Contents lists available at GrowingScience

Journal of Future Sustainability

homepage: www.GrowingScience.com/jfs

Effects of hybrid non-linear feature extraction method on different data sampling techniques for liver disease prediction

Rubia Yasmin[a], Ruhul Amin[a] and Md. Shamim Reza[b*]

[a]*Masters of Science, Department of Statistics, Pabna University of Science and Technology, Pabna, Bangladesh*
[b]*Associate Professor, Department of Statistics, Pabna University of Science and Technology, Pabna, Bangladesh*

C H R O N I C L E

Article history:
Received: January 2, 2022
Received in revised format: June 18, 2022
Accepted: August 15, 2022
Available online: September 15, 2022

Keywords:
Liver Disease
Imbalanced Data
Non-linear Feature Extraction
Prediction

A B S T R A C T

Liver disease indicates inflammatory condition of the liver, liver cirrhosis, cancer, or an overload of toxic substances. A liver transplant may reinstate and extend life if a patient has severe liver disease. In the last few years, machine learning (ML) based diagnosis systems have played a vital role in assessing liver patients which eventually leads to proper treatment and saves human life. In this study, we try to predict liver patients by adopting a hybrid feature extraction method to enhance the performance of the ML algorithm. Medical data frequently exhibits non-linear patterns and class imbalances. This is undesirable for the majority of ML algorithms and degrades performance. Here, we present a hybrid feature space that combines t-SNE, Isomap nonlinear features, and kernel principal components that can explain 90% of the variation in the data as a solution to this issue. Before feeding the ML model, data preprocessing techniques including class balancing, identifying outliers, and impute missing values are used. A simulation study and ensemble learning also conducted to justify the proposed prediction performances. Our suggested hybrid non-linear feature exhibits a 2-20 % improvement over existing studies and the ensemble classifier achieved an ideal and outstanding accuracy of 91.33 %.

1. Introduction

Liver Disease is any disturbance of liver function that causes illness (Stoppler & MD, n.d.). Some common liver diseases are Liver cancer, liver cirrhosis, liver failure, and hepatitis A, hepatitis B, and hepatitis C. Liver Cirrhosis is the 8-the leading cause of death in lower-middle-income countries according to the statistics of the World Health Organisation (*WHO | World Health Organization*, n.d.). Viruses, drug overdoses, alcohol abuse, diabetes, immune system abnormality, and so on are the major causes of liver diseases. Liver disease does not always show noticeable signs and symptoms. Without a properly functioning liver, a person cannot survive. If liver diseases are diagnosed at the first stage, then it is possible to manage. Disease classification is a challenging issue for medical diagnosis and prediction. Each liver disease will have its specific treatment rehabilitation. Blood tests, CT (computerized axial tomography) scans, and MRI have commonly advised tests from doctors to determine liver disease. To confirm a specific diagnosis liver biopsy is required (Benjamin Wedro, MD, FACEP, n.d.).

Machine learning techniques can be used in medical problems so we can easily predict the disease and the cost of diagnosis can also be reduced. These also help doctors in making accurate decisions on patients. However, most of the medical datasets including the Indian liver patients' disease (IPLD) dataset have class imbalance problems. Imbalanced classification is a challenging task because the traditional machine learning models and evaluation metrics assume a balanced class distribution. Different real-world datasets have non-linearity between their features. This non-linear pattern of the features decreases the classification accuracy. Different non-linear feature extraction techniques can reduce the non-linearity problem. Kernel PCA, t-SNE, and Iso-map are the most commonly used non-linear feature extraction techniques (Leon-Medina et al., n.d.).

* Corresponding author.
E-mail address: shamim.reza@pust.ac.bd (Md. S. Reza)

ISSN 2816-8151 (Online) - ISSN 2816-8143 (Print)

doi: 10.5267/j.jfs.2022.9.005

1.1 Dataset description

ILPD and simulated imbalanced datasets are used here. Firstly, the Indian Liver Patient Dataset (ILPD) obtained from the UCI repository is used here to classify liver disease (*Index of /Ml/Machine-Learning-Databases/00225*, n.d.). This dataset contains 11 features with one target feature. ILPD dataset consists of 583 samples where 416 (71.4%) are liver disease patients and 167 (28.6%) are non-liver disease patients. This indicates an imbalanced dataset. There are 142 female patients and 441 male patients. The imbalance ratio (IR), which in this dataset is approximately 2.49, is calculated using the formula shown below:

$$\text{IR} = \frac{\text{Majority Samples}}{\text{Minority Samples}}$$

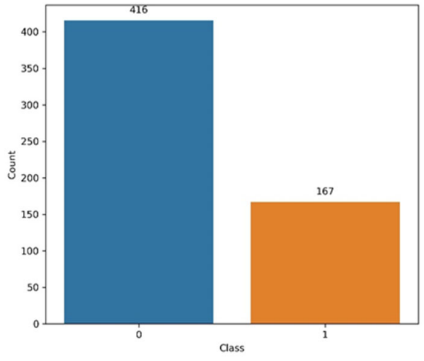


Fig. 1. Patients Class Based on Disease

That indicates the majority class is almost 2.49 times bigger than the minority class. Here the majority class is liver disease patients and the minority class is non-liver disease patients.

We use the imbalanced Synthetic Dataset with 1000 samples and 20 features. It contains 696 non-cases and 304 cases. The imbalanced ratio (IR) is about 2.5 here.

2. Materials and methods

We tried to resample the samples in the dataset thus we get a balanced dataset. This section represents brief information about the materials and methods which we used in this study. We propose a hybrid non-linear feature extraction technique for the prediction of liver disease.

2.1 Data pre-processing

In Machine Learning, data pre-processing means preparing, cleaning, and organizing the raw data into an appropriate format so the machine learning models can easily handle them (Kang & Tian, 2018). Almost all the real-world datasets are noisy, incomplete, and inconsistent for the machine learning algorithm. In the ILPD dataset 'Gender' feature is categorical so we convert it to a number by using the Scikit-learn package OrdinalEncoder. Here, we found only 4 missing values in the Globulin Ratio feature that has been replaced by the robust measure median. We check the outliers by using the absolute Z-Score value and the total values filtered by setting a 3-sigma threshold label. If any values were greater than this threshold label it recognized outliers. Then it has been replaced by the median for each feature. Thus, we can avoid the problem of loss of potential information from the dataset. So, the number of samples remains the same after pre-processing. We applied Standard scalar transformation using the Scikit-learn package StandardScaler for scaling the features processing. We applied Standard scalar transformation using the Scikit-learn package StandardScaler for scaling the features.

2.2 Data Sampling Techniques

The main goal of this study is to accurately determine liver patients. Random over-sampling (ROS), random under-sampling (RUS), synthetic minority oversampling (SMOTE), Tomek Links (TL), and SMOTE+ENN are used here to balance the dataset. Random oversampling (ROS) aims to improve the class imbalance by increasing the size of the minority class. Minority class samples are randomly reproduced until the optimum class ratio is attained (Chkirbene et al., 2021). The random Under-sampling technique randomly decreases the size of majority class events without losing any data in minority class events. Losing majority class information is the main disadvantage of this technique.

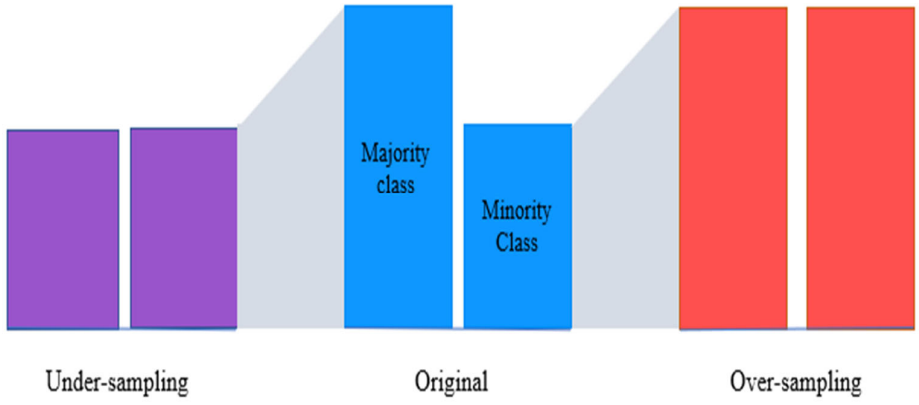


Fig. 2. Random under-sampling and over-sampling technique.

SMOTE stands for Synthetic Minority Oversampling Technique that generates synthetic samples of the dataset's minority class. The minority class is oversampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors (Junsomboon & Phienthrakul, 2017). SMOTE has a process for creating synthetic data by using the given equation:

$$Z_n = Z_i + (Z_k - Z_i)\,\mu$$

Here, Z_n is a new synthetic data

Z_i is sample data selected in the minority class

Z_k is sample data that selected 1 from 3 KNN of Z_i in the minority class

μ is a random constant range from 0 to 1.

Tomek link is a special type of under-sampling algorithm which is a refinement of the Condensed Nearest Neighbor (CNN) technique. This technique eliminates the boundary instances which have more chances of misclassification. Tomek link is a pair of patterns that are the closest neighbors and belong to separate classes (Kumar & Thakur, 2021).

SMOTE + ENN is a hybrid technique that removes a larger number of observations from the sample space where ENN stands for Edited Nearest Neighbors. Any example whose class label varies from the class of at least two of its three closest neighbors is removed by ENN (Batista et al., 2004). The noisy samples can be removed as borderline samples using this procedure. As a consequence, the class distinction is more apparent and straightforward.

2.3 Non-linear Feature Extraction Techniques

The feature extraction technique produces new features by selecting, transforming, and manipulating the raw features to enhance prediction model's accuracy. Without losing significant information, this technique compresses the data into optimal quantities for algorithms to process by creating artificial features.

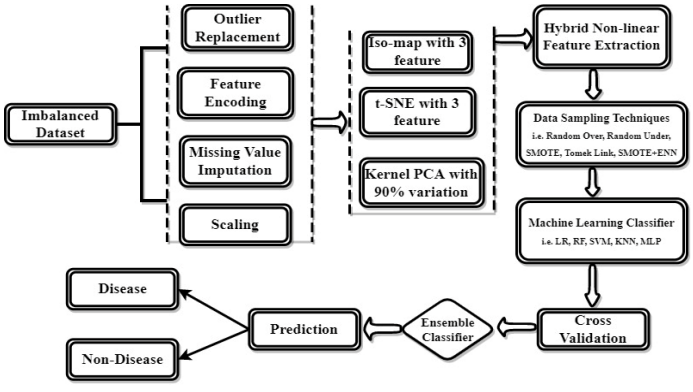


Fig. 3. Proposed hybrid non-linear feature extraction technique workflow

2.3.1 Kernel PCA

Kernel PCA refers to the non-linear variant of normal PCA(CESARSOUZA, n.d.). In kernel PCA, many types of kernels are employed, such as 'linear,' 'rbf,' 'polynomial,' and 'sigmoid'. The kernelized version can capture effectively a subspace that decreases the dimension of the original feature space by changing the feature space x_k into a higher dimension space $\Omega(x)$ (Ezukwoke et al., 2019). We summarize the procedure for kernel PCA as follows:

Algorithm: **Input**: $\Omega(x) \in \theta(x, x_k)$ where

$\Omega(x) \in \mathbb{R}^d$

Suppose, $\Theta = \theta\ (x_j, x_k)$

begin

Select a kernel θ;

Construct Gram Matrix

$\hat{\Theta} = \Theta - 1_{1/N}\,\Theta - \Theta\,1_{1/N} + 1_{1/N}\,\Theta\,1_{1/N}$;

Solve eigen problem $\Theta\,\varepsilon_k = \lambda\,\varepsilon_k$;

Project data in new space

$\hat{x} = \sum_{i=1}^{N} \varepsilon_k\,\theta$

end

Output: $\hat{x} \in \mathbb{R}^{\theta}$ where $\theta \ll d$.

2.3.2 t-SNE

t-SNE or t-distributed Stochastic Neighbor Embedding is mostly used to visualize high-dimensional data by embedding it in a 2D or 3D space. The t-SNE technique starts by generating a probability distribution that reflects the mutual distance connections between the points in the high-dimensional space (Rodrigues, n.d.). It can be represented as:

$$p_{j,i} = \frac{exp\,(-\|zi-zj\|^2/2\sigma i^2)}{\sum exp\,(-\|zi-zk\|^2/2\sigma i^2)} \quad where\ k \neq i$$

Then, this algorithm finds similar relations between the data points and creates a low dimensional space. It can be represented as:

$$q_{j,i} = \frac{exp\,(-\|wi-wj\|^2)}{\sum exp\,(-\|wi-wk\|^2)} \quad where\ k \neq i$$

To optimize the problem, t-SNE minimizes the sum of Kullback-Leiber divergence of overall data points using a gradient descent method.

2.3.3 Iso-map

Isometric mapping, also abbreviated as Iso-map, is a variation of metric multidimensional scaling that uses geodesic distance to describe nonlinear data. Based on spectral theory, its technique attempts to preserve the geodesic distance in low-dimensional space. It builds a neighborhood network first, then utilizes the graph to calculate the geodesic distance between all data points. The dataset is then low-dimensionally embedded via eigen value decomposition of the geodesic distance's matrix (Yousaf et al., 2020).

3. Results and Discussion

In this study, we use the ensemble machine learning algorithm for the prediction of liver disease patients. Ensemble learning algorithms strategically combine different predictions from multiple learning algorithms for improving their performance. For the ensemble classifier in this study, we employ the Logistic Regression, Random Forest, K-Nearest Neighbor, Support Vector Machine, and Multilayer Perceptron algorithm. 10-fold cross-validation is used for every classifier to avoid biased results. Different evaluation metrics such as Accuracy, Precision, Recall, F1-Score, and G-mean score are used here for result comparison. Another pictorial way of comparing the performance is the receiver-operating characteristic (ROC) curve. These measures are calculated from the confusion matrix as illustrated in Table: 1 (for 2 class problems).

Table 1
Confusion- matrix

		Predict
Actual	Disease (Positive)	No-disease (Negative)
Positive	TP	FP
Negative	FN	TN

Where TP is the number of correctly classified positive instances, FP is the number of misclassified positive instances, FN is the number of misclassified negative instances, TN is the number of correctly classified negative instances.

$$\text{Accuracy} = \frac{TP+TN}{(TP+TN+FP+FN)}, \qquad \text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}, \qquad \text{F-1 score} = \frac{2*TP}{2*TP+FP+FN}$$

$$\text{G-mean score} = \sqrt{\text{Precision} * \text{Recall}}$$

For this experiment, we use five different data sampling techniques to balance the two imbalanced datasets. This section presents the specific results of the classifier performance. According to Table 2, the ensemble classifier produces the best results with an accuracy of 89.78% for the ILPD dataset after utilizing the SMOTE+ENN data sampling approach. Table 3 demonstrates enhanced performance when using the suggested hybrid non-linear feature extraction strategy on the ILPD dataset. After the data have been processed using the suggested way, we can see that several data sampling techniques including random oversampling (ROS), Tomek Link, and SMOTE+ENN functioned well. The recently created data sampling method SMOTE+ENN produced the best results when used with the suggested strategy. This gives 91.33 % accuracy, 87.97 % precision, 98.87 % recall score, 93.03 % F1-score and 89.17 % G-mean score.

Table 2
Classifier's performance without proposed technique on ILPD dataset

Data Sampling Techniques	Accuracy %	Precision %	Recall %	F1-Score %	G-Mean %
ROS	80.75	74.33	93.99	83.01	79.68
RUS	65.27	62.44	76.65	68.82	64.27
SMOTE	78.49	73.94	87.98	80.35	77.91
Tomek Link	69.59	51.02	14.97	23.15	37.45
SMOTE+ENN	**89.78**	**87.54**	**98.66**	**92.76**	**84.39**

Table 3
Classifier's performance with proposed technique on ILPD dataset

Data Sampling Techniques	Accuracy	Precision	Recall	F1-Score	G-Mean
ROS	80.79	75.86	89.90	82.29	80.12
RUS	63.78	62.92	67.07	64.93	63.69
SMOTE	76.68	71.94	87.52	78.96	75.92
Tomek Link	72.25	65.57	23.95	35.09	47.52
SMOTE+ENN	**91.33**	**87.97**	**98.87**	**93.09**	**89.17**

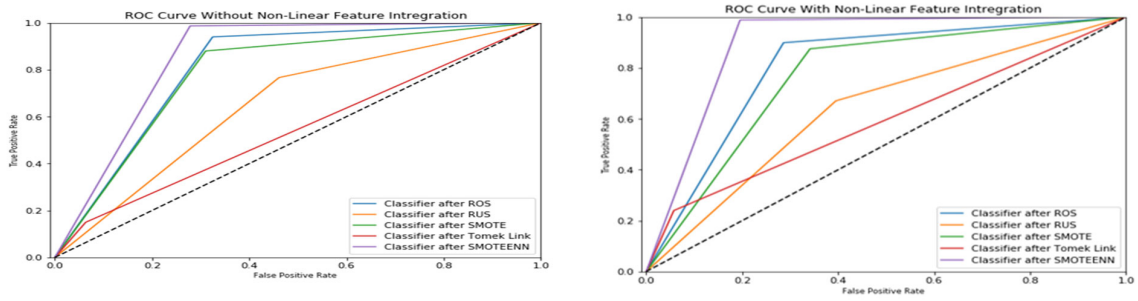
An imbalanced synthetic dataset with 1000 sample observations is treated as a second dataset. We apply the proposed technique to this dataset and observe the effects of the classifier's performance. Results are shown in Table 4 and Table 5. From Table 4, we get an ensemble learning algorithm that gives better performance after resampling the dataset with the SMOTE+ENN technique with 91.36 % accuracy. We apply the proposed technique to the synthetic dataset and get random under-sampling (RUS), Tomek Link, and SMOTE+ENN sampling shows their better performance in prediction. RUS gives 67.27 % accuracy and Tomek Link gives 76.38 % accuracy, SMOTE+ENN gives 93.31% accuracy.

Table 4
Classifiers Performance without Proposed Technique on Synthetic Dataset

Data Sampling Techniques	Accuracy	Precision	Recall	F1-Score	G-Mean
ROS	84.27	81.67	88.36	84.84	84.17
RUS	66.94	68.33	63.15	65.64	66.83
SMOTE	80.24	78.79	82.76	80.73	80.21
Tomek Link	76.38	74.01	43.09	54.47	63.21
SMOTE+ENN	**91.36**	**91.54**	**94.63**	**93.06**	**90.32**

Table 5
Classifier's performance with proposed technique on a synthetic dataset

Data Sampling Techniques	Accuracy	Precision	Recall	F1-Score	G-Mean
ROS	79.38	79.51	79.17	79.34	79.38
RUS	67.27	68.82	63.19	65.87	67.14
SMOTE	77.66	78.10	76.87	77.48	77.65
Tomek Link	76.38	74.72	43.75	55.19	63.65
SMOTE+ENN	**93.31**	**94.46**	**93.52**	**93.28**	**93.27**



(a) (b)

Fig. 4. ROC curve (a) without and (b) with the proposed technique

Fig. 4 shows the ROC curve for the classifier before and after applying the proposed technique. The false positive rate is on the X-axis and the True positive rate is on the Y-axis. We can observe that, after resampling the dataset with SMOTE+ENN, the ensemble classifier gives a better result.

In Figure 5, the Ensemble classifier shows higher values for the SMOTE+ENN technique with the proposed technique than others. That is, if we use this proposed hybrid non-linear feature extraction technique jointly with the data sampling method then this gives a better result in the prediction of an imbalanced classification problem.

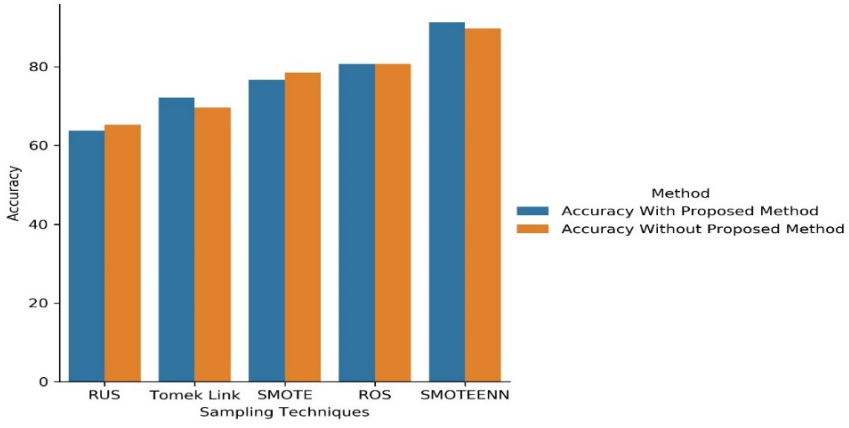


Fig. 5. Bar diagram for accuracy comparison of ILPD dataset

Table 6
Comparison between the proposed technique and other's research on ILPD

Reference No.	Algorithm	Data Splitting Method	Accuracy
(Bahramirad et al., 2013)	Logistic Regression	WEKA-tool	73.39
(Kumar & Thakur, 2019)	KNN	10-fold cv	74.67
(Fathi et al., 2020)	G-SVM	10-fold cv	90.90
(Gulia et al., 2014)	Random Forest	WEKA-tool	71.87
(Kumar & Thakur, 2021)	Variable-NWFKNN	10-fold cv	87.71
(Singh et al., 2020)	Logistic Regression	10-fold cv	74.36
This Study	**Hybrid Feature Integra-**	**10-fold cv**	**91.33**

Table 6 shows that the proposed non-linear feature integration method gives the best result with an accuracy of 91.33% for the ILPD dataset. The proposed method can diagnose liver disease more accurately.

4. Conclusion

In medical diagnosis, accurate and effective liver disease categorization is crucial. The correct categorization of liver illness can assist clinicians in developing effective treatment plans. Using kernel PCA, t-SNE, and Iso-map, we suggested a unique hybrid feature extraction approach. This approach for extracting non-linear features improves prediction accuracy. In a 10-fold cross-validation technique, our suggested non-linear feature extraction exhibits a 2-20 % improvement over existing studies. Logistic Regression, Random Forest, SVM, K-NN, and MLP classifiers are used to build an ensemble learning classifier. The ensemble classifier produced an excellent and optimum accuracy of 91.33 percent for our proposed hybrid non-linear feature extraction technique. It would be interesting in the future to see the performance of this method for image data classification in such a framework that is cost-effective and viable in recognizing disease in the medical dataset.

References

Bahramirad, S., Mustapha, A., & Eshraghi, M. (2013, September). Classification of liver disease diagnosis: a comparative study. In *2013 Second International Conference on Informatics & Applications (ICIA)* (pp. 42-46). IEEE.

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, *6*(1), 20–29. https://doi.org/10.1145/1007730.1007735

Benjamin Wedro, MD, FACEP, F. (n.d.). *Liver Disease: Early Signs, Symptoms, Treatment, Stages, Types & Diet*. Retrieved April 26, 2022, from https://www.medicinenet.com/liver_disease/article.htm

CESARSOUZA. (n.d.). *Kernel Functions for Machine Learning Applications – César Souza*. Retrieved May 4, 2022, from http://crsouza.com/2010/03/17/kernel-functions-for-machine-learning-applications/

Chkirbene, Z., Erbad, A., Hamila, R., Gouissem, A., Mohamed, A., Guizani, M., & Hamdi, M. (2021). A Weighted Machine Learning-Based Attacks Classification to Alleviating Class Imbalance. *IEEE Systems Journal*, *15*(4), 4780–4791. https://doi.org/10.1109/JSYST.2020.3033423

Ezukwoke, K., Zareian, S., & Regression, L. (2019). *KERNEL METHODS FOR PRINCIPAL COMPONENT ANALYSIS (PCA) A comparative study of classical and kernel pca . December*. https://doi.org/10.13140/RG.2.2.17763.09760

Fathi, M., Nemati, M., Mohammadi, S. M., & Reza, A.-K. (2020). *A MACHINE LEARNING APPROACH BASED ON SVM FOR CLASSIFICATION OF LIVER DISEASES*. *32*(2), 1–9. https://doi.org/10.4015/S1016237220500180

Gulia, A., Vohra, R., & Rani, P. (2014). *Liver Patient Classification Using Intelligent Techniques*. *5*(4), 5110–5115.

Index of /ml/machine-learning-databases/00225. (n.d.). Retrieved May 1, 2022, from https://archive.ics.uci.edu/ml/machine-learning-databases/00225/

Junsomboon, N., & Phienthrakul, T. (2017). Combining over-sampling and under-sampling techniques for imbalance dataset. *ACM International Conference Proceeding Series*, *Part F1283*(1), 243–247. https://doi.org/10.1145/3055635.3056643

Kang, M., & Tian, J. (2018). *Machine Learning : Data Pre-processing*. 111–130.

Kumar, P., & Thakur, R. S. (2019). *Early Detection of the Liver Disorder from Imbalance Liver Function Test Datasets*. *4*, 179–186.

Kumar, P., & Thakur, R. S. (2021). Liver disorder detection using variable- neighbor weighted fuzzy K nearest neighbor approach. *Multimedia Tools and Applications*, *80*(11), 16515–16535. https://doi.org/10.1007/s11042-019-07978-3

Leon-Medina, J. X., Anaya, M., Pozo, F., & Tibaduiza, D. (n.d.). *Nonlinear Feature Extraction Through Manifold Learning in an Electronic Tongue Classification Task*. https://doi.org/10.3390/s20174834

Rodrigues, G. (n.d.). *Automatic feature extraction with t-SNE | by Gonçalo Rodrigues | Jungle Book | Medium*. Retrieved May 4, 2022, from https://medium.com/jungle-book/automatic-feature-extraction-with-t-sne-62826ce09268

Singh, J., Bagga, S., & Kaur, R. (2020). Software-based Prediction of Liver Disease with Feature Selection Software-based Prediction of Liver Disease with Feature Selection and Classification Techniques and Classification Techniques. *Procedia Computer Science*, *167*(2019), 1970–1980. https://doi.org/10.1016/j.procs.2020.03.226

Stoppler, M. C., & MD. (n.d.). *Liver Disease Symptoms, Signs & Causes*. Retrieved April 23, 2022, from https://www.medicinenet.com/liver_disease_symptoms_and_signs/symptoms.htm

WHO | World Health Organization. (n.d.). Retrieved April 26, 2022, from https://www.who.int/

Yousaf, M., Rehman, T. U., & Jing, L. (2020). An Extended Isomap Approach for Nonlinear Dimension Reduction. *SN Computer Science*, *1*(3). https://doi.org/10.1007/s42979-020-00179-y