3
Classifier results with all features of the ILPD balanced data using train-test method.

Classifiers	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
Logistic	51.92	54.71	52.72	53.70	51.87
K-nearest Neighbors	73.55	70.67	85.45	77.36	72.82
Random Forest	69.23	66.91	82.77	73.98	68.40
SVM	46.15	48.27	25.45	33.33	47.42
MLP	**85.50**	**80.62**	**94.54**	**87.02**	**84.51**
Ensemble Classifier	72.11	70.30	81.81	75.63	71.52

6.2. Experimental result and discussion

The results have been derived from both simulated and real data. Table 2 shows simulation study results using the proposed feature integration method.

On the simulated dataset, 7 PCs explained 95% variation of the data out of 10 simulated features, where the ensemble classifier achieved the best accuracy and AUC of 91.00%. In the factor analysis, 7 latent factors were found with the help of a scree plot, the MLP classifiers achieved an accuracy of 91.70% and AUC of 91.69%. On the other hand, one (Class-1) LDA component feature provided an accuracy of 77.90% and an AUC

Table 4
Classifier results with all features of the ILPD balanced data using the 10-fold cross-validation method.

Classifiers	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
Logistic	54.94	54.40	60.81	57.43	54.92
K-nearest Neighbors	72.95	71.17	77.16	74.04	72.95
Random Forest	**87.78**	**83.69**	**93.75**	**88.43**	**87.74**
SVM	70.07	65.15	86.29	74.25	70.07
MLP	84.97	79.75	93.75	86.18	84.97
Ensemble Classifier	81.49	75.00	97.47	83.61	81.49

Table 5
Proposed integration strategy classifier results of the ILPD balanced data using train-test method.

Classifiers	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
Logistic	50.96	50.92	79.67	70.74	84.84
K-nearest Neighbors	74.03	73.62	71.09	84.25	77.11
Random Forest	70.19	69.44	65.75	88.88	75.59
SVM	57.69	57.40	58.33	64.81	61.40
MLP	**84.61**	**84.29**	**80.64**	**92.59**	**88.20**
Ensemble Classifier	67.30	66.81	65.15	79.62	71.66

Table 6
Proposed integration strategy classifier results of the ILPD balanced data using 10-fold cross-validation method.

Classifiers	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
Logistic	55.40	54.63	63.70	58.82	55.40
K-nearest Neighbors	67.90	70.89	72.59	71.73	71.39
Random Forest	**88.10**	**85.33**	**92.30**	**88.68**	**88.20**
SVM	67.90	63.09	86.29	72.89	67.90
MLP	83.53	79.12	91.10	84.69	83.53
Ensemble Classifier	55.40	54.63	63.70	58.82	55.40

of 77.91% by SVM. Our proposed feature integration provided the highest accuracy of 92.00%, and an AUC score of 91.99% in the KNN. Additionally, the evaluation measures used in the proposed technique considerably enhanced the detection of liver disease, as demonstrated in Table 2.

The ILPD data classification results utilizing a different protocol are shown in Table 3 through Table 5. We analyzed the ILPD data in two different protocols. Table 3 displayed the train-test method and the 10-fold cross-validation is shown in Table 4. Table 3 shows that the MLP classifier provides the highest accuracy of 85.50%, and AUC of 84.51% which is a higher value than other evaluation measures in the train-test split protocol without any feature extraction of the ILPD data. Table 4 shows the 10-fold cross-validated result, the RF classifier gives an accuracy of 87.78%, and an AUC score of 87.74% which is much better compared to the train-test split method. We then compared the considered statistical features integration methods using several classifiers with the train-test method shown in Table 5 and the 10-fold cross-validation method shown in Table 6.

In our proposed model, the MLP classifier reduced the detection rate of liver diseases by 0.89% compared to the train test scheme without feature extraction, as shown in Table 7. In addition, precision, recall, F1, and AUC score are increased by 2–3%. As can be seen from Table 5, the RF classifier outperforms all other classification models in all evaluation matrices and it exhibits accuracy, precision, recall, F-1, and AUC score values of 88.10%, 85.33%, 92.30%, 88.68%, and 88.20%, respectively. Table 8 compares the proposed method with recent existing studies on the ILPD dataset. The final results of the research have been improved through the use of the feature integration technique and achieved an accuracy of 88.10%, and an AUC score of 88.20% which is significantly better for liver patient classification than that of the existing work.

In our proposed approach, Fig. 6 shows that the ROC curve for RF classifiers covers more than 88% of the area covered by the ROC curves for the other classifiers, indicating that the RF classifier is superior to the others. Three factors contribute to the increased performance: (a) proper treatment of outliers and replacement of missing values with the median; (b) data balancing strategy; and (c) feature integration approaches that improved liver disease classification accuracy. As a result, of integrated features and proper management of outliers and missing values, the suggested method of liver patient recognition performance has improved.

6.3. Run time comparison of the proposed ML model

We calculated the time taken by each ML algorithm, which is another illuminating performance measurement, to contrast the strength of the proposed model with the existing models. In the two evaluation protocols, by utilizing the train-test method with all features the MLP algorithm achieved the highest accuracy of 85.50% and AUC of 84.5% in 0.018 s of run time, and with the integrated features the MLP algorithm achieved the highest accuracy of 84.61% and AUC of 88.2% in 0.634 s run time. And through the 10-fold cross-validation method with all features the RF algorithm achieved the highest accuracy of 87.78% and AUC of 87.74% in 0.026 s run time, and with the integrated features the RF algorithm achieved the highest accuracy of 88.1% and AUC of 88.2% in 3.337 s run time. Finally, we measure the performance of simulation data, which obtained supreme accuracy of 92.00% and AUC of 91.99% in 0.016 s, the run time of our proposed feature integration method. In the performance comparison of the recent studies, most of the existing research work for this dataset, the authors do not calculate the execution run time for the ML algorithm. Authors, J. Singh et al. [9] and Kuzhippallil et al. [11] only considered the execution run time in their research work whereas J. Singh et al. used the LR model and achieved the highest accuracy of 74.36% and took the execution run time of 0.01 after the feature selection and Kuzhippallil et al. XGBoost and Light GBM achieved the same accuracy of 86.00% with an execution run time of 0.191 s and 0.0059 s after the feature selection method. The framework proposed in this paper utilizes feature extraction. In most cases, feature

Table 7
Compare the performance analysis of the proposed integrated method on the ILPD dataset.

Evaluation Protocol	ML Classifiers	Performance metrics					
		With all feature			With proposed integrated feature		
		Accuracy (%)	AUC (%)	Time (sec)	Accuracy (%)	AUC (%)	Time (sec)
Train-Test	LR	51.92	51.87	0.022	50.96	84.84	0.036
	KNN	73.55	72.82	0.018	74.03	77.11	0.038
	RF	69.23	68.40	0.022	70.19	75.59	2.215
	SVM	46.15	47.42	0.015	57.69	61.40	0.047
	MLP	**85.50**	**84.51**	**0.018**	**84.61**	**88.20**	**0.634**
	Ensemble	72.11	71.52	0.017	67.30	71.66	1.941
10-fold cross-validation	LR	54.94	54.92	0.021	55.40	55.40	0.018
	KNN	72.95	72.95	0.027	67.90	71.39	0.024
	RF	**87.78**	**87.74**	**0.026**	**88.10**	**88.20**	**3.337**
	SVM	70.07	70.07	0.021	67.90	67.90	0.019
	MLP	84.97	84.97	0.022	83.53	83.53	0.020
	Ensemble	81.49	81.49	0.015	55.40	55.40	0.092

Table 8
Performance comparison of the recent studies with the proposed method on the ILPD dataset.

Author & Reference	Year	Protocol	Classifiers	Accuracy (%)	AUC (%)
K. Gupta et al. [36]	2022	Train-Test	LR	57.00	–
			GNB	54.00	–
			DT	61.00	–
			RF	63.00	–
			AdaBoost	62.00	–
			GB	60.00	–
			Extreme Boosting	60.00	–
			Light GB	63.00	–
			KNN	57.00	–
			XGBoost	60.00	–
			Stacking	62.00	–
I. Altaf et al. [37]	2022		XG Boost	70.69	–
			Bagging Meta Estimator	70.11	–
			MLP	70.11	–
			Voting Ensemble	73.56	–
J. Singh et al. [9]	2020	10-fold cross-validation	LR	74.36	–
			NB	55.90	–
			SMO	71.36	–
			IBk	67.41	–
			J48	70.67	–
			RF	71.87	–
Sreejith et al. [10]	2020	Train-Test	RF without OSMOTE & CMVO	69.43	–
			RF with OSMOTE & without CMVO	82.62	–
			RF (OSMOTE + MVO-FS)	82.46	–
Gan et al. [18]	2020	Train-Test	AdaC-TANBN	68.23	69.53
			TANBN	71.74	60.23
			BN	69.03	62.12
			SVM	67.57	63.59
Kuzhippallil et al. [11]	2020	Train-Test	MLP	82.00	–
			K-NN	79.00	–
			LR	76.00	–
			DT	84.00	–
			RF	88.00	–
			Gradient Boosting	84.00	–
			AdaBoost	83.00	–
			XGBoost	86.00	–
			Light GBM	86.00	–
			Stacking Estimator	85.00	–
P. Kumar et al. [22]	2021	10-fold cross-validation	NWKNN	72.31	68.08
			Fuzzy-NWKNN	76.31	74.48
			Variable-NWKNN	87.71	82.41
Amare et al. [20]	2019	Train-Test	CNB	60.85	–
			K-NN	68.61	–
			NB	71.36	–
M. Babu et al. [21]	2016	10-fold Cross-validation	NBC	56.00	–
			K-NN	64.00	–
			C 4.5	69.00	–
Our proposed method		10-fold cross-validation	Logistic	55.40	55.40
			K-NN	67.90	71.39
			Random Forest	**88.10**	88.20
			SVM	67.90	67.90
			MLP	83.53	83.53
			Ensemble	82.09	55.40

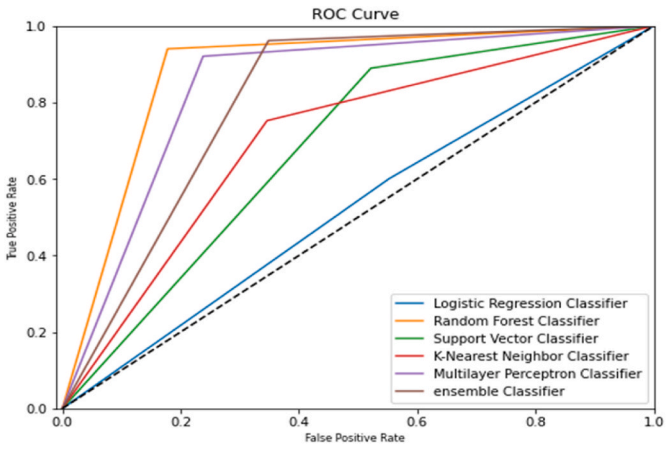

Fig. 6. ROC curves for the ILPD data using several approaches.

extraction takes longer than feature selection. The accuracy and AUC of our proposed work are 88.10% and 88.20%, respectively, outperforming other works. However, the execution time of our algorithm is a little bit longer than that of J. Singh and Kuzhipallil's work due to the feature extraction technique.

6.4. *Benchmarking of the proposed integration model*

The proposed feature integration model is compared to recent existing work on liver patient identification, both with and without feature selection/extraction using the ILPD dataset. We take into account the key criterion accuracy to identify liver patients using our proposed integration model. In our research, another statistic called AUC is also considered to be a crucial assessment in the medical field used to detect liver diseases. We have used the proposed statistical feature integration model using the six-classification algorithm to the ILPD dataset and contrasted our results with the most recent ILPD research, which is shown in Table 8. K. Gupta et al. [36] achieved the highest accuracy of 63.00% with the Light GB and RF classifier, which is 25.1% lowest than our proposed model. With the help of voting ensemble classifiers, I. Altaf et al. [37] acquired the greatest performance of 73.56% which is 14.54% worse than our proposed model. J. Singh et al. [9] obtained the maximum accuracy of 74.36% by the LR classifier, which is 13.74% less than our proposed model. Sreejith et al. [10] proposed a method, which combines the Chaotic Multi-Verse Optimizations (CMVO) algorithm for feature selection with the Orchard-enhanced Synthetic Minority Over-sampling Technique (OSMOTE) for data balancing, which is applied to the ILPD dataset, the RF classifier achieves an accuracy of 82.46%, which is 5.64% less than our proposed model.

To flawlessly diagnose liver patients, Gan et al. [18] utilize both accuracy and AUC metrics, and they reached the best performance for accuracy of 71.74% and AUC of 69.53% where both accuracy and AUC score is 16.36%, and 18.67% less than our proposed model. By using an RF classifier, Kuzhipallil et al. [11] achieved the greatest accuracy of 88.00%, which is 0.10% less than our proposed strategy. P. Kumar et al. [22] found that Variable-NWKNN received the greatest performance for detecting liver disorders, with an accuracy of 87.71% and AUC of 82.41% a difference of 0.39% and 5.79% less than our suggested model. Amare et al. [20] and Babu et al. [21], who both worked on the ILPD liver diseases dataset, obtained the best accuracy of 71.36% by the NB classifier and an accuracy of 69.00% by C4.5, respectively. In comparison to the feature integration model we proposed, these results are, respectively, 16.74% and 19.1% less precise. However, our proposed model's highest and lowest levels of accuracy developed is 34.1% and 0.10%, respectively, when compared to the results of recent studies.

7. Conclusion & future work

In this paper, we have explored improved feature extraction systems for liver patient classification using statistical machine learning techniques by adopting dimensionality reduction approaches such as PCA, FA, and LDA. The system extracted an improved feature space that accounts for the maximum variation in the data, the covariance between the observed variables, and a linear combination of observed variables that maximizes the class separation. Additionally, different robust statistical measures were used to handle missing values, outliers, and data balancing performed to avoid overfitting and bias. We have also performed a simulation study to reproduce the result using the proposed approach and achieved an average accuracy of 91.40% in the ensemble classification algorithm. Using the proposed method on the challenging ILPD benchmark dataset, the recognition rate has improved between 1% and 18.5%, or nearly 89% accuracy, and AUC by RF ML algorithm based on cross-validation protocol compared to given reference-based approaches. The proposed method is advantageous when we have a massive amount of data and want to reduce the number of features without losing any important information. Due to time limitations, we are not able to investigate whether the proposed method reduced or not the dimensionality of deep-transfer learning features from a pre-trained model or features obtained from a different layer of the convolutional neural network on image data.

To improve this method, some more directions can be investigated in the future. These include investigating the non-linear dimensionality reduction method and evaluating the proposed method on an automated feature extraction system to further confirm this method.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.imu.2022.101155.

References

[1] Lin RH. An intelligent model for liver disease diagnosis. Artif Intell Med 2009;47(1):53–62. https://doi.org/10.1016/j.artmed.2009.05.005.
[2] Collaborators C. Articles the global, regional, and national burden of cirrhosis by cause in 195 countries and territories, 1990 – 2017 : a systematic analysis for the Global Burden of Disease Study 2017. 2017. p. 245–66. https://doi.org/10.1016/S2468-1253(19)30349-8.
[3] Harshpreet Kaur GS. The diagnosis of chronic liver disease using machine learning techniques. Inf. Technol. Ind. 2021;9(2):554–64. https://doi.org/10.17762/itii.v9i2.382.
[4] Tapper EB, Parikh ND. Mortality due to cirrhosis and liver cancer in the United States, 1999-2016: observational study. BMJ Jul. 2018;362:2817. https://doi.org/10.1136/BMJ.K2817.
[5] Joloudari JH, Saadatfar H, Dehzangi A, Shamshirband S. Computer-aided decision-making for predicting liver disease using PSO-based optimized SVM with feature selection. Inform Med Unlocked 2019;17(August). https://doi.org/10.1016/j.imu.2019.100255.
[6] Jacob J, Mathew JC, Mathew J, Issac E. Diagnosis of liver disease using machine learning techniques. 2018.
[7] Ullah S, Awan MD, Sikander Hayat Khiyal M. Big data in cloud computing: a resource management perspective. Sci Program 2018. https://doi.org/10.1155/2018/5418679.
[8] Stone JV. Principal component analysis and factor analysis. Indep. Compon. Anal. 2018:2–3. https://doi.org/10.7551/mitpress/3717.003.0017.
[9] Singh J, Bagga S, Kaur R. Software-based prediction of liver disease with feature selection and classification techniques. Procedia Comput Sci 2020;167(2019):1970–80. https://doi.org/10.1016/j.procs.2020.03.226.
[10] Sreejith S, Khanna Nehemiah H, Kannan A. Clinical data classification using an enhanced SMOTE and chaotic evolutionary feature selection. Comput Biol Med 2020;126(February):103991. https://doi.org/10.1016/j.compbiomed.2020.103991.
[11] Kuzhippallil MA, Joseph C, Kannan A. Comparative analysis of machine learning techniques for Indian liver disease patients. In: 2020 6th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2020; 2020. p. 778–82. https://doi.org/10.1109/ICACCS48705.2020.9074368.
[12] Ali M, Shiaeles S, Bendiab G, Ghita B. Malgra: machine learning and N-GRAM malware feature extraction and detection system. Electronics (Switzerland) 2020;9(11):1–20. https://doi.org/10.3390/electronics9111777.
[13] Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. J. Big Data 2019;6(1). https://doi.org/10.1186/s40537-019-0192-5.
[14] Pasyar P, Mahmoudi T, Kouzehkanan SM. Informatics in Medicine Unlocked Hybrid classification of diffuse liver diseases in ultrasound images using deep convolutional neural networks. Inform Med Unlocked 2021;22(December 2020):100496. https://doi.org/10.1016/j.imu.2020.100496.
[15] Hassannataj J, Saadatfar H, Dehzangi A, Shamshirband S. Informatics in Medicine Unlocked Computer-aided decision-making for predicting liver disease using PSO-based optimized SVM with feature selection. Inform Med Unlocked 2019;17(August):100255. https://doi.org/10.1016/j.imu.2019.100255.
[16] Pasha SJ, Mohamed ES. Novel Feature Reduction (NFR) model with machine learning and data mining algorithms for effective disease risk prediction. IEEE Access 2020;8:184087–108. https://doi.org/10.1109/ACCESS.2020.3028714.
[17] Pasha SJ, Mohamed ES. Advanced hybrid ensemble gain ratio feature selection model using machine learning for enhanced disease risk prediction. Inform Med Unlocked 2022;32(June):101064. https://doi.org/10.1016/j.imu.2022.101064.
[18] Gan D, Shen J, An B, Xu M, Liu N. Integrating TANBN with cost-sensitive classification algorithm for imbalanced data in medical diagnosis. Comput Ind Eng 2020;140(January):106266. https://doi.org/10.1016/j.cie.2019.106266.
[19] Abdar M, Yen NY, Hung JCS. Improving the diagnosis of liver disease using multilayer perceptron neural network and boosted decision trees. J Med Biol Eng 2018;38(6):953–65. https://doi.org/10.1007/s40846-017-0360-z.
[20] Anagaw A, Chang YL. A new complement naïve Bayesian approach for biomedical data classification. J Ambient Intell Hum Comput 2019;10(10):3889–97. https://doi.org/10.1007/s12652-018-1160-1.
[21] Babu MSP, Ramjee M, Katta S, Swapna K. Implementation of partitional clustering on ILPD dataset to predict liver disorders. Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS 2016:1094–7. https://doi.org/10.1109/ICSESS.2016.7883256. 0.
[22] Kumar P, Thakur RS. Liver disorder detection using variable- neighbor weighted fuzzy K nearest neighbor approach. Multimed Tool Appl 2021;80(11):16515–35. https://doi.org/10.1007/s11042-019-07978-3.
[23] Straw I, Wu H. Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction. BMJ Heal. Care Informatics Apr. 2022;29(1):100457. https://doi.org/10.1136/BMJHCI-2021-100457.
[24] UCI machine learning repository: ILPD (Indian liver patient dataset) data set. https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset. [Accessed 14 September 2022].
[25] Krawczyk B. Learning from imbalanced data: open challenges and future directions. Prog. Artif. Intell. 2016;5(4):221–32. https://doi.org/10.1007/s13748-016-0094-0.
[26] Li L, Liu S, Peng Y, Sun Z. Overview of principal component analysis algorithm. Optik 2016;127(9):3935–44. https://doi.org/10.1016/j.ijleo.2016.01.033.
[27] Tharwat A. Principal component analysis - a tutorial. Int. J. Appl. Pattern Recognit. 2016;3(3):197. https://doi.org/10.1504/ijapr.2016.10000630.
[28] Pedroli E, et al. Exploratory and confirmatory factor analysis of the 9-item utrecht work engagement scale in a multi-occupational female sample. A Cross-Sectional Study 2019. https://doi.org/10.3389/fpsyg.2019.02771.
[29] Berg RI. Factor analysis. Am Statistician 1972;26(4):59. https://doi.org/10.1080/00031305.1972.10477372.
[30] Johnson RA, Wichern DW. Factor analysis and inference for structured covariance matrices. Appl. Multivar. Stat. Anal. 2002:477–529.
[31] Li K, Tang P. An improved linear discriminant analysis method and its application to face recognition. Appl Mech Mater 2014;556–562:4825–9. https://doi/org/10.4028/www.scientific.net/AMM.556-562.4825.
[32] Cai Z, Poulos RC, Liu J, Zhong Q. Machine learning for multi-omics data integration in cancer. iScience 2022;25(2):103798. https://doi.org/10.1016/j.isci.2022.103798.
[33] Maniruzzaman M, et al. Accurate diabetes risk stratification using machine learning: role of missing value and outliers. J Med Syst 2018;42(5):1–17. https://doi.org/10.1007/s10916-018-0940-7.
[34] Nkechinyere EM, I IA. 1. In: Comparison of Different Methods of Outlier Detection in Univariate Time Series Data; 2015. p. 54–82.
[35] DeLeo JM. Receiver operating characteristic laboratory (ROCLAB): software for developing decision strategies that account for uncertainty. Proc. - 2nd Int. Symp.

Uncertain. Model. Anal. ISUMA 1993 1993;(2):318–25. https://doi.org/10.1109/ISUMA.1993.366750.

[36] Gupta K, Jiwani N, Afreen N, Divyarani D. Liver disease prediction using machine learning classification techniques. Proc. - 2022 IEEE 11th Int. Conf. Commun. Syst. Netw. Technol. CSNT 2022 2022;(August):221–6. https://doi.org/10.1109/CSNT54456.2022.9787574.

[37] Altaf I, Butt MA, Zaman M. Hard voting meta classifier for disease diagnosis using mean decrease in impurity for tree models. Rev Comput Eng Res 2022;9(2):71–82. https://doi.org/10.18488/76.v9i2.3037.