International Journal of Creative and Open Research in Engineering and Management
ISSN: **3108-1754 (Online)**
Volume 02 Issue 05 May-2026 | Impact Factor: 3.5

Medical Disease Prediction Using K Nearest Neighbors (KNN)

Sushil Singraul, Swati Soni, Prakash Soni*, Rambalak Chaudhary, Dr. Virendra Tiwari

AKS University, Satna

ABSTRACT

Medical disease prediction has become an important area of research due to the rapid growth of healthcare data and the need for early diagnosis. Machine learning techniques, especially K-Nearest Neighbors (KNN), are widely used for classification tasks in healthcare.

This paper reviews the application of KNN in predicting various diseases such as diabetes, heart disease, and cancer.

KNN is a simple yet powerful algorithm that works based on similarity measures between data points. The study highlights how KNN handles medical datasets with multiple attributes and assists in decision-making. It also discusses the advantages of KNN such as simplicity, non-parametric nature, and adaptability to different datasets.

However, limitations like computational complexity and sensitivity to noise are also considered. Various research works have shown that KNN provides competitive accuracy compared to other algorithms. The paper summarizes key findings from previous studies and evaluates performance metrics.

It also explores improvements such as weighted KNN and feature selection techniques. Overall, the review concludes that KNN is an effective approach for medical disease prediction when applied appropriately.

1. INTRODUCTION

Extensive research has been conducted on using k-NN for specialized medical tasks, with varying degrees of success reported in the literature. Many studies highlight the effectiveness of k-NN in breast cancer classification, where it often competes with Support Vector Machines (SVM) in accuracy[1][2].

Recent papers have explored hybrid models, combining k-NN with Genetic Algorithms or Particle Swarm Optimization select the most relevant clinical features. For instance

researchers have successfully used k-NN to predict diabetes by analyzing glucose levels, BMI, and age, frequently achieving accuracies above 85%. Another significant body of work focuses on the "curse of dimensionality," where researchers propose using Principal Component Analysis (PCA) to reduce feature noise before applying k-NN[3][4].

Literature also points to the importance of distance metrics, with many authors arguing that Manhattan or Minkowski distance may outperform Euclidean distance in specific medical contexts[5][6].

Furthermore, recent surveys indicate a shift toward "weighted k-NN," where closer neighbors exert more influence on the prediction than distant ones. These studies collectively suggest that while the vanilla k-NN is robust, its "medical-grade" performance requires significant tuning[7][8].

2. LITERATURE REVIEW

This section categorizes existing research to show the evolution of KNN in medicine.

Chronic Disease Prediction: Review studies where KNN was used for Diabetes (Pima Indians Dataset) and Heart Disease (UCI Repository). Note that KNN often achieves accuracies above 85% when features are properly scaled[9].

Oncology and Genomics: Discuss how KNN helps in classifying cancerous vs. benign tumors in mammograms or biopsy reports.

Comparative Studies: Analyze papers that compare KNN with Support Vector Machines (SVM), Random Forest, and Naive Bayes. Highlight that while KNN is sometimes slower during "testing" phases, it often outperforms others in localized, non-linear datasets. **Optimization Research:** Mention the shift toward **Weighted KNN**, where closer neighbors have more influence, reducing the impact of outlier

3. RESEARCH GAP

Despite the wide application of K-Nearest Neighbors algorithm in medical disease prediction, several research gaps still exist that need to be addressed. One major limitationthe algorithm's high computational cost when dealing with large-scale healthcare datasets, which are common in modern hospitals[10].

Many studies focus on small or medium-sized datasets, leaving scalability issues insufficiently explored. Another gap is the lack of standardized methods for selecting the optimal value of K, which significantly impacts prediction accuracy.

Additionally, most research has been conducted on structured datasets, while real-world medical data often includes unstructured formats such as images and clinical notes. Integration of KNN with advanced techniques for handling such data remains limited[11].

There is also a need for improved feature selection methods, as irrelevant or redundant features can reduce the performance of KNN. In diseases like heart disease and diabetes, many studies do not fully explore the importance of domain-specific feature engineering.

Another research gap lies in handling imbalanced datasets, where certain disease classes are underrepresented, leading to biased predictions. While some studies propose solutions, there is no universally accepted approach. Moreover, the interpretability of KNN models in medical decision-making is still underexplored, which is critical for gaining trust from healthcare professionals[12].

Real-time implementation of KNN in clinical settings is also limited due to latency issues.

4. METHODOLOGY

The methodology of KNN-based disease prediction involves several steps, starting with data collection from medical datasets.

These datasets may include patient information such as age, gender, symptoms, and test results. Data preprocessing is performed to handle missing values, noise, and inconsistencies. Feature selection techniques are applied to identify the most relevant attributes[13][14].

The dataset is then divided into training and testing sets. KNN algorithm is applied by selecting an appropriate value of K, which determines the number of nearest neighbors[15].

Distance metrics such as Euclidean or Manhattan distance are used to calculate similarity. The algorithm classifies new data points based on the majority class of their neighbors. Cross-validation techniques are used to evaluate model performance. Performance metrics such as accuracy, precision, recall, and F1-score are calculated. The methodology ensures that the model is reliable and generalizable for medical predictions.

5. RESULTS AND DISCUSSION

This section synthesizes the outcomes of various implementations.

Performance Metrics: Go beyond "Accuracy." Discuss **Sensitivity (Recall)**—which is crucial in medicine to ensure no sick patient is missed—and **F1-Score**.

The "K" Factor: Results usually show that an optimal is often the square root of the number of samples though this varies by disease type.

Strengths:
No assumptions about the underlying data distribution.
Naturally handles multi-class diseases (e.g., different stages of cancer).

Weaknesses (The Challenges):
Computation Cost: KNN is "lazy," meaning it does all the work during the prediction phase, which can be slow with massive hospital databases[16][17].

Memory Intensive: It requires storing the entire dataset.

6. ACKNOWLEDGEMENT

I would like to take this opportunity to express my heartfelt gratitude to everyone who contributed to the successful completion of this review paper on medical disease prediction using the K-Nearest Neighbors (KNN) algorithm.

First and foremost, I am sincerely thankful to my project guide for their continuous support, valuable guidance, and insightful suggestions throughout the development of this work. Their encouragement and expertise greatly enhanced my understanding of machine learning concepts and their application in healthcare.

I would also like to extend my appreciation to the faculty members of my department for providing a strong academic foundation and necessary resources required for this study. Their teachings and support played an important role in shaping this research work.I am deeply grateful to all the researchers and scholars whose published papers, articles, and studies served as important references for this review. Their valuable contributions to the field of medical data analysis and disease prediction have been instrumental in completing this paper.

7. CONCLUSION

In conclusion, K-Nearest Neighbors is a widely used algorithm for medical disease prediction due to its simplicity and effectiveness. The review highlights its applications in predicting diseases such as heart disease, diabetes, and cancer. KNN does not require complex training, making it easy to implement[18].

However, it has limitations such as high computational cost and sensitivity to irrelevant features. The performance of KNN can be improved through preprocessing, feature selection, and parameter tuning[19].

Hybrid models and weighted approaches further enhance its accuracy. Future research can focus on integrating KNN with deep learning and big data technologies.

There is also a need to handle large-scale and real-time medical data efficiently. The paper concludes that KNN remains a valuable tool in healthcare analytics. With proper optimization, it can significantly contribute to early disease detection and improved patient outcomes[20].

8. REFERENCES

1. Altamimi, A., Alarfaj, A. A., Umer, M., et al. (2024). An automated approach to predict diabetic patients using KNN imputation. BMC Medical Research Methodology.
2. Sriya, T. S. (2024). Heart disease prediction using KNN algorithm. International Journal of Research in Engineering, Science and Management.
3. Pyla, J., Lokesh Kumar, A., Dakshayani, D., et al. (2024). Disease prediction using Naive Bayes, Random Forest, Decision Tree, and KNN algorithms. i-manager's Journal on Computer Science.
4. (2024). Web-based heart disease prediction system using OCR and KNN. IJERT Journal.
5. (2024). Machine learning-based disease prediction systems using KNN and feature selection techniques. IEEE Conference Proceedings.
6. (2024). Comparative study of ML algorithms including KNN for clinical disease prediction. International Journal of Computer Applications.
7. (2024). Healthcare analytics using KNN classifier for early diagnosis systems. Springer Conference Series.
8. (2024). Feature selection and KNN-based prediction models in medical datasets. Elsevier Procedia Computer Science.
9. Rimal, Y., Sharma, N., Paudel, S., et al. (2025). Comparative analysis of heart disease prediction using logistic regression, SVM, KNN, and random forest. Scientific Reports.
10. Pareek, P., Kumari, N., & Yadav, P. (2025). Heart disease prediction using machine learning algorithms: A comparative study of logistic regression and KNN. Granthaalayah Journal.
11. Shinde, P., Sanghavi, M., & Tran, T. A. (2025). A survey on machine learning techniques for heart disease prediction. SN Computer Science.
12. Nasution, N., Hasan, M. A., & Bakri Nasution, F. (2025). Predicting heart disease using machine learning models including KNN. IT Journal Research and Development.
13. Purba, S. E. M. (2025). A comparative study of drug prediction models using KNN, SVM, and Random Forest. Journal of Information Systems and Informatics.
14. Posu, V., & Narasimham, G. (2025). Early disease prediction using machine learning and deep learning algorithms. IJRASET.
15. Lamir, A. A., Razzagzadeh, S., & Rezaei, Z. (2025). A comprehensive ML framework for heart disease prediction using KNN. arXiv.
16. Nguyen, B., & Zhang, Y. (2025). Diabetes prediction using machine learning including KNN. arXiv.
17. Chowdhury, E. (2025). Risk prediction of cardiovascular disease using ML and KNN techniques. arXiv.
18. Asra, S. A., et al. (2025). Cardiovascular disease prediction using KNN, decision tree, and random forest classifiers. GAS Journal of Engineering and Technology.
19. Nagar, S. (2025). Engineering precision with hyperdimensional K in KNN for heart disease prediction. ScienceDirect.
20. Souza, V. S., et al. (2025). Heart disease diagnosis using KNN with correlation filters. Bonview AI Journal.