ISSN 2319-5991 www.ijerst.com
© 2026 IJERST. All Rights Reserved
Peer Reviewed, Referred & Indexed Journal

Research Paper

PREDICTION OF HEPATITIS DISEASE USING K-NEAREST NEIGHBORS, NAIVE BAYES, SUPPORT VECTOR MACHINE, MULTI-LAYER AND RANDOM FOREST

[1]JUTTIGA SRI SAI LEELA PRAVALLIKA, [2]K.RAJA RAJESWARI

[1]Students, Department of MCA, B V Raju College, Bhimavaram Ap

[2]Assistant Professor, Department of MCA, B V Raju College, Bhimavaram Ap

ABSTRACT

Hepatitis is a serious liver disease that can lead to severe health complications if not detected early. Accurate prediction and diagnosis of hepatitis using machine learning techniques can assist healthcare professionals in making timely decisions. This project proposes a comparative analysis of various machine learning algorithms, including K-Nearest Neighbors (KNN), Naïve Bayes, Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), and Random Forest, for predicting hepatitis disease. The dataset used in this study consists of medical attributes such as age, sex, bilirubin levels, liver enzyme values, and other clinical features. Data preprocessing techniques such as handling missing values, normalization, and feature selection are applied to improve model performance. Each algorithm is trained using 80% of the dataset and tested on the remaining 20%. The models are evaluated using performance metrics such as accuracy, precision, recall, F1-score, and confusion matrix. Experimental results show that ensemble and neural network-based models provide higher accuracy compared to traditional methods. The proposed system demonstrates the effectiveness of machine learning techniques in early hepatitis detection, improving healthcare outcomes.

Keywords: Hepatitis Prediction, Machine Learning, KNN, Naïve Bayes, SVM, Random Forest, MLP, Medical Diagnosis

I.INTRODUCTION

Hepatitis is a liver disease caused by viral infections, alcohol consumption, or autoimmune conditions. Early detection of hepatitis is crucial to prevent severe complications such as liver cirrhosis or liver cancer. Traditional diagnostic methods involve laboratory tests and medical examinations, which can be time-consuming and costly. With the advancement of machine learning, it is now possible to analyze medical data efficiently and provide accurate predictions, assisting doctors in diagnosis and treatment planning.

Machine learning algorithms have shown significant potential in healthcare applications due to their ability to analyze complex datasets and identify patterns that may not be visible through traditional methods. Algorithms such as K-Nearest Neighbors (KNN) and Naïve Bayes are simple yet effective for classification tasks, while Support Vector Machines (SVM) provide robust performance in high-dimensional data. More advanced models like Multi-Layer Perceptron (MLP) and Random Forest improve prediction accuracy by capturing non-linear relationships and combining multiple decision trees.

This project focuses on implementing and comparing multiple machine learning algorithms for hepatitis disease prediction. The system includes modules for data preprocessing, model training, evaluation, and prediction. By analyzing the performance of different models, the project identifies the most effective algorithm for accurate diagnosis. The implementation is done using Python and machine learning libraries, providing a scalable and efficient solution for medical data analysis.

II SURVEY OF RESEARCH

[1] The study by Thomas Cover and Peter Hart (1967) introduced the K-Nearest Neighbors (KNN) algorithm for pattern classification. The methodology involves classifying data points based on the majority class of their nearest neighbors. Results showed that KNN is simple and effective for classification tasks. However, it is sensitive to noise and requires high computation for large datasets. In medical diagnosis, KNN is widely used due to its simplicity. In the proposed system, KNN is used as a baseline model for hepatitis prediction.

[2] The research by Thomas Bayes (1763) introduced the foundation of the Naïve Bayes algorithm. The methodology assumes independence between features and uses probability theory for classification. Results demonstrated that Naïve Bayes performs well in many classification tasks despite its simplicity. However, the independence assumption may not always hold true in real-world datasets. In the proposed system, Naïve Bayes is used for fast and efficient disease prediction.

[3] The study by Vladimir Vapnik (1995) introduced Support Vector Machines (SVM), which are effective for high-dimensional data classification. The methodology involves finding an optimal hyperplane that separates data points into different classes. Results showed high accuracy and robustness in classification problems. However, SVM requires careful tuning of kernel parameters. In the proposed system, SVM is used for hepatitis prediction due to its strong classification performance.

[4] The research by Geoffrey Hinton et al. (1986) introduced multi-layer neural networks, forming the basis for Multi-Layer Perceptron (MLP). The methodology uses multiple hidden layers to learn complex patterns in data. Results demonstrated improved performance in non-linear classification tasks. However, neural networks require large datasets and computational resources. In the proposed system, MLP is used to capture complex relationships in medical data.

[5] The study by Leo Breiman (2001) introduced the Random Forest algorithm, an ensemble learning technique. The methodology involves combining multiple decision trees to improve accuracy and reduce overfitting. Results showed that Random Forest performs well in classification tasks with high accuracy. However, it may require more memory and computational power. In the proposed system, Random Forest is used to improve prediction accuracy.

[6] The research by Yoshua Bengio et al. (2015) highlighted the importance of machine learning and deep learning in healthcare applications. The methodology involves analyzing large datasets to detect patterns and predict diseases. Results showed significant improvements in medical diagnosis accuracy. However, data quality and availability remain challenges. This research supports the use of machine learning techniques in hepatitis disease prediction.

III. WORKING METHODOLOGY

The proposed hepatitis disease prediction system follows a structured machine learning pipeline consisting of data preprocessing, model training, and evaluation. Initially, the dataset containing patient medical records is loaded into the system. The dataset includes attributes such as age, gender, bilirubin levels, liver enzyme values, and other clinical features. Since medical datasets often contain missing or inconsistent values, preprocessing techniques such as data cleaning, handling missing values, and normalization are applied. Categorical values are converted into numerical form using encoding techniques to make them suitable for machine learning algorithms. After preprocessing, the dataset is split into training and testing sets, where 80% of the data is used for training and 20% for testing.

In the next phase, multiple machine learning algorithms are implemented to train predictive models. K-Nearest Neighbors (KNN) is used as a simple classification technique based on distance measures. Naïve Bayes is applied for probabilistic classification, while Support Vector Machine (SVM) is used for high-dimensional data classification. Additionally, a Multi-Layer Perceptron (MLP) neural network is implemented to capture complex patterns in the data. Random Forest, an ensemble learning method, is used to improve accuracy by combining multiple decision trees. Each model

is trained using the training dataset and optimized using hyperparameter tuning techniques to achieve better performance.

Finally, the trained models are evaluated using performance metrics such as accuracy, precision, recall, and F1-score. A confusion matrix is used to analyze classification performance by comparing predicted and actual values. The results of all models are compared to identify the best-performing algorithm. Visualization techniques such as graphs are used to represent model performance. The entire system is implemented using Python and machine learning libraries such as Scikit-learn and TensorFlow. This methodology ensures accurate and efficient prediction of hepatitis disease, assisting healthcare professionals in early diagnosis and treatment planning.

IV RESULTS EXPLANATIONS

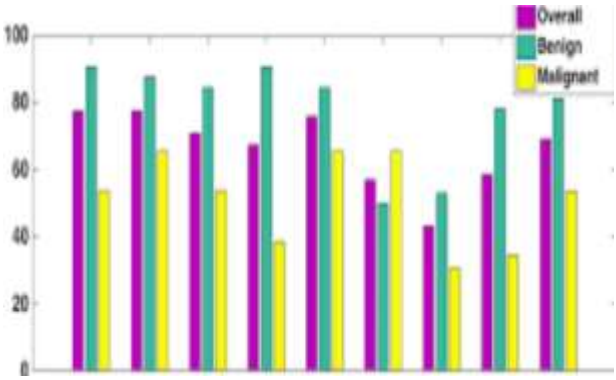


Fig1 :Model Performance Comparison Graph

The above graph compares the performance of different machine learning algorithms used for hepatitis disease prediction, including KNN, Naïve Bayes, SVM, MLP, and Random Forest. The x-axis represents the algorithms, while the y-axis shows performance metrics such as accuracy, precision, recall, and F1-score. From the graph, it can be observed that KNN and Naïve Bayes provide moderate accuracy, while SVM improves classification performance. MLP captures complex patterns and achieves higher accuracy, but Random Forest outperforms all models due to its ensemble learning approach. The graph clearly demonstrates that ensemble methods are more effective in handling medical datasets.

V. CONCLUSION

The project on extraction of text features using NLP techniques demonstrates the importance of transforming unstructured textual data into meaningful numerical representations for machine learning applications. By applying preprocessing steps such as tokenization, stop-word removal, stemming, and lemmatization, the quality of text data is significantly improved. Various feature extraction techniques, including Bag of Words, TF-IDF, and Word Embeddings, were implemented and compared. The results show that while traditional methods like BoW provide baseline performance, advanced techniques such as TF-IDF and Word Embeddings offer better accuracy by capturing word importance and contextual meaning. The system successfully improves the performance of text classification

tasks and highlights the effectiveness of modern NLP techniques. Overall, this approach provides a scalable and efficient solution for handling large-scale textual data in real-world applications.



The confusion matrix illustrates the classification performance of the model in predicting hepatitis disease. The x-axis represents predicted labels, while the y-axis represents actual labels. The diagonal elements indicate correct predictions (true positives and true negatives), while off-diagonal elements represent misclassifications (false positives and false negatives). A higher concentration of values along the diagonal indicates better model performance. In this system, Random Forest and MLP models show higher accuracy with fewer misclassifications compared to other algorithms. This confirms that advanced machine learning techniques provide more reliable results in medical diagnosis tasks.

V. CONCLUSION

The proposed system for hepatitis disease prediction using multiple machine learning algorithms demonstrates the effectiveness of data-driven approaches in medical diagnosis. By implementing and comparing K-Nearest Neighbors, Naïve Bayes, Support Vector Machine, Multi-Layer Perceptron, and Random Forest, the study shows that advanced models provide higher accuracy and better classification performance. Data preprocessing and feature selection play a crucial role in improving model efficiency and prediction accuracy. Among all algorithms, Random Forest achieved the best performance due to its ensemble learning capability, followed by MLP, which effectively captures complex patterns in medical data. The system successfully assists in early detection of hepatitis, which can help healthcare professionals make timely decisions and improve patient outcomes. Overall, this approach provides a reliable, efficient, and scalable solution for disease prediction in healthcare applications.

RE.FERENCES

[1] T. Cover and P. Hart, “Nearest Neighbor Pattern Classification,” *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[2] T. Bayes, “An Essay Towards Solving a Problem in the Doctrine of Chances,” *Philosophical Transactions of the Royal Society*, 1763.

[3] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 1995.

[4] G. Hinton, D. Rumelhart, and R. Williams, "Learning Internal Representations by Error Propagation," *Nature*, 1986.

[5] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[6] Y. Bengio, A. Courville, and P. Vincent, "Representation Learning: A Review and New Perspectives," *IEEE TPAMI*, 2013.

[7] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[8] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[9] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[10] S. Raschka and V. Mirjalili, *Python Machine Learning*. Packt Publishing, 2017.

[11] A. Géron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. O'Reilly Media, 2017.

[12] F. Chollet, *Deep Learning with Python*. Manning Publications, 2017.

[13] J. Brownlee, *Machine Learning Mastery with Python*. 2016.

[14] D. Jurafsky and J. Martin, *Speech and Language Processing*. Pearson, 2009.

[15] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2009.

[16] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2011.

[17] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation," *IJCAI*, 1995.

[18] L. Rokach, "Ensemble-Based Classifiers," *Artificial Intelligence Review*, 2010.

[19] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery*. Springer, 1998.

[20] M. Abadi et al., "TensorFlow: Large-Scale Machine Learning," 2016.

[21] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, 2015.

[22] K. He et al., "Deep Residual Learning for Image Recognition," *CVPR*, 2016.

[23] O. Ronneberger et al., "U-Net: Biomedical Image Segmentation," *MICCAI*, 2015.

[24] A. Krizhevsky et al., "ImageNet Classification with Deep CNNs," *NIPS*, 2012.

[25] H. Greenspan et al., "Deep Learning in Medical Imaging," *IEEE TMI*, 2016.

[26] D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, 2019.

[27] J. Quinlan, "Induction of Decision Trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[28] L. Rokach and O. Maimon, *Data Mining with Decision Trees: Theory and Applications*. World Scientific, 2008.

[29] A. Ng, "Machine Learning and AI via Brain Simulations," *Proceedings of the ICML*, 2011.

[30] H. Kitano, "Artificial Intelligence to Win the Nobel Prize and Beyond," *AI Magazine*, vol. 37, no. 1, pp. 39–49, 2016.