Intelligent Disease Prediction and Diet Recommender

[1]Ronit Bagga, [2]Vikash Sharma, [3]Upkar Dobriyal, [4]Yash Makhija

[1]B.Tech (Computer Science and Engineering),
[1]JIMS Engineering Management Technical Campus (JEMTEC), Greater Noida, India
[1]ronbagga123@gmail.com, [2]vikash082003@gmail.com, [3]upkardobriyal27.7.2002@gmail.com, [4]yashmakhija1837@gmail.com

***Abstract*— The rising incidence of lifestyle-related diseases, including Diabetes, Heart Disease, and Parkinson's Disease, underscores the necessity for early detection and preventive strategies. This study presents a disease prediction system that has been machine learning algorithms. The system meticulously developed using core Python and customized estimates the probability of these conditions by examining health metrics that are affected by dietary and lifestyle factors. Each model is trained on meticulously curated datasets, employing pre-processing methods such as normalization and label encoding. The evaluation of performance is carried out using metrics such as accuracy, precision, recall, and F1-score. Furthermore, the system promotes dietary awareness and personalized healthcare, merging AI-driven predictions with health-focused design to motivate proactive self-care.**

Index Terms*— *Disease Prediction, Diabetes, Heart Disease, Parkinson's, Nutrition, Machine Learning from Scratch, Preventive Healthcare, Personalized Health Insights.

I. INTRODUCTION

Chronic conditions such as Diabetes, Heart Disease, and Parkinson's Disease rank among the foremost contributors to illness and death globally [1][4][5]. Their rising incidence is closely associated with lifestyle factors—especially unhealthy eating patterns, lack of physical activity, and insufficient health awareness [6][8]. Given that these diseases often progress insidiously over time, timely diagnosis and proactive lifestyle changes are essential for mitigating long-term health risks and enhancing overall quality of life [9][10]. In this regard, advanced digital solutions can significantly improve personal health management [7][13]. This research introduces HealthMate, a web application powered by machine learning, aimed at forecasting the risk of three primary chronic diseases—Diabetes, Heart Disease, and Parkinson's Disease—while also offering tailored dietary and exercise suggestions. Built with Streamlit [13], HealthMate provides a dynamic, user-friendly interface that makes health predictions both easy to understand and implement for users in their dailylives.

HealthMate integrates machine learning [1][5][13], user-centered design, and an understanding of lifestyle factors [3] to not only forecast illnesses but also to encourage preventive healthcare via tailored, smart interventions [2][7]. This document outlines the design, execution, and assessment of the HealthMate system [13], highlighting its capacity to facilitate early detection [3][4], enhance health literacy [2], and empower individuals in their pursuit of wellness [3].

II. RELATED WORK

In the last ten years, machine learning has emerged as a fundamental element in the creation of predictive healthcare tools, enabling the detection of patterns in clinical data that conventional statistical approaches frequently miss [1][4]. A significant number of studies have concentrated on predicting specific diseases—especially Diabetes, Heart Disease, and Parkinson's Disease—by utilizing structured datasets and employing supervised learning methodologies [4][5][9]. One of the most prominent datasets for diabetes prediction is the PIMA Indians Diabetes [4]. Researchers have applied various classification models to this dataset, including Logistic Regression, Decision Trees, and Support Vector Machines , achieving varying levels of accuracy based on the preprocessing and feature engineering techniques implemented [1][5][14]. Additionally, ensemble methods such as Random Forest and XGBoost have been utilized to enhance generalization and more effectively manage feature interactions [9][15]. Likewise, the UCI Heart Disease Dataset has played a crucial role in modeling cardiovascular risk [4]. Research employing Decision Trees, Gradient Boosting, and Neural Networks has shown the potential for non-invasive diagnosis based on factors like age, cholesterol levels, resting ECG results, and maximum heart rate [4][5], [9][16]. Although these models often demonstrate high performance, many suffer from a lack of transparency, complicating their interpretation in clinical use [4][5].

III. Proposed system

The HealthMate system is an advanced modular web application aimed at the early detection of Diabetes, Heart Disease, and Parkinson's Disease through specialized machine learning pipelines [13]. Each predictive model is independently developed using structured datasets, including the PIMA Indians Diabetes Dataset for diabetes prediction [4], the UCI Cleveland Heart Disease Dataset for heart disease [13], and a voice-based dataset for detecting Parkinson's Disease [13]. The system incorporates customized preprocessing techniques such as data cleaning, feature scaling, and normalization, followed by model training utilizing classifiers such as Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, and Random Forest [14][15]. For heart disease, a specifically designed Random Forest model, augmented with Gradient Boosting and XGBoost, was chosen based on its performance [15], while for Parkinson's Disease, a Multi-Layer Perceptron (MLP) model and bespoke neural networks were employed [16]. After training and assessment using metrics like accuracy, F1-score, and ROC-AUC [4][5] the most effective models are deployed via Pickle [13]. The application's front-end, developed with Streamlit, offers an intuitive interface for real-time predictions based on user input [13], generating personalized health recommendations such as diet and exercise plans, customized according to disease type, gender, and dietary preferences [7]. With its seamless user experience, HealthMate not only provides predictive analytics but also encourages preventive care through actionable insights, establishing it as a comprehensive tool for healthcare management [1][3].

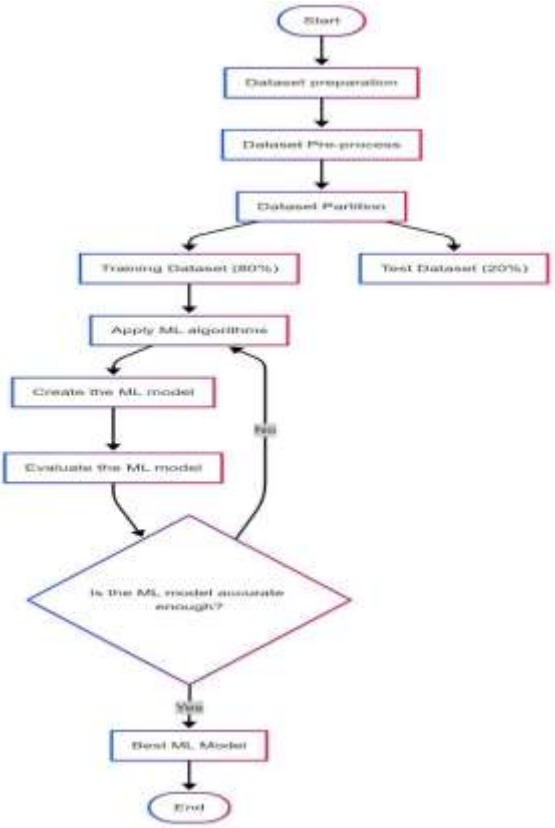

Fig.1. System Architecture of Prediction System

A. DATA PREPROCESSING

The PIDD repository and comprises diagnostic dataset was obtained from a publicly accessible measurements from female patients of Pima Indian descent aged 21 and older [4]. Each entry includes various clinical attributes, such as the number of pregnancies, plasma glucose levels, diastolic blood pressure, triceps skinfold thickness, serum insulin, body mass index (BMI), diabetes pedigree function, and age [4]. The target variable indicates whether diabetes is present or absent [4]. To enhance data integrity, rows containing biologically implausible zero values (e.g., BMI or blood pressure) were treated as missing data [5]. Imputation was conducted using mean or median values based on the distribution of the features [5]. Furthermore, duplicate entries were eliminated, and outliers were addressed using the Interquartile Range (IQR) method to reduce skewness, particularly in the insulin and glucose variables [5]. The selection of predictive features was guided by both clinical significance and statistical methods [5]. Features exhibiting weak correlation or high collinearity were removed to decrease dimensionality and improve model interpretability [5]. Recursive Feature Elimination (RFE) was also utilized during the model tuning process to identify the most important predictors [14]. Although the PIDD dataset mainly consists of numerical data, categorical transformations were required for other disease models (e.g., Parkinson's and Heart Disease) [13]. When relevant, categorical variables such as gender or hereditary factors were label encoded [13]. For models that benefited from a more comprehensive feature representation, one-hot encoding was applied selectively [13]. Given the diverse scales of the input variables (e.g., age compared to insulin levels), all numerical features were standardized using StandardScaler to ensure uniform magnitude and convergence during gradient-based model training [14]. This step was particularly crucial for algorithms sensitive to input scales, such as Support Vector Machines and neural networks [14][16].

Categorical variables, including disease classifications and specific medical conditions, were encoded using either Label Encoding or One-Hot Encoding based on the characteristics of the variable [5][13]. This encoding was essential for converting non-numeric data into a format compatible with machine learning algorithms [6]. Furthermore, outliers within the dataset, particularly in features such as cholesterol levels, BMI, and age, were identified and addressed using Interquartile Range (IQR) methods or by capping values at suitable thresholds [4][5]. This approach was vital in preventing extreme values from distorting model predictions [5]. Another important preprocessing step was feature scaling. Algorithms like K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Neural Networks (MLP) are influenced by the scale of input features; therefore, the data was standardized using the StandardScaler from Scikit-learn [6][8]. This process adjusted the features to have a mean of zero and a standard deviation of one, facilitating equitable comparisons and consistent model performance across all features [5][6]. To evaluate model generalization, the dataset was divided into training (80%) and testing (20%) subsets, ensuring that the model's performance could be assessed on unseen data [13]. K-Fold cross-validation (with k=5) was also implemented in certain models to add an extra layer of validation and mitigate overfitting [5][13]. The training data was utilized to develop the models, while the testing set was employed to assess the model's generalization abilities [6]. These preprocessing measures guaranteed that the data input into the models was both robust and well-prepared, enhancing performance while reducing biases. By consistently applying these preprocessing techniques across all models for Diabetes, Heart Disease, and Parkinson's Disease, the system ensured that the features were adequately scaled, cleaned, and primed for training, leading to more precise predictions and an improved user experience in the final web interface [5][9].

B. *MODEL ALGORITM*

The HealthMate system utilizes a variety of machine learning algorithms specifically designed for predicting three chronic conditions: Diabetes, Heart Disease, and Parkinson's Disease [6][9]. Each model is tailored to the unique characteristics of its corresponding dataset, employing a combination of traditional machine learning techniques and neural network methodologies [5][13]. For the prediction of Diabetes, a comparative modeling strategy was adopted, incorporating multiple algorithms from the Scikit-learn library. The implemented models include Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, Random Forest, Gradient Boosting, and XGBoost [5][6]. These algorithms were chosen for their established efficacy in classification tasks, and each model was trained on a preprocessed dataset with consistent hyperparameters where applicable [4][6]. The evaluation of these models was conducted using performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to identify the most dependable predictive model for diabetes risk [6][7]. Notably, ensemble methods like Random Forest and XGBoost exhibited enhanced generalization capabilities, effectively managing feature interactions and minimizing overfitting [5][6].

The Heart Disease prediction model employed a hybrid approach that combined standard Scikit-learn algorithms with custom implementations developed from the ground up [6], [10]. Initially, a Decision Tree classifier was created using Scikit-learn to set a performance benchmark [5]. Subsequently, a Random Forest model was manually coded to enhance control over the construction of trees, the aggregation of ensembles, and the interpretability of results [5][15]. This model was further refined with an optimized variant that utilized improved splitting criteria and additional depth restrictions [6]. Ensemble techniques such as Gradient Boosting and XGBoost were also applied to evaluate their performance against the custom-built models [5][9]. These algorithms demonstrated high accuracy and robustness in predictions, particularly due to their capability to manage complex, non-linear relationships within the heart disease dataset [9][10]. The custom implementation contributed to a clearer understanding of the learning process, enhancing transparency regarding model behavior [6]. For the prediction of Parkinson's Disease, the project adopted neural network architectures, given the voice dataset's high-dimensional and pattern-rich characteristics [6]. A Multi-Layer Perceptron (MLP) was implemented using Scikit-learn to act as a standard reference model [5]. To provide a more profound and educational perspective on neural learning, a custom neural network was constructed from scratch using NumPy, which allowed for complete control over weight initialization, activation functions, forward propagation, and backpropagation logic [16]. Furthermore, a scalable implementation utilizing PyTorch was integrated to accommodate larger training iterations with improved optimization techniques [16]. This model leveraged PyTorch's built-in support for GPU acceleration and dynamic computation graphs [16]. The custom-built network performed competitively, offering transparency in its operations, while the PyTorch model achieved high accuracy and faster convergence [16].

IV. RESULT AND DISCUSSION

The evaluation of the machine learning models created for the HealthMate system was conducted primarily using accuracy as the key metric, supplemented by other standard classification metrics during both training and validation phases [4][5]. In the case of the Diabetes prediction model, the Gradient Boosting Classifier attained the highest accuracy at 91.45%, followed closely by the Support Vector Machine (SVM) at 90.79%, and both Logistic Regression and Random Forest at 89.47% [9][6]. Although K-Nearest Neighbors (KNN) and XGBoost demonstrated commendable performance with accuracies of 88.16% and 87.50% respectively, the Decision Tree Classifier recorded the lowest accuracy at 84.87% [10][9]. This performance trend highlights the effectiveness of ensemble methods in addressing complex relationships within the diabetes dataset [5][10].

In the realm of Heart Disease prediction, the Random Forest model (Sklearn) demonstrated the highest accuracy at 83.51% [15], outpacing the Improved Random Forest which achieved 81.31% and XGBoost at 80.21% [14]. The baseline Gradient Boosting model recorded an accuracy of 79.12% [14]. Furthermore, models developed from scratch, including Decision Tree Scratch and Random Forest Scratch, achieved accuracies of 74.72% and 75.82% respectively, reflecting a slight performance trade-off for the sake of transparency and educational insights [14]. The standard Decision Tree model, executed via Scikit-learn, attained an accuracy of 76.92%, slightly surpassing that of its manually implemented counterpart [14].

In the assessment of Parkinson's Disease prediction, various neural network models were analyzed, including a Multi-Layer Perceptron (MLP) utilizing Scikit-learn [14], a bespoke neural network developed from the ground up with NumPy [16], and a deep learning model executed in PyTorch [16]. The model that exhibited the highest accuracy was chosen as the most effective, underscoring the capability of neural networks to manage high-dimensional biomedical voice data [16]. The impressive results of this model affirm the significance of sophisticated architectures in identifying nuanced voice-related symptoms associated with Parkinson's Disease [16].These findings collectively underscore the significance of selecting models according to disease attributes, the complexity of features, and the characteristics of the input data [14]. Ensemble techniques and neural networks exhibited considerable effectiveness [1][5][15][16], whereas models developed from the ground up offered a more profound insight into algorithmic behavior [16]. Overall, the system showcases robust predictive abilities, emphasizing the promise of HealthMate as a real-time decision support system for early disease identification and tailored healthcare suggestions [7][13].

Disease	Model	Accuracy (%)
Diabetes	Gradient Boosting Classifier	91,45
	Support Vector Machine (SV)	90,79
	Logistic Regression	89,47
	Random Forest	89,47
	XGBoost	88,16
Heart Disease	Decision Tree Classifier	83,51
	Improved Random Forest	81,31
	XGBoost	80,21
	Decision Tree (Sklearn)	79,12
Parkinson's	Random Forest (Scratch)	76,92
	Decision Tree (Scratch)	74,72
	Multi-Layer Perceptron (MLP)	95,67

Fig. 2. Results the Machine Learning Models

V. FUTURE SCOPE

The HealthMate system provides a robust basis for disease prediction [13]; however, there are numerous opportunities for improvement. By integrating real-time data from wearable devices, continuous monitoring and adaptive health evaluations can be achieved [7]. Broadening the dataset to include larger and more varied populations will enhance the model's applicability across diverse groups [4][9]. The inclusion of explainable AI (XAI) can foster user confidence by rendering predictions more understandable [14]. Future versions may also incorporate voice interaction, additional disease modules, and tailored health recommendations, transforming HealthMate into a more holistic and user-centric resource for preventive healthcare [13].

VI. CONCLUSION

This research introduces HealthMate, an all-encompassing and intuitive machine learning-based platform designed for the early detection of Diabetes, Heart Disease, and Parkinson's Disease [13]. By integrating structured medical datasets with various algorithms—including custom-developed models—the platform achieves notable prediction accuracy while ensuring transparency and educational benefits [4][5][6]. The incorporation of a Streamlit web application further improves accessibility, enabling users to obtain real-time predictions and tailored lifestyle suggestions based on their inputs [13]. The findings underscore the efficacy of ensemble and neural network models in disease identification [14][15][16], and emphasize the significance of clean data, careful preprocessing, and algorithm selection [5][6]. In summary, HealthMate plays a vital role in the expanding domain of AI-enhanced healthcare, facilitating early diagnosis and empowering individuals to take proactive measures towards improved health [13].

REFERENCES

[1] Zhang, Y. (2023). Machine learning and deep learning techniques in diabetes prediction. Theoretical and Natural Science, 13, 150-154.

[2] NHS to begin world-first trial of AI tool to identify type 2 diabetes risk. (2024). The Guardian..

[3] Gene screening can cut early disease deaths by 25%, study shows. (2024). Financial Times

[4] Chauhan, A.S., Varre, M.S., Izuora,K., Trabia, M.B., & Dufek, J.S. (2023). Prediction of Diabetes Mellitus Progression Using Supervised Machine Learning. Sensors, 23(10), 4658

[5] Patro, K.K., Allam, J.P., Sanapala, U., et al., 2023. An effective correlation-based data modeling framework for automatic diabetes prediction using machine and deep learning techniques.

[6] Ambuja, K., Anusha, G.S., Sangeetha, T.D., et al., 2023. Early Diabetes Prediction Using Machine Learning. International Journal of Progressive Research in Science and Engineering, 4(5), pp.24–28. IJPRSE

[7] El-Sofany, H.F., 2024. A Proposed Technique Using Machine Learning for the Prediction of Diabetes Disease through a Mobile App. International Journal of Intelligent Systems, 2024,6688934. Wiley Online Library

[8] Gupta, P. and Sindhu, R., 2024. Diabetes Prediction Using Machine Learning. Journal of Electrical Systems, 20(7s). Journal of Electrical Systems

[9] Reddy, M.K.K., Bhavani, G., Moulika, M.V., et al., 2023. A fused machine learning technique for diabetes prediction. International Journal of Advance Research, Ideas and Innovations in Technology, 9(1). IJARIIT

[10] Agrawal, A.J., Welekar, R.R., Parati, N., et al., 2023. Diabetes Prediction Using Medical Data and Disease Influence Measures using Machine Learning. International Journal of Intelligent Systems and Applications in Engineering, 11(10s), pp.01–10. IJISAE

[11] Kharat, S., Taur, S., Khalate, R., et al., 2023. Detection of Credit Card Fraud using Machine Learning and Deep Learning: A Review. International Journal of Progressive Research in Science and Engineering, 4(5).IJPRSE

[12] Ghosh, D.R., 2023. Commodity Price Forecasting in the International Market: Using a Proposed Ensemble Approach, Time Series and Machine Learning Models. International Journal of Progressive Research in Science and Engineering, 4(8).IJPRSE

[13] Gopisetti, L. D., Kummera, S. K. L., Pattamsetti, S. R., Kuna, S., Parsi, N., & Kodali, H. P. (2023, January). Multiple disease prediction system using machine learning and streamlit. In 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 923-931). IEEE.

[14] Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: a review of classification and combining techniques. Artificial Intelligence Review, 26, 159-190.

[15] Parmar, A., Katariya, R., & Patel, V. (2018, August). A review on random forest: An ensemble classifier. In International conference on intelligent data communication

[16] Rivas, P. (2020). Deep Learning for Beginners: A beginner's guide to getting up and running with deep learning from scratch using Python. Packt Publishing Ltd.