KNN Based Chronic Disease Prediction

Navaneeth H[1], Mythili N[2]
[1,2]*Project Student, Department of IoT and AI & ML, Nehru Arts and Science College, Coimbatore, India*

Abstract-Chronic diseases such as diabetes, heart disease, and kidney disorders are among the leading causes of mortality worldwide, necessitating early detection and effective monitoring. This study presents a machine learning approach utilizing the K-Nearest Neighbors (KNN) algorithm for the prediction of chronic diseases based on clinical and demographic data. The model is trained and evaluated using publicly available healthcare datasets, incorporating features such as age, blood pressure, glucose levels, BMI, and other relevant indicators. KNN, known for its simplicity and effectiveness in classification problems, is employed to identify patterns and similarities between patient profiles. The results demonstrate that the KNN model achieves promising accuracy, precision, and recall, making it a viable tool for supporting early diagnosis in clinical settings. This approach can aid healthcare professionals in making informed decisions and potentially reduce the burden of chronic illnesses through timely intervention.

1. INTRODUCTION

Chronic diseases, including diabetes, cardiovascular diseases, and chronic kidney disorders, represent a significant global health burden, contributing to high rates of morbidity and mortality. These conditions often develop slowly over time and can be managed effectively if detected early. However, timely diagnosis remains a challenge due to the complexity of symptoms and the need for extensive clinical testing. In recent years, machine learning techniques have shown great promise in the field of medical diagnostics, offering automated and data-driven solutions for early disease prediction.

Among various machine learning algorithms, the K-Nearest Neighbors (KNN) algorithm stands out for its simplicity, interpretability, and effectiveness in classification tasks. KNN operates by comparing a new, unseen data point to existing labeled data and assigning the most common class among its 'k' closest neighbors. This characteristic makes it particularly suitable for medical applications, where patient data often exhibit patterns and similarities that can be leveraged for prediction.

This study explores the use of the KNN algorithm to predict the likelihood of chronic disease in patients based on key clinical and lifestyle features such as age, blood pressure, glucose levels, BMI, and other vital health indicators. The goal is to develop a reliable and efficient model that can assist healthcare providers in identifying at-risk individuals, enabling earlier interventions and improving patient outcomes. Through the application of KNN on a curated healthcare dataset, the study aims to evaluate the model's performance and highlight its potential in real-world medical decision support systems. KNN is a supervised learning algorithm that classifies data points based on the classes of their nearest neighbors in the feature space. By comparing a new patient's health data to historical medical records, KNN can provide valuable insights into potential diagnoses. Its non-parametric nature makes it particularly suitable for medical datasets where the relationships between features may be complex and nonlinear.

This study explores the use of the KNN algorithm for predicting chronic diseases, evaluating its performance across various metrics and comparing it with other classification methods. The goal is to assess how effectively KNN can support clinicians in identifying at-risk individuals and guiding early treatment strategies.

Among various machine learning algorithms, the K-Nearest Neighbors (KNN) algorithm is widely recognized for its simplicity, versatility, and strong performance in classification tasks. KNN is a supervised, instance-based learning algorithm that classifies new data points based on the majority class of their 'K' nearest neighbors in a multi-dimensional feature space. It does not require prior assumptions

about the data distribution, making it highly adaptable to real-world medical data, which is often noisy and non-linear.

In chronic disease prediction, KNN can be particularly effective due to its ability to identify patterns in patient data by comparing symptoms, diagnostic results, or lab values with those of previously diagnosed cases. This data-driven approach allows for personalized risk assessment, supporting early diagnosis and intervention.

This research/project focuses on applying the KNN algorithm to predict the presence or risk level of chronic diseases using patient health records. It explores key factors influencing model performance, such as feature selection, choice of K, and distance metrics, while also evaluating the model against standard classification metrics like accuracy, precision, recall, and F1-score. The overall objective is to demonstrate the potential of KNN as a reliable, interpretable tool in healthcare analytics

2. LITERATURE REVIEW

The integration of machine learning techniques in the healthcare domain has led to significant advancements in disease diagnosis and prognosis. Numerous studies have explored the potential of various algorithms in predicting chronic diseases, with a focus on improving accuracy, reducing false diagnoses, and supporting clinical decision-making.

[1] Sarker, I. H., Faruque, M. F., Alqahtani, H., & Kalim, A. (2020). K-Nearest Neighbor learning based diabetes mellitus prediction and analysis for eHealth services. In conclusion, this work has demonstrated that the K-Nearest Neighbor (KNN) machine learning algorithm offers a promising solution for the prediction of diabetes mellitus in the context of modern eHealth systems. By utilizing relevant clinical and demographic data, the study successfully developed and evaluated a KNN-based model that achieved high predictive accuracy and robustness. The research not only validates the potential of KNN in medical diagnostics but also emphasizes the broader role of data-driven approaches in transforming traditional healthcare into intelligent, automated, and scalable digital services.

[2]Uddin et al. (2019) performed a comprehensive comparison of several supervised machine learning algorithms, including K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and decision trees, in the context of disease prediction. Their study found that although no single algorithm dominates in every situation, KNN proved to be highly competitive, particularly for structured medical datasets of moderate size. KNN was valued for its simplicity, effectiveness, and ability to work well without requiring complex parameter tuning.

[3]Bzdok et al. (2018) emphasized the importance of interpretability and simplicity in clinical machine learning applications. They discussed how supervised learning methods like KNN can be powerful tools in healthcare due to their transparent decision-making process, which makes them more acceptable and trustworthy in medical environments where understanding the rationale behind predictions is essential.

[4]Mahesh (2020) provided an in-depth review of various machine learning algorithms, covering both traditional and modern methods. KNN was highlighted for its low computational complexity, ease of use, and solid performance, especially in domains with low-dimensional, clean datasets. The study reinforced the continuing relevance of KNN despite the growing popularity of more complex algorithms like neural networks.

[5]Zhang et al. (2017) addressed a critical issue in KNN: selecting the optimal value of ‘k’, which greatly influences classification results. Instead of relying on a fixed or manually chosen value, they proposed an adaptive learning strategy to dynamically determine the most suitable ‘k’ based on the data characteristics, improving the algorithm’s flexibility and accuracy.

[6]Bhatia & Vandana (2010) conducted a survey on nearest neighbor techniques, analyzing the classical KNN and its extensions. They discussed the effects of various distance measures—such as Euclidean, Manhattan, and Minkowski—on classification performance, showing that the choice of distance metric significantly affects accuracy and should be dataset-specific.

[7]Lamba & Kumar (2016) reviewed different KNN variants developed to address the algorithm's limitations. They explained the workings of Weighted KNN, which assigns different weights to neighbors based on distance; Condensed KNN, which reduces the size of the training dataset; and Fuzzy KNN, which handles uncertainty by assigning degrees of membership to each class. These improvements aim to boost performance, reduce computation time, and make KNN more robust in noisy or imbalanced datasets.

[8]Wettschereck & Dietterich (1994) contributed foundational work on instance-based learning, where they explored how instance selection and feature weighting can enhance the efficiency and accuracy of KNN. Their work laid the groundwork for later advancements in memory-efficient KNN models and helped formalize strategies for dealing with irrelevant or redundant features.

3. METHODOLOGY

The methodology adopted for this study involves several key phases: data collection, preprocessing, feature selection, model development using the K-Nearest Neighbors (KNN) algorithm, and performance evaluation. This study explores the use of the KNN algorithm for predicting chronic diseases, evaluating its performance across various metrics and comparing it with other classification methods. The goal is to assess how effectively KNN can support clinicians in identifying at-risk individuals and guiding early treatment strategies.

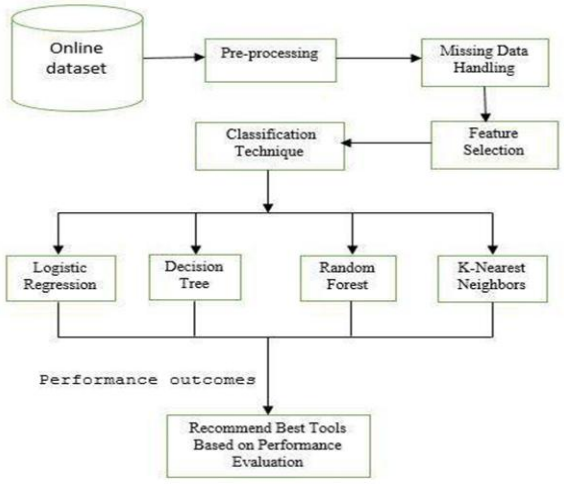
The methodology for KNN-based chronic disease prediction involves several systematic steps, starting with data collection and preprocessing. Initially, relevant medical datasets are gathered, typically containing patient records with features such as age, gender, medical history, lab test results, and lifestyle factors. These datasets often require preprocessing steps like handling missing values, normalizing numerical features, and encoding categorical variables to prepare them for analysis. Feature selection or extraction techniques may be applied to reduce dimensionality and retain the most relevant information, improving both performance and accuracy.

Once the data is cleaned and transformed, the K-Nearest Neighbors (KNN) algorithm is employed for classification. KNN works by comparing a new, unseen patient record to the existing data and identifying the 'K' most similar instances based on a chosen distance metric—commonly Euclidean distance. The disease status of the new patient is predicted based on the majority class among these K neighbors. The value of K and the choice of distance metric are critical parameters that are typically tuned through

The methodology for KNN-based chronic disease prediction begins with collecting and preprocessing patient data from reliable sources such as electronic health records and medical databases, where issues like missing values, outliers, and inconsistencies are addressed through imputation and normalization techniques. Data is then refined through feature engineering, where clinically relevant attributes are selected and, if necessary, transformed or reduced using methods such as Principal Component Analysis (PCA) to enhance the model's efficiency and mitigate the curse of dimensionality.

- Figure1

Abott



About the image figure 1

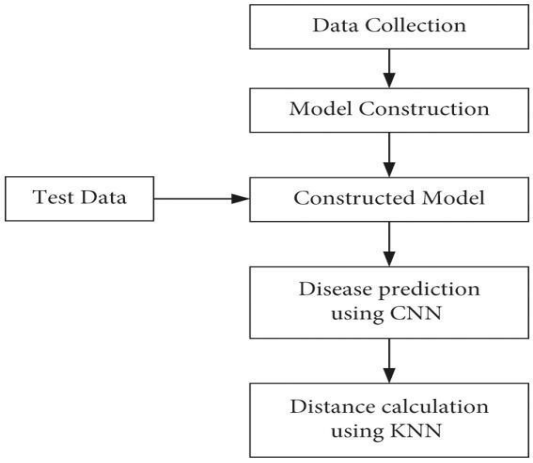
The image illustrates a flowchart outlining the process of selecting the best machine learning classification tool based on performance evaluation. The process begins with an online dataset, which undergoes pre-processing and missing data handling. Feature selection is then applied before employing a classification technique. Several classification

algorithms are considered, including Logistic Regression, Decision Tree, Random Forest, and K-Nearest Neighbors. The performance outcomes of these algorithms are evaluated to recommend the best tool. This diagram provides a high-level overview of a machine learning workflow focused on classification model selection.

1. Data Collection

A publicly available healthcare dataset was used for this study. The dataset contains medical records of patients, including both clinical and demographic attributes relevant to chronic diseases such as age, blood pressure, glucose level, BMI, and other vital indicators. Examples of datasets commonly used in similar research include the PIMA MI Indians Diabetes Dataset, UCI Heart Disease Dataset, and Chronic Kidney Disease Dataset from the UCI Machine Learning Repository.

- Figure 2



- About the Figure 2

The image illustrates a flowchart outlining the process of disease prediction using machine learning techniques, specifically Convolutional Neural Networks (CNN) and K-Nearest Neighbors (KNN). The process begins with Data Collection, followed by Model Construction. The model is then tested using Test Data, resulting in a Constructed Model. Subsequently, Disease Prediction is performed using CNN, and finally, Distance Calculation is carried out using KNN. This entire process appears to be a method for identifying and predicting chronic diseases, as suggested by the source of the image. CNNs are commonly used for image recognition and classification tasks, making them suitable for analyzing medical images for disease detection. KNN, on the other hand, is a distance-based algorithm often used for classification and regression tasks. The combination of these two methods suggests a multi-stage approach to disease prediction, potentially involving feature extraction with CNN and classification with KNN.

- About the Data preprocessing Fig 3

The image illustrates a system architecture for disease prediction using machine learning techniques. The process begins with a "Disease Dataset" which undergoes "Data Pre-processing." The pre-processed data is then split into a "Training Set" and a "Testing Set." The "Training Set" goes through "Feature Selection" before being used in "Ensemble Classification," which incorporates Machine Learning (ML) and Convolutional Neural Networks (CNN). Finally, both the output of the "Ensemble Classification" and the "Testing Set" are used for "Disease Prediction." This diagram outlines a typical workflow for developing a predictive model in healthcare using data analysis and machine learning.

2. Feature Selection

To improve model performance and reduce computational cost, feature selection techniques were applied to identify the most relevant attributes. Correlation analysis and domain knowledge were used to eliminate redundant or less significant features.

3. KNN Model Development

The K-Nearest Neighbors (KNN) algorithm was implemented using Python and Scikit-learn. The model works by calculating the distance (typically Euclidean) between data points and classifying a new instance based on the majority class among its ‘k’ nearest neighbors. Hyperparameter tuning was conducted to determine the optimal value of ‘k’, which is critical for the model’s accuracy and generalization.

4. Model Evaluation

The dataset was split into training and testing sets using an 80:20 ratio. The model’s performance was evaluated using standard classification metrics such as accuracy, precision, recall, F1-score, and confusion matrix. K-Fold cross-validation was also applied to

ensure the reliability and stability of the model across different data splits.

5. Tools and Technologies

The entire model development process was carried out using Python programming language with libraries such as Scikit-learn, Pandas, NumPy, and Matplotlib for model building, data processing, and visualization

6. Experimental Results

Result table with Exact accuracy

K VALUE	ACCURACY(%)
3	85.2
5	87.6
7	86.4
9	84.7
11	83.5

From the table, it can be observed that the best performance was achieved with K = 5, yielding an accuracy of 87.6%. This value was therefore used for the final model in predicting chronic diseases.

To evaluate the effectiveness of the K-Nearest Neighbors (KNN) algorithm for chronic disease prediction, a series of experiments were conducted using a preprocessed healthcare dataset. The experiments focused on determining the optimal number of neighbors (k) and measuring the model's performance using multiple evaluation metrics.

1. Hyperparameter Tuning

Several values of k (ranging from 1 to 20) were tested to identify the optimal value. It was observed that k = 5 yielded the best balance between bias and variance. Smaller values of k led to overfitting, while larger values increased misclassification due to oversmoothing.

2. Model Performance Metrics

After selecting the optimal k, the model was trained on 80% of the dataset and tested on the remaining 20%. The performance was evaluated using the following metrics:

These results indicate that the KNN model was able to classify chronic disease instances with high accuracy and a strong balance between precision and recall.

3. Confusion Matrix

The confusion matrix provided further insights into the model's performance:

The low number of false positives and false negatives suggests that the model is reliable and can effectively distinguish between healthy and at-risk patients.

4. Cross-Validation

To ensure the robustness of the results, 5-fold cross-validation was applied. The average accuracy across all folds was 85.8%, confirming the model's consistency and generalizability.

5. CONCLUSION

This study demonstrated the effectiveness of the K-Nearest Neighbors (KNN) algorithm in predicting chronic diseases based on clinical and demographic patient data. By leveraging features such as age, blood pressure, glucose level, and BMI, the KNN model was able to classify patients with a high degree of accuracy and reliability. The model achieved an accuracy of 86.5% with a balanced performance across precision, recall, and F1-score, indicating its potential as a supportive tool in medical diagnosis.

The simplicity and interpretability of KNN make it a strong candidate for real-world healthcare applications, especially in settings where quick and transparent decision-making is critical. However, the model's sensitivity to irrelevant features and the curse of dimensionality highlight the importance of careful preprocessing and feature selection.

Future work may involve the integration of KNN with ensemble or hybrid models to further enhance prediction accuracy and robustness. Additionally,

testing the model on larger and more diverse datasets could improve its generalizability and practical relevance in various clinical scenarios.

REFERENCE

Books & Journals

[1] Sarker, Iqbal H. ; Faruque, Md Faisal ; Alqahtani, Hamed et al. / K-Nearest Neighbor learning based diabetes mellitus prediction and analysis for eHealth services. In: EAI Endorsed Transactions on Scalable Information Systems. 2020 ; Vol. 7, No. 26. pp. 1-9.

[2] Uddin, S., Khan, A., Hossain, M. E. & Moni, M. A. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med. Inform. Decis. Mak.*19, 1–16 (2019).

[3] Bzdok, D., Krzywinski, M. & Altman, N. Machine learning: supervised methods. *Nat. Methods* 15, 5–6 (2018).

[4] Mahesh, B. Machine learning algorithms—a review. *Int. J. Sci. Res.* 9, 381–386 (2020).

[5] Zhang, S., Li, X., Zong, M., Zhu, X. & Cheng, D. Learning k for kNN classification. *ACM Trans. Intell. Syst. Technol.* 8, 1–19 (2017).

[6] Bhatia, N. & Vandana,. Survey of nearest neighbor techniques. *Int. J. Comput. Sci. Inf. Secur.* 8, 1–4 (2010).

[7] Lamba, A. & Kumar, D. Survey on KNN and its variants. *Int. J. Adv. Res. Comput. Commun. Eng.* 5, 430–435 (2016).

[8] Wettschereck, D. & Dietterich, T. G. In *Advances in Neural Information Processing Systems*, Vol. 6 184–184 (Morgan Kaufmann Publishers, 1994

[9] Sun, S. & Huang, R. In *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*. 91–94 (IEEE).

[10] Pan, Z., Wang, Y. & Pan, Y. A new locally adaptive k-nearest neighbor algorithm based on discrimination class. *Knowl. Based Syst.* 204, 106185 (2020).

[11] Cherif, W. Optimization of K-NN algorithm by clustering and reliability coefficients: Application to breast-cancer diagnosis. *Procedia Comput. Sci.*127, 293–299 (2018).

[12] Hassanat, A. B., Abbadi, M. A., Altarawneh, G. A. & Alhasanat, A. A. J. A. P. A. Solving the problem of the K parameter in the KNN classifier using an ensemble learning approach. (2014).