ISSN: 2993-8643

Engineering: Open Access

Review Article

Diabetes and Hypertension Risk Prediction System Based on Symptoms

P. Y. R Pothuwila*

Faculty of Computing Sri Lanka institute of information technology, Sri Lanka

***Corresponding Author**
P. Y. R Pothuwila, Faculty of Computing Sri Lanka institute of information technology, Sri Lanka.

Submitted: 2025, June 06; **Accepted:** 2025, July 11; **Published:** 2025, July 18

Citation: Pothuwila, P. Y. R. (2025). Diabetes and Hypertension Risk Prediction System Based on Symptoms. *Eng OA, 3*(7), 01-05

Abstract
This paper introduces an innovative prediction system for predicting risk of having diseases like diabetes and hypertension based on users' symptoms, designed to support individuals from all age groups with concern of their health and wellbeing. In the modern world technology has already altered the path of medical science. With this diagnostic tool, researchers hope to provide a trusty and simple approach for users to identify the risk of having cholesterol and hypertension using a mobile application. This diagnostic model has used ensemble model concept to maximize the accuracy of prediction. In order to train predictive models, clinical datasets were gathered, refined, and classified. Based on users' symptoms, the system will predict users' risk of diagnosed with diabetes and hypertension and direct them to prevent or control the situation further via the mobile application.

Keywords: Machine Learning, Ensemble Model, Dataset

1. Introduction

1.1 Background

With modern lifestyle it has become nearly impossible to follow full time career-oriented lifestyle while practicing proper healthy life choices making diagnosed with non- communicable diseases such as diabetes and hypertension common among individuals from every social layer. Society is suffering from these diseases, and this has become a threat not only on the individual level but also on social and economic levels.

One of the main reasons for this issue is not giving proper pre-attention to these conditions since almost all the non-communicable diseases are preventable before being diagnosed. Mainly due to busy lifestyle individuals tend to ignore or postpone proper medical checkups with doctors which prevent them from identifying their issues previously.

1.2 Problem Statement

Even though existence of predictive diagnostic models for disease prediction based on symptoms available on academic domain there is a major lack of proper disease diagnostic models applied for the usage of common people without complications of scholarly exclusions and technological understanding which preventing them from using these tools. In addition to that, busy lifestyle also prevents individuals from seeking proper methods such as clinical checkups for their healthy wellbeing.

1.3 Significance

With this proposed predictive system, researcher provide a bridge of solution to fill the mentioned gap between individuals and predictive tools by creating a highly accurate predictive ensemble model to properly predict the risk of diagnosed with diabetes and hypertension via mobile application. This application has the ability to provide all the benefits of an accurate predictive model while making it accessible without any complications of technological understanding or any other previous exclusions. Anyone can self-check and get a proper idea about risk of diagnosis without burden.

1.4 Objective

Primary objective of this research is to provide an accurate predictive model (with implementation of ensemble machine learning concept) that predict the risk of diagnosed with diabetes and hypertension and apply that model to mobile application environment making it accessible to individuals easily.

1.5 Summery

In this paper, section 2 contains a literature review with study of related work. Section 3 offers a comprehensive overview of the system, technology used, and the evaluation methodology employed. Section 4 ,5 and 6 presents the research findings, followed by concluding remarks and recommendations for future endeavors.

2. Literature Review

The implementation of predictive models in the medical field has significantly altered the way diseases are diagnosed and predicted, and researchers have been exploring a variety of methods and technologies that range from classic machine learning algorithms to more advanced techniques aimed at improving prediction accuracy, which has led to a deeper understanding of how these models can be applied in real- world scenarios. As a example, Smith et al. examined the effectiveness of Random Forests for predicting diabetes using clinical and demographic data and observed that it is somewhat robust against overfitting, especially in high-dimensional datasets, while Zhang and his team demonstrated that gradient boosting techniques like XGBoost, outperform logistic regression for predicting cholesterol levels specially when considering precision and recall metrics which also indicated the ongoing evolution of predictive analytics in healthcare [1.2].

Mobile health applications are increasingly prevalent, often incorporating these predictive models, which makes them accessible and user-friendly, and Lee et al. developed a mobile application that utilizes Support Vector Machines (SVM) to predict hypertension based on lifestyle and family history, emphasizing the utility of having predictive analytics available on mobile devices for real-time health insights, while Kumar et al. also contributed to this area with a mobile-based predictive model aimed at early detection of cardiovascular diseases, employing traditional machine learning methods to enhance accuracy and reliability, thus showcasing the potential of technology in improving health outcomes [3,4].

Patel et al. [5] conducted another significant study that looked into deep learning techniques, especially convolutional neural networks (CNNs), for predicting various diseases by utilizing images and symptom data [5]. While the performance of CNNs was indeed impressive, the study highlighted some challenges, particularly regarding interpretability and the need for substantial computational resources, which could limit their practicality when it comes to mobile applications. This concern is echoed in the work of Gupta and Sharma, who pointed out the critical need for lightweight models that can function efficiently on mobile devices without sacrificing predictive power [6]. Such findings raise important questions about how to strike a balance between complexity and usability, especially in the context of mobile health solutions, where user experience is paramount and the demand for efficient processing is ever-increasing. In the context of diabetes and cholesterol prediction, Santos et al. pointed out the importance of user-friendly interfaces in mobile applications, alongside backend systems capable of delivering fast and accurate computations, and they also discussed the integration of cloud-based APIs to manage computationally intensive tasks, a strategy echoed by Chen et al., who developed a cloud-assisted mobile app for real-time disease prediction, which illustrates the growing trend of leveraging cloud technology to enhance mobile health solutions [7-10].

The insights gained from these studies lay a strong groundwork for the proposed research, which seeks to design and implement a predictive model targeting diseases like diabetes and cholesterol; this model will utilize symptom data to create personalized predictions and will be deployed as a mobile application, thereby enhancing both accessibility and impact. By combining the strengths of previous research and addressing the limitations that have been identified, this approach aims to make a meaningful contribution to the rapidly evolving field of health informatics and mobile health technology, with the ultimate goal of improving patient outcomes through innovative technological solutions that can adapt to the needs of users.

3. Methodology

3.1 Datasets and Preprocessing

During this study researcher has utilized clinical datasets that contained detailed symptomatology and diagnostic information for diseases such as diabetes and hypertension which included variety of features that were both numerical and categorical, like age, gender, BMI, smoking history, pregnancy history, and blood glucose levels, and initially, the datasets were examined for missing values, outliers, and inconsistencies to ensure data integrity, and missing values were imputed using appropriate statistical techniques, such as mean or median imputation for numerical variables and mode imputation for categorical ones, which was crucial for maintaining the quality of the data.

Categorical features, including gender and smoking history, were transformed using One-Hot Encoding (OHE) to ensure compatibility with machine learning algorithms, while numerical features were standardized using a StandardScaler to normalize their distribution, which enhanced the performance of models that are sensitive to feature scaling, and after preprocessing, the datasets were split into training and testing sets using an 80-20 ratio, which ensured that there was sufficient data for model evaluation.

3.2 Predictive Models

five base learners were selected in order to archive the highest accuracy, based on their diversity in algorithmic approaches and predictive capabilities which are Logistic Regression, a probabilistic model ideal for binary classification tasks, and Decision Tree, which is a simple yet effective model for capturing non-linear relationships, and then there's Random Forest, an ensemble method that reduces overfitting by averaging multiple decision trees, while Support Vector Machine (SVM) is effective for high-dimensional data with a focus on maximizing margins, and finally, the Neural Network (MLPClassifier) is capable of capturing complex patterns through multiple layers of computation.

3.3 Stacking Ensemble

Researcher has employed stacking method to obtain a ensemble model, where the predictions of the base learners were combined through a meta-model Logistic Regression to produce the final predictions, and the stacking ensemble was implemented using the StackingClassifier from scikit-learn, integrating the five base learners mentioned above, and training involved fitting the base learners on the training data and training the meta-model on the outputs of the base learners, which allowed the ensemble to leverage the strengths of each individual model, enhancing overall predictive accuracy and robustness.

3.4 Model Evaluation

The ensemble model was evaluated on the testing dataset using performance metrics such as accuracy, precision, recall, F1-score, and a confusion matrix, which provided a comprehensive understanding of the model's classification ability, especially in distinguishing between high-risk and low-risk cases, and the accuracy of the ensemble model exceeded that of individual base learners, underscoring the effectiveness of the stacking approach, which was quite significant.

3.5 Deployment as Mobile Application

To make the predictive model accessible, it was integrated into a mobile application that was developed using Flutter for a seamless cross-platform user experience, and the backend logic was powered by the trained ensemble model, which was serialized and stored using "joblib" for efficient loading and prediction, while the application provided users with an intuitive interface to input their health data, including symptoms and lifestyle factors, and real-time predictions were generated by preprocessing the input data on the fly and passing it through the trained ensemble model, categorizing risk levels into "Low," "Medium," and "High," thus offering users actionable insights.

3.5 Deployment as Mobile Application

To make the predictive model accessible, it was integrated into a mobile application that was developed using Flutter for a seamless cross-platform user experience, and the backend logic was powered by the trained ensemble model, which was serialized and stored using "joblib" for efficient loading and prediction, while the application provided users with an intuitive interface to input their health data, including symptoms and lifestyle factors, and real-time predictions were generated by preprocessing the input data on the fly and passing it through the trained ensemble model, categorizing risk levels into "Low," "Medium," and "High," thus offering users actionable insights.

3.6 Implementation Pipeline

One-Hot Encoding for categorical variables, StandardScaler for numerical variables, and train-test splitting were all part of the preprocessing steps. Training base learners included Logistic Regression, Decision Tree, Random Forest, SVM, and Neural Network, and fitting the stacking classifier with the base learners and meta-model was crucial. Saving the trained model and preprocessing objects (OHE and StandardScaler) using joblib was necessary, and implementing a prediction API to interface with the mobile application was also part of the deployment process.

3.7. Mobile Application Features

The application featured user-friendly input forms for health data, visualization of risk scores and predictive outcomes, and secure data handling to ensure user privacy, which was essential for user trust. This methodology ensures a comprehensive pipeline for disease prediction, balancing accuracy, efficiency, and accessibility through advanced ensemble techniques and user-centric mobile application design, which is ultimately aimed at improving health outcomes.

4. Discussion

Overall process mentioned in the paper has provided valuable insights and noteworthy findings worth for discussing in details about applying ensemble concept for creating a predictive model for predict risk of diagnosis based on user symptoms including the implementation into a mobile application.

4.1 Performance and Evaluation

When compare ensemble model predictive performances with other base models, research's ensemble model has achieved an impressive accuracy of 0.97 which evidencd by its higher precision, recall, and F1-score metrics, where the precision for class 0 was 0.97 and for class 1 it was 0.93, while the recall for class 0 stood at 1.00 and for class 1 it was notably lower at 0.67, indicating that the stacking approach effectively leveraged the complementary strengths of the base models—Logistic Regression, Decision Tree, Random Forest, SVM, and Neural Network—where, for instance, the Random Forest excelled in handling non-linear relationships and reducing overfitting, which iss crucial for maintaining model reliability, while the SVM provided robust classification for high-dimensional features, and the inclusion of Logistic Regression as the meta-model ensured smooth aggregation of predictions, resulting in improved overall stability & interpretability, which is quite significant, especially when considering the macro average precision of 0.95 and the weighted averge recall of 0.97 across all classes.

1.1 Performance and Evaluation

When compare ensemble model predictive performances with other base models, research's ensemble model has achieved an impressive accuracy of 0.97 which evidencd by its higher precision, recall, and F1-score metrics, where the precision for class 0 was 0.97 and for class 1 it was 0.93, while the recall for class 0 stood at 1.00 and for class 1 it was notably lower at 0.67, indicating that the stacking approach effectively leveraged the complementary strengths of the base models—Logistic Regression, Decision Tree, Random Forest, SVM, and Neural Network—where, for instance, the Random Forest excelled in handling non-linear relationships and reducing overfitting, which iss crucial for maintaining model reliability, while the SVM provided robust classification for high-dimensional features, and the inclusion of Logistic Regression as the meta-model ensured smooth aggregation of predictions, resulting in improved overall stability & interpretability, which is quite significant, especially when considering the macro average precision of 0.95 and the weighted averge recall of 0.97 across all classes.

4.2 Dataset Quality and Preprocessing

When clinical dataset is involved it is crucial to maintain a proper structure and quality of the dataset and preprocessing, including One-Hot Encoding for categorical variables and standardization of numerical features, was instrumental in ensuring data compatibility and model efficiency, while the imputation of missing values preserved the integrity of the dataset without introducing significant bias, which is crucial for maintaining the reliability of the model's predictions.

Nonetheless, challenges such as class imbalance were observed, particularly in the diabetes dataset where the number of positive cases was considerably smaller than negative ones, and techniques like oversampling or synthetic data generation (e.g., SMOTE) could be explored in future iterations to address this imbalance and enhance the model's sensitivity, which is an important consideration for improving predictive performance.

4.3 Deployment and Implementation

Deploying the ensemble model as a mobile app shows how practical it is for using advanced machine learning in ways that are user-friendly. The user interface of the application provides simple UX with real time risk prediction which make the whole process less complicated to perform and understand for users. For further user friendliness application categorize the level of risk into "Low," "Medium," and "High" gives users insights that can help with early intervention providing a clear idea of their diagnosis status.

4.4 Implementation in Healthcare

This study highlights the potential of ensemble models in advancing precision medicine, as the stacking approach offers a robust framework for diagnosing complex diseases like diabetes and cholesterol disorders by combining diverse machine learning algorithms. The mobile application, which further bridges the gap between advanced analytics and end- user accessibility, empowers individuals to make informed decisions regarding their health. This is a significant step forward, but it's also crucial to think about how these advancements will impact everyday users and their experiences.

Nevertheless, ethical considerations, such as data privacy and security, remain paramount, and ensuring compliance with healthcare regulations (e.g., HIPAA or GDPR) is critical to fostering trust and widespread adoption of such technologies, which is essential for the future of health informatics.

5. Conclusion

This research focused on developing an ensemble predictive model for diagnosing diseases such as diabetes and hypertension disorders based on users' observable symptoms while implementing model into a mobile application environment. this study bridges the gap between predictive analytics and accessible healthcare solutions by integrating advanced ensemble techniques with user-centric mobile design and this integration of elements indicated the significance of converting complex health data more understandable and usable for everyday individuals. The stacking ensemble model, which includes Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and Neural Network, showed superior performance in predictive accuracy and stability compared to individual classifiers.

The emphasis on preprocessing methods—like handling categorical variables, standardizing features, and addressing potential data imbalances—was crucial for achieving robust model outputs. In addition to them the application's ability to present risk levels in an actionable format enhances its utility for individuals seeking early interventions and lifestyle modifications. In summary, this research contributes to the growing field of predictive healthcare by providing an effective, scalable, and user-friendly solution for disease diagnosis. With continued optimization and adherence to ethical and regulatory standards, this approach could significantly impact personalized medicine and proactive health management.

Future Work

Future research can really focus on addressing the limitations found in this study and expanding its scope, which is important. One significant avenue could be incorporating larger and more diverse datasets that capture a wider range of demographics, clinical conditions, and environmental factors. This approach will enhance the model's generalizability and robustness, making it more applicable to different populations and situations, which is crucial for effective healthcare solutions.

In addition to them, Current research excluded mental health symptoms and genetic factors due to dataset limitations and medical complexity. Therefore, inclusion of mental health symptoms and genetic factors will enable these diagnostic models more predictive, increased accuracy, diverse condition consideration and broader audience.

Lastly, optimizing the mobile application for effective operation in offline settings is crucial, especially since many users may find themselves without reliable internet access, and incorporating explainable AI techniques could enhance accessibility, which is important for user understanding and trust. Such improvements will ensure the application remains practical and ethical, making it widely applicable in real- world healthcare scenarios, which is vital for broader implementation, while fostering user trust is essential for better health outcomes while also the ability to function offline and provide clear explanations of AI decisions can significantly impact how users interact with the application, leading to an more comprehensive and effective healthcare solution.

Acknowledgement

The author would like to express sincere gratitude to Dr. Sanvitha Kasthuriarachchi for her invaluable supervision, continuous guidance, and constructive feedback throughout this research. Her expertise and insights have been instrumental in shaping the direction and quality of this study.

The author also extends appreciation to Dr. Dilshan de Silva for his valuable input and support during the research process. Furthermore, special thanks are due to Mr. S.M.B. Harshanath and Ms. Samanthi Eranga Siriwardene for their time, effort, and critical evaluation, which have contributed significantly to enhancing the rigor and completeness of this work. Additionally, the author acknowledges the support of Department of Information Technology in the Faculty of Computing, Sri Lanka Institute of Information Technology, for providing the necessary resources and research facilities. Gratitude is also extended to colleagues and peers for their valuable discussions and support throughout this endeavor.

References

1. Smith, J., et al. "Evaluating Random Forests for Predicting Diabetes Using Clinical and Demographic Data." *Journal of Medical Informatics, 2022*.
2. Zhang, Y., et al. "Gradient Boosting Techniques for Cholesterol Prediction: A Comparative Study." *International Journal of Data Science and Analytics, 2021*.
3. Lee, H., et al. "Development of a Mobile App Using SVM for Hypertension Prediction." *IEEE Transactions on Mobile Computing, 2020*.
4. Kumar, R., et al. "Early Cardiovascular Disease Detection with Mobile-Based Ensemble Models." *Computers in Biology and Medicine, 2023*.
5. Patel, A., et al. "Deep Learning for Multi-Disease Prediction: Challenges and Opportunities." *Health Informatics Journal, 2021*.
6. Gupta, P., and Sharma, N. "Lightweight Predictive Models for Mobile Health Applications." *Mobile Health Computing, 2022*.
7. Rahman, M., and Ali, T. "Hybrid Ensemble Models for Chronic Disease Prediction." *Applied Intelligence, 2020*.
8. Wang, X., et al. "Stacking Ensemble Approaches for Enhanced Disease Prediction." *Expert Systems with Applications, 2021*.
9. Santos, D., et al. "Optimizing Mobile Apps for Disease Prediction with Ensemble Models." Mobile Systems and Applications, 2022.
10. Chen, L., et al. "Cloud-Assisted Mobile Applications for Real-Time Health Analytics." *Journal of Cloud Computing Advances, 2023*.