AI-driven Predictive Health and Disease Risk Modeling

Abstract-- Artificial Intelligence has revolutionized healthcare by enabling predictive modeling for early disease detection and risk assessment. This study explores the development of AI-driven models designed to analyze clinical, genomic, and lifestyle data for diseases such as heart conditions, diabetes, asthma, and sleep disorders. Using sophisticated machine learning methods, such as ensemble methods like Random Forest, Gradient Boosting, and AdaBoost, the models are highly accurate and interpretable, allowing for timely interventions. The research incorporates explainable AI tools like LIME and SHAP to foster trust among clinicians and patients. Comprehensive preprocessing, feature engineering, and hyperparameter tuning ensure robust a model performance. Results indicate the high capability of ensemble models, with notable accuracy improvements in predicting heart disease (91.3%) and diabetes (97%). Additionally, models tailored for sleep disorders and other conditions emphasize computational efficiency and ethical considerations. This work contributes to a data-informed healthcare future, promoting proactive medical care and enhanced patient outcomes.

Keywords-- Artificial Intelligence, Predictive Modeling, Machine Learning, Disease Risk Assessment, Heart Disease Prediction, Diabetes Prediction, Explainable AI, Ensemble Methods, Gradient Boosting, Data-Driven Healthcare, Proactive Medical Care, Healthcare Analytics.

INTRODUCTION

Artificial Intelligence has become a transformative technology in healthcare, bridging critical gaps in early disease detection and personalized medical interventions. By leveraging advanced machine learning and deep learning algorithms, AI-driven predictive health systems analyze vast datasets comprising clinical, genomic, and lifestyle information to identify potential health risks proactively. Because of this capability, the accuracy of diagnosing patients is improved while also allowing the caregiver accuracy to implement timely and effective treatment strategies.

The potential of such systems lies in their ability to process real-time health metrics, such as heart rate, blood pressure, and oxygen levels, from wearable devices and sensors. These metrics, when analyzed through robust algorithms Random Forest, Support Vector Machine (SVM), and Gradient

Boosting, can be used to glean insights into health conditions, such as cardiovascular diseases, respiratory failures, and neurological disorders. Furthermore, tools like Explainable AI (XAI), including LIME and SHAP, ensure model transparency, fostering trust among medical professionals and patients.

However, challenges persist in scalability, dataset diversity, and the computational demands of these systems. This study explores the development of an AI-driven predictive model, focusing on accuracy, interpretability, and clinical applicability. By incorporating ensemble techniques and hyperparameter tuning, the research focuses on improving the predictive reliability while addressing ethical considerations and privacy concerns in healthcare data usage. This work contributes to a vision of a healthier, data-informed society, where proactive healthcare interventions become the norm.

RELATED WORK

Artificial Intelligence has revolutionized healthcare by enabling early disease detection and improving predictions with advanced forms of machine learning and deep learning techniques. Researchers have developed reliable predictive systems for various health conditions using diverse methodologies. G. K. Thakur et al. [1] proposed a model integrating clinical, genomic, and lifestyle data with Random Forest, Gradient Boosting Techniques, and Deep Learning, achieving high accuracy while emphasizing the importance of diverse datasets and computational resources. W. P. Siraskar et al. [2] utilized Support Vector Machines (SVM) for cardiovascular disease detection, achieving a 96.5% accuracy with robust diagnostic metrics. Z. Zhu et al. [3] introduced a Knowledge-based Attention Network (KAN) that combined Graph Neural Networks (GNN) and knowledge graphs to address feature sparsity in Electronic Medical Records (EMRs). S. Srivastava et al. [4] developed a Bagging-Fuzzy-GBDT model for heart disease severity classification, highlighting its multi-class classification capability but noting scalability challenges. L. Dewangan et al. [5] employed Random Forest and SVM for coronary heart disease risk prediction, focusing on interpretability and key risk factors such as age and cholesterol. A. Sharma et al. [6] investigated symptom-driven disease risk assessment using SVM and Long Short-Term Memory (LSTM),

advancing personalized treatment and early diagnosis. P. D. K and M. S. Abirami [7] developed an AI Clinical Decision Support System (AI-CDSS) leveraging CNNs and RNNs for cardiovascular disease management, stressing ethical and regulatory compliance. K. Mittal et al. [8] explored classifiers like K-Nearest Neighbors (KNN) and Random Forest for predicting Parkinson's disease, where KNN achieved the highest accuracy at 94%. C. S. Ganesh et al. [9] used ML models - Random Forest, and Linear Regression for disease prediction and lifestyle monitoring, achieving high accuracy in predicting diabetes and breast cancer. A. Sethi et al. [10] emphasized Explainable AI (XAI) techniques such as LIME and SHAP for heart disease prediction, achieving a 96.07% accuracy and enhancing interpretability for clinical applications. These studies collectively demonstrate AI's transformative potential in improving predictive accuracy, personalized treatment, and proactive healthcare management while highlighting challenges in scalability, interpretability, and ethical considerations.

METHODOLOGY

Data Collection:

This research used publicly available health datasets, mainly compiled for a range of medical conditions such as heart failure, diabetes, asthma, sleep disorders, among others, that are publicly available through Kaggle. These datasets consist of structured data comprising clinical, demographic, and behavioural health risk factors to make predictions about disease presence and other health outcomes.

Common Features Across Datasets:

Age: Patient?s age in years.

Gender: Biological sex (Male/Female/Other).

BMI: Body Mass Index indicating body fat.

Physical Activity: Weekly physical activity in hours.

Smoking History: Smoking habits categorized into levels.

Stress Level: Subjective stress rating (scale of 1-10).

Disease-Specific Features:

Heart Attack: Features like chest pain type, cholesterol levels, blood pressure, and heart rate.

Diabetes: Features including hypertension, blood glucose level, and HbA1c level.

Asthma: Features like family history of asthma, allergies, wheezing, and shortness of breath.

Sleep Disorders: Features including sleep duration, quality of sleep, daily steps, and blood pressure.

Data Preprocessing

Handling Missing Values:

Identification of Missing Values:

Initially, missing entries are detected by scanning each column in the dataset for null or undefined values. Missing values are detected by checking for NaN or null entries within each column.

Imputation Strategy:

Numerical variables such as BMI, age, and HbA1c are imputed using median values to maintain distribution balance

For categorical features like smoking history, had missing values imputed with the mode (most frequent value) to maintain consistency with the majority category.

Categorical Encoding:

Encoding Categorical Features:

All categorical features like gender, smoking history, and others are encoded to make them interpretable by machine learning models. For binary features like gender, label encoding is used where categories are represented as 0 and 1.

For features with multiple categories, like smoking history, one-hot encoding is applied which will create a separate binary column for each category.

Normalization/Scaling of Numerical Features:

Min-Max Scaling:

For numerical features such as BMI, HbA1c levels, and age, Min-Max scaling is applied. This scaling technique transforms the values into a fixed range, typically between 0 and 1. The formula used is: Scaled Value=X?min(X)max(X)?min(X)\text{Scaled Value} = \frac{X - \text{min}(X)}{\text{max}(X) - \text{min}(X)}Scaled Value=max(X)?min(X)X?min(X)?

Model Training and Evaluation

Model Selection:

We employed a variety of algorithms in order to compute clinical risk for disease, choosing specific methods that were appropriate for binary classifications and which could identify complex relationships that existed among the data. The following models were utilized:

Logistic Regression:

A baseline model that offers interpretability and performs well on linearly separable data. It serves as a benchmark to compare other complex models.

Decision Trees:

Selected for their simplicity and interpretability. Decision Trees are adept at capturing non-linear relationships in the data, although they are prone to overfitting.

Support Vector Machines (SVM):

We explored multiple kernels to enhance the effectiveness of SVM:

Linear Kernel: For linearly separable data.

Radial Basis Function (RBF) Kernel: To capture complex non-linear relationships.

Polynomial Kernel: To model interactions and complex patterns in the data.

Random Forest:

An ensemble learning algorithm to reduce overfitting by combining many Decision Trees trained on random samples of the data. It is great for working with imbalanced data and detecting feature interactions.

K-Nearest Neighbors (KNN):

A non-parametric method included for its simplicity and effectiveness in identifying patterns based on proximity in feature space.

Boosting Algorithms:

Advanced ensemble methods were applied to enhance predictive accuracy. The following boosting techniques were explored:

Gradient Boosting: Used for its iterative approach to minimizing errors and improving predictions.

XGBoost: Known for its speed and performance, XGBoost (Extreme Gradient Boosting) builds on Gradient Boosting with optimizations such as tree pruning, regularization, and parallel processing. Its scalability and ability to handle missing data make it a robust choice for boosting tasks.

AdaBoost: Focuses on misclassified instances in each iteration, making it effective for reducing bias.

Hyperparameter Tuning

To ensure the best performance, hyperparameter tuning was conducted on each model using two approaches:

Grid Search:

An exhaustive search through a specified parameter space to identify the best-performing combination. This method ensures a thorough exploration of potential hyperparameter settings.

Randomized Search:

A randomized sampling approach, allowing for quicker exploration of hyperparameters while maintaining effectiveness in discovering optimal configurations.

Ensemble Learning:

Post-hyperparameter tuning, ensemble methods were applied to further enhance predictive accuracy. If the ensemble model demonstrated higher accuracy or other performance improvements, it was selected as the final model. Otherwise, the original best-performing model is retained.

Performance Metrics:

To comprehensively evaluate the models, we used a combination of the following metrics:

Accuracy: Measures the proportion of correctly predicted instances.

Precision: Assesses the number of true positives among predicted positives, critical for reducing false positives.

Recall (Sensitivity): Measures how well the model catches actual positive cases, crucial in healthcare for capturing all relevant cases.

F1-Score: Combines precision and recall into one score and is useful in situations with class imbalance.

Model Selection

The last model for every disease prediction task was chosen depending on its performance on the major evaluation metrics such as F1-score, precision, recall, and accuracy. These metrics offered a holistic analysis of the capacity of the model to balance right classifications, address class imbalances, and keep false positives or false negatives low.

The model demonstrating the highest overall performance was selected. In cases where combining the predictions of two or more models through ensemble methods showed a meaningful improvement, the ensemble model was chosen. Otherwise, the original best-performing standalone model was retained to ensure simplicity and efficiency.

RESULT

Heart Disease Prediction Results:

TABLE I. Classification Model Evaluation Metrics

Fig. 1. Interpretation of outcomes using LIME

Fig. 2. Interpretation of results using SHAP for feature importance(global)

The Random Forest model obtained the highest accuracy at 90%, while both Logistic Regression and AdaBoost (before tuning) showed an accuracy of 89%. After hyperparameter tuning, AdaBoost's accuracy increased to 91%, with notable improvements in precision (0.93) and recall (0.92) for predicting heart disease (Class 1).

The ensemble model, which combined the three algorithms, resulted in a slight improvement in accuracy to 91.3%, with 94% recall for Class 1, indicating its strong ability to detect heart disease. The balanced performance across Class 1 and Class 0 makes the ensemble model the best approach for heart disease prediction.

Diabetes Prediction Results:

TABLE II. Classification Model Evaluation Metrics

Fig. 3. Interpretation of outcomes using LIME

Fig. 4. Interpretation of results using SHAP for feature importance(global)

The Gradient Boosting model yielded an accuracy of 97%, comparable to other top-performing

models like Random Forest and Decision Tree. It showed excellent precision (0.99), minimizing false positives, which is essential in diabetes prediction. The model also had a solid F1-score (0.81),which indicated a good balance between precision and recall.The recall was slightly lower at (0.69) compared to models like XGBoost and KNN, the overall performance of Gradient Boosting made it the most reliable choice for diabetes prediction. Despite exploring hyperparameter tuning and ensemble methods, the original Gradient Boosting model was retained due to its strong results.

Sleep Disorder Prediction Results:

TABLE III. Classification Model Evaluation Metrics

Fig. 5. Interpretation of outcomes using LIME

Fig. 6. Interpretation of results using SHAP for feature importance(global)

The Random Forest model achieved one of the highest accuracies at 89%, along with Gradient Boosting and XGBoost, all demonstrating strong and consistent performance. The recall and F1-score for these models (0.84 and 0.85 respectively) indicate a reliable ability to detect sleep disorders. Despite evaluating hyperparameter tuning, the tuned Random Forest model showed minimal gains over the default configuration, suggesting that the original model was already well-optimized. The consistent performance across key metrics, combined with its simplicity and interpretability, makes the default Random Forest model the most effective and efficient choice for sleep disorder prediction in this study.

PCOS Prediction Results:

TABLE IV. Classification Model Evaluation Metrics

Fig. 7. Interpretation of outcomes using LIME

Fig. 8. Interpretation of results using SHAP for feature importance(global)

The Gradient Boosting model yeilded the highest accuracy of 90%, along with strong precision (0.86) and recall (0.83), leading to an F1-score of 0.85 for detecting PCOS (Class 1). Random Forest and XGBoost performed well with accuracies of 89% and 88% respectively, but showed slightly lower recall and F1-scores compared to Gradient Boosting. In contrast, models like SVM and KNN showed significantly lower recall and F1-scores, limiting their effectiveness. Overall, the

Gradient Boosting model was deemed the most effective approach to predict PCOS, largely due to its superior and balance performance across all metrics presented.

CONCLUSION

This work indicates the effectiveness of AI models to enhance early identification and forecast multiple health issues such as heart conditions, diabetes, PCOS, and sleeping problems. With various supervised machine learning classifiers, including Random Forest, Logistic Regression, Gradient Boosting, and ensemble models being used, our work realized accuracy, precision, and recall measures in all of the clinical disciplines

Among the models tested, ensemble and tuned classifiers consistently delivered strong predictive performance, especially for heart disease and diabetes. In addition, the use of explainable AI tools like SHAP and LIME further enabled transparent and interpretable predictions, making the models more trustworthy for real-world medical applications.

Our research has demonstrated how Artificial Intelligence can augment healthcare by supporting clinicians with data-driven insights and early diagnosis capabilities. Future work can focus on integrating real-time patient data, expanding to more diseases, and incorporating deep learning models for further accuracy enhancement.

Ultimately, AI-powered predictive systems represent a promising direction for preventive healthcare, offering personalized risk assessments and contributing to timely medical interventions.

REFERENCES

G. K. Thakur, N. Khan, H. Anush, and A. Thakur, "AI-Driven Predictive Models for Early Disease Detection and Prevention," 2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS), Chikkaballapur, India, 2024, pp. 1-6, doi: 10.1109/ICKECS61492.2024.10616851.

W. P. Siraskar et al., "Enhancing Cardiovascular Health through AI-Driven Heart Disease Detection," 2024 International Conference on Electrical Electronics and Computing Technologies (ICEECT), Greater Noida, India, 2024, pp. 1-6, doi: 10.1109/ICEECT61758.2024.10738991.

Z. Zhu et al., "Development of AI-Driven Disease Prediction Model for Medical Education

Management," 2024 6th International Conference on Internet of Things, Automation and Artificial Intelligence (IoTAAI), Guangzhou, China, 2024, pp. 355-358, doi: 10.1109/IoTAAI62601.2024.10692974.

S. Srivastava et al., "AI-Driven Prediction and Hierarchical Classification of Heart Disease Severity," 2024 International Conference on Electronics, Computing, Communication and Control Technology (ICECCC), Bengaluru, India, 2024, pp. 1-6, doi: 10.1109/ICECCC61767.2024.10593947.

L. Dewangan et al., "Machine Learning-Based Risk Prediction for Coronary Heart Disease Using Clinical Data," 2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI), Raipur, India, 2023, pp. 1-6, doi: 10.1109/ICAIIHI57871.2023.10488969.

A. Sharma et al., "Analysis on Symptoms Driven Disease Risk Assessment using Artificial Intelligence Approach," 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2024, pp. 1-7, doi: 10.1109/ICRITO61523.2024.10522221.

P. D. K and M. S. Abirami, "AI Clinical Decision Support System (AI-CDSS) for Cardiovascular Diseases," 2023 International Conference on Computer Science and Emerging Technologies (CSET), Bangalore, India, 2023, pp. 1-7, doi: 10.1109/CSET58993.2023.10346885.

K. Mittal et al., "Artificial Intelligence Assisted Classifier?s and Neural Network Based Prediction and Classification of Parkinson?s Disease," 2024 4th International Conference on Innovative Practices in Technology and Management (ICIPTM), Noida, India, 2024, pp. 1-6, doi: 10.1109/ICIPTM59628.2024.10563233.

C. S. Ganesh et al., "Data-Driven Disease Prediction and Lifestyle Monitoring System," 2024 1st International Conference on Cognitive, Green and Ubiquitous Computing (IC-CGU), Bhubaneswar, India, 2024, pp. 1-8, doi: 10.1109/IC-CGU58078.2024.10530759.

A. Sethi et al., "Explainable Artificial Intelligence (XAI) Approach to Heart Disease Prediction," 2024 3rd International Conference on Artificial Intelligence For Internet of Things (AIIoT), Vellore, India, 2024, pp. 1-6, doi: 10.1109/AIIoT58432.2024.10574635.