IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

PROACTIVE HEALTHCARE MACHINE LEARNING IN JAUNDICE RISK PREDICTION

[1]DURGA K, [2]KANNIKANTI CHETAN KUMAR, [3]RAHUL PORIA, [4]V.GOWRI

[1]Student, [2]Student, [3]Student, [4]Assistant Professor
[1]Computer Science and Engineering,
[1]SRM Institute of Science and Technology, Chennai, India

Abstract : When predicting the yellow addiction phase using monitored machine learning, a model is developed in which patients are based on different stages of yellow skin based on the relevant features extracted from medical data. Yellow und is a disease characterized by yellow colouring of the skin and eyes due to increased levels of bilirubin in the blood, and can be reached at various stages. By using machine learning monitored techniques such as classification algorithms, this study aims to accurately predict the stage of yellowing in a patient using input features such as bilirubin levels, liver function tests, patient sensory statistics. The summary of this study title describes the importance of predicting yellow poisoning stages for timely interventions and treatment planning, how predictive models are developed, and the potential impact of healthcare providers on patient care optimization and management This study focuses on using monitored machine learning to classify patients at different stages based on clinical and diagnostic data. Yellowish is usually identified by the skin and yellow and eyes due to increased bilirubin levels, altering the severity and requiring a timely diagnosis for effective treatment. The purpose of this study is to create predictive models using algorithms such as decision-making, tree support, vector machine support, or logistics regression to analyse key features such as liver enzyme mirror, bilirubin concentration, patient age, and medical background. By accurately identifying the stages , this model can support clinicians with the fact that it can support diagnostic efficiency and enabling more targeted treatment strategies. Research highlights the importance of machine learning to improve diagnostic accuracy in hepatics and toned patient care.

IndexTerms - Proactive healthcare, machine learning, junction risk prediction, predictive analytics, healthcare, early diagnosis, clinical decision support, risk assessment, medical data mining, health informatics, patient monitoring, preventive medicine, monitored learning, distinctive selection

I.INTRODUCTION

Jaundice is a major clinically significant condition characterized by yellow discoloration of the skin, sclera, and mucosa due to elevated concentrations of bilirubin in the bloodstream. It reflects a wide range of underlying pathologies, including liver dysfunction, haemolytic disorders, and biliary tract obstruction, and remains a critical public health issue, particularly in regions with limited access to diagnostic tools and healthcare infrastructure. Timely detection, accurate staging, and proper management of jaundice are essential to prevent severe complications such as liver failure, chronic liver disease, and systemic infections, which contribute to significant morbidity and mortality. The advancement of computer technology, particularly machine learning (ML), has introduced new opportunities in medical diagnosis and disease prediction. With its ability to process large, complex clinical datasets, recognize hidden patterns, and construct predictive models, ML shows considerable promise in enhancing traditional diagnostic pathways and supporting proactive healthcare delivery. Among ML methodologies, supervised learning algorithms, which are trained on labelled datasets to predict outcomes for unseen data, have demonstrated efficacy in disease classification and risk prediction tasks. This study aims to leverage supervised machine learning techniques to predict the stages and risk levels associated with jaundice by utilizing clinical features such as laboratory test results, patient demographics, symptomatic presentations, and historical medical records. Robust feature selection techniques and model optimization are implemented to improve predictive accuracy and interpretability. The goal is to assist healthcare professionals in making timely and informed clinical decisions, thereby improving patient outcomes, reducing healthcare costs, and advancing the integration of artificial intelligence into routine clinical practice. Early diagnosis and intervention, supported by computational models, are critical steps in mitigating the consequences of jaundice and improving the quality of care. The success of machine learning applications in jaundice risk prediction not only underscores the potential of AI to transform healthcare delivery but also highlights the increasing importance of data-driven approaches in addressing complex medical challenges in the evolving global health landscape.

II. NEED OF THE STUDY

Yellowish appearance remains a critical health problem around the world, especially in environments where diagnostic skills and access to timely health care are often limited. The disease manifests itself by increased bilirubin levels and yellow color of the skin and eyes, and can be attributed to a variety and serious underlying causes such as liver disease, hemolytic anemia, and bile disorders. If this is diagnosed immediately

and not managed, yellowish can lead to life-threatening complications such as liver failure and systemic infections. Traditional diagnostic methods often rely on the knowledge of clinical experts and may not be easily accessible in all regions. In particular, the monitored learning algorithms show significant possibilities for identifying disease patterns and predicting health outcomes using patient data. These techniques should be used to develop predictive models, support clinical decisions, optimize treatment strategies, and promote early intervention. This study addresses this need by using monitored methods for machine learning to predict the level and severity of yellowing based on relevant clinical characteristics. Integrating such a data-controlled approach into health practices could improve patient outcomes, reduce diagnostic delays, and lead to efficient

.

2.1Population and Sample

The population of this study includes people whose yellowing is diagnosed or suspected, including cases of various stages and causes. Samples were from an anonymized clinical data set containing detailed patient information such as age, gender, symptoms, clinical laboratory results (such as bilirubin levels, liver enzymes) and associated medical history. Patients were locked up based on the completeness and accuracy of the records, particularly the results needed for monitored machine learning. Data records with missing data, double entries, or unrelated liver conditions are excluded. This curated sample was used to train and evaluate predictive models to assess the level and severity of yellowing.

2.2 Data and Sources of Data

Data from this study were obtained from anonymized clinical records and published medical data records containing information about yellowed patients. These data records include essential characteristics such as patient demographics, symptom representation, clinical tests (including bilirubin level and liver function parameters), and stages or outcomes of the yellowing. Sources can include electronic hospital records (ELHRS), state health databases, or open access platforms such as Kaggle or UCI repository for machine learning. Data is processed to ensure the integrity, consistency, and relevance of learning goals and serve as the basis for training and validating machine learning models.

2.3 Theoretical framework

This study is based on the theoretical foundations of monitored machine learning, and focuses on classification techniques to predict the level of yellowing and risk. Selected algorithms Naive Bayes, Support Vector Machines (SVMs), and Random Forest are rooted in different statistical and arithmetic learning theories. Naive Bayes works according to the principle of probabilistic classification using Bayes' theorem, which assumes distinctive independence. SVM is based on the concept of finding the best separation of optimal hyper levels in high-dimensional space. An ensemble learning method, Random Forest sets outcomes and aggregates results to build up several decisions, improve prediction accuracy, and reduce over adaptation. These algorithms are applied to clinical data such as laboratory values, symptoms, and demographic information to identify patterns associated with the Junge addiction phase. By using these models, the framework supports reliable risk prediction and supports members of health occupations with timely data-based decisions.

III. RESEARCH METHODOLOGY

The proposed study uses machine-based methodology to design an intelligent and predictive framework for early detection of yellowing and risk assessment. The research process is systematically organized at several key stages beginning with data collection, in which a comprehensive dataset of clinical sources such as Kaggle is compiled. This data record includes important clinical variables such as bilirubin level, LEF results (live liver function test), patient demographics, medical history, and clinically validated yellowed classification. This includes handling missing values, normalizing data distribution, removing outliers, and converting categorical data into a machine-readable format. This ensures the quality and consistency of the data fed into the machine learning model. The elimination of redundant or ineffective variables makes the model more efficient and interpretable, and at the same time with high prediction accuracy. This step is very important for reducing computational compensation and improving model transparency. This algorithm specializes in outlet patterns that can exhibit serious or abnormal symptoms of yellow, allowing early identification of high-risk patients that traditional classifiers may overlook. In particular, scenarios with high dimensional data contribute significantly to overall classification performance. Generalization and resistance to excessive adaptation. Various power metrics are calculated, including accuracy, accuracy, recall, F1 score, and AUC-ROC. Tools such as the confusion matrix and recall curves provide deeper insight into the predictive behavior of all models and their ability to distinguish between yellowing and different levels of risk. This interface ensures that the system not only acts as an accurate diagnostic support tool, but also becomes accessible and usable in real health environments.

3.1 Statistical tools and econometric models

- **Complement Naive Bayes (CNB):** A probabilistic classification algorithm particularly suited for imbalanced datasets. It applies Bayes' theorem under the assumption of feature independence and is effective in handling skewed data distributions, making it ideal for early-stage jaundice detection where certain risk categories may be underrepresented.
- **One-Class Support Vector Machine (SVM):** An unsupervised anomaly detection model used to identify rare or extreme cases by learning the boundary of normal data. It is employed in this study to detect outliers that may signify severe or atypical forms of jaundice progression.
- **Random Forest Classifier:** An ensemble learning method based on constructing multiple decision trees. It is utilized for its high accuracy, robustness, and ability to manage both linear and non-linear patterns in data. Additionally, it offers insights into feature importance, which enhances the interpretability of the model.
- **Cross-Validation (K-Fold):** A statistical method used to evaluate model performance by dividing the dataset into 'k' subsets. Each subset is used as a test set while the remaining subsets are used for training, helping to prevent overfitting and ensure the generalizability of results.
- **Confusion Matrix:** A tool for evaluating classification performance by comparing predicted labels against actual labels. It provides detailed insights into true positives, false positives, true negatives, and false negatives.

- **Receiver Operating Characteristic (ROC) Curve and Area Under the Curve (AUC):** These are used to measure and visualize the trade-off between sensitivity and specificity across various thresholds, offering a graphical interpretation of model discrimination ability.
- **Recall Curve:** This curve tracks the sensitivity of models across different classification thresholds, highlighting the model's ability to correctly identify positive cases (i.e., actual jaundice risks).

3.1.1 Descriptive Statistics

In this research, several machine learning models were employed to predict the risk and stages of jaundice. Three primary methodologies are explored in this section: **Naive Bayes**, **Support Vector Mchine** and **Random Forest**. These models aim to predict jaundice risk based on various clinical features, including bilirubin levels and liver function tests.

3.1.2 Naive Bayes Model

Naive Bayes is a stochastic classification algorithm based on the Bayes theorem, which accepts independence under the properties of class labels. This is particularly suitable for medical diagnostic tasks where functions such as liver function markers and patient history are used.

This study uses the Naive Bayes classification algorithm to predict jaundice risk predictions based on clinical data. Naive Bayes is a stochastic machine learning model based on the Bayes theorem with basic assumptions of conditional independence between properties. Despite this simplification, it has been proven to be highly effective in medical applications where symptoms and diagnostic marker patterns are consistent with stochastic thinking.

The data records used included patient files with important attributes such as serum image mirrors, liver enzyme measurements, and symptomatic indicators such as age, sex, fatigue, skin discoloration, and anorexia. Prior to model training, data were subjected to processing procedures that included handling missing values, coding categorical variables using label coding, and normalization of continuous features to ensure uniformity in scaling. The cleaned data records were then split into training and test rates in an 80:20 split.

Given the continuity of the characteristics, the Gaussian Naive Bayes classifier was selected. This model was trained with training data and evaluated on test data using standard performance indicators. The naive Bayes model achieved a general accuracy of **89.04%**. This demonstrates a powerful ability to identify people too correctly at the risk of yellowing. The high accuracy combined with the simplicity and efficiency of the algorithm supports use in real-time environments where early diagnosis is extremely important, with low resources.

3.1.3 Support Vector Machine (SVM)

A support vector machine (SVM) classifier was implemented to assess its effectiveness in predicting yellow risk based on clinical parameters. SVM is a monitored learning algorithm that creates the best hyperplanes for high-dimensional spaces for separate instances of different classes. This is particularly effective for binary classification tasks, and works well with complex, nonlinear relationships between properties.

For SVM classifiers, the same pre-machined data records were used. This was used in the naive Bayesian model to ensure a consistent evaluation framework. Input functions included clinical indicators such as serum bilirubin levels, liver function tests, age, gender, observable symptoms such as yellowing and fatigue discoloration. Characteristic normalization was used to scale the data evenly to improve model convergence.

The SVM model used the kernel for the radial-based functions (RBF) to process nonlinear patterns in the data. A split of the train test at 80:20 was used and hyperparameters were set in cross-validation to improve generalization. In the evaluation, the SVM classifier achieved a total accuracy of **88.67%,** indicating strong predictive capabilities. Although the SVM was slightly lower than the naive Bayes model, it showed competitive performance and enhanced the applicability of machine learning models in aggressive health systems for early detection of yellow.

3.1.4 Random Forest

Random forest algorithm was also applied to the jaundice risk prediction task to assess performance against other classifiers. Random Forest is an ensemble learning technology that builds a collection of decisions during training and issues its prediction type. This approach reduces excessive adaptation or overfitting and increases robustness by combining some weak learners with powerful learners.

The same data records used in previous models were used for characteristics such as bilirubin mirror, liver adaptation, age, gender, and symptom-related indicators. The preprocessing procedure included handling missing values, coding categorical features, and normalization of numbers to maintain consistency across all models implemented.

This model was trained in a split of 80:20 train tests. The number of trees and maximum depth were set by cross-validation to ensure optimal performance. When evaluating, the random forest classifier achieves **100%** accuracy and indicates the complete classification of the test set. This result suggests that the model was recorded very effectively when recording characteristic patterns related to yellow. However, further testing on external or large datasets is recommended to validate the generalizability of the model and eliminate adaptive capabilities.

3.1.5 Comparison of the Models

Algorithm	Metric	Previous Study (%)	Current Study (%)
Naive Bayes	Accuracy	85.20	89.04
	Precision	82.50	87.80
	Recall	83.00	88.50
	F1-Score	82.75	88.14
SVM	Accuracy	84.75	88.67
	Precision	83.00	86.90
	Recall	84.00	87.60
	F1-Score	83.50	87.24
Random Forest	Accuracy	95.50	100.00
	Precision	94.20	100.00
	Recall	95.00	100.00
	F1-Score	94.60	100.00

Table 3.1 Comparison of previous and current accuracy level obtained

The Naive Bayes classifier achieved an accuracy of 89.04%. This demonstrates strong performance in addressing the stochastic relationships between clinical features. Its simplicity and computational efficiency make it suitable for real-time health applications, especially in low resource settings. Effectively learn nonlinear patterns within data records and provide competitive results. The nature of SVM requires more voice and arithmetic resources compared to the naive Bayes, but makes it robust to surpass in higher-dimensional rooms.

The Random Forest Classifier has significantly surpassed other models and achieved 100% accuracy on the test set. This result suggests that ensemble learning approaches in recording complex interactions between characteristics are highly effective. However, such perfect accuracy indicates the possibility of excessive adaptation, as it must be further evaluated using external validation data records to confirm generalizability.

IV. RESULTS AND DISCUSSION

4.1 Results

The following screenshots illustrate the functional interface and output results of the developed jaundice risk prediction website. These images demonstrate the user input process, system-generated predictions, and the comparative performance of the implemented machine learning models.

Table 4.1: Results of the Accuracy particularly obtained

Algorithm	Accuracy (%)
Naive Bayes	89.04%
Support Vector Machine	88.67%
Random Forest	100.00%

Table 4.1 refers to the Random Forest Classifier surpassed other models by achieving full accuracy on the test dataset. This shows that ensemble learning approaches to combine several decisions are highly effective when recording complex patterns and relationships within clinical characteristics. However, such high scores indicate excessive adaptation, especially when the model is not validated with external or invisible data. To check generalization capabilities, further testing of large or independent data sets is recommended.
The naive Bayes algorithm showed strong performance with an accuracy of 89.04% and used stochastic relationships between characteristics. Its simplicity, low computing costs and simple implementation make it particularly suitable for real applications and resource-related environments such as rural clinics and mobile health platforms.

The support vector machine with 88.67% accuracy showed competitive performance. Nonlinear boundaries were modeled using the RBF kernel (radial fundamental function). However, SVMs usually require more arithmetic resources and hyperparameter adjustments, which can limit scalability in actual health preparation. Naive Bayes offers interpretability and speed, making it practical for embedded or bold health systems. SVM provides a balance between performance and model complexity. Random Forest offers a high level of accuracy, but care must be taken to ensure overadaptation and external validity.

V. ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to SRM Institute of Science and Technology, Department of Computer Science for providing the necessary support and resources to carry out this research. The authors also appreciate the support of faculty and peers who contributed to the completion of this work.

REFERENCES

[1] Nainika Saini, Ashok Kumar, Preeti Khera, "Non-Invasive Bilirubin Detection Technique for Jaundice Prediction Using Smartphones," International Journal of Computer Science and Information Security (IJCSIS), Vol. 14, No. 8, August 2016.

[2] M. Penhaker, V. Kasik, B. Hrvolova, "Advanced Bilirubin Measurement by a Photometric Method," Elektronika Ir Elektrotechnika, Vol. 19, 2014, pp. 47-50.

[3] Smriti Shrivastava, "Diagnosis of Neonatal Jaundice Using Artificial Neural Networks," International Indexed & Refereed Research Journals, Vol. 4, 2013, pp. 69-72.

[4] James W. Kronberg, "Optical Transcutaneous Bilirubin," US Patent 5 259 382, November 9, 1993.

[5] Tai-Wing Wu, Rochester, N.Y., "Assay for Bilirubin," United States Patent 4 069 016, January 17, 1978, pp. 1-32.

[6] Mark McEwen, Karen Reynolds, "Noninvasive Detection of Bilirubin Using Pulsatile Absorption."

[7] David P. Dewitt, Robert E. Hannemann, John F. Wiechel, "Method for Determining Bilirubin Concentration from Skin Reflectance," United States Patent 4 029 085, June 4, 1977, pp. 1-8.

[8] Steven L. Jacques, David G. Oelberg, Iyad Saidi, "Method and Apparatus for Optical Measurement of Bilirubin in Tissue," United States Patent 5 353 790, 1994, pp. 1-79.

[9] Gagan Mahajan, "Transcutaneous Bilirubinometer in Assessment of Neonatal Jaundice in Northern India," Indian Pediatrics, Vol. 42, 2005, pp. 41-45.

[10] G. Vreman, L. Wong, D. Stevenson, "Noninvasive Bilirubin Monitoring in Neonates," Journal of Perinatology, Vol. 24, 2004, pp. 555-565.

[11] K. Maisels, "A Comparison of Transcutaneous and Serum Bilirubin Measurements," Pediatrics Journal, Vol. 122, 2008, pp. 872-877.

[12] S. Subramanian, R. Somasundaram, "Machine Learning Approach for Jaundice Prediction in Neonates," IEEE Transactions on Biomedical Engineering, Vol. 67, 2020, pp. 1258-1267.

[13] J. Gourlay, B. Matheson, "Smartphone-Based Jaundice Detection Using AI Algorithms," Journal of Medical Informatics, Vol. 10, 2019, pp. 321-329.

[14] T. Smith, A. Johnson, "Bilirubin Detection Using Optical Sensors," IEEE Sensors Journal, Vol. 15, 2017, pp. 45-50.

[15] R. Alzahrani, J. Wong, "Deep Learning in Neonatal Hyperbilirubinemia Prediction," Artificial Intelligence in Medicine, Vol. 27, 2021, pp. 78-90.

[16] L. O'Brien, R. Dunne, "Use of Wearable Devices for Monitoring Neonatal Jaundice," Medical Engineering & Physics, Vol. 20, 2018, pp. 15-23.

[17] J. Patel, M. Anand, "Bilirubin Estimation Using Spectral Analysis," Journal of Optical Engineering, Vol. 34, 2016, pp. 1124-1131.

[18] K. Mukherjee, "Development of a Portable Bilirubin Analyzer for Rural Healthcare," International Journal of Biomedical Engineering, Vol. 14, 2020, pp. 89-97.

[19] B. Fernandez, P. Lucas, "Photonic Sensors for Neonatal Jaundice Diagnosis," Optics and Lasers in Engineering, Vol. 42, 2015, pp. 421-430.

[20] S. Li, W. Tang, "Noninvasive Spectrophotometric Detection of Jaundice in Adults," Clinical Biochemistry, Vol. 50, 2019, pp. 765-772.