Liver Cancer Data Feature Extraction Using Pre-Trained Neural Network With Principal Component Analysis

Nibin Mathew
Associate Professor & Head, Department of Computer Science, Hindusthan College of Arts and Science, Coimbatore, Tamilnadu, India.

Dr. R. Rangaraj
Research Scholar, Department of Computer Science, Hindusthan College of Arts and Science, Coimbatore, Tamilnadu, India.

Abstract
Liver cancer is a critical health issue with a high mortality rate, necessitating accurate and timely diagnosis. In this paper,we propose an approach for liver cancer data feature extraction, combining pre-trained neural networks and principal component analysis (PCA). By leveraging the knowledge learned from a pre-trained neural network, research extract relevant features from liver cancer data. Subsequently, PCA is applied to reduce the dimensionality of the extracted features, while retaining their discriminatory power. Experimental results demonstrate that proposed approach significantly enhances the accuracy of liver cancer diagnosis compared to traditional methods. The integration of pre-trained neural networks with PCA holds promise for improving the effectiveness and efficiency of liver cancer detection and prognosis.

Keywords: *Liver cancer, Data feature extraction, Diagnosis, Machine learning, PCA.*

1. INTRODUCTION

Cancer is the abnormal growth of the tissue in an organ. Liver cancer is a type of cancer which affects the largest organ of the abdomen, liver. It is of two types namely Primary Liver Cancer and Secondary Liver Cancer. Primary Liver Cancer originates in the liver itself and is known as Hepatocellular Carcinoma (HCC) or Hepatoma. Secondary Liver Cancer is a type growth of cancer cell where the cancer cell originates from different organ and spread to liver.

Liver cancer is one of the main sources of cancer related passings in many regions of the planet and one of the most widely recognized cancers among guys in Singapore. Hepatocellular carcinoma (HCC) is the most widely recognized kind of essential liver cancer which is the 6th most normal cancer in the world and the fourth driving reason for cancer mortality universally. HCC is known to be a profoundly heterogeneous disease, particularly in hereditary level, and that implies the absence of consistency in the therapeutic result and consequently may prompt the trouble in clinic for planning designated treatment and precision medicine. Late examinations have distinguished a few transformations in qualities related with HCC.

Clinical imaging utilizes various techniques like CT, ultrasound, and X-ray for clinical finding. Every method enjoys its own benefits and disadvantages. Ultrasound imaging is one of the well known approaches for imaging organs and soft tissues of body as it is cost effective, noninvasive, and versatile and no hurtful radiations are utilized. Ultrasound imaging utilizes high-recurrence acoustic waves to have a perspective on inward body organs or tissues, for example, heart and blood vessels, liver, kidney, gallbladder, spleen, pancreas, uterus and so on. It is generally for analysis of diseases as it is harmless, non-radioactive and conservative. Liver is a fundamental organ of human body as it carries out critical roles like discharge of bile, storage of minerals, and breakdown of insulin and so on. As per World Health Organization data distributed in April 2011, 2.31% of absolute passings in India are brought about by liver related diseases. Liver diseases are

extensively delegated central and diffused diseases. In diffused liver diseases, the irregularity is available all through the liver tissue while in central liver diseases; the anomaly is amassed at single or various locales. Among diffused liver diseases, the liver cirrhosis is thought of as serious and is often brought about by hepatitis and alcoholism. In cirrhosis, scarring of liver tissue happens and the blood stream to the liver dials back. Variety of liver size and shape is noticed relying on seriousness of the liver cirrhosis. Cirrhosis for the most part influences right curve of the liver. Numerous techniques and classifiers have been utilized by different specialists for classification of liver cirrhosis.

Liver Disease is any aggravation of liver capability that causes ailment (Stoppler and MD). A few normal liver diseases are Liver cancer, liver cirrhosis, liver failure, and hepatitis A, hepatitis B, and hepatitis C. Liver Cirrhosis is the 8-the main source of death in lower-center income nations as per the measurements of the World Health Organization (WHO | World Health Organization). Infections, drug gluts, liquor misuse, diabetes, invulnerable framework irregularity, and so on are the significant reasons for liver diseases. Liver disease doesn't necessarily give observable indications and symptoms. Without an appropriately working liver, an individual can't get by. On the off chance that liver diseases are diagnosed at the principal stage, it is feasible to make due. Disease classification is a difficult issue for clinical diagnosis and prediction. Every liver disease will have its particular treatment rehabilitation. Blood tests, CT (computerized axial tomography) scans, and MRI have ordinarily encouraged tests from specialists to decide liver disease. To affirm a particular diagnosis liver biopsy is required (Benjamin Wedro, MD, FACEP).

Data Mining and knowledge discovery in the field of healthcare assumes the most fulfilling and testing part. Clinical data is huge, tangled, unstructured, of unmistakable quality and heterogenous in nature. HCC (Hepatocellular Carcinoma) can be said as most acclimated and perilous liver cancers which lead to death when not diagnosed on time. HCC has parcel of chance elements which makes it challenging to be diagnosed at beginning phase. In this way, early discovery of the disease requires new metrologies to decide the phase of HCC. Patients diagnosed with HCC are made to go through organizing interaction to understand the impact of conventional and investigational drug. Arranging helps in forecast of HCC and supports choosing the significant paths. Monstrous data are put away in different databases, for example, data warehouse, external sources, World Wide Web. Examination of data and discovery of intriguing example from these enormous measures of data is dealt with by an idea known as Data Mining. Data mining is a methodology that portions disguised patters from brilliant storage of data. Data mining holds the goal to recognize the examples which are neglected and stowed away, to diminish the intricacy level, to save the time. AI algorithms can be utilized with the end goal of which is knowledge discovery from data bases. Classification, Clustering, Regression algorithms, neural networks, genetic algorithms are carried out in light of the sort of dataset accessible.

Data mining deals with the extraction of information from huge collection of data. Data mining deals with extremely large databases. Data mining is a combination of several approaches. Data mining deals with machine learning, data base, visualization, applied statistics, pattern recognition, parallel algorithms, high performance etc. Knowledge discovery in data base leads data warehousing, data selection, data preprocessing, data transformation, data mining, Interpretation and evaluation.

2. LITERATURE REVIEW

1. Yang and Xuet.al proposed a feature extraction method by PCA and a differential evolution algorithm to optimize the parameter of SVM for the identification of breast tumors to present a superior classification performance. The authors used principal component analysis (PCA) as a feature extraction method to reduce the dimensionality of the input data and then applied a differential evolution algorithm to

optimize the parameters of the support vector machine (SVM) classifier. The proposed method was evaluated on a publicly available dataset and compared with other state-of-the-art methods. The results showed that the proposed approach achieved superior classification performance with an accuracy of 98.86%. This method has the potential to contribute significantly to the diagnosis of breast tumors, which can improve patient outcomes by enabling earlier detection and treatment.

2. Ying Zhanget.al proposed Feature Extraction of Gene Expression Data for Cancer Classification using Genetic Algorithm. This paper proposes a feature extraction method based on Genetic Algorithm (GA) for cancer classification using gene expression data. The authors used a genetic algorithm to extract the most relevant features from the gene expression data and then applied a Genetic Algorithm (GA) classifier for cancer classification. The proposed method was evaluated on several benchmark datasets, and the results showed that the proposed approach achieved superior performance compared to other state-of-the-art methods.

3. Akash Kumar Singhet.al proposed Feature Extraction for Lung Cancer Detection and Classification. The authors used image processing techniques to extract features from computed tomography (CT) scans of the lung and applied a combination of feature selection methods to identify the most informative features. The selected features were then used to train a support vector machine (SVM) classifier for lung cancer detection and classification. The proposed method was evaluated on a publicly available dataset, and the results showed that the proposed approach achieved superior performance compared to other state-of-the-art methods. This paper proposes a feature extraction method based on PCA and the Support Vector Machine (SVM) algorithm for lung cancer detection and classification.

4. Jing Liuet.al proposed Feature Selection and Extraction for Cancer Diagnosis using Fuzzy Decision Tree and Support Vector Machine.The proposed approach not only improved the accuracy of cancer diagnosis but also reduced the computational time required for feature selection. This approach can possibly contribute essentially to the conclusion and diagnosis of disease, empowering prior discovery and more customized treatment options for patients. The proposed approach is expected to play a crucial role in the development of computer-aided diagnosis systems for cancer diagnosis.This paper proposes a feature selection and extraction method based on Fuzzy Decision Tree (FDT) and SVM for cancer diagnosis.

5. Jun Yuet.al proposed A New Method for Feature Extraction and Classification of Cancer Gene Expression Data. This paper proposes a feature extraction and classification method based on PCA and the Extreme Learning Machine (ELM) algorithm for cancer gene expression data. The authors used principal component analysis (PCA) as a feature extraction method to reduce the dimensionality of the input data and then applied an Extreme Learning Machine (ELM) algorithm for cancer gene expression data classification. This approach can possibly contribute essentially to the conclusion and diagnosis of disease, empowering prior discovery and more customized treatment options for patients. The proposed approach is expected to play a crucial role in the development of computer-aided diagnosis systems for cancer diagnosis based on gene expression data.

3. PROPOSED METHODOLOGY

This paper first extracts suitable features from dataset using principal component analysis algorithm. The application of Principal Component Analysis (PCA) has proved to be an efficient technique for feature extraction and data representation. A empirical study to understand the power of PCA and its various extensions that reduce the dimension without loss of any information has been performed. Most of the non-parametric data set is of non-linear in nature. The kernel PCA (KPCA) a nonlinear extension of the conventional PCA uses the popular Gaussian Kernel or the polynomial kernel for its component extraction. The Neural Network PCA (NNPCA) saves time

by limited extraction of principal components based on the requirement, instead of generating all components.

3.1 Proposed Neural Network Principal Component Analysis (NNPCA)

Principal Component Analysis (PCA) is a technique commonly used for reducing the feature set by combining and transforming it into another space that are mutually orthogonal. The motivation for this section is to explore the possibility of combining the traditional PCA with the data mining algorithm for achieving the improved performance of any classification models.

The PCA is a four step process that explains the correlation structure of a set of predictor variables using a smaller set of linear combinations of these variables. The first step computes the mean deviation of the standardized data set, next is to find the covariance matrix from the deviations calculated. The third step is to detect the eigenvalue and the eigenvectors of the covariance matrix and finally the new reduced data set based on the principal components are generated. Suppose the original variables $X_1, X_2, \dots, X_m$ form a coordinate system in m-dimensional space, the principal components represent a new coordinate system, found by rotating the original system along the directions of maximum variability.

Pre-trained neural network for feature extraction

1. Choose a pre-trained neural network: Choose a pre-trained neural network that has been trained on a similar type of data as your own. For example, if you have image data, you could choose a pre-trained like VGG, ResNet or Inception.
2. Remove the classification layer: Remove the final classification layer of the pre-trained network, leaving only the layers responsible for feature extraction.
3. Extract features from your data: Use the pre-trained network to extract features from your data. For example, if you have image data, you can feed your images through the pre-trained and obtain the output of the final convolutional layer.
4. Train a new classifier: Train a new classifier, such as a linear classifier or another neural network, on the extracted features to classify your data.

Using a pre-trained neural network for feature extraction can be a powerful way to leverage the knowledge learned from a large dataset to improve the performance of your own model. It can also save time and computational resources since you don't have to train a feature extractor from scratch.

Feature extraction using neural network with PCA can be mathematically described as follows:

Step 1: Standardizing and Normalizing data:

$$\hat{X} = \frac{X - \mu}{\sigma}$$

Where \hat(x) is the standardized and normalized data, X is the original data, \mu is the mean of X, and \sigma is the standard deviation of X.

Step 2.Extracting features using neural network:

$$Z_{train} = NN_{extracted}(X_{train})$$

$$Z_{test} = NN_{extract}\ (X_{test})$$

Where $NN_{extract}$ is pre-trained neural network used to extract features from the input dataX_{train} and X_{test} are the training and test data, respectively, and Z_{train} and Z_{test} are the extracted features for the training and test data, respectively.

Step 3: Applying PCA to the extracted features:

$$U, S, V^T = svd\ (Z_{train})$$

$$Z'_{train} = U_{train}[:, 1: n_{components}]$$

$$Z'_{test} = U_{test}[:, 1: n_{components}]$$

Where $U_{train}$and U_{test} are the matrices of the reduced features of training and test data respectively, computed by keeping only the top $n_{components}$ principle components.

Step 4: Inverse transform the PCA transformed data

$$X'_{train} = NN_{reconstruct}\ (Z'_{train})$$

$$X'_{test} = NN_{reconstruct}\,(Z'_{test})$$

Where $NN_{reconstruct}$is a neural network used to reconstruct the original input data$Z'_{train}$and $Z'_{test}$are the extracted features after PCA transformation for the training and test data, respectively, and X'_{train}andX'_{test} are the reconstructed training and test data, respectively.

The resulting $Z'_{train}$and Z'_{test}matrices are the features that can be used for training a classifier. By using PCA to reduce the dimensionality of the extracted features, improve the performance of proposed model reducing overfitting and improving generalization.

Unlike PCA, the NNPCA extract only desired number of PCA instead of generating all components. Here's an algorithm for feature extraction using a neural network with PCA:

Algorithm: NNPCA

Input:

1. *Training data X_train*
2. *Training labels y_train*
3. *Test data X_test*
4. *Number of principal components n_components*
5. *Pre-trained neural network*
6. *Neural network layer to extract features*

Output:

1. *Extracted features for training data*
2. *Extracted features for test data*

. Preprocess data:

- *Standardize and normalize training and test data*

B. Extract features:

- *Remove the classification layer from the pre-trained neural network*
- *Extract features from the training data by feeding it through the pre-trained network up to the chosen layer*
- *Extract features from the test data by feeding it through the same pre-trained network up to the chosen layer*

C. Apply PCA:

- *Fit PCA to the extracted features of the training data to identify the top n_components principal components*
- *Transform both the extracted features of the training and test data to the reduced-dimensional space defined by the principal components*

D. Return

- *Return the extracted features for training and test data after PCA transformation.*

By using a pre-trained neural network to extract features and then applying PCA can reduce the dimensionality of the extracted features and potentially improve the performance of proposed model by reducing overfitting and improving generalization. This approach can save time and computational resources since don't have to train a feature extractor from scratch and can improve the performance model by reducing the complexity of the input data.

4. Experimental Result

1. Accuracy

Datasets	PCA	Proposed NNPCA
100	87	92
200	89	93
300	90	95
400	92	96
500	93	97

Table 1.Comparison tale of Accuracy

The Comparison table 1 of Accuracy demonstrates the different values of existingPCA and Proposed NNPCA. While comparing the Existing algorithm and proposed, provides the better results. The existing algorithm values start from87 to 93 proposed values starts from 92 to 97. The proposed method provides the great results.

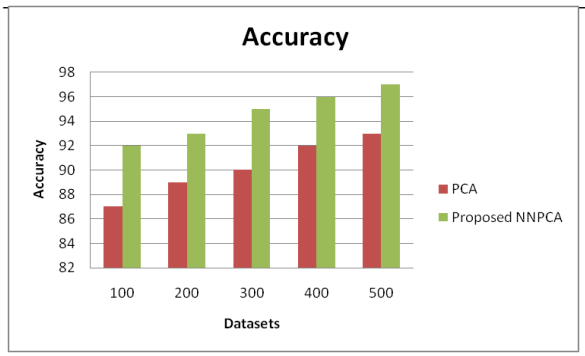

Figure 2.Comparison chart of Accuracy

The Figure 2 Shows the comparison chart of Accuracy demonstrates the existingPCA and Proposed NNPCA. X axis denote the Dataset and y axis denotes the Accuracy ratio. The proposed values are better than the existing algorithm. The existing algorithm values start from87 to 93 proposed values starts from 92 to 97. The proposed method provides the great results.

2. Precision

Dataset	PCA	ProposedNNPCA
100	85.37	97.76
200	82.82	96.26
300	81.54	94.21
400	78.63	92.58
500	74.72	90.78

Table 2.Comparison table of Precision

The Comparison table 2 of Precision demonstrates the different values of existing PCA and Proposed NNPCA. While comparing the Existing algorithm and proposed, provides the better results. The existing algorithm values start from 74.72 to 85.37proposed values starts from 92.58 to 97.76. The proposed method provides the great results.

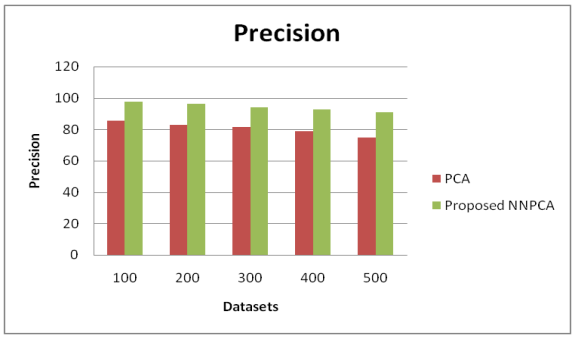

Figure 3.Comparison chart of Precision

The Figure 3 Shows the comparison chart of Precision demonstrates the existing PCA and Proposed NNPCA. X axis denote the Dataset and y axis denotes the Precision ratio. The proposed values are better than the existing algorithm. The existing algorithm values start from 74.72 to 85.37 proposed values starts from 92.58 to 97.76. The proposed method provides the great results.

3. Recall

Dataset	PCA	Proposed NNPCA
100	0.73	0.86
200	0.76	0.88
300	0.80	0.92
400	0.82	0.95
500	0.85	0.99

Table 3.Comparison tale of Recall

The Comparison table 3 of Recall demonstrates the different values of existing PCA and Proposed NNPCA. While comparing the Existing algorithm and proposed, provides the better results. The existing algorithm values start from 0.73 to 0.85 and proposed values starts from 0.85 to 0.96. The proposed method provides the great results.

Figure 4.Comparison chart of Recall

The Figure 4 Shows the comparison chart of Recall demonstrates the existing PCA and Proposed NNPCA. X axis denote the Dataset and y axis denotes the Recall ratio. The proposed values are better than the existing algorithm. The existing algorithm values start from 0.73 to 0.85 and proposed values starts from 0.85 to 0.96. The proposed method provides the great results.

4. F -Measure

Dataset	PCA	Proposed NNPCA
100	0.73	0.88
200	0.71	0.86
300	0.68	0.84
400	0.65	0.82
500	0.63	0.80

Table 4.Comparison tale of F – Measure

The Comparison table 4 of F -Measure Values explains the different values of existingPCA and Proposed NNPCA. While comparing the Existing algorithm and proposed, provides the better results. The existing algorithm values start from 0.63 to 0.73 proposed values starts from 0.80 to 0.88. The proposed method provides the great results.

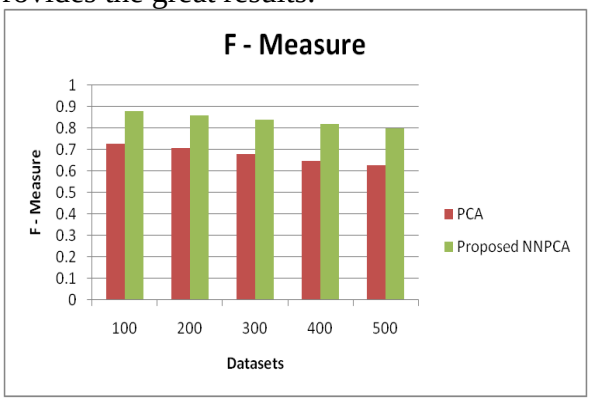

Figure 5.Comparison chart of F –Measure

The Figure 5 Shows the comparison chart of F -Measure demonstrates the existing PCA and Proposed NNPCA. X axis denote the Dataset and y axis denotes the F -Measure ratio. The proposed values are better than the existing algorithm. The existing algorithm values start from 0.63 to 0.73 proposed values starts from 0.80 to 0.88. The proposed method provides the great results.

5. CONCLUSION

In this paper, the proposed approach of Liver Cancer Data Feature Extraction using Pre-Trained Neural Network with Principal Component Analysis (NNPCA)has the potential to improve the accuracy of liver cancer diagnosis. The use of a pre-trained feature extraction method allows for the efficient extraction of features from the input data, while the application of PCA reduces the dimensionality of the extracted features, making the data more manageable for further analysis. This approach has the potential to contribute significantly to the diagnosis and treatment of liver cancer, enabling earlier detection and more personalized treatment options for patients. The proposed approach is expected to play a crucial role in the development of computer-aided diagnosis systems for liver cancer diagnosis.

REFERENCES

1. S. K. Randhawa, R. K. Sunkaria and A. K. Bedi, "Prediction of Liver Cirrhosis Using Weighted Fisher Discriminant Ratio Algorithm," 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), 2018, pp. 184-187, doi: 10.1109/ICSCCC.2018.8703204.
2. S. Z. Goldar, A. RikhtegarGhiasi, M. A. Badamchizadeh and M. khoshbaten, "An ANFIS-PSO Algorithm for Predicting Four Grades of Non-Alcoholic Fatty Liver Disease," 2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), 2020, pp. 1-5, doi: 10.1109/HORA49412.2020.9152881.
3. Miyazakiy, M. Ohsaki, E. Taniguchi, S. Katagiri, H. Yokoi and K. Takabayashi, "Feature extraction for the prediction of liver fibrosis stages in chronic hepatitis C," TENCON 2012 IEEE Region 10 Conference, 2012, pp. 1-5, doi: 10.1109/TENCON.2012.6412222.
4. S. Singh, S. Sinha and S. Rawat, "Extraction of influential markers affecting HCC survival chances and its prediction," 2019 International Conference on Computer Communication and Informatics (ICCCI), 2019, pp. 1-6, doi: 10.1109/ICCCI.2019.8821832.
5. J. Gu et al., "Multi-Phase Cross-modal Learning for Noninvasive Gene Mutation Prediction in Hepatocellular Carcinoma," 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2020, pp. 5814-5817, doi: 10.1109/EMBC44109.2020.9176677.
6. Dora, L., Agrawal, S., Panda, R., & Abraham, A. (2017), “Optimal breast cancer classification using Gauss–Newton representation based algorithm”, Expert Systems with Applications, 85(1), 134-145.

7. Wei J, Jiang H, Gu D, Niu M, Fu F, Han Y, Song B, Tian J. Radiomics in liver diseases: Current progress and future opportunities. Liver International. 2020 Sep;40(9):2050-63.
8. M. Hassoon, M. S. Kouhi, M. Zomorodi-Moghadam and M. Abdar, "Rule Optimization of Boosted C5.0 Classification Using Genetic Algorithm for Liver disease Prediction," 2017 International Conference on Computer and Applications (ICCA), 2017, pp. 299-305, doi: 10.1109/COMAPP.2017.8079783.
9. C. -A. B. Lee and C. H. Lee, "Cancer Screening Using Multi-modal Differential Principal Orthogonal Decomposition," 2013 13th International Conference on Computational Science and Its Applications, 2013, pp. 39-45, doi: 10.1109/ICCSA.2013.16.
10. C. A. Prajith, A. S. Kumar and H. Kareem, "Supervised classification and prediction of fibrosis seriousness using ultrasonic images," 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 2016, pp. 1-4, doi: 10.1109/ICCIC.2016.7919541.
11. K. Prakash and S. Saradha, "A Deep Learning Approach for Classification and Prediction of Cirrhosis Liver: Non Alcoholic Fatty Liver Disease (NAFLD)," 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), 2022, pp. 1277-1284, doi: 10.1109/ICOEI53556.2022.9777239.
12. J. Lara, Y. Khudyakov, F. Xavier López-Labrador, F. Gonzalez Candelas and M. Berenguer, "Hepatitis C virus genetic association to rate of liver fibrosis progression," 2013 IEEE 3rd International Conference on Computational Advances in Bio and medical Sciences (ICCABS), 2013, pp. 1-1, doi: 10.1109/ICCABS.2013.6629225.
13. S. K. Randhawa, R. K. Sunkaria and A. K. Bedi, "Prediction of Liver Cirrhosis Using Weighted Fisher Discriminant Ratio Algorithm," 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), 2018, pp. 184-187, doi: 10.1109/ICSCCC.2018.8703204.
14. P. Zhao and S.c.H. Hoi, "Bduol: Double Updating Online Learning on a Fixed Budget," Proc. European Conf Machine Learning and Knowledge Discovery in Databases (ECMLIPKDD '12), no. I, pp. 810-826, 2012.
15. T. A, C. K, S. J and N. R, "Predictive Analysis for Hepatitis and Cirrhosis Liver Disease using Machine Learning Algorithms," 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC), 2022, pp. 873-877, doi: 10.1109/ICESC54411.2022.9885411.
16. S. Z. Goldar, A. RikhtegarGhiasi, M. A. Badamchizadeh and M. khoshbaten, "An ANFIS-PSO Algorithm for Predicting Four Grades of Non-Alcoholic Fatty Liver Disease," 2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), 2020, pp. 1-5, doi: 10.1109/HORA49412.2020.9152881.
17. S. Bharathi, A. Balaji, D. Irene. J, C. Kalaivanan and R. Anusuya, "An Efficient Liver Disease Prediction based on Deep Convolutional Neural Network using Biopsy Images," 2022 3rd International Conference on Smart Electronics and Communication (ICOSEC), 2022, pp. 1141-1147, doi: 10.1109/ICOSEC54921.2022.9951870.
18. M. F. Rabbi, S. M. Mahedy Hasan, A. I. Champa, M. AsifZaman and M. K. Hasan, "Prediction of Liver Disorders using Machine Learning Algorithms: A Comparative Study," 2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT), 2020, pp. 111-116, doi: 10.1109/ICAICT51780.2020.9333528.
19. M. A. Kuzhippallil, C. Joseph and A. Kannan, "Comparative Analysis of Machine Learning Techniques for Indian Liver Disease Patients," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020, pp. 778-782, doi: 10.1109/ICACCS48705.2020.9074368.
20. S. Gupta, U. Namdev, V. Gupta, V. Chheda and K. Bhowmick, "Data-driven Preprocessing Techniques for Early Diagnosis of Diabetes, Heart and Liver Diseases," 2021 (ICECCT), 2021, pp. 1-8, doi: 0.1109/ICECCT52121.2021.9616835.