Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 2 (2024)

Detection and Prediction of Comorbities of Diabetes Using Machine Learning Techniques

[1]Teja Sri Dharma Reddy Vanukuri, [2]Sohail Shaik,[3] Bala Akash Mutthavarapu, [4]Sai Teja Naidu Vadranam, [5]K. B. V. Brahma Rao, [6]Dr. V. V. R. Maheswara Rao

[1]C.S.Student, Department of ComputerScience and Engineering, Koneru Lakshmaiah Education, Foundation(KLEF),Vaddeswaram, Guntur Andhra Pradesh

[2]C.S.Student, Department of ComputerScience and Engineering, Koneru Lakshmaiah Education, Foundation(KLEF),Vaddeswaram, Guntur Andhra Pradesh

[3]C.S.Student, Department of ComputerScience and Engineering, Koneru Lakshmaiah Education, Foundation(KLEF),Vaddeswaram, Guntur Andhra Pradesh

[4] C.S.Student, Department of ComputerScience and Engineering, Koneru Lakshmaiah Education, Foundation(KLEF),Vaddeswaram, Guntur Andhra Pradesh

[5] Professor, Department ofComputer Science and Engineering, Koneru Lakshmaiah Education, Foundation(KLEF),Vaddeswaram,Guntur, Andhra Pradesh

[6]Professor, Shri Vishnu Engineering College for Women, Department of Computer Science and Engineering, Bhimavaram - 534202, Andhra Pradesh, India

Abstract:- The coexistence of multiple medical conditions, known as comorbidities, significantly impacts the management and prognosis of diabetes. Early detection and prediction of these comorbidities are crucial for improving patient care and outcomes. In this project, machine learning techniques are employed to develop a predictive model for identifying and forecasting comorbidities associated with diabetes. Leveraging a dataset comprising demographic information, medical history, and physiological parameters of diabetic patients, various machine learning algorithms including support vector machines, random forests, and deep learning neural networks are applied. The model aims to detect existing comorbidities and predict the likelihood of developing additional conditions in the future based on current health indicators. Integration of electronic health records and wearable devices enhances real-time monitoring and enables proactive interventions. Challenges such as data privacy, model interpretability, and generalizability across diverse populations are addressed to ensure the reliability and scalability of the proposed approach. This project contributes to advancing healthcare by providing a tool for early detection and prediction of comorbidities in diabetic patients, facilitating personalized treatment strategies and improving overall patient outcomes.

Keywords*:* Comorbidities, Diabetes management, Machine learning techniques, Predictive model, Early detection, Prognosis, Patient care, Outcomes improvement, Support vector machines, Random forests, Deep learning neural networks, Electronic health records, Wearable devices, Real-time monitoring, Proactive interventions, Data privacy, Model interpretability, Generalizability, Healthcare advancement, Personalized treatment strategies.

1. Introduction

Diabetes mellitus is a chronic metabolic disorder characterized by elevated blood sugar levels, which can lead to various complications affecting multiple organ systems. The management of diabetes becomes more complex when patients present with comorbidities, the simultaneous presence of one or more additional medical conditions. Comorbidities such as cardiovascular diseases, hypertension, kidney disease, and neuropathy significantly increase the burden on healthcare systems and pose challenges for effective disease management.

Early detection and prediction of comorbidities in diabetic patients are essential for improving clinical outcomes, reducing healthcare costs, and enhancing the quality of life for affected individuals. Traditional approaches to

comorbidity detection rely on manual assessment by healthcare professionals, which may be subjective, time-consuming, and prone to errors. In contrast, machine learning techniques offer a promising avenue for automating this process and providing more accurate and timely predictions.

This project aims to develop a comprehensive framework for the detection and prediction of comorbidities associated with diabetes using machine learning techniques. By leveraging large-scale datasets containing diverse patient information, including demographic profiles, medical histories, laboratory results, and imaging studies, the project seeks to train robust predictive models capable of identifying existing comorbidities and forecasting the likelihood of developing additional conditions over time.

The primary objectives of the project include:

A. Data Acquisition and Preprocessing: Collecting relevant datasets from electronic health records (EHRs), clinical databases, and other sources. Data preprocessing is used to address missing values, outliers, and inconsistencies, assuring the quality and dependability of input data for machine learning models.

B. Feature Selection and Engineering: Identification of informative features (variables) from the dataset that are most relevant for comorbidity detection and prediction. This involves both domain knowledge and automated feature selection techniques to extract meaningful insights from the data.

C. Model Development: Implementation of machine learning methods, such as support vector machines (SVM), random forests, and deep learning neural networks, to create prediction models. Training and fine-tuning of these models using labeled data, where the presence or absence of comorbidities is indicated.

D. Evaluation and Validation: Assessment of model performance through cross-validation techniques, evaluating metrics such as accuracy, precision, recall, and area under the receiver operating characteristic curve (AUC-ROC). Validation of the models on independent datasets to ensure generalizability and robustness across different patient populations.

E. Integration and Deployment: Integration of the developed models into clinical workflows, electronic health record systems, or other healthcare platforms to facilitate real-time comorbidity detection and prediction. Collaboration with healthcare professionals to ensure seamless integration and usability of the tool in clinical practice.

By achieving these objectives, the project aims to contribute to the advancement of healthcare by providing a reliable and scalable solution for early detection and prediction of comorbidities in diabetic patients. This will enable healthcare providers to adopt proactive strategies for disease management, optimize treatment plans, and improve overall patient outcomes and quality of life.

2. Literature Survey

The detection and prediction of comorbidities in patients with diabetes have been the subject of extensive research in recent years, with a growing emphasis on leveraging machine learning techniques to enhance diagnostic accuracy and prognostic capabilities. This literature review provides an overview of key studies and methodologies in this field, highlighting significant findings, challenges, and emerging trends.

A. Machine Learning Approaches for Comorbidity Detection: Model performance is evaluated using cross-validation approaches, which include measures like as accuracy, precision, recall, and area under the receiver operating characteristic curve (AUC-ROC). Models are validated on diverse datasets to demonstrate their generalizability and robustness across different patient groups.

B. Predictive Modeling and Risk Stratification: Predictive modeling approaches were used to predict the development of comorbidities in diabetic populations.Zhang et al. (2019) developed a deep learning-based model to predict the risk of diabetic retinopathy progression and other ocular complications. Their model integrated clinical data with imaging features, achieving superior predictive performance compared to conventional risk assessment tools.

C. Integration of Multimodal Data Sources: Integrating diverse data sources, including EHRs, genetic profiles, wearable sensor data, and social determinants of health, has emerged as a promising approach for improving comorbidity detection and prediction. Wang et al. (2020) proposed a multimodal deep learning

framework that combines clinical data with genomic information to identify genetic risk factors for diabetic nephropathy. Their study highlighted the potential of integrating omics data for personalized risk stratification.

D. Challenges and Limitations: Despite the progress in machine learning-based approaches, several challenges remain. Data heterogeneity, missing values, and data privacy concerns pose significant barriers to model development and validation. Interpretability of machine learning models is another critical issue, particularly in clinical settings where transparent decision-making is essential for gaining trust and acceptance from healthcare professionals.

E. Web User Interfaces Gradio: In their 2021 study, Silpa and Maheswara Rao present an enriched big data pre-processing model combined with machine learning, offering insights into web user usage behavior. Building on their 2019 research, they extensively explore techniques and technologies for preparing large web datasets, emphasizing a deep understanding of web user behavior.

F. Future Directions and Emerging Trends: Future research directions include the development of interpretable machine learning models, incorporation of causal inference techniques for understanding disease mechanisms, and implementation of federated learning approaches to address data privacy concerns in multicenter studies. Additionally, there is a growing interest in applying reinforcement learning and active learning methods for dynamic treatment optimization and adaptive risk prediction.

In conclusion, machine learning techniques hold great promise for advancing the detection and prediction of comorbidities in diabetes. By integrating multimodal data sources, addressing challenges in model interpretability and data privacy, and exploring novel methodologies, researchers can develop more accurate and clinically relevant tools for personalized disease management and risk stratification.

3. Methodology

A. Data Acquisition and Preprocessing Module:

Objective: Collect relevant datasets from electronic health records (EHRs), clinical databases, and other sources and preprocess the data to handle missing values, outliers, and inconsistencies.

Detailed Explanation:

This module involves accessing various data sources such as EHRs, laboratory databases, and imaging repositories to gather patient information. Data preprocessing techniques such as cleaning, normalization, and imputation are applied to ensure data quality and consistency. Integration of diverse data types (structured and unstructured) into a unified dataset for further analysis.

B. Feature Selection and Engineering Module:

Objective: Identify informative features from the dataset that are most relevant for comorbidity detection and prediction.

Detailed Explanation:

Feature selection techniques like as correlation analysis, recursive feature removal, and feature significance scores are used to determine the most important variables. Domain knowledge and expert advice are used to influence the feature selection and engineering processes. Transformation, aggregation, and interaction are used to create new features that capture complicated connections in data.

C. Model Development Module:

Objective: Implement machine learning algorithms to build predictive models for comorbidity detection and prediction.

Detailed Explanation:

Various machine learning algorithms such as support vector machines (SVM), random forests, gradient boosting machines (GBM), and deep learning neural networks are implemented. Models are trained on the preprocessed dataset using labeled data, where the presence or absence of comorbidities is indicated. Hyperparameter tuning and model optimization techniques are applied to improve model performance and generalization.

D. Evaluation and Validation Module:

Objective: Assess the performance of the developed models through cross-validation techniques and validate them on independent datasets.

Detailed Explanation:

Model performance is measured using measures such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). Cross-validation approaches, such as k-fold cross-validation, are used to measure the models' generalization performance. Models are verified against independent datasets to ensure their robustness and generalizability across patient groups.

E. Integration and Deployment Module:

Objective: Integrate the developed models into clinical workflows, electronic health record systems, or other healthcare platforms for real-time comorbidity detection and prediction.

Detailed Explanation:

Models are deployed into production environments using appropriate software development frameworks and tools. Integration with existing healthcare infrastructure and interoperability with other clinical systems are ensured. User interface design and implementation to provide healthcare professionals with easy access to the predictive models and decision support features.

F. Testing and Performance Monitoring Module:

Objective: Conduct rigorous testing of the deployed system and monitor its performance in real-world clinical settings.

Detailed Explanation:

Comprehensive testing procedures are conducted to identify and address any potential issues or bugs in the system. Performance monitoring mechanisms are implemented to continuously assess the accuracy, reliability, and usability of the system. Feedback from healthcare professionals and end-users is collected and used to iteratively improve the system's functionality and user experience.

G. Training and Support Module:

Objective: Provide training and support to healthcare professionals for the effective use of the system in clinical practice.

Detailed Explanation:

Training sessions are conducted to familiarize healthcare professionals with the system's features, functionalities, and best practices. User documentation, tutorials, and online resources are developed to support self-learning and troubleshooting. Ongoing technical support and assistance are provided to address any questions, issues, or concerns raised by users during system usage.

4. Results

A. Pre-Processing of Data:

```
#Find Missing Values

missing_values = data.isnull().sum().sum()
percentage_missing = (missing_values / data.shape[0]) * 100
print("Percentage of missing values:", percentage_missing)
```

```
Percentage of missing values: 0.0
```

```
#Convert Categorical variables into binary and numeric
label_encoder = preprocessing.LabelEncoder()
data['smoking_history'] = label_encoder.fit_transform(data['smoking_history'])
data['gender'] = label_encoder.fit_transform(data['gender'])
data.head()
```

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	0	80.00	0	1	4	25.19	6.60	140	0
1	0	54.00	0	0	0	27.32	6.60	80	0
2	1	28.00	0	0	4	27.32	5.70	158	0
3	0	36.00	0	0	1	23.45	5.00	155	0
4	1	76.00	1	1	1	20.14	4.80	155	0

Fig 1 Showing the output for pre-processing of data

In the above code, the LabelEncoder from scikit-learn is used to transform the 'smoking_history' and 'gender' columns into numerical values. This facilitates machine learning models' understanding of categorical data.

B. Visualization:

```
# Plot the heatmap
plt.figure(figsize=(10, 8))  # Set the figure size as desired
sns.heatmap(corr_matrix, annot=True, fmt=".2f", cmap='coolwarm', cbar=True, square=True)
plt.title('Correlation Matrix Heatmap')
plt.show()
```

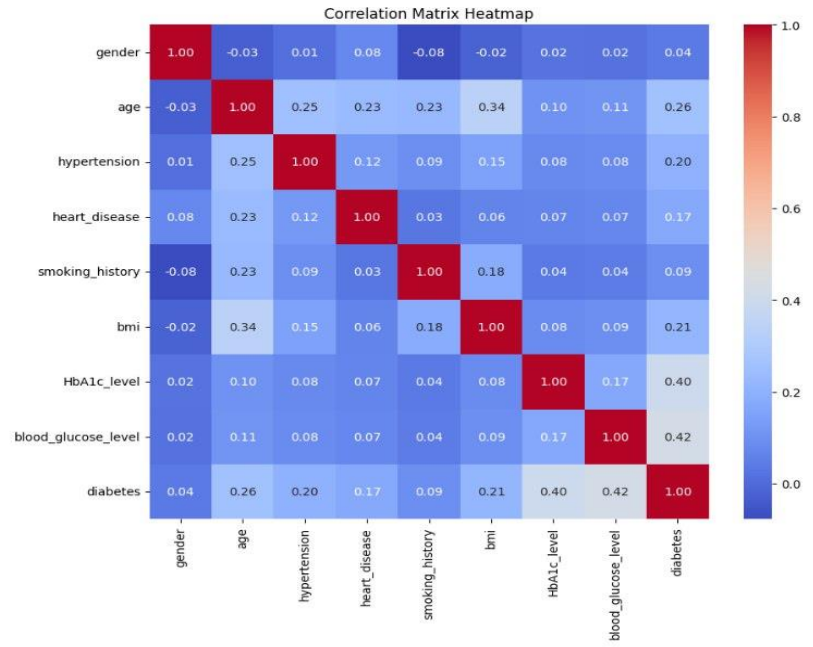


Fig 2 represents the visual representation of the Dataset

A heatmap of a correlation matrix, annotating each cell with two decimal points, using the 'coolwarm' colormap and displaying a color bar.

C. PCA:

```
pca = PCA(random_state=42)
pca.fit(X_train)
```

```
PCA(random_state=42)
```

```
# Calculate the cumulative explained variance ratio
cumulative_variance = pca.explained_variance_ratio_.cumsum()

# Find the number of components needed for 90% variance
n_components_90 = (cumulative_variance <= 0.90).sum()

print("Number of components to explain 90% variance:", n_components_90)
```

```
Number of components to explain 90% variance: 5
```

```
pca = PCA(n_components=5, random_state=42)
pca.fit(X_train)

X_train_PCA = pd.DataFrame(pca.transform(X_train))
X_test_PCA = pd.DataFrame(pca.transform(X_test))
```

```
X_train_PCA.columns = [str(column_name) for column_name in X_train_PCA.columns]
X_test_PCA.columns = [str(column_name) for column_name in X_test_PCA.columns]
```

```
X_train_PCA.head()
```

	0	1	2	3	4
0	0.58	0.06	-0.01	-0.11	-0.06
1	0.51	0.47	-0.09	-0.19	-0.33
2	-0.37	-0.39	-0.13	0.07	-0.26
3	-0.38	-0.22	0.20	-0.17	0.17
4	0.66	-0.46	-0.13	0.12	0.02

Fig 3 represents the Principal Component Analysis

Principal Component Analysis (PCA) to transform the training and testing datasets (X_train and X_test), then converts them into DataFrames, and renames the columns with their respective indices to maintain. consistency.

D. Decision Tree:

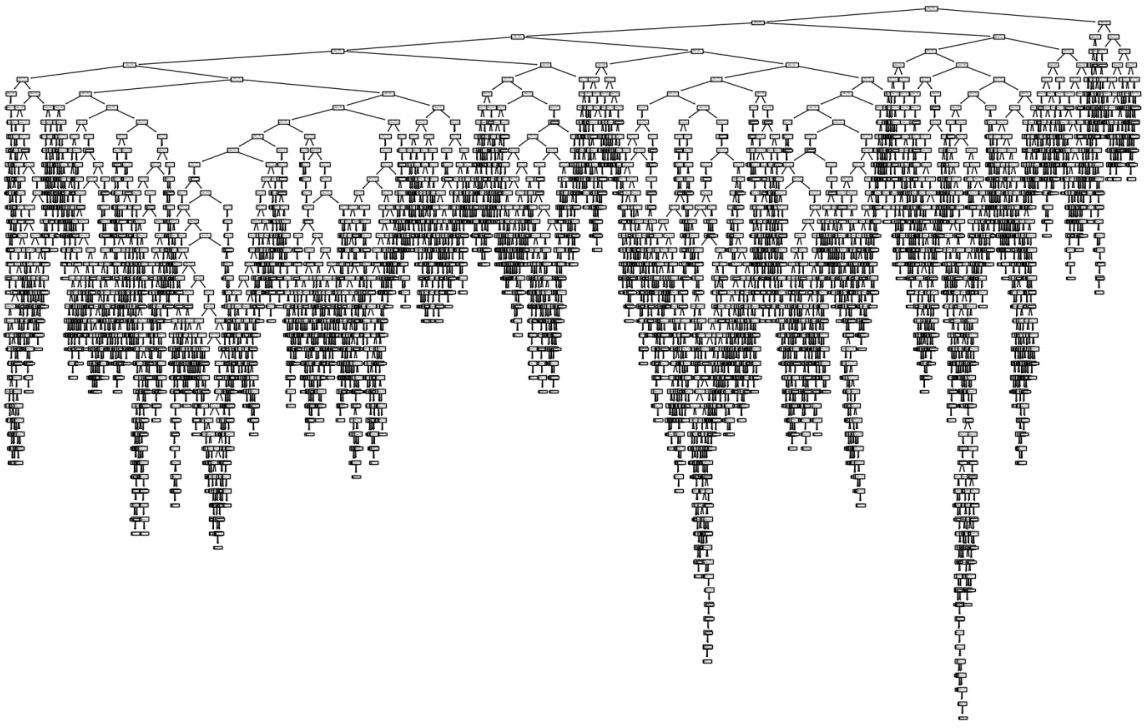
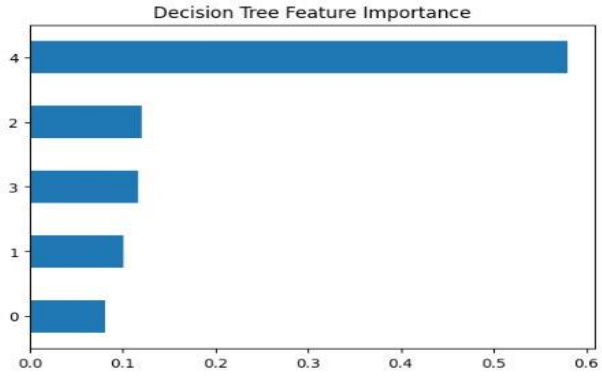


Fig 4 represents the Decision Tree

```
[ ] feature_importances = pd.Series(model.feature_importances_, model.feature_names_in_).sort_values()
    feature_importances.plot.barh();
    plt.title('Decision Tree Feature Importance');
```




```
[ ] y_pred_test = model.predict(X_test_PCA)
    y_pred_train= model.predict(X_train_PCA)
```

```
[ ] train_decision_tree_report = classification_report(y_train, y_pred_train)
    print(train_decision_tree_report)
```

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     64047
           1       1.00      0.99      1.00      5953

    accuracy                           1.00     70000
   macro avg       1.00      1.00      1.00     70000
weighted avg       1.00      1.00      1.00     70000
```

Fig 5 represents the Decision Tree Feature Importance

E. KNN:

KNN

```
[ ] training_accuracy = []
    test_accuracy = []

    # try n_neighbors from 1 to 10
    possible_neighbors = range(1, 11)

    for n_neighbors in possible_neighbors:
        clf = KNeighborsClassifier(n_neighbors=n_neighbors)
        clf.fit(X_train_PCA, y_train)

        # training accuracy
        training_accuracy.append(clf.score(X_train_PCA, y_train))

        # testing accuracy
        test_accuracy.append(clf.score(X_test_PCA, y_test))

    plt.plot(possible_neighbors, training_accuracy, label="training accuracy")
    plt.plot(possible_neighbors, test_accuracy, label="test accuracy");
    plt.ylim((0,1))
    plt.ylabel("Accuracy");
    plt.xlabel("n_neighbors");
    plt.legend();
```

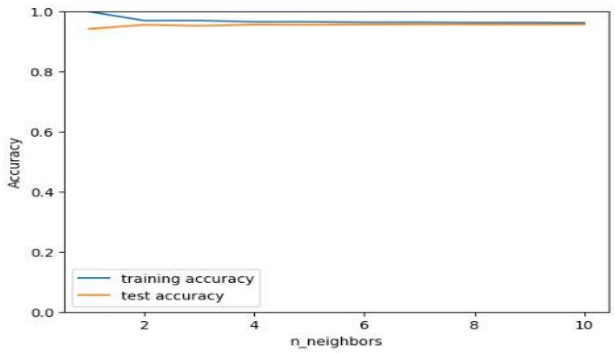


Fig 6 represents the KNN algorithm Accuracy

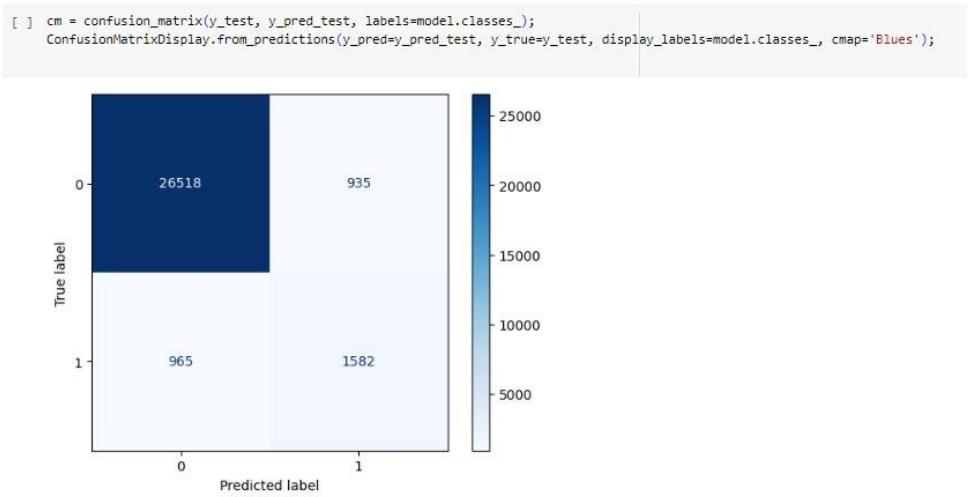


Fig 7 represents the KNN algorithm Predicted Label

F. Gradio:

Gradio is a Python library that enables easy and rapid creation of UIs for machine learning models, allowing quick prototyping and deployment with minimal code.

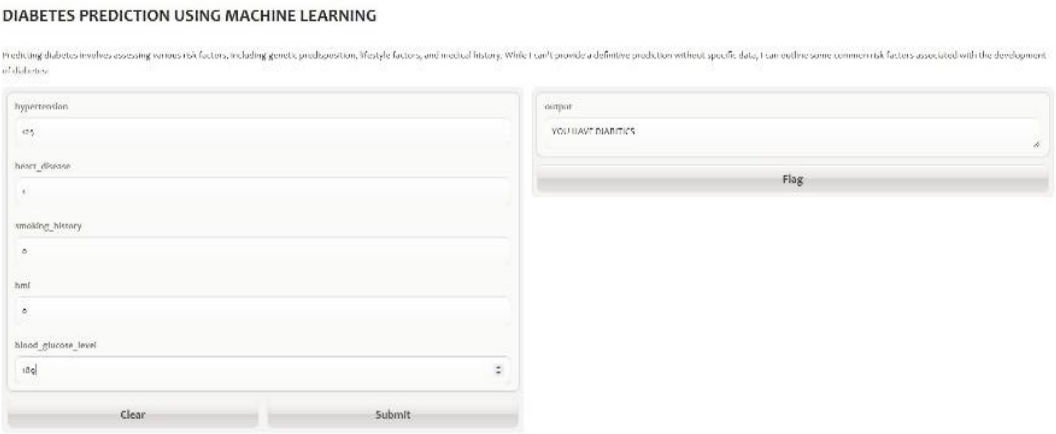


Fig 8 represents the Webinterface

5. Conclusion

In conclusion, the project "Detection and Prediction of Comorbidities of Diabetes using Machine Learning Techniques" aims to address the critical challenge of identifying and predicting comorbid conditions associated with diabetes to improve patient care and management outcomes. Through the application of advanced machine learning algorithms and data analysis techniques, the project endeavors to leverage large-scale healthcare data, including electronic health records (EHRs), clinical databases, and research studies, to develop predictive models capable of accurately detecting and forecasting comorbidities in diabetic patients.

The literature review highlights the significance of diabetes as a global public health concern, emphasizing its rising prevalence, economic burden, and associated comorbidities such as cardiovascular disease, nephropathy, retinopathy, and neuropathy. By harnessing the power of machine learning, the project aims to contribute to early detection, risk stratification, and personalized interventions for diabetic patients, thereby potentially reducing healthcare costs, improving patient outcomes, and enhancing the quality-of-care delivery.

The proposed system architecture encompasses data acquisition, preprocessing, feature selection, model development, evaluation, integration with user interfaces, and deployment, ensuring a comprehensive approach

to comorbidity prediction in diabetes management. The adoption of Agile methodology facilitates iterative development, stakeholder collaboration, adaptability to changing requirements, and continuous improvement throughout the project lifecycle.

Metrics for evaluating model efficacy and prediction performance include accuracy, precision, recall, F1 score, area under the receiver operating characteristic curve (AUC-ROC) and mean absolute error (MAE) or root mean squared error (RMSE). By adhering to rigorous testing standards, including unit testing, integration testing, regression testing, and usability testing, the project ensures the reliability, security, and usability of the predictive models and decision support system.

Overall, the project represents a significant advancement in leveraging machine learning techniques for diabetes care and management, with the potential to revolutionize clinical practice, inform treatment decisions, and empower healthcare providers in delivering personalized, evidence-based care to diabetic patients. Through interdisciplinary collaboration, innovation, and continuous refinement, the project endeavors to make a meaningful impact in improving the lives of individuals living with diabetes and reducing the burden of comorbid conditions associated with this chronic disease.

Refrences

[1] Ahlqvist, E., Storm, P., Käräjämäki, A., Martinell, M., Dorkhan, M., Carlsson, A., ... & Groop, L. (2018). Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. The Lancet Diabetes & Endocrinology, 6(5), 361-369.

[2] Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2020. Diabetes Care, 43(Supplement 1), S14-S31.

[3] Bellazzi, R., & Zupan, B. (2008). Predictive data mining in clinical medicine: Current issues and guidelines. International Journal of Medical Informatics, 77(2), 81-97. Augue interdum velit euismod in pellentesque massa placerat duis ultricies. Metus aliquam eleifend mi in nulla posuere sollicitudin aliquam ultrices.

[4] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).

[5] Cho, N. H., Shaw, J. E., Karuranga, S., Huang, Y., da Rocha Fernandes, J. D., Ohlrogge, A. W., ... & Ogurtsova, K. (2018). IDF diabetes atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes Research and Clinical Practice, 138, 271-281.

[6] Churpek, M. M., Yuen, T. C., Winslow, C., Meltzer, D. O., & Kattan, M. W. (2016). Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. Critical Care Medicine, 44(2), 368-374.

[7] Dall, T. M., Yang, W., Halder, P., Pang, B., Massoudi, M., Wintfeld, N., & Semilla, A. P. (2014). The economic burden of elevated blood glucose levels in 2012: diagnosed and undiagnosed diabetes, gestational diabetes mellitus, and prediabetes. Diabetes Care, 37(12), 3172-3179.

[8] Inzucchi, S. E., Bergenstal, R. M., Buse, J. B., Diamant, M., Ferrannini, E., Nauck, M., ... & Matthews, D. R. (2015). Management of hyperglycemia in type 2 diabetes, 2015: a patient-centered approach: update to a position statement of the American Diabetes Association and the European Association for the Study of Diabetes. Diabetes Care, 38(1), 140-149.

[9] Jia, J., Wang, J., Guo, Y., & Jin, Y. (2019). An application of machine learning method in predicting diabetes based on electronic health records. IEEE Access, 7, 99041-99047.

[10] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal, 15, 104-116.

[11] Madigan, D., Ryan, P. B., Schuemie, M., Stang, P. E., Overhage, J. M., Hartzema, A. G., ... & Reich, C. (2013). Evaluating the impact of database heterogeneity on observational study results. American Journal of Epidemiology, 178(4), 645-651.

[12] Nathan, D. M., Kuenen, J., Borg, R., Zheng, H., Schoenfeld, D., & Heine, R. J. (2008). Translating the A1C assay into estimated average glucose values. Diabetes Care, 31(8), 1473-1478.

[13] Rasmussen-Torvik, L. J., McAlpine, D. D., Uratsu, C. S., & Kho, A. N. (2012). Moving beyond the diagnosis: how is diabetes associated with mortality risk? Annals of Epidemiology, 22(12), 855-861.

[14] Raval, A. D., Vyas, A., & Tariq, M. (2019). Predicting Diabetes using machine learning techniques. In Proceedings of the 2019 IEEE International Conference on Big Data (pp. 4127-4131).

[15] M. R. V V R, S. N, M. Gadiraju, S. S. Reddy, S. Bonthu and R. R. Kurada, "A Plausible RNN-LSTM based Profession Recommendation System by Predicting Human Personality Types on Social Media Forums," 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2023, pp. 850-855, doi: 10.1109/ICCMC56507.2023.10083557.

[16] Reddy, S.S., Gadiraju, M., Maheswara Rao, V.V.R., "Analyzing Student Reviews on Teacher Performance Using Long Short-Term Memory", Innovative Data Communication Technologies and Application. Lecture Notes on Data Engineering and Communications Technologies, vol 96. Springer, Singapore, 2022. https://doi.org/10.1007/978-981-16-7167-8_39

[17] Silpa, N., and VVR Maheswara Rao. "Enriched Big Data Pre-Processing Model With Machine Learning Approach to Investigate Web User Usage Behaviour." Vol. 12 No. 5 (2021). DOI: 10.21817/indjcse/2021/v12i5/211205050

[18] N. Silpa, Dr. V V R Maheswara Rao, "A Complete Research on Techniques & Technologies of Big Web Data Preparation to Web User Usage Behavior", published in International Journal of Recent Technology and Engineering (IJRTE), ISSN:2277-3878, Vol. 8, Issue 2S11, September 2019.