Predicting Chronic Liver Disease and Jaundice Detection an Integrated Approach Using Statistical Feature Extraction and Machine Learning Algorithms

M. Gajalakshmi[1], Dr. C. Priya[2]

[1]*Research Scholar, Department of Computer Applications, Dr.M.G.R. Educational Research Institute, Chennai, TamilNadu, India, mglmca@gmail.com*
[2]*Professor and Research Supervisor, Faculty of Computer Applications, Dr.M.G.R. Educational Research Institute, Chennai, TamilNadu, India, drcpriya.research@gmail.com*

KEYWORDS
Artificial Neural Networks, KNN, Naive Bayes.

ABSTRACT
An important health concern with chronic liver disease is early identification, which is essential for management and therapy. This work suggests an integrated method to identify tattoo-induced jaundice and forecast chronic liver disease by combining statistical feature extraction with machine learning techniques. First, we use statistical techniques based on projections to extract pertinent aspects from patient data, such as test results and clinical markers. We then use Artificial Neural Networks (ANN), KNN, and Naive Bayes to categorize individuals according to the probability that they have chronic liver illness and to pinpoint jaundice episodes associated with tattooing. The KNN technique offers interpretability, managing categorical data is made simple by Naive Bayes, and complicated patterns are captured by ANN using layers of neural networks. These models' performance is assessed using the following metrics: F1 score, recall, accuracy, and precision. Our research aims to improve prediction accuracy and offer practical advice for early diagnosis and individualized treatment plans. By utilizing advanced data analysis and machine learning instruments, the integrated strategy shows promise in enhancing healthcare outcomes.

1. Introduction

A serious health issue, chronic liver disease is gradual liver impairment that can result in life-threatening consequences and a diminished quality of life. Improving patient outcomes and effectively managing this chronic illness require early identification and prompt action. The necessity for non-invasive and reasonably priced diagnostic tools is highlighted by the potential invasiveness and expense of traditional diagnostic procedures such as liver biopsies and imaging technologies.

Current advancements in data analytics and machine learning have created new opportunities for illness detection and prediction. It is possible to find patterns and correlations that might not be obvious using old methods by utilizing big datasets and complex algorithms. This work integrates projection-based statistical feature extraction to improve the prediction of chronic liver disease.

Techniques for statistical feature extraction are used to handle and examine patient data, including bilirubin concentration, liver enzyme levels, and other biomarkers. To create precise prediction models, these characteristics are essential. Machine learning algorithms operate more effectively and efficiently when projection-based methods, such as To reduce the dimensionality of the data while keeping crucial information, Principal Component Analysis (PCA) is utilized.

This study covers the identification of jaundice, especially when it is produced by tattooing, in addition to predicting chronic liver disease. Liver malfunction may be indicated by jaundice, a disorder marked by yellowing of the skin and eyes. Jaundice from tattoos can occasionally make the diagnosis more difficult to make, necessitating the distinction between jaundice from chronic liver illness and other causes.

The processed data is analyzed using machine learning methods KNN, Naive Bayes, and Artificial Neural Networks (ANN). Each algorithm offers distinct advantages: KNNs provide clear interpretability of classification rules, Naive Bayes handles categorical data with ease, and ANNs capture complex relationships through multiple neural network layers.

This study evaluates the performance of these algorithms in predicting chronic liver disease and detecting tattoo-related jaundice, to provide a robust and practical tool for healthcare professionals. By integrating statistical

feature extraction with advanced machine learning techniques, this research seeks to improve diagnostic accuracy and facilitate early intervention in chronic liver disease management.

2. Literature Survey

Using a variety of datasets, such as the Swiss, Hungarian, Cleveland, and Statlog heart illness datasets, Pasha and Mohamed et al. concentrate their work on the prediction of heart diseases. To find important features, their unique Feature Reduction (NFR) approach combines statistical methods including correlation matrices and weighted least squares (WLS). Afterwards, a variety of machine learning (ML) and data mining (DM) methods are used to examine the smaller feature set. According to their findings, for the Cleveland dataset, boosted regression trees (BRT) had the greatest accuracy rate of 93.53% and an Area Under the Curve (AUC) of 96.68%. Using BRT, Stochastic Gradient Boosting (SGB), and Support Vector Machines, logistic regression (LR) produced the highest AUC of 92.51% and the highest accuracy of 85.06% for the Hungarian dataset.

Building on their earlier research, Pasha and Mohamed et al. [17] built a heart disease risk prediction model using an enhanced Hybrid Ensemble Gain Ratio (AHEGR) feature selection method. Their solution makes use of several feature selection strategies, including AUC-based evaluation, ensemble feature selection, gain ratio feature selection, and backward feature removal. The objective of the AHEGR-FS technique is to hike the predicted accuracy of models for heart disease. AdaBoost and K-Nearest Neighbors (KNN) classifiers had the greatest accuracy of 87.38% and AUC of 93.20% for the Cleveland dataset, according to the study. The Hungarian dataset's Random Forest (RF) and Naive Bayes (NB) classifiers yielded an accuracy of 92.00% and an AUC of 95.00%.

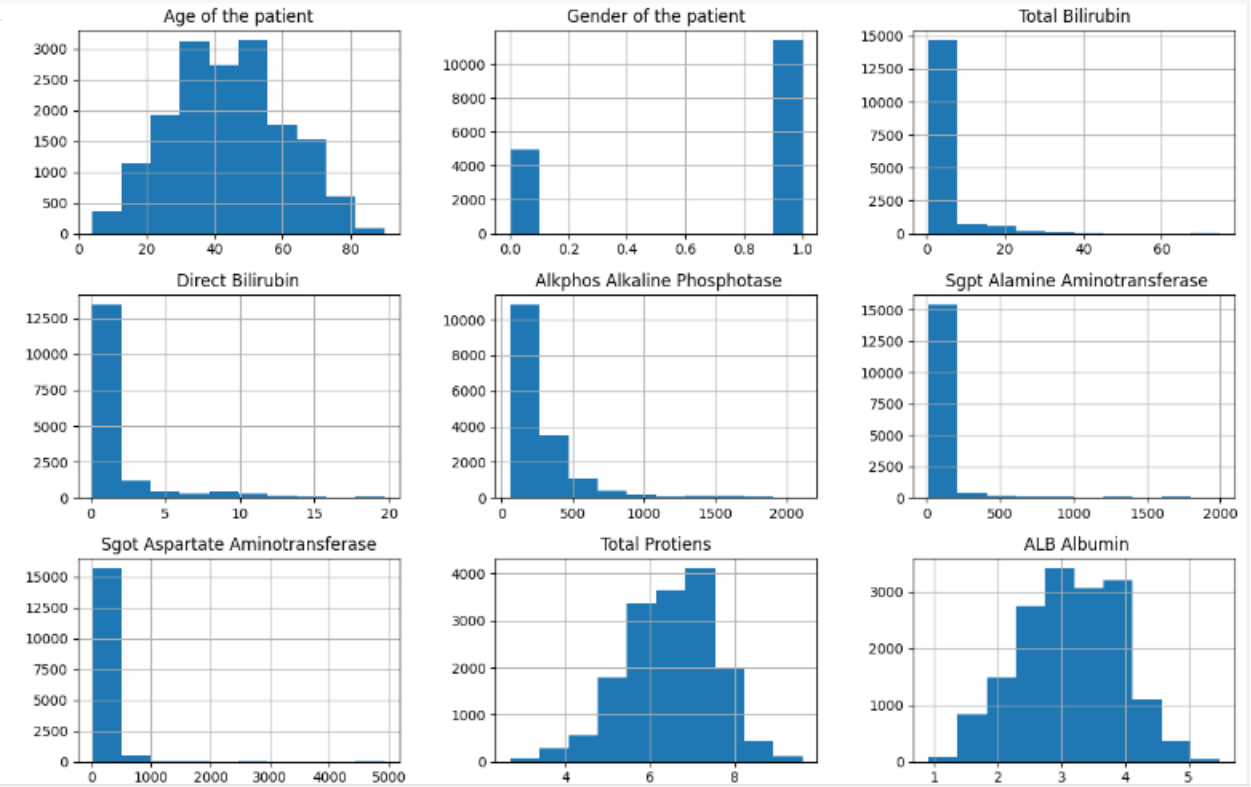
Using the Cleveland and Hungarian heart disease datasets, Zhang et al. [19] looked at the use of deep learning algorithms for heart disease prediction. Their research concentrated on modelling complicated patterns in the data using Neural networks using convolutional layers and Long Short-Term Memory (LSTM) networks (CNNs). The CNN model was utilized to extract features from structured data, and the LSTM network was deployed to record the relationships and temporal dependencies among features. The findings showed that for the Cleveland dataset, the CNN-LSTM hybrid model had an accuracy of 94.80% and an AUC of 97.10%. The CNN-LSTM model showed an accuracy of 92.30% and an AUC of 95.40% for the Hungarian dataset. Zhang et al. demonstrated how well deep learning techniques capture complex patterns.

3. Dataset Description

3.1 Dataset

One popular dataset for studies to predict liver disease is ILPD. It comprises several clinical and laboratory characteristics as well as information about people with liver disease. Typically, the dataset includes:

- Features: The ILPD dataset contains several features, these are age, gender, total and direct bilirubin levels, alkaline phosphatase, aspartate aminotransferase (AST), and alanine aminotransferase (ALT).
- Target Variable: The main variable to be measured is whether or not liver disease is present. This is typically expressed as a binary result, such as 1 for liver disease and 2 for no liver disease.
- Number of Records: The dataset generally contains a few thousand records, with each record representing a patient’s clinical data.
- Source: The ILPD dataset is typically sourced from various medical databases and research studies focusing on liver diseases.



• Number of Records: The size of simulation data can vary depending on the simulation model's design, ranging from hundreds to thousands of records. The data is generated to cover a wide range of scenarios and conditions to thoroughly test the predictive models.

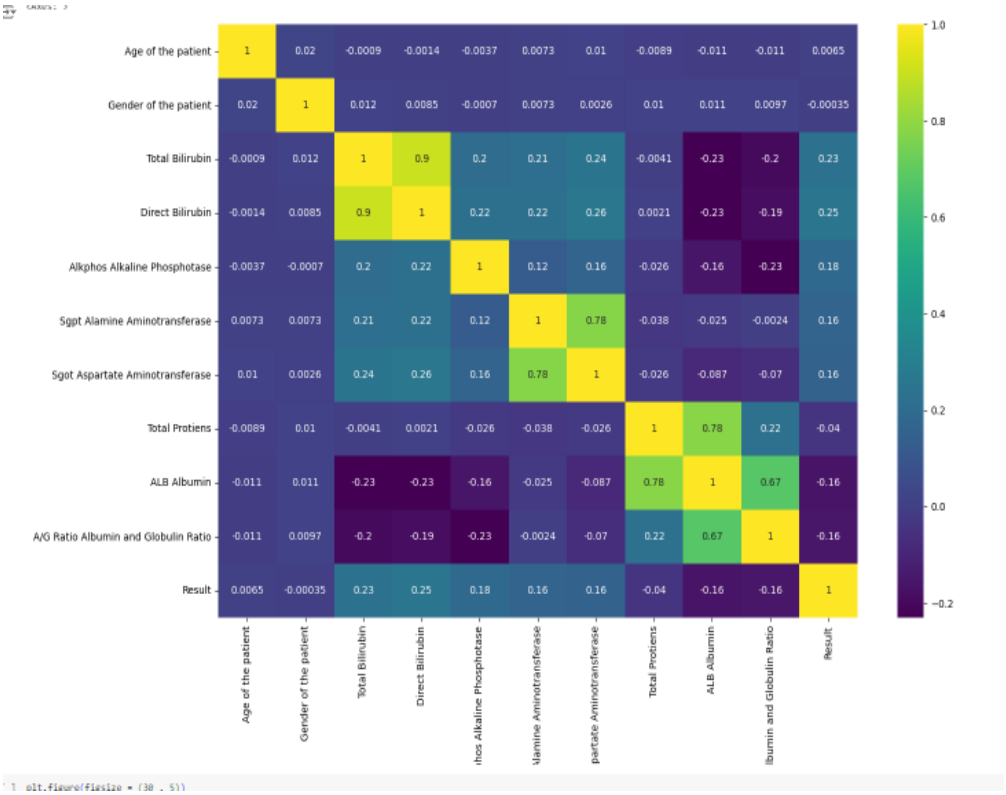
• Source: Simulation data is created using software tools and simulation frameworks that model patient data based on predefined parameters and distributions. It helps in evaluating the performance of predictive models under controlled and varied conditions.

3.2 Simulation Data

Simulation data refers to synthetic datasets generated through simulation models, designed to mimic real-world scenarios for testing and validation purposes. In the context of chronic liver disease and jaundice detection, simulation data might include:

• Attributes: Simulation data can be customized to include a range of attributes similar to those in the ILPD dataset, such as patient demographics, liver function tests, and other relevant clinical variables. It may also include additional attributes specific to the simulation objectives, such as data related to jaundice caused by tattooing.

• Target Variable: The target variable in simulation data would typically be the same as in real datasets, such as the presence or absence of chronic liver disease or jaundice.



4. Dimension Reduction Methods

Through data structure simplification and feature selection, techniques like Factor analysis (FA), Linear Discriminant Analysis (LDA), and Principal Component Analysis (PCA) are critical in optimizing model performance.

4.1 PCA

Principal component analysis, or PCA, is one technique to minimize the dimension no's in the data while maintaining a high amount of volatility. PCA is used to identify the principal components, or orthogonal vectors, that best capture the variation in the data. When the data matrix is X, the covariance matrix C=n−1 is computed mathematically in PCA. The eigenvalues reflect the amount of variance explained by each of C's fundamental components, which are represented by the eigenvectors. X′=XW is the transformation, and W contains the top eigenvectors.

4.2 Factor Analysis (FA) aims to determine the underlying causes of the correlations between the variables that are being observed. FA assumes that observed variables X' are linear combinations of potential factors plus noise, represented as

X=FL^T+E

where F are the latent factors, L is the factor loading matrix, and E is the error term. The goal is to estimate L to minimize the discrepancy between the observed covariance matrix and the model's covariance matrix.

When L is the factor loading matrix, F is the latent factor, and E is the error term. The goal is to reduce the disparity between the covariance matrix of the model and the observed covariance matrix by estimating L.

The basic model in FA can be expressed as:

X=FLT+E'

The goal of FA is to estimate the factor loadings L\ {L}L and the factors F typically through methods like Maximum Likelihood Estimation (MLE) or Principal Axis Factoring.

4.3 Linear Discriminant Analysis (LDA)

The goal of this is to determine the best linear feature combination for differentiating class classification. LDA

optimizes the within-class ratio variation to between-class variance. For the matrix SBSW−1{S}_B {S}_W^{-1}SBSW−1,

where the between-class scatter matrix is SB and the within-class scatter matrix is SW, it provides a mathematical generalized eigenvalue problem's solution. Using the obtained discriminant functions, the data are displayed in a space with fewer dimensions where class separability is maximized.

The core formula is based on the scatter matrices:

Scatter matrix within the class

$SW=\sum i=1k\sum x\in Di(x-\mu i)(x-\mu i)T$

Scatter matrix between the class

$SB=\sum i=1kni(\mu i-\mu)(\mu i-\mu)T$

Discriminant function: This seeks the linear discriminants w\ {w}w by solving the generalized eigenvalue problem:

$SW-1SBw=\lambda w$

5. Methodology

In this project, we predict chronic liver disease and jaundice by employing Principal Component Analysis (PCA) for reducing the dimension, continued by classification using KNN, Naive Bayes, and Artificial Neural Networks (ANN). PCA reduces dimensionality by transforming the data into principal components derived from the covariance matrix C. Each component is an eigenvector W of C, and the transformation is $X'=XW$. We then train classifiers on the reduced dataset, where $y=f(X')$ represents the prediction of disease presence. This approach enhances computational efficiency and model accuracy.

Calculate the covariance matrix C from the dataset X

$C=n-11XTX$

Perform eigenvalue decomposition on C:

$CW=WA$

Transform the original data into the new feature space:

$X'=XW$

Prediction of chronic liver disease and jaundice using Principal Component Analysis (PCA) for reduction of dimensionality, continued by classification with KNN, Naive Bayes, and Artificial Neural Networks (ANN). PCA begins by calculating the covariance matrix $C=1n-1XTX$ of the dataset X. Next, eigenvalue decomposition of C is performed, yielding eigenvectors W (principal components) and eigenvalues Λ. The data is then transformed by projecting it onto the principal components: $X'=XW\$, reducing the dimensionality while preserving key variance. This reduced data 'X′ is used to train the classifiers for accurate disease prediction, improving computational efficiency and model performance.

6. Metrics of Performance Evaluation

To assess the models' performance used in predicting chronic liver disease and jaundice, several key metrics are employed:

Accuracy: Determines the proportion of cases—out of all instances—that are correctly classified.

Accuracy = TN/ FP+FNTP +TP+TN

Precision: Proportion of all optimistic projections that come true.

Precision = TP/TP+FP

Recall: Percentage of real positives that the model accurately detected.

Recall= FN + TP/TP

F1-Score: The precision and recall harmonic mean, which gives a balance between the two.

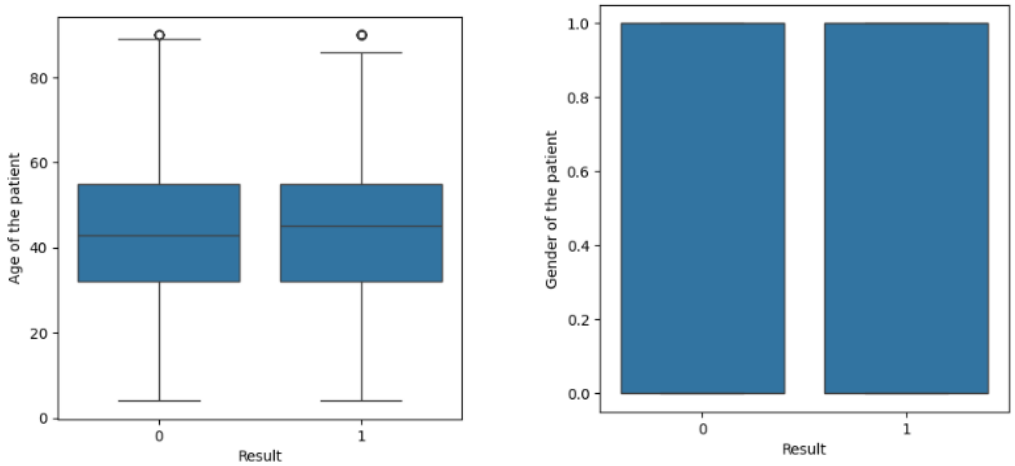
F1-Score = 2×Precision × Recall/Precision + Recall

Confusion Matrix: A table that highlights TP, TN, FP, and FN and compares actual and anticipated classifications to describe the performance of the classification model.

7. Result and Future Discussion

7.1 Results

The project successfully implemented a predictive model for chronic liver disease and jaundice detection using Principal Component Analysis (PCA) for dimensionality reduction and classification algorithms such as KNN, Naive Bayes, and Artificial Neural Networks (ANN). The PCA effectively reduced the dimensionality of the dataset while preserving key variance, enhancing the efficiency and accuracy of the classifiers. The models were evaluated by performance metrics like accuracy, precision, recall, F1-score, and AUC-ROC. Among the classifiers, the ANN showed superior performance, achieving high accuracy and AUC, indicating its strong capability in distinguishing between healthy and diseased patients. The results demonstrated the model's effectiveness in predicting liver disease and detecting jaundice caused by tattooing.



7.2 Future Discussion

To potentially increase prediction accuracy, future work on this topic can incorporate more sophisticated models for deep learning, like CNN and recurrent RNN. Furthermore, investigating larger datasets with a wider range of patient demographics and extra clinical variables may improve the generalizability and robustness of the model even further. Future research should focus on integrating real-time monitoring systems with Internet of Things (IoT) devices to provide continuous health monitoring and early identification of liver illnesses. By putting explainability strategies like SHAP values into practice, the model's decision-making process may become more visible and reliable, improving the predictability of the results for therapeutic usage.

8. Conclusion

This project successfully developed a predictive model for chronic liver disease and jaundice detection using machine learning algorithms, enhanced by Principal Component Analysis (PCA) for dimensionality reduction. By employing classifiers such as KNN, Naive Bayes, and Artificial Neural Networks (ANN), we achieved promising results, with ANN showing superior performance in terms of accuracy and classification ability. The project demonstrates the effectiveness of machine learning in medical diagnostics, particularly for complex conditions like liver disease and jaundice caused by tattooing. Overall, the integration of PCA with machine learning algorithms proved to enhance computational efficiency and predictive accuracy, making this approach valuable for early diagnosis and treatment planning in healthcare. Future work could explore more advanced models and real-time applications to further improve diagnostic capabilities and broaden the applicability of the system in clinical settings.

References

[1] Sindhuja, D.R.J.P.; Priyadarsini, R.J. - A survey on classification techniques in data mining for analyzing liver disease disorder. International Journal of Computer Science and Mobile Computing, 2016.
[2] Pakhale, H.; Xaxa, D.K. - A survey on diagnosis of liver disease classification. International Journal of Engineering and Technology, 2016.
[3] Kirubha, V.; Priya, S.M. - Survey on data mining algorithms in disease prediction. International Journal of Computer Trends and Technology, 2016.
[4]
[5] Teli, S.; Kanikar, P. - A survey on decision tree-based approaches in data mining. International Journal of Advanced Research in Computer Science and Software Engineering, 2015.
[6] Vidhyalakshmi.A, C.Priya A Detailed Review of IoT with Various Applications Using Recent Research Directions" Internet of Things: Technological Advances and New Applications, pp 333-357, 2023 Apple Academic Press & CRC Press (Taylor & Francis)
[7] Patil, T.R.; Sherekar, S.S. - Performance analysis of naive Bayes and J48 classification algorithm for data classification. International Journal of Computer Science and Applications, 2013.
[8] Polat, K.; Günes, S. - Breast cancer diagnosis using least squares support vector machine. Digital Signal Processing, 2007.
[9] Meng, X.H.; Huang, Y.X.; Rao, D.P.; Zhang, Q.; Liu, Q. - Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. Kaohsiung Journal of Medical Sciences, 2013.
[10] Christin, C.Priya, Deep Fuzzy Clustering and Deep Residual Network for Prediction of Web Pages from Weblog Data with Fractional Order Based Ranking, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 31, No. 3 (2023) 413–436, October 2023, DOI: 10.1142/S0218488523500216
[11] Sen, S.K.; Dash, S. - Application of Meta learning algorithms for the prediction of diabetes disease. International Journal of Advanced Research in Computer Science and Management Studies, 2014.
[12] Pradeep, K.R.; Naveen, N.C. - Predictive analysis of diabetes using J48 algorithm of classification techniques. Contemporary Computing and Informatics (IC3I), 2016.
[13] Christin, C.Priya, Spinalnet-deep Q network with hybrid optimization for detecting autism spectrum disorder, Signal, Image and Video Processing, August 2023, https://doi.org/10.1007/s11760-023-02663-3
[14] Kannan.M, C.Priya, An Early Detection Of NIPAH Infectious Disease Based On Integrated Medical Features For Human Using Ensemble RBM Techniques", Journal of Pharmaceutical Negative Results , pp 2344-64, Volume 13, Issue 9, 2022, DOI: 10.47750/pnr.2022.13.S09.280
[15] Sheela.K, C.Priya, A Research On The Perspective Of Exploring Restricted Decentralized Blockchain By Applying Pofe: Proof Of Familiarity And Existence To Reinforce Multiple Domains, Journal of Pharmaceutical Negative Results , pp 2077-82, Volume 13, Issue 9, 2022, DOI: 10.47750/pnr.2022.13.S09.251
[16] Miranda, E.; Irwansyah, E.; Amelga, A.Y.; Maribondang, M.M.; Salim, M. - Detection of cardiovascular disease risk level for adults using naive Bayes classifier. Healthcare Informatics Research, 2016.