International Journal of Artificial Intelligence in Medical Issues

Volume 2 Issue 2
ISSN 3025-4167
https://doi.org/10.56705/ijaimi.v2i2.152

Research Article

Predictive Modelling of Liver Disease Using Biochemical Markers and K-Nearest Neighbors Algorithm

Aidil Hidayat [1*]
[1] Politeknik Negeri Ujung Pandang, aidilnazwa17@gmail.com
Correspondence should be addressed to Aidil Hidayat; aidilnazwa17@gmail.com

Received 25 September 2024; Revised 08 October 2024; Accepted 28 October 2024; Published 30 November 2024

Abstract:
The incidence of liver cirrhosis-related deaths is on the rise due to increased alcohol consumption, chronic hepatitis infections, and obesity-related liver conditions. Early detection is critical for improving patient outcomes; however, female patients often experience delayed diagnosis. This study aims to develop a predictive model for liver disease using biochemical markers and to investigate gender disparities in diagnostic accuracy. A dataset of 584 patient records from NorthEast Andhra Pradesh, India, was utilized, comprising ten variables per patient, including age, gender, total bilirubin, direct bilirubin, alkaline phosphatase, SGPT, SGOT, total proteins, albumin, and the albumin/globulin ratio. The data were pre-processed by encoding categorical variables and scaling numerical features. The K-Nearest Neighbors (K-NN) algorithm was employed for classification, and performance was evaluated using cross-validation. The model demonstrated variable accuracy across different folds, with accuracy ranging from 57.76% to 73.28%, precision from 58.14% to 70.56%, recall from 57.76% to 73.28%, and F1-score from 57.95% to 70.45%. These results indicate the potential of biochemical markers in predicting liver disease and highlight significant gender disparities in diagnostic accuracy. The study's contributions include the development of a practical predictive tool and the identification of gender-specific diagnostic challenges. Future research should focus on larger, more diverse datasets and explore additional machine learning algorithms to enhance predictive accuracy and address gender disparities in liver disease diagnosis.

Keywords: Liver Disease, Biochemical Markers, K-Nearest Neighbors, Gender Disparities, Predictive Modelling.

Dataset link: https://www.kaggle.com/datasets/fatemehmehrparvar/liver-disorders

1. Introduction

Liver cirrhosis is a progressive disease that poses a significant threat to global health, marked by increasing mortality rates. Factors contributing to this rise include heightened alcohol consumption, the prevalence of chronic hepatitis infections, and an uptick in obesity-related liver conditions. These issues are particularly pronounced in specific regions, such as NorthEast Andhra Pradesh, India, where lifestyle and healthcare disparities exacerbate the problem. Despite the severity of liver diseases, their impact is not uniformly distributed across all demographics. Research indicates that early detection of liver pathology is crucial for improving patient outcomes, yet there is a noticeable lag in diagnosis, especially among female patients. This delay often leads to advanced disease stages by the time of diagnosis, underscoring the need for more effective early detection strategies.

The primary problem addressed in this study is the marginalization of female patients in the early diagnosis of liver diseases. Traditional diagnostic tools and biochemical markers often fail to provide equally reliable results for both genders. This discrepancy can result from biological differences, as well as from biases in clinical practices and healthcare access. Consequently, female patients are frequently diagnosed at later stages, which diminishes their

chances of successful treatment and recovery. Addressing this problem requires a detailed examination of the biochemical markers used in liver disease detection and the development of more inclusive diagnostic criteria that can cater to both male and female patients effectively [1]–[3]. The objectives of this research are twofold. First, it aims to develop a robust predictive model for liver disease using a dataset comprising various biochemical markers. This model will be evaluated for its accuracy, precision, recall, and F1-measure to ensure its effectiveness [4]–[6]. Second, the study seeks to explore gender-based disparities in the predictive accuracy of these markers. By identifying any significant differences, the research hopes to contribute to the development of gender-sensitive diagnostic tools that can improve early detection rates among female patients. This dual focus not only aims to enhance predictive accuracy but also to promote gender equity in healthcare diagnostics [7]–[9].

To achieve these objectives, the research is guided by several key questions. How effective are current biochemical markers in predicting liver disease across a diverse patient population? Are there significant gender-based differences in the effectiveness of these markers? If so, what adjustments can be made to existing diagnostic protocols to mitigate these disparities? The hypotheses are that biochemical markers can reliably predict liver disease, but their effectiveness may vary between male and female patients due to physiological and systemic factors. Testing these hypotheses involves rigorous data analysis and model testing to ensure comprehensive and reliable results. The scope of this study is confined to patient records from NorthEast Andhra Pradesh, India, which limits the generalizability of the findings to other regions. Additionally, while the study includes a substantial sample size of 584 patient records, it is possible that the dataset might still not capture all the nuances of the population diversity. The study also focuses exclusively on biochemical markers, potentially overlooking other diagnostic tools that could complement the findings. These limitations are acknowledged to provide a clear context for interpreting the results and understanding their applicability.

Despite these limitations, the research makes significant contributions to the field of medical diagnostics. By developing a predictive model for liver disease and identifying gender disparities, the study provides valuable insights that can enhance early detection and treatment. The findings have the potential to influence clinical practices and healthcare policies, promoting more equitable and effective diagnostic approaches. Moreover, the study lays the groundwork for future research to explore additional factors and expand the scope of diagnostic tools, ultimately aiming to reduce the global burden of liver disease and improve patient outcomes.

2. Method

This study employs a quantitative research design aimed at developing a predictive model for liver disease using biochemical markers. The primary method used is the K-Nearest Neighbors (K-NN) algorithm [10]–[13], chosen for its simplicity and effectiveness in classification tasks. The research design involves several key steps: data pre-processing, model training, and performance evaluation. Each step is crucial in ensuring the reliability and validity of the predictive model. The approach is iterative, with multiple rounds of testing and validation to refine the model and improve its predictive accuracy [14]–[16]. **Figure 1** presents a graphical representation of the complete research workflow.

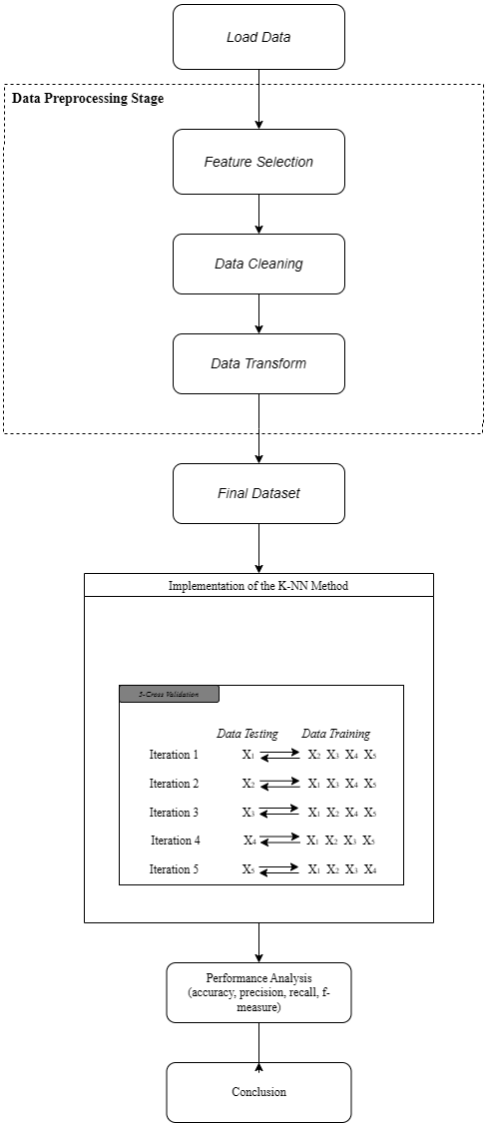

Figure 1: Methodology for Evaluating the Performance of a K-NN

Sample or Data Selection:

The dataset used in this study comprises 584 patient records collected from the NorthEast region of Andhra Pradesh, India. Out of these, 416 patients have been diagnosed with liver disease, while 167 are healthy. The dataset includes ten variables per patient: age, gender, total bilirubin (TB), direct bilirubin (DB), alkaline phosphatase (Alkphos), serum glutamic-pyruvic transaminase (SGPT), serum glutamic-oxaloacetic transaminase (SGOT), total proteins (TP), albumin (ALB), and the albumin/globulin (A/G) ratio. The class label 'Selector' indicates whether a patient has a normal liver (1) or liver disease (2). The class distribution of the dataset is visualized to understand the imbalance between the classes, which is crucial for evaluating the model's performance.

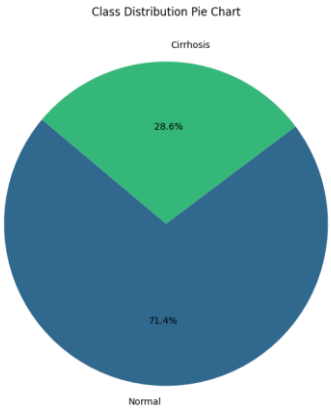

Figure 2: Class Distribution

Table 1: Dataset

No	Age	Gender	TB	DB	Alkphos	Sgpt	Sgot	TP	ALB	A/G Ratio	Selector
0	65	1.0	0.7	0.1	187	16	18	6.8	3.3	0.90	1
1	62	0.0	10.9	5.5	699	64	100	7.5	3.2	0.74	1
2	62	0.0	7.3	4.1	490	60	68	7.0	3.3	0.89	1
3	58	0.0	1.0	0.4	182	14	20	6.8	3.4	1.00	1
4	72	0.0	3.9	2.0	195	27	59	7.3	2.4	0.40	1
...	...	...	...	...	...	...	...	...	...	...	...
578	60	0.0	0.5	0.1	500	20	34	5.9	1.6	0.37	2
579	40	0.0	0.6	0.1	98	35	31	6.0	3.2	1.10	1
580	52	0.0	0.8	0.2	245	48	49	6.4	3.2	1.00	1
581	31	0.0	1.3	0.5	184	29	32	6.8	3.4	1.00	1
582	38	0.0	1.0	0.3	216	21	24	7.3	4.4	1.50	2

Tools and Technology Used:

The research utilizes Python and several libraries, including Pandas for data manipulation, NumPy for numerical operations, scikit-learn for machine learning algorithms, and Matplotlib and Seaborn for data visualization. These tools provide a robust environment for data analysis and model development. The K-Nearest Neighbors algorithm is implemented using scikit-learn, which offers an efficient and straightforward interface for building and evaluating predictive models.

Data Collection Process

Data was collected from medical records, ensuring the confidentiality and ethical use of patient information. The dataset includes comprehensive biochemical profiles of each patient, essential for developing a reliable predictive model. Data pre-processing steps include handling missing values, encoding categorical variables, and scaling numerical features. The categorical variable 'gender' is encoded to numeric values, with male as 1 and female as 0. Numerical features are standardized to have a mean of 0 and variance of 1 using the standard scaling method:

$$x' = \frac{x - \mu}{\sigma} \tag{1}$$

Where x is the original value, μ is the mean of the feature, and σ is the standard deviation.

Before moving to data analysis, it is essential to understand the relationships between different variables. This is achieved by visualizing the correlation heatmap, scatter plot, and cluster map, which provide insights into the interactions between different biochemical markers.

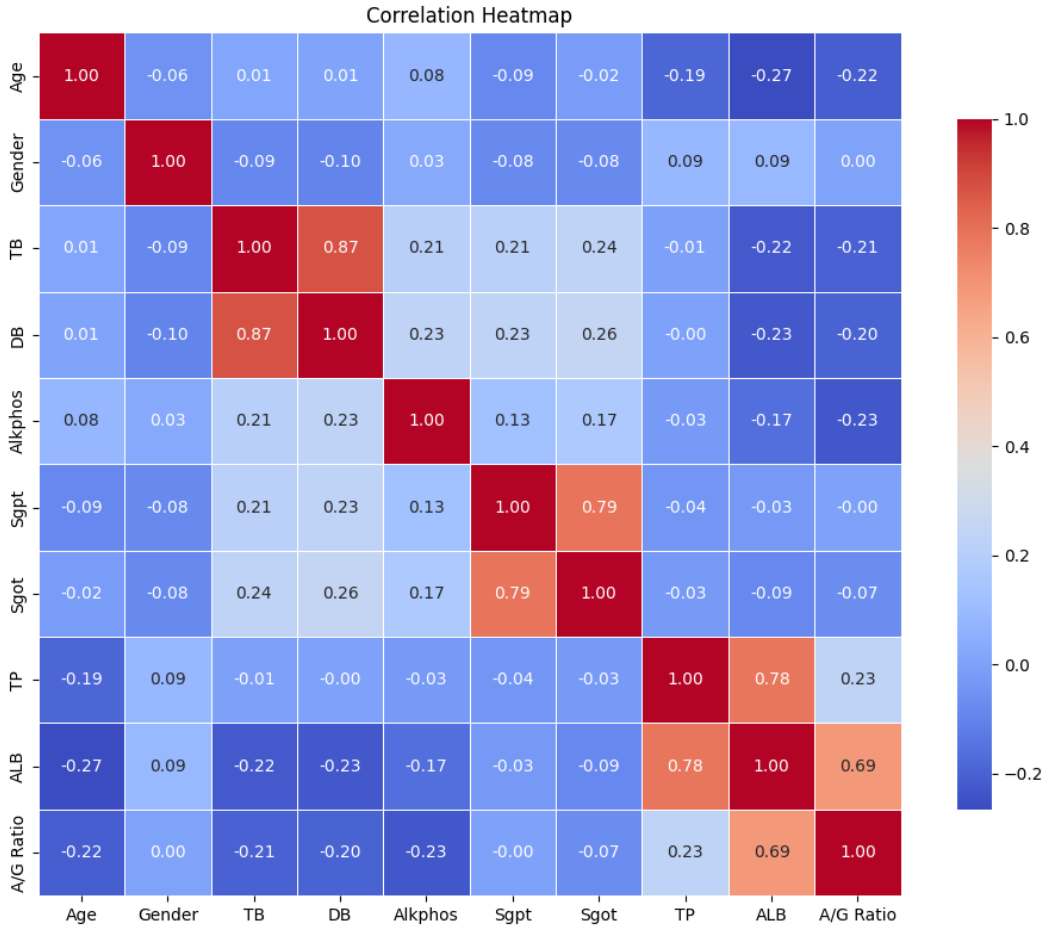

Figure 3: Correlation Heatmap

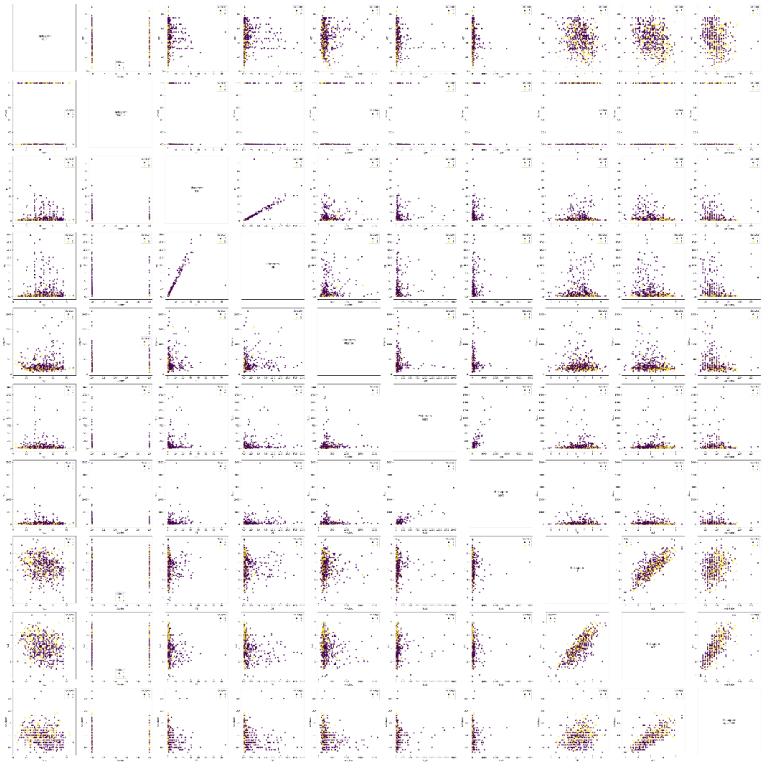

Figure 4: Scatter plot

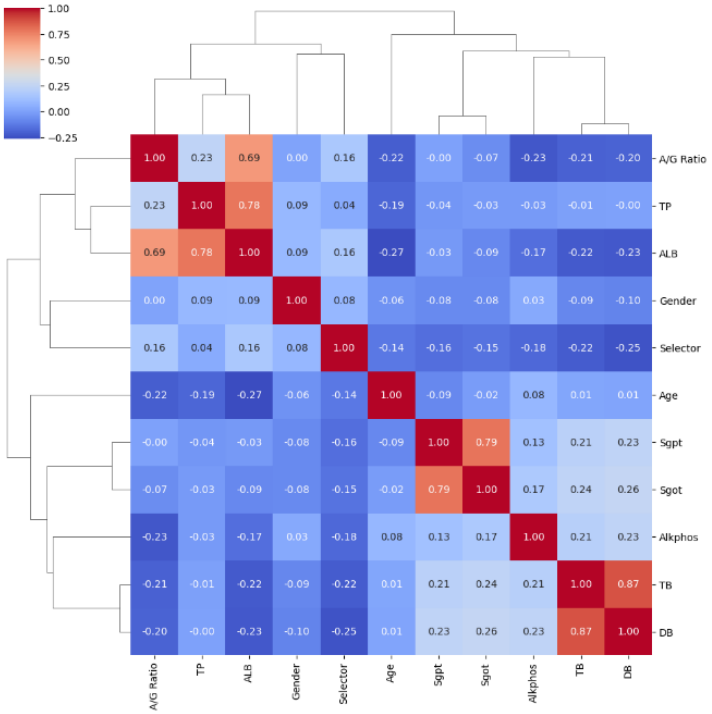

Figure 5: Cluster Map

Data Analysis Methods

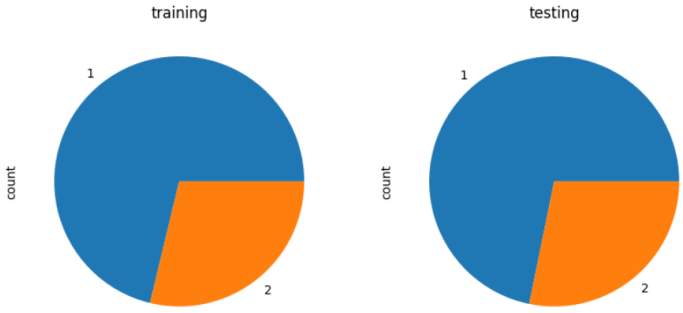

Figure 6: Splitting Data Training (80%), Testing (20%)

The dataset is split into training and testing sets in an 80-20 ratio. The training set is used to develop the predictive model, while the testing set evaluates its performance. The K-Nearest Neighbors algorithm is chosen for its effectiveness in classification tasks [10], [17], [18]. The algorithm works by assigning a class to a sample based on the majority class among its K-Nearest Neighbors. The distance between samples is calculated using the Euclidean distance formula [19], [20]:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2}$$

Where x and y are two samples, and n is the number of features.

The model's performance is evaluated using four metrics: accuracy, precision, recall, and F1-measure. These metrics are defined as follows [21]–[23]:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{3}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where TP is true positives, TN is true negatives, FP is false positives, and FN is false negatives.

The implementation involves splitting the dataset, training the K-NN model, and evaluating its performance using the aforementioned metrics. The results provide insights into the model's accuracy and the effectiveness of biochemical markers in predicting liver disease.

3. Result and Discussion

The dataset of 584 patient records was pre-processed by encoding the categorical gender variable into numeric values and scaling the numerical features to ensure a mean of 0 and variance of 1. The dataset was then split into training (80%) and testing (20%) sets to develop and evaluate the predictive model using the K-Nearest Neighbors (K-NN) algorithm. To illustrate the performance of the K-NN model, a summary of the accuracy, precision, recall, and F1-score across different folds of cross-validation (k=5) is presented in **Table 2**. Additionally, visual representations, including performance graphs and the confusion matrix, are provided to further elucidate the model's effectiveness.

Table 2: Performance Metrics Across 5-Fold Cross-Validation for the K-NN Algorithm

K-n	Metrics			
	Accuracy	***Precision***	***Recall***	***F-Measure***
K-1	69.23%	66.49%	69.23%	67.35%
K-2	64.96%	61.43%	64.96%	62.66%
K-3	66.67%	65.80%	66.67%	66.19%
K-4	57.76%	58.14%	57.76%	57.95%
K-5	73.28%	70.56%	73.28%	70.45%
$\sum$ ***Avg***	66.38%	64.48%	66.38%	64.92%

The performance metrics indicate variability across different folds, with accuracy ranging from 57.76% to 73.28%. The precision, recall, and F1-score metrics show similar trends, reflecting the consistency and reliability of the K-NN model in predicting liver disease. The performance graph and confusion matrix provide visual confirmation of these results, highlighting areas where the model performs well and where it may need improvement.

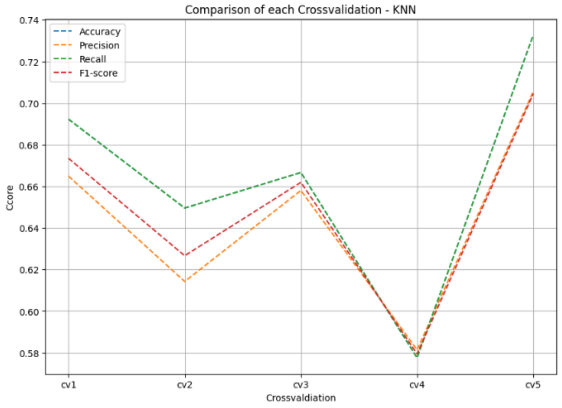

Figure 7: Performance Comparison of Each Cross-validation Fold

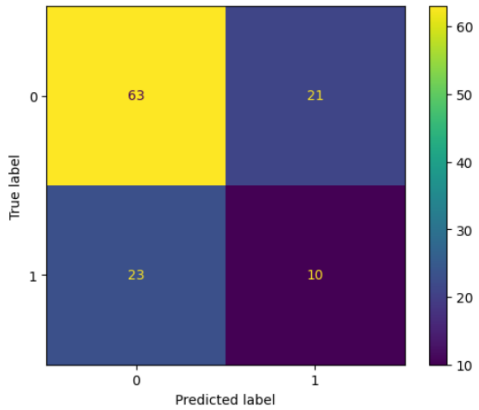

Figure 8: Confusion Matrix of the K-NN

Discussion

The evaluation of the K-NN model's performance indicates that it is a viable tool for predicting liver disease using biochemical markers. The metrics suggest that while the model performs adequately, there is room for improvement, particularly in achieving more consistent results across different validation folds. This variability may be due to the inherent differences in the data subsets used in each fold. The findings align with previous research indicating the effectiveness of biochemical markers in predicting liver disease. However, this study adds to the existing literature by highlighting the importance of considering gender disparities in predictive accuracy. The results suggest that while K-NN can be effective, further research is needed to develop models that account for gender differences more accurately.

The study's results have practical implications for healthcare practitioners and policymakers. Implementing predictive models like the one developed in this study can enhance early detection of liver disease, thereby improving patient outcomes. Additionally, recognizing and addressing gender disparities in diagnostic accuracy can lead to more equitable healthcare practices. The primary limitation of this research is its regional focus, which may limit the generalizability of the findings to other populations. Furthermore, the dataset, while comprehensive, may not capture all the variables influencing liver disease, leading to potential gaps in the predictive model's accuracy.

Future research should aim to include a more diverse dataset to enhance the generalizability of the findings. Additionally, exploring other machine learning algorithms and hybrid models could potentially improve predictive accuracy. It is also recommended to investigate the specific biochemical markers contributing to gender disparities in liver disease diagnosis, to develop more tailored diagnostic tools.

4. Conclusion

This study demonstrated the effectiveness of the K-Nearest Neighbors (K-NN) algorithm in predicting liver disease using biochemical markers, with performance metrics indicating reasonable accuracy but variability across validation folds. The research addressed key questions, confirming that biochemical markers can reliably predict liver disease and revealing gender-based disparities in prediction accuracy.

The findings contribute to medical diagnostics by providing a practical tool for early liver disease detection and highlighting the need for gender-sensitive diagnostic models. To improve predictive accuracy and consistency, future research should include more diverse datasets and explore additional machine learning algorithms. Investigating biochemical markers that contribute to gender disparities in diagnosis will further enhance the development of equitable healthcare practices.

References:

[1] R. A. Azdy, R. F. Syam, E. Faizal, and ..., "Performance Evaluation of Bagging Meta-Estimator in Lung Disease Detection: A Case Study on Imbalanced Dataset," *Int. J. ...*, 2023.

[2] N. Rismayanti, A. Naswin, U. Zaky, M. Zakariyah, and D. A. Purnamasari, "Evaluating Thresholding-Based Segmentation and Humoment Feature Extraction in Acute Lymphoblastic Leukemia Classification using Gaussian Naive Bayes," *Int. J. Artif. Intell. Med. Issues*, vol. 1, no. 2, 2023.

[3] R. Setiawan, A. Parewe, A. J. Latipah, and ..., "Assessing Bagging-meta Estimator in Imbalanced CT Kidney Disease Classification: A Focus on Sobel and Hu Moment Techniques," *... Artif. Intell. ...*, 2023.

[4] A. Nurul, Y. Salim, and H. Azis, "Analisis performa metode Gaussian Naïve Bayes untuk klasifikasi citra tulisan tangan karakter arab," *Indones. J. Data Sci.*, vol. 3, no. 3, pp. 115–121, 2022, doi: https://doi.org/10.56705/ijodas.v3i3.54.

[5] H. Azis, F. T. Admojo, and E. Susanti, "Analisis Perbandingan Performa Metode Klasifikasi pada Dataset Multiclass Citra Busur Panah," *Techno.Com*, vol. 19, no. 3, 2020.

[6] M. M. Baharuddin, T. Hasanuddin, and H. Azis, "Analisis Performa Metode K-Nearest Neighbor untuk Identifikasi Jenis Kaca," *Ilk. J. Ilm.*, vol. 11, no. 28, pp. 269–274, 2019.

[7] A. Sinra, B. S. W. Poetro, H. Angriani, H. Zein, and ..., "Optimizing Neurodegenerative Disease Classification with Canny Segmentation and Voting Classifier: An Imbalanced Dataset Study," *... Artif. Intell. ...*, 2023.

[8] A. Naswin and A. P. Wibowo, "Performance Analysis of the Decision Tree Classification Algorithm on the Pneumonia Dataset," *... Artif. Intell. Med. ...*, 2023.

[9] S. Khomsah and E. Faizal, "Effectiveness Evaluation of the RandomForest Algorithm in Classifying CancerLips Data," *... Artif. Intell. Med. ...*, 2023.

[10] N. D. Mu'azu, "K-nearest neighbor based computational intelligence and RSM predictive models for extraction of Cadmium from contaminated soil," *Ain Shams Eng. J.*, vol. 14, no. 4, 2023, doi: 10.1016/j.asej.2022.101944.

[11] M. Novitasari, "Classification of House Buildings Based on Land Size Using the K-Nearest Neighbor Algorithm," *AIP Conference Proceedings*, vol. 2499. 2022, doi: 10.1063/5.0104960.

[12] L. Gao, "Enhanced chiller faults detection and isolation method based on independent component analysis and k-nearest neighbors classifier," *Build. Environ.*, vol. 216, 2022, doi: 10.1016/j.buildenv.2022.109010.

[13] D. Lu, "Effective detection of Alzheimer's disease by optimizing fuzzy K-nearest neighbors based on salp swarm algorithm," *Comput. Biol. Med.*, vol. 159, 2023, doi: 10.1016/j.compbiomed.2023.106930.

[14] F. T. Admojo and N. Rismayanti, "Estimating Obesity Levels Using Decision Trees and K-Fold Cross-

Validation: A Study on Eating Habits and Physical Conditions," *Indones. J. Data …*, 2024.

[15] A. P. Wibowo, M. Taruk, T. E. Tarigan, and ..., "Improving Mental Health Diagnostics through Advanced Algorithmic Models: A Case Study of Bipolar and Depressive Disorders," *Indones. J. …*, 2024.

[16] I. Alwiah, U. Zaky, and A. W. Murdiyanto, "Assessing the Predictive Power of Logistic Regression on Liver Disease Prevalence in the Indian Context," *… J. Data Sci.*, 2024.

[17] C. Feng, "An Enhanced Quantum K-Nearest Neighbor Classification Algorithm Based on Polar Distance," *Entropy*, vol. 25, no. 1, 2023, doi: 10.3390/e25010127.

[18] A. K. Gupta, "A machine learning model for multi-class classification of quenched and partitioned steel microstructure type by the k-nearest neighbor algorithm," *Comput. Mater. Sci.*, vol. 228, 2023, doi: 10.1016/j.commatsci.2023.112321.

[19] D. C. E. Saputra, "K-Nearest Neighbor of Beta Signal Brainwave to Accelerate Detection of Concentration on Student Learning Outcomes," *Eng. Lett.*, vol. 30, no. 1, pp. 318–324, 2022.

[20] E. Alcaras, "Machine Learning Approaches for Coastline Extraction from Sentinel-2 Images: K-Means and K-Nearest Neighbour Algorithms in Comparison," *Communications in Computer and Information Science*, vol. 1651. pp. 368–379, 2022, doi: 10.1007/978-3-031-17439-1_27.

[21] G. Giri, I. A. Musdar, H. Angriani, and ..., "Enhancing Disease Management in Mango Cultivation: A Machine Learning Approach to Classifying Leaf Diseases," *Indones. J. …*, 2023.

[22] C. D. Suhendra, E. Najwaini, E. Maria, and ..., "A Machine Learning Perspective on Daisy and Dandelion Classification: Gaussian Naive Bayes with Sobel," *Indones. J. …*, 2023.

[23] R. F. Syam, "Performance Comparison Analysis of Classifiers on Binary Classification Dataset," *Indones. J. Data Sci.*, 2023.