BioPhecy: Early Prediction of Lifestyle Diseases

[1]Santhosh Kumar K L, [2]Payal Dayanand,[2]Grace Evangeline,[2]Mayurikaa Sivakumar,[2]Sanjiv M V

[1]Assistant Professor, [2] B.Tech Students
[1 &2] School of Computer Science and Engineering,
[1&2]Presidency University, Bengaluru, India

Abstract - Developing a web-based system for predicting diseases based on user-input symptoms presents an opportunity to offer preliminary guidance to those seeking medical advice. This platform could prove beneficial, especially for individuals with limited access to immediate healthcare services, aiding them in identifying potential health issues. However, several critical considerations must be addressed. The accuracy and reliability of the system hinge on the quality and diversity of datasets used for training the prediction model. Moreover, it's essential to emphasize that while this tool can provide initial guidance, it cannot replace a thorough diagnosis from a qualified healthcare professional. Safeguarding user data and adhering to ethical and legal regulations are also paramount. Therefore, balancing the system's role as a support tool while encouraging users to seek professional medical assistance for confirmed diagnoses is crucial. Continuous updates and user education about the system's limitations are vital aspects to ensure its effectiveness and responsible use. Machine learning empowers the system to adapt and refine its predictions based on real-time information, contributing to more precise and timely guidance for users. However, ensuring transparency in the system's learning process is crucial, allowing users to understand the basis of the predictions made.

Index Terms - AI Diagnosis, Predictive Machine, Symptom Analysis, Machine Learning Models

I. INTRODUCTION

The software initiative, BioPhecy, endeavors to proactively predict diseases before their onset, utilizing a suite of machine learning models for development, rigorous testing, and validation. Recent findings published in The Lancet and its affiliated publications shed light on a concerning escalation of diseases such as stroke, diabetes, heart attacks, AIDS, and dengue across India in the past 25 years. Comprehensive assessments conducted across various Indian states between 1990 and 2016 reveal a consistent upward trend in the prevalence of cardiovascular diseases, diabetes, chronic respiratory ailments, cancer, and suicides. Notably, incidences of heart diseases and strokes have surged by over 50% in each state during this period, significantly contributing to the overall mortality and disease burden, effectively doubling their impact over the past quarter-century. Within this study, we propose a predictive model that integrates Logistic Regression, Naive Bayes, and SVM (Support Vector Machines) classifiers to forecast illnesses based on diverse symptomatology. This amalgamated predictive system operates on a web-based platform, leveraging a collection of health-related datasets sourced from various repositories. The primary aim is to address scenarios where immediate medical consultation becomes impractical due to prior commitments or unavailability of healthcare professionals. In such exigencies, our automated system assumes a pivotal role, offering urgent guidance by analyzing user-input symptoms through a graphical user interface (GUI). Employing sophisticated data processing techniques, the system accurately identifies potential diseases aligned with patient-specific details. Resultantly, users gain insights to seek further medical interventions from appropriate disease specialists based on the system's prognostications. This research endeavor strives to provide accessible healthcare guidance.

II. LITERATURE SURVEY

In [1] the author says, in the realm of biomedical research, the integration of pattern detection and machine learning (ML) has emerged as a promising avenue to enhance the reliability of disease detection and approach. This amalgamation not only upholds the impartiality in decision-making processes but also offers a robust framework for automating and refining algorithms to analyze high-dimensional and multi-modal biomedical data. This survey paper delves into an in-depth comparative analysis of various ML algorithms employed in the detection of prevalent diseases such as heart disease and diabetes. The focus of this study revolves around assembling a compendium of algorithms and techniques utilized in ML for disease detection, thereby illuminating the decision-making processes within this domain.

In [2] the author is utilizing Machine Learning for Disease Prediction involves a system designed to anticipate diseases based on user-provided symptoms. This system processes symptoms as inputs and generates disease probabilities as outputs. The predictive modeling relies on the Naïve Bayes classifier, a supervised machine learning algorithm, to calculate disease probabilities. The exponential growth in biomedical and healthcare data underscores the importance of precise medical data analysis, facilitating early disease identification and improved patient care. Employing techniques such as linear regression and decision trees, this study focuses on predicting diseases such as Diabetes, Malaria, Jaundice, Dengue, and Tuberculosis.

In [3] the authors mention about the utilization of Machine learning in disease prediction systems is pivotal in forecasting ailments based on user-entered symptom data, providing reliable insights derived from this information. This system serves individuals seeking to ascertain their ailment status when not facing immediate danger, offering suggestions and preventive measures to maintain good health. It serves as a tool for users to identify potential illnesses through this predictive mechanism. The increasing diversity of diseases coupled with a limited doctor-patient ratio has amplified the adoption of specific disease prediction technologies, intensifying concerns about healthcare. Our focus lies in furnishing users with prompt and accurate disease prognoses, accounting for symptom severity and utilizing the best algorithms, complemented by potential doctor consultations.

In [4] Contemporary environmental conditions and lifestyle choices expose humans to a multitude of diseases. The early identification and prediction of these ailments hold immense significance in mitigating their severity. Manual disease identification by physicians often faces challenges in accuracy. This paper aims to discern and forecast prevalent chronic illnesses in patients. Leveraging advanced machine learning techniques becomes pivotal in establishing a robust categorization for identifying individuals with chronic diseases reliably. Disease prediction poses a formidable challenge, wherein data mining assumes a pivotal role. The proposed system endeavors to provide comprehensive disease prognoses based on patient symptoms, employing machine learning algorithms such as Convolutional Neural Network (CNN) for automated feature extraction and disease prediction, and K-Nearest Neighbor (KNN) for precise distance calculation within the dataset, culminating in accurate disease predictions.

In [5] the correlation between a balanced lifestyle and the prevention of lifestyle-related diseases has been well-established. Early prediction of the likelihood of such diseases can significantly impact preventive measures, reducing treatment costs, and fostering better quality of life. This research endeavors to develop a health assessment and disease prediction system geared towards averting lifestyle-related diseases by evaluating individuals' health status and lifestyle choices. Employing Decision Tree methodology, hence the author's study aims to construct an intelligent system capable of predicting potential diseases in users and providing tailored suggestions for adopting a healthier lifestyle.

In [6] the authors introduce an innovative application of genetic algorithms for function extraction, employing a system that evaluates individuals by combining SVM classifiers tailored for each function. Each primitive is associated with an SVM classifier, enabling a selection process within the proposed method. The approach, validated on chronic disease data sourced from diverse smart devices and utilizing different classifier types, demonstrates robustness. Accurate analysis of medical data, facilitated by data mining, enhances early patient care. This classification methodology achieved precise results for diseases like cancer, heart disease, kidney disease, and others, indicating a potential one-in-28 incidence rate among individuals in India based on a dataset containing 569 rows and 32 columns.

In [7], the author mention's that their study underscores the effective utilization of data mining techniques and artificial intelligence to enhance disease prediction methodologies. Focusing initially on prevalent diseases, the scope extends toward encompassing high-fatality conditions like various cancers. Employing sophisticated algorithms such as K-means, kernel K-means, Gaussian mixture models, clustering, and K-nearest neighbor for precise clustering, the objective is to facilitate early prognostication and intervention. This proactive approach aims to curtail fatality rates, particularly in severe illnesses like cancer, thereby fostering long-term economic benefits.

In [8], the authors describe that their study introduces the Machine Learning based Cardiovascular Disease Diagnosis (MaLCaDD) framework, tailored for accurate cardiovascular disease prediction. Addressing missing data using mean replacement and tackling data imbalance via Synthetic Minority Over-sampling Technique (SMOTE) are initial steps. Feature Importance techniques are then applied for feature selection using the Logistic Regression and K-Nearest Neighbor (KNN) across three benchmark datasets i.e., Framingham, Heart Disease, Cleveland. Comparative analysis demonstrates MaLCaDD's superior accuracy with a reduced feature set compared to existing state-of-the-art methods, making it a reliable tool for early cardiovascular disease diagnosis in real-world scenarios.

In [9] author says that their research delves into symptom-based decision prediction, utilizing user-provided symptoms as inputs for disease prognosis. Highlighting the significance of data diversity in accuracy, the study underscores the crucial need to evaluate models using varied machine learning algorithms. Emphasizing the dynamic nature of these techniques and their potential for discrepancies, the paper stresses the pivotal role of accuracy evaluation in disease prediction models. It emphasizes the utilization of Machine Learning's experiential decision-making paradigm in this context.

In [10] author proposes a study that develops and validates a predictive model for visual impairment among older adults. Analyzing 586 participants, key risk factors including age, blood pressure, physical activity, diabetes, ocular history, and education level were identified using logistic regression. Cataract emerged as the primary cause of visual impairment. This model presents a promising tool for early intervention to mitigate visual impairment in the elderly.

III. PROPOSED METHOD

With reference to Fig 3.1, it depicts the proposed method for the early prediction of lifestyle diseases. Every machine learning algorithm during the process of development and deployment of models have to undergo various steps such as:

3.1 ***Variable Identification and Data Validation:*** This involves employing validation techniques to gauge model error rates, especially when datasets may not fully represent broader populations. The process entails identifying missing values, duplicates, and comprehending data types, distinguishing between float or integer variables. Data Validation, Cleaning, and preparing stages involve thorough variable analysis, detecting and rectifying missing or duplicate values, and leveraging a distinct validation dataset for impartial model assessment. Cleaning procedures encompass column renaming and eliminating redundant data, ultimately bolstering data quality for more refined analytics. Pre-processing steps are integral, as they precede analysis by converting raw data into a format compatible with machine learning algorithms. This stage ensures data alignment with model requirements, such as handling null values for specific algorithms like Random Forest, thus optimizing overall execution.

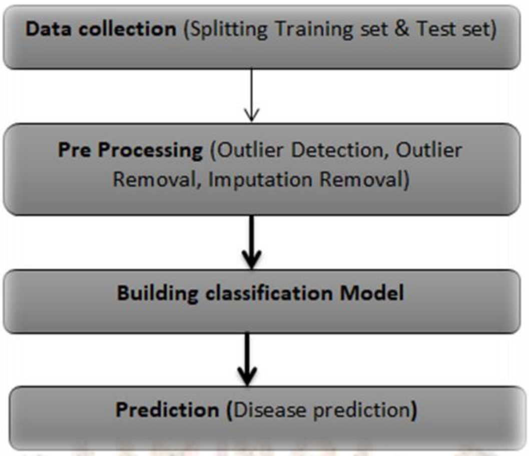
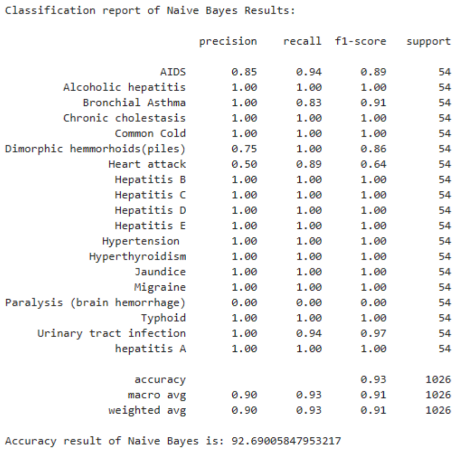

Data collection (Splitting Training set & Test set)
Pre Processing (Outlier Detection, Outlier Removal, Imputation Removal)
Building classification Model
Prediction (Disease prediction)

Fig.3.1 The Proposed Method.

3.2 ***Exploratory Data Analysis (EDA) through Visualization*****:** It is a crucial aspect in statistics and machine learning. Data Visualization, vital for qualitative insights, uncovers patterns, outliers, and intricacies often missed by numerical analysis alone. Proficiency in various plotting techniques—line plots for time series, bar charts for categories, scatter plots for variable relationships—is essential for summarizing distributions and understanding complex data relationships. Managing outliers is critical, as they skew attribute values and distort data representations, impacting machine learning accuracy. Employing cross-validation techniques ensures model robustness, testing its ability to generalize beyond training data for accurate pattern recognition.

3.3 ***Building Classification Models:*** In the development of a classification model, our research spans several crucial methodologies within machine learning. This includes foundational techniques like Logistic Regression and Decision Trees, pivotal for establishing predictive models and segmenting datasets based on attribute values, respectively. Additionally, Support Vector Machines (SVM) and Random Forest techniques play a significant role, with SVM defining optimal hyper-planes and Random Forest countering decision tree over fitting through ensemble learning. Furthermore, K-Nearest Neighbor (KNN) and Naive Bayes methods contribute their distinct strengths: KNN's reliance on neighboring data for predictions and Naive Bayes' efficiency in handling conditional probabilities, making it adept with large datasets. These methodologies collectively contribute to the construction of a comprehensive classification model, ensuring enhanced predictive accuracy and model reliability**.**

3.4 ***Comprehensive Evaluation and Comparison of machine learning algorithms:*** Here we focus on centering on critical metrics such as Sensitivity, Specificity, and prediction accuracy. This evaluation encompasses prominent algorithms—Logistic Regression, Random Forest, Decision Trees, and Support Vector Machines (SVM)—assessing their performance through key metrics derived from confusion matrices, including True Positive, True Negative, False Positive, and False Negative rates. Sensitivity, gauging the model's capability to identify true positives, is juxtaposed with Specificity, reflecting accuracy in predicting true negatives for example refer to (Fig 3.2).Central to this module is the emphasis on uniform evaluation methodologies, employing techniques like cross-validation and visual aids to discern the most suitable algorithms. A standardized test framework is advocated, ensuring equitable and consistent assessments across different models on identical datasets. This phase underscores the necessity of methodological rigor to effectively evaluate and compare machine learning algorithms, paving the way for informed model selection in real-world applications.



```
Classification report of Naive Bayes Results:

                               precision    recall  f1-score   support

                          AIDS       0.85      0.94      0.89        54
            Alcoholic hepatitis       1.00      1.00      1.00        54
              Bronchial Asthma       1.00      0.83      0.91        54
           Chronic cholestasis       1.00      1.00      1.00        54
                   Common Cold       1.00      1.00      1.00        54
  Dimorphic hemmorhoids(piles)       0.75      1.00      0.86        54
                  Heart attack       0.50      0.89      0.64        54
                   Hepatitis B       1.00      1.00      1.00        54
                   Hepatitis C       1.00      1.00      1.00        54
                   Hepatitis D       1.00      1.00      1.00        54
                   Hepatitis E       1.00      1.00      1.00        54
                 Hypertension        1.00      1.00      1.00        54
                Hyperthyroidism       1.00      1.00      1.00        54
                      Jaundice       1.00      1.00      1.00        54
                      Migraine       1.00      1.00      1.00        54
  Paralysis (brain hemorrhage)       0.00      0.00      0.00        54
                       Typhoid       1.00      1.00      1.00        54
       Urinary tract infection       1.00      0.94      0.97        54
                   hepatitis A       1.00      1.00      1.00        54

                      accuracy                           0.93      1026
                     macro avg       0.90      0.93      0.91      1026
                  weighted avg       0.90      0.93      0.91      1026

Accuracy result of Naive Bayes is: 92.69005847953217
```

Fig.3.2 Classification Report for Naïve Bayesian.

3.5 ***Prediction /GUI:*** Upon concluding the metrics evaluation and comparative analysis of algorithms based on accuracy, known as the classification report, our research endeavors culminate in the strategic selection of the Naive Bayesian machine learning algorithm for final disease prediction within the Graphical User Interface (GUI). This GUI interface comprises a list of 40 symptoms, allowing users to select any number of symptoms as desired. Upon selection, users can proceed by clicking the predict button, facilitating the display of development of Graphical User Interfaces (GUIs). Employing Tkinter, we architect UI applications, crafting windows and diverse graphical components pivotal for user interaction. Its integration as a standard Python package not only guarantees accessibility but also underscores its role in fortifying security measures for individual users or accountants.

IV. RESULTS

Let us look into the results obtained from each particular stage of our disease prediction system from the initial to the final stage. Under the topic exploratory analysis for visualization, below we have (Fig 4.1) representing a bar graph portraying the presence of one particular symptom itching. The graph shows the presence of itching as a symptom and not a symptom.

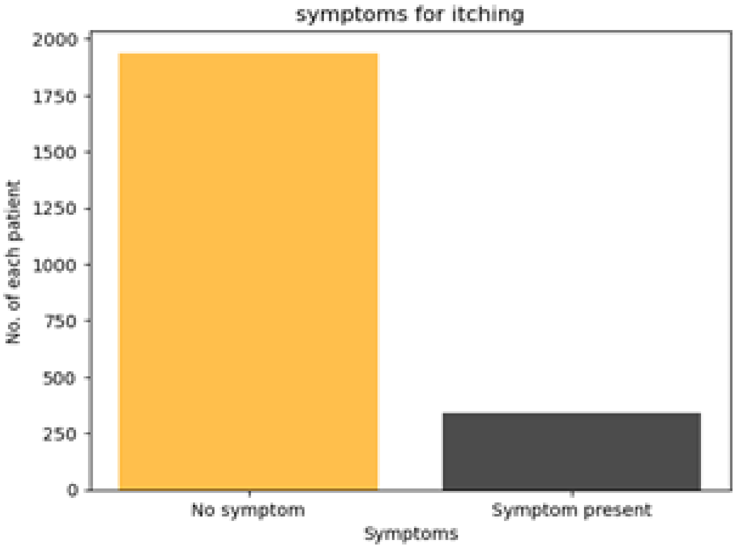


Fig.4.1 Bar Graph for Symptom Present vs. No Symptom Present for Itching

Let us look into the main aspect which is the accuracy of each model from which we decide which particular model has to be chosen for prediction of the disease. Refer Table 1 for the accuracy percentage.

Table 1. Accuracy of Algorithms

SN.	Algorithm	Accuracy (%)
1	Logistic Regression	92.69005
2	SVM	93.12865
3	Naïve Bayesian	95.42854

With reference to (Table 1) we can see that out of 6 machine learning algorithms(refer pt. 3.3) that we have trained our model on we have selected only 3 machine learning algorithms which are logistic regression , Support Vector Machine (SVM) and Naïve Bayesian and as for the final prediction Naïve Bayesian has been selected since it has comparatively high accuracy.

With reference to Fig 4.2 we can observe the look of our User-Interface, and Fig 4.3 shows us the prediction obtained after the user gives an input of few symptoms to get a predicted output.

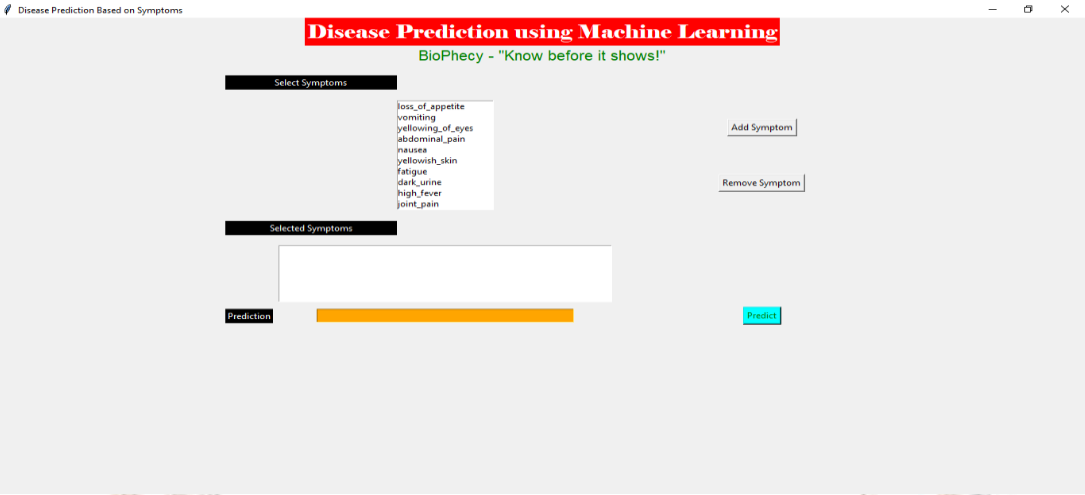
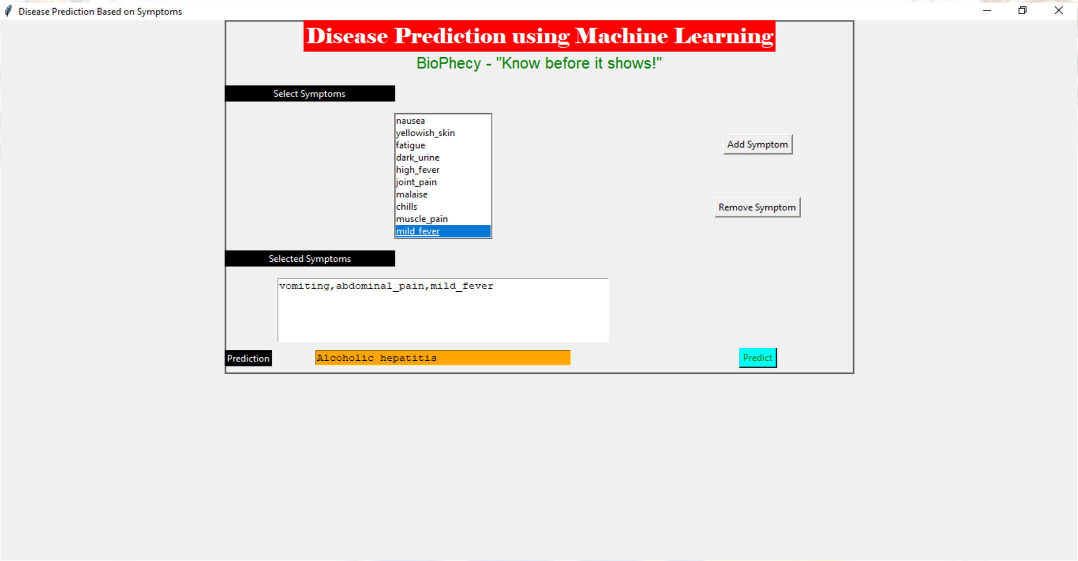


Fig.4.2 User-Interface



Selected Symptoms: vomiting,abdominal_pain,mild_fever

Prediction: Alcoholic hepatitis

Fig.4.3 User-Interface, Prediction of Disease Given Symptom

V. CONCLUSION

In conclusion, this study embarked on a comprehensive exploration of machine learning applications in disease prediction. From meticulous data validation and algorithm selection to the integration of a user-friendly Graphical User Interface (GUI) utilizing Tkinter, this research delineated a robust framework for healthcare applications. The iterative phases traversed in this research underscored the significance of methodological rigor in evaluating and selecting the Naive Bayes algorithm as the optimal choice for disease prediction within the GUI. This culmination, along with the utilization of Tkinter, a proficient Python library for GUI development, highlights the critical fusion of advanced machine learning techniques with accessible interfaces in healthcare. By bridging the gap between sophisticated algorithms and user-friendly interfaces, this study paves the way for practical implementations in healthcare settings. As technology evolves, the integration of cutting-edge algorithms with intuitive interfaces not only augments predictive accuracy but also ensures seamless user interactions and heightened security in healthcare applications. This research sets the stage for future advancements, emphasizing the pivotal role of machine learning in fostering efficient, secure, and accessible healthcare solutions.

VI. REFERENCES

[1] Hamsagayathri, P., and S. Vigneshwaran. "Symptoms based disease prediction using machine learning techniques." 2021 Third international conference on intelligent communication technologies and virtual mobile networks (ICICV). IEEE, 2021.

[2] Gomathy, C. K., and Mr. A. Rohith Naidu. "The prediction of disease using machine learning." International Journal of Scientific Research in Engineering and Management (IJSREM) 5.10 (2021).

[3] Patil, Kajal, et al. "Multiple Disease Prognostication Based On Symptoms Using Machine Learning Techniques." ITM Web of Conferences. Vol. 44. EDP Sciences, 2022.

[4] Alanazi, Rayan. "Identification and prediction of chronic diseases using machine learning approach." Journal of Healthcare Engineering 2022 (2022).

[5] Parab, Aditi, Prachiti Gholap, and Vijaya Patankar. "DiseaseLens: A Lifestyle related Disease Predictor." 2022 5th International Conference on Advances in Science and Technology (ICAST). IEEE, 2022.

[6] NagabhushanaRao, M., et al. "PREDICTION OF CHRONIC DISEASES AT AN EARLY PHASE USING MACHINE LEARNING APPROACH." Turkish Journal of Physiotherapy and Rehabilitation 32:3.

[7] Sakshi Gaur, Sarvesh Sharma and Ayush Tripathi,"Early Prediction and Prevention of Lifestyle Diseases",EasyChair Preprint no. 5702, June 4, 2021

[8] Rahim, Aqsa, et al. "An integrated machine learning framework for effective prediction of cardiovascular diseases." IEEE Access 9 (2021): 106575-106588.

[9] Kanamarlapudi, Sriya, et al. "Comparison and Analysis of Various Machine Learning Algorithms for Disease Prediction." 2023 7th International Conference on Computing Methodologies and Communication (ICCMC). IEEE, 2023.

[10] Zhao, Yue, and Aiping Wang. "Development and validation of a risk prediction model for visual impairment in older adults." International journal of nursing sciences 10.3 (2023): 383-390