International Research Journal of Education and Technology
Peer Reviewed Journal
ISSN 2581-7795

Investigation on Hepatitis With Its Severity Grading Using Machine Learning Algorithm

SRI HARI VISWANATH S
Dept. of Artifical Intelligence and Machine Learning
Bannari Amman Institute of Technology

VIJAY KUMAR S
Dept. of Artifical Intelligence and Machine Learning
Bannari Amman Institute of Technology

AMOKAA V A
Dept. of Artifical Intelligence and Machine Learning
Bannari Amman Institute of Technology

GURUPREETHA V
Dept. of Computer Science and Engineering
Bannari Amman Institute of Technology

Abstract—**Hepatitis is a liver illness that can be life-threatening if not diagnosed and treated in a timely manner. The severity of the disease must be graded with precision so as to inform the right treatment for better patient outcomes. This study examines the use of machine learning agents- K-Nearest Neighbors (KNN), Convolutional Neural Networks (CNN), and Logistic Regression (LR) – to classify patients depending on the severity of hepatitis. According to the dataset that included the clinical diagnosis, blood tests, and patient characteristics, the models acquired were used to separate the ascribed cases of hepatitis into; mild, moderate, and severe ones. K-Nearest Neighbors was used due to its ease in pattern searching on the feature space through proximity, CNN on the other hand due to its capability of finding important features in complex data without supervision, that conduct enhancement to LR which is mainly practiced for two categories. The models were assessed on the parameters of accuracy, precision, recall, and F1-score. It was found in the initial research that CNN outperformed KNN and LR in the classification of the severity of hepatitis infection with the effective use of complex information about the disease, which is a rare ability. KNN and LR also appeared to be competitive immediately after diagnosing even early lesions. This study shows the power of machine learning, and more specifically CNN, increasing the accuracy of grading hepatitis severity, providing an opportunity for physicians to enhance their diagnostic accuracy and the management of their patients.**

Keywords— Hepatitis, Severity Grading, Machine Learning, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression (LR), Medical Diagnostics, Disease Prediction, Liver Disease Classification, Clinical Features, etc.

I. INTRODUCTION

One of the biggest global health issues is hepatitis, which is characterized by inflammation of the liver. It can result from many causes such as viral infections, alcoholism, and some autoimmune diseases among many others. The progressing hepatitis shall bring a crippling condition of the liver, such as cirrhosis or cancer. Early diagnosis and correct severity grading are crucial in effective treatment. Traditional diagnosis processes are sometimes time-consuming and not very specific, and thus more attention has been given to the applications of machine learning in improving the accuracy and effectiveness of the diagnosis and decision-making on the diseases' severities.

In this study, we will be discussing the application of three most commonly used algorithms that are K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Logistic Regression (LR) in the classification of the grades of liver hepatitis. Patient data-including clinical features like a level of liver enzymes, bilirubin concentration, and patient demographics-will then be analyzed to compare the obtained models' predictive performance regarding disease severity. The outcomes of this study would reveal the best algorithm, which may, in turn enhance the diagnostic power and aid the clinicians in taking proper early decision treatment.

II. APPLICATION OF MACHINE LEARNING TO HEPATITIS DISEASE DETECTION

Detection of hepatitis disease through the application of the machine learning (ML) involves the use of algorithms to detect trends in data and make forecasts on the status or course of the disease. Hepatitis is a disease that involves the liver and can be caused by several viruses, making it important to detect the disease as early as possible. Machine-learning techniques will improve the accuracy of diagnosis through the effective handling of complex datasets including, medical records, blood test results, and imaging.

III. STEPS IN MACHINE LEARNING IN DIAGNOSING HEPATITIS

A. Data Collection:

Such machine learning models depend highly on previous patients' information such as liver enzymes, bilirubin's, demographics, and clinical records. The datasets are expected to be cleaned at this stage, with the irrelevant and missing data being removed through Data preprocessing.

B. Feature Selection:

The training model captures key factors which determine the diagnosis of hepatitis. These features might include specific biomarkers, age, gender, and other relevant factors that help the model make accurate predictions.

C. Model Training:

Among the languages SVM, Decision Trees, Random Forest and Neural networks are utilized frequently. Such models are provided with labeled data where the variable (disease: present or present) has been determined and the machine is then exposed to patterns relevant to hepatitis.

D. Installation as and Validation of the Model:

Test data is used for evaluation of the performance of the model that has been built and includes accuracy, sensitivity and specificity. The cross-validation techniques are used to ensure most of the subsets of the data population are adequately covered by the model.

E. Predicting and supporting decision making:

The physician, after some time, can use the model to categorize a new chronic hepatitis patient at risk of the infection. Thus, such predictions can be used by Clinicians as tools that assist diagnosis enhancing the speed and accuracy of the diagnosis of the patients.

I. Objectives

A. Improve Diagnostic Accuracy

- Develop and evaluate a K-Nearest Neighbors (KNN) model for grading hepatitis severity into mild, moderate, and severe categories with high precision and reliability.
- Leverage KNN's proximity-based pattern recognition to identify subtle trends in clinical features such as liver enzyme levels, bilirubin concentrations, and patient demographics.

B. To Address Limitations of Traditional Methods

- Mitigate inefficiencies in traditional diagnostic approaches, which are often time-intensive and prone to variability in severity grading.
- Introduce a data-driven, automated method to assist clinicians in making faster and more accurate severity assessments.

C. To Automate Hepatitis Severity Classification

- Utilize KNN to classify patients' hepatitis severity based on clinical datasets, enhancing the speed and consistency of diagnoses.
- Demonstrate KNN's effectiveness in handling medical data and simplifying the diagnostic process.

D. To Evaluate KNN's Performance Against Other Models

- Compare the predictive performance of KNN with other machine learning models such as CNN and Logistic Regression in terms of accuracy, precision, recall, and F1-score.
- Highlight the scenarios where KNN excels, particularly in early-stage hepatitis detection and smaller datasets.

E. To Provide Clinician Support

- Offer an interpretable machine learning model like KNN to assist healthcare professionals in decision-making, ensuring better patient outcomes through precise and timely treatment.

II. SAFETY AND COMPLIANCE

1. *Data Privacy and Patient Confidentiality* - The dataset used in this study complies with established data privacy regulations, ensuring the protection of patient information. Ethical guidelines, such as those outlined in HIPAA (Health Insurance Portability and Accountability Act) or GDPR (General Data Protection Regulation), were adhered to during data collection and processing.
2. *Ethical Considerations* - The study ensured that the data usage was approved for research purposes, avoiding any misuse of sensitive medical data.
3. *Regulatory Compliance* **-** The research aligns with the ethical standards prescribed by
4. *Safety in Application* **-** Machine learning models were designed with caution to avoid overgeneralization or misclassification, which could lead to incorrect clinical decisions.
5. *Transparency and Accountability* **-** The methodology and model performance metrics are fully documented to ensure reproducibility and accountability. Clinicians using these models are provided with clear explanations of the decision-making process, especially in the case of algorithms like KNN, to ensure interpretability.
6. *Security Measures* **-** For datasets stored or shared during the research, encryption and secure data transfer protocols were employed.
7. *Mitigation of Risks* **-** Potential risks, such as model bias or errors in severity classification, were minimized through rigorous testing and validation.

III. Hyperparameter Tuning for KNN

A. Key Hyperparameters in KNN

1. ***Number of Neighbors (k):*** *Defines the number of nearest neighbors considered for classification.*
2. ***Distance Metrics:*** *Determines how the "closeness" of neighbors is measured. Common metrics include:*
3. ***Euclidean Distance****: Straight-line distance, suitable for continuous features.*
4. ***Manhattan Distance****: Sum of absolute differences, useful for high-dimensional datasets.*
5. ***Minkowski Distance****: Generalized form that includes both Euclidean and Manhattan.*

B. Optimization Strategies

1. ***Grid Search:*** *Systematically testing combinations of kkk values and distance metrics to find the best configuration.*
2. ***Cross-Validation:*** *Splitting the dataset into multiple folds to validate performance and select optimal hyperparameters without overfitting.*
3. ***Feature Scaling:*** *Normalizing or standardizing features to ensure equal contribution to distance computations.*
4. ***Weighted KNN:*** *Assigning weights to neighbors based on their distance, giving closer neighbors higher influence on predictions.*

IV. Visualization and Insights

Effective visualization is essential for presenting and interpreting the findings of machine learning models, especially in medical diagnostics like hepatitis severity classification. Below, we describe some of the visual tools you can use to present insights from your KNN model and help in decision-making:

A. *Scatter Plots*

Scatter plots are useful for visualizing the relationship between key features in the dataset and how they correlate with different severity levels of hepatitis.

- **X-axis**: Bilirubin concentration
- **Y-axis**: ALT levels (liver enzyme)
- **Color**: Different colors for different severity classes (e.g., red for severe, orange for moderate, green for mild)

This visualization helps identify patterns and potential thresholds for severity categorization based on clinical features.

B. *Heatmaps*

Heatmaps are useful for visualizing the correlation between different features in the dataset. A heatmap helps identify how different variables (e.g., liver enzyme levels, bilirubin concentration) are related to each other and to hepatitis severity.

X-axis and Y-axis: Different features (e.g., age, bilirubin, ALT, AST)

This helps identify multicollinearity (correlation between features), which may affect the KNN model’s performance. Features that are highly correlated can be combined or one of them can be removed to improve the model's efficiency.

C. *Confusion Matrix*

A confusion matrix is one of the most valuable tools for assessing the performance of a classification model. It shows how the predicted severity categories (mild, moderate, severe) compare to the actual categories.

- **X-Axis:** Predicted Severity (Mild, Moderate, Severe)
- **Y-Axis:** Actual Severity (Mild, Moderate, Severe)
- **Diagonal Elements:** True Positives (Correct Classifications).
- **Off-Diagonal Elements:** False Positives and False Negatives.

D. *KNN Decision Boundaries*

To visualize how KNN classifies hepatitis severity based on feature values, you can plot decision boundaries that show how the model separates different classes in the feature space.

- **X-axis**: Bilirubin concentration
- **Y-axis**: ALT levels
- **Color coding**: Red for severe, orange for moderate, green for mild
- **Boundary lines**: Shows how the KNN model classifies a new instance based on the closest neighbors.

E. *ROC-AUC Curve*

The Receiver Operating Characteristic (ROC) curve illustrates the model's ability to distinguish between the different severity classes, while the Area Under the Curve (AUC) quantifies this ability.

- **X-axis**: False Positive Rate (FPR)
- **Y-axis**: True Positive Rate (TPR)
- **Multiple curves**: You can plot ROC curves for each severity class (e.g., mild vs. moderate, mild vs. severe) to see how well the model distinguishes between these categories.

V. Challenges and Limitations

A. *Challenges in Data Collection and Preprocessing*

- Missing values in key clinical features can reduce the dataset size or introduce bias during imputation.
- Disparities in the number of patients across severity grades can skew the model towards dominant classes.
- Outliers or inconsistencies in clinical measurements can affect the model’s performance.
- Large numbers of features may increase computational complexity and lead to overfitting in KNN.

B. *Limitations of KNN*

1. **Sensitivity to Noise:** KNN is highly susceptible to noisy data, which can mislead the proximity-based classification.
2. **Scalability Issues:** The computational cost increases significantly with large datasets due to the need to compute distances for all points.
3. **Class Overlap:** In cases where severity grades overlap in feature space, KNN may struggle to differentiate between classes effectively.
4. **Feature Dependence:** KNN's performance is highly dependent on feature scaling and selection. Poorly scaled features can distort distance calculations.

VI. Future Scope for Addressing Challenges

1. **Advanced Preprocessing:**
   - Employ robust methods for handling missing data and noise, such as advanced imputation techniques or outlier removal algorithms.
2. **Dimensionality Reduction:**
   - Use techniques like Principal Component Analysis (PCA) or feature selection to reduce complexity and improve scalability.
3. **Hybrid Models:**
   - Combine KNN with other algorithms (e.g., ensemble methods) to mitigate its limitations while leveraging its strengths.
4. **Parallel Computation:**
   - Implement parallel processing or approximate nearest neighbor algorithms to speed up computations for large datasets.
5. **Larger and More Diverse Datasets:**
   - Incorporate more patient data from varied demographics to reduce bias and enhance generalizability.

References

[1] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

[2] World Health Organization (WHO). (2023). *Global Hepatitis Report 2023*. Available: [URL]

[3] Cover, T., & Hart, P. (1967). Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, 13(1), 21–27.

[4] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437.

[5] Jiang, F., Jiang, Y., Zhi, H., et al. (2017). Artificial intelligence in healthcare: past, present, and future. *Stroke and Vascular Neurology*, 2(4), 230–243.

[6] Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13, 281–305.

[7] Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a “right to explanation”. *AI Magazine*, 38(3), 50–57.