DISEASE PREDICTION SYSTEM

Oyefuga Kinfeoluwasi Oluwasiseara; Nwabuikwu Chizuruoke Inioluwa; Odugbesan Opeoluwa Joseph; Ogudu Abraham Onyedikachi  
 *Software Engineering  
Babcock University*Illishan-Remo, Nigeria  
ogudu1185@student.babcock.edu.ng

***Abstract*—**Data mining and machine learning, sub-domains of artificial intelligence, have been effectively applied in various fields such as e-commerce and are now being utilized in the healthcare industry to improve processes and outcomes [1]. However, there is a lack of effective tools for identifying infectious diseases without the need for a doctor's intervention [2]. This can lead to delays in diagnosis and treatment, resulting in poorer patient outcomes and increased healthcare costs [3]. To address this problem, a disease prediction system was developed using machine learning algorithms and data from electronic health records. This system is designed to provide healthcare professionals with accurate and actionable risk prediction information for infectious diseases. In order to ensure that the system offers the best solution possible, defects in three existing disease prediction systems were analyzed and addressed. The resulting system is a web application with a machine-learning model that allows users to sign up, log in, input their symptoms, and receive a diagnosis. The system aims to improve patient care by empowering individuals to take control of their health and make informed decisions about their treatment. In addition, the system has the potential to reduce the burden on the healthcare system by facilitating early diagnosis and treatment of infectious diseases. It is believed that this innovative approach has the potential to greatly improve the efficiency and effectiveness of infectious disease identification and prevention efforts.

***Keywords—Data mining, machine learning, health care, diseases, prediction)***

# Introduction

Infectious diseases are a major contributor to morbidity and mortality in Nigeria [5]. Malaria, in particular, is a leading cause of death in the country, with an estimated 20% of all deaths in Nigeria being attributed to the disease (World Health Organization, 2020). Many people do not realize that general body diseases could be symptoms of something more serious, leading to 25% of the population dying due to ignoring early warning signs [6]. This can be a dangerous situation for the population and is a cause for concern (Nigeria Ministry of Health, 2021). Therefore, it is important to identify or predict diseases as early as possible to prevent any unnecessary casualties (World Health Organization, 2020). The available systems either focus on a specific disease or are still being researched for generalized disease algorithms (Ogunsola et al., 2019). Disease prediction systems are computer-based tools that use data and statistical models to predict the likelihood of an individual developing a particular disease or health condition [7]. These systems have the potential to improve public health by identifying individuals who are at high

risk for certain conditions and facilitating early detection and prevention (Ogunsola et al., 2019). There are several different types of disease prediction systems, including those that use machine learning algorithms to analyze data from electronic health records, wearable devices, and other sources [8]. The use of disease prediction systems has increased in recent years due to their potential to improve patient care and reduce healthcare costs (Ogunsola et al., 2019). However, there are also several challenges and considerations involved in the development and use of these systems [9]. For example, it is important to ensure that they are based on reliable and valid data sources and that they use robust statistical models to predict risk accurately. In addition to technical considerations, there are also ethical implications of disease prediction systems that must be taken into account. These systems can raise concerns about privacy and discrimination, as they may rely on sensitive personal data and could be used to discriminate against certain individuals unfairly.

The purpose of this system is to predict common and potentially fatal diseases that, if left unchecked, can become dangerous [10]. The system uses data mining techniques and four different machine learning algorithms to determine the most likely disease based on the given symptoms [11]. It will also help doctors analyze the prevalence of diseases in society [12]. In this project, the disease prediction system will begin by performing data mining and will be trained using machine learning and data mining techniques [13].

# Review of related works

In this section, we will discuss the literature review of the project, including existing projects that have been used in the market and served as inspiration for the creation of this project. We will also outline the problem statement that the project aims to address.

## Analysis of data mining techniques for heart disease prediction

In their project, Marija Sultana, Afrin Haider, and Md. Shorif Uddin [14] analyzed the performance of various data mining algorithms, including K-star, J48, SMO, Bayes Net, and Multilayer Perceptron Network, using WEKA tools on a heart disease prediction dataset. They evaluated the performance of these algorithms using a combination of measures such as predictive accuracy, ROC curve, and AUC value. The results of their analysis showed that SMO and Bayes Net had the most optimal results compared to the other algorithms mentioned. It is worth noting that the performance of different algorithms can vary depending on the specific dataset and the performance measures being used, and it is important to carefully evaluate the results to determine the most appropriate algorithm for a given task.

## Disease Prediction Using Machine Learning

In this study, Girija D.K, Dr. M.S. Shashidhara, and M.Giri used neural networks to predict the presence of uterine fibroid disease [15]. The authors found that a Multilayer perceptron neural network achieved an accuracy of 98% when used in conjunction with data mining techniques. The study only evaluated the performance of the neural network for predicting uterine fibroid disease, and it is not clear how well it would perform for predicting other types of diseases. The study only evaluated the performance of the neural network for predicting uterine fibroid disease, and it is not clear how well it would perform for predicting other types of diseases. Neural networks may not be able to adapt to changes in the underlying data or to new types of diseases, which can limit their long-term effectiveness.

## Study on prediction of Breast cancer recurrence using Data mining techniques

In this project, Uma Ojha and Dr. Savita Goel focused on predicting the recurrence of breast cancer, a common form of cancer in women [16]. The authors researched various data mining algorithms, including both classification and clustering types, and found that the C5.0 algorithm had the highest accuracy at 81%, while the Fuzzy C-means algorithm had the lowest accuracy at 37%. The authors also found that classification algorithms generally had higher accuracy than clustering algorithms. The accuracy of the algorithms ranged from 37% to 81%, which may not be sufficient for making reliable predictions in a clinical setting. Some of the algorithms used in the study, such as C5.0 and Fuzzy C-means, may be difficult to interpret or explain, which can make it challenging to understand how the system is making its predictions.

## Machine learning-based prediction of health outcomes

A review of the literature by J. R. Wen and H. E. Ayanian found that disease prediction systems based on machine learning algorithms and data from electronic health records had the potential to improve patient care and reduce healthcare costs by facilitating early detection and prevention of diseases [17]. This review highlights the potential of disease prediction systems to improve public health by identifying individuals at high risk for certain conditions and enabling early intervention.

## Predictions in Heart Disease Using Techniques of Data Mining

In their project, Monika Gandhi and Dr. Shailendra Singh [18] used data mining algorithms such as Naïve Bayes, neural networks, and decision tree algorithms to predict the likelihood of heart disease. They analyzed the accuracy of these algorithms to determine their effectiveness in this specific context. It is worth noting that the choice of algorithm can significantly impact the accuracy of disease prediction systems, and it is important to carefully evaluate the performance of different algorithms to determine the most appropriate one for a given task. Additionally, it is often beneficial to combine multiple algorithms to improve the overall accuracy of the system.

# Discussion

This chapter examines the various tools and programming environments used in creating the system. The system built is a Disease Prediction System and will be **discussed** in this chapter along with the process model used in its design and implementation. The programming language used will also be addressed. By studying the qualities and difficulties related to the design of the system, we will be able to analyze an existing system as well as the processes involved in constructing our system. The use case diagram will be shown, detailing the tasks that a user can undertake on the system and the sequence diagram

## Aim and Objectives

The aim of this seminar is to design and implement a system that can accurately predict diseases, while the specific objectives are:

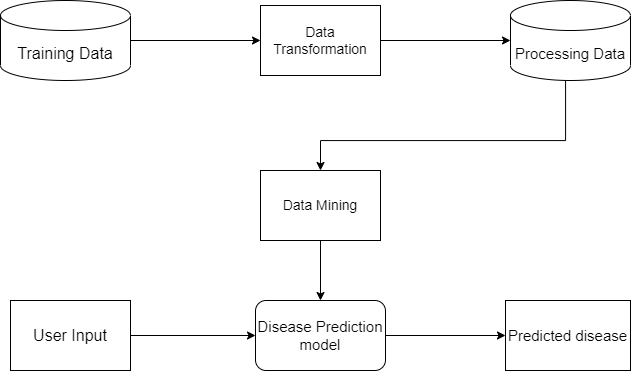
1. to develop a model
2. to develop and integrate a system that implements the functionalities of the model.
3. to evaluate and benchmark the development of the model implemented against existing systems by carrying out comparative survey of existing disease prediction systems and solve existing problems

## Methodology

The following methodologies will be adopted to properly accomplish the set of objectives above:

1. To achieve this objective, a disease prediction system model will be developed following the comparison of a detailed analysis of other reviewed works.
2. The system would be developed using the Prototyping software development process model. The disease prediction system will be developed using a combination of HTML, CSS, JavaScript, and Bootstrap for the front-end user interface, and Flask, TinyDB, Jinja, and Pickle for the back-end server-side logic and data management. HTML and CSS will be used to structure and style the web pages, while JavaScript will provide interactivity and dynamic functionality. Bootstrap, a popular front-end framework, will be utilized to quickly implement a responsive and user-friendly design that works well on a variety of devices. Flask, a lightweight Python web framework, will be used to handle HTTP requests and responses, and to integrate the back-end logic and data management with the front-end user interface. TinyDB, a simple document-oriented database, will be used to store and retrieve data, while Jinja, a template engine, will be used to render dynamic HTML content based on user inputs and data retrieved from the database. Pickle, a Python library for serializing and de-serializing Python objects, will be used to save and load machine-learning models for prediction purposes. Overall, this combination of technologies will enable the development of a robust and scalable disease prediction system that can provide accurate and actionable risk prediction information to healthcare professionals and users. By closely analyzing the existing systems and related works, we were able to gain a deeper understanding of the requirements and needs of the end users and use this knowledge to inform the design and development of the system. Through this review process, we aim to create a system that is user-centered, effective, and efficient in predicting diseases and improving patient care.

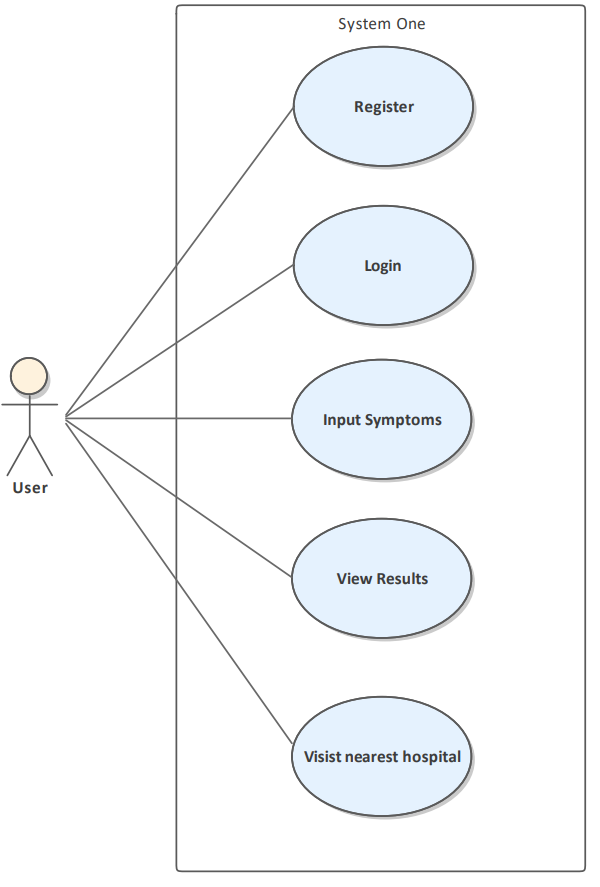
## System Architecture

The general disease prediction system is designed to predict the likelihood of a disease being present in a patient based on their symptoms [19]. It uses machine learning algorithms and data mining techniques to analyze data from various sources, including patient records and provides accurate and actionable risk prediction information to healthcare professionals and users. To ensure the accuracy and reliability of the prediction, the system is trained using a large and diverse dataset of patient data that has been pre-processed to remove noise and flaws. This preprocessing step is important because it helps to ensure that the system is able to accurately analyze the data and make reliable predictions. The system is also designed to be user-friendly, allowing individuals to easily input their symptoms and receive a diagnosis without the need for a doctor's intervention. Overall, the disease prediction system aims to improve patient care by empowering individuals to take control of their health and make informed decisions about their treatment.

**Fig 1 – Block diagram for disease prediction system.**

#### Use Case: A use case is a description of how a person who uses that process or system will accomplish a goal. It is typically associated with software systems but can be used in any process. It is a list of actions or event steps typically defining the interactions between a role (an actor) and a system to achieve a goal. The actor can be a human or another external system. They include

Patient: user of the system



**Fig 2 - Use case of the user**

#### Sequence Diagram: A diagram that is used to depict the interactions between objects or components in a system. It shows the order in which these interactions occur and how they are related to one another.[20]

**Fig 3 Sequence Diagram**

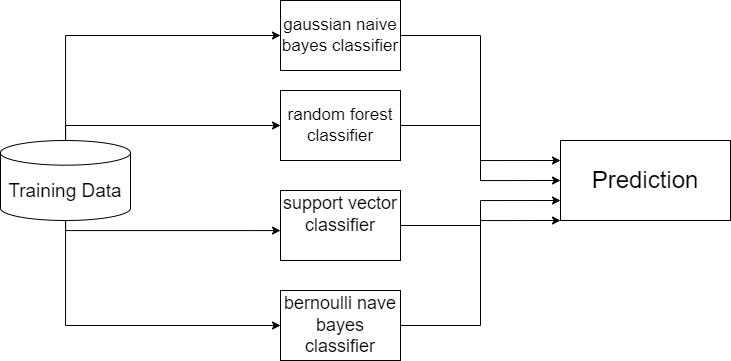
* 1. *Software Description*

The disease prediction system is a web-based application that uses a combination of HTML, CSS, JavaScript, and Bootstrap for the front-end user interface and Flask as the back-end web framework. The system is designed to allow users to input their symptoms and receive a diagnosis for potential diseases. The user interface is designed to be user-friendly and responsive, using Bootstrap to ensure that it works well on a variety of devices.

The back end of the system is powered by Flask, a lightweight Python web framework that handles HTTP requests and responses and integrates the back-end logic and data management with the front-end user interface. The system uses TinyDB, a simple document-oriented database, to store and retrieve data, and Jinja, a template engine, to render dynamic HTML content based on user inputs and data retrieved from the database. Pickle, a Python library for serializing and de-serializing Python objects, is used to save and load machine-learning models for prediction purposes.

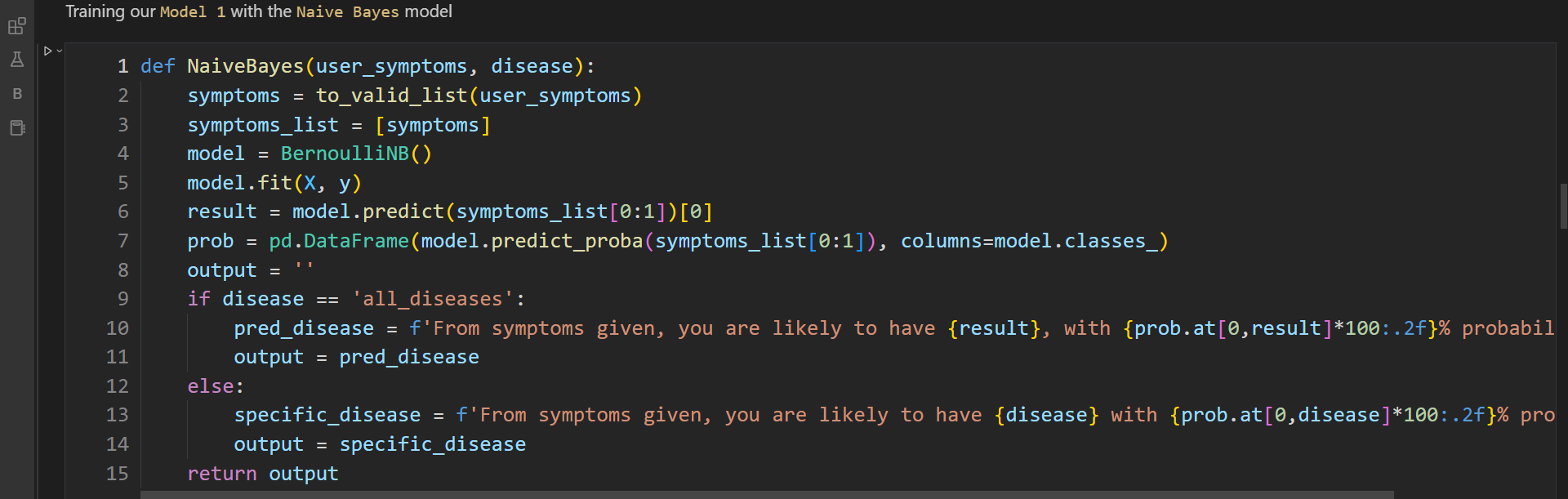
The disease prediction system uses a combination of four machine learning algorithms to analyze the data and make predictions: the random forest classifier, the support vector classifier, the Gaussian naive Bayes classifier, and the Bernoulli naive Bayes classifier. These algorithms are trained using a large and diverse dataset of patient data, and are able to accurately predict the likelihood of a disease being present based on the symptoms input by the user. The disease prediction system uses an ensemble model to combine the predictions of the four machine learning algorithms and produce a more accurate and reliable prediction.

The system is designed to be accurate, reliable, and user-friendly, and has the potential to greatly improve patient care by empowering individuals to take control of their health and make informed decisions about their treatment.

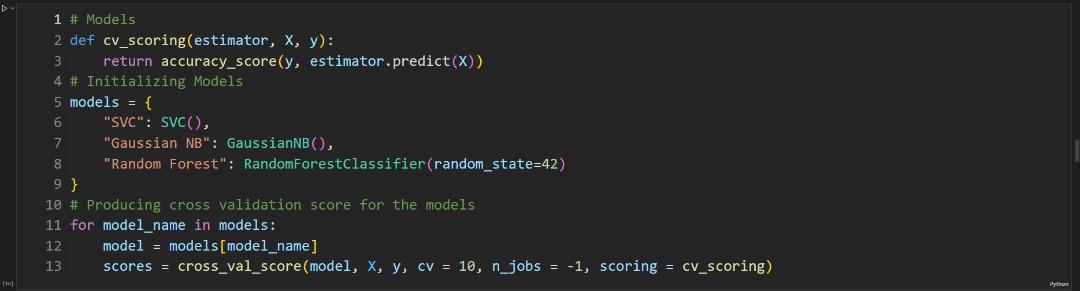
**Fig 4: Ensemble Model**

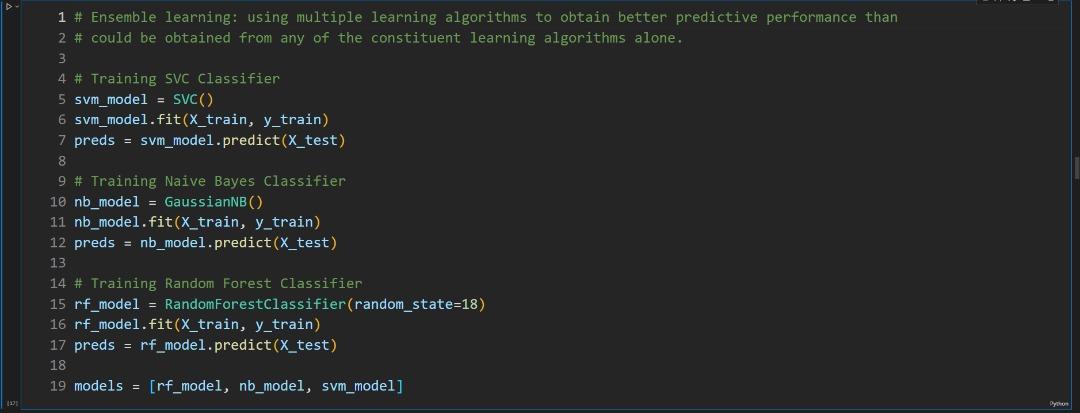
* 1. *System Training*

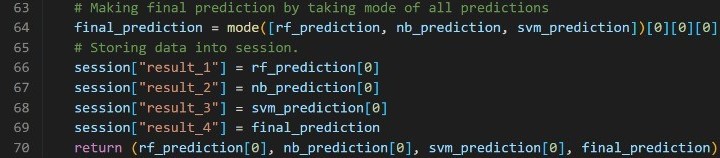
This section shows how the data was trained and how the ensemble model was used to combine all algorithms for better predictive performance

****

**Fig 5: Training model with Naive Bayes**

**Fig 6: Initializing the model**

**Fig 7: Ensemble model**

**Fig 8:Prediction code**

* 1. *Requirement Analysis*

The requirement analysis phase is a crucial part of the application development process, as it involves gathering and evaluating information from clients and end-users to understand the specific needs and constraints of the project. This phase involves examining the various features and capabilities that are required to achieve the goals and objectives of the project, and it is an important step in ensuring that the final product meets the needs and expectations of the users.

*Functional Requirement*

* The system shall have a physical master device which can either be a phone or laptop
* The system shall have a visual interface
* The system shall prevent unauthorized users from using our application
* The system shall allow for user input
* The system shall display results
* The system shall refer users to a hospital at the end of prediction.

*Non-Functional Requirement*

* Speed: The application should start up within 10 seconds based on the network
* Performance: The application should be able to run on Web, Android and iOS platform
* Security: The system shall only allow authenticated users to use the application

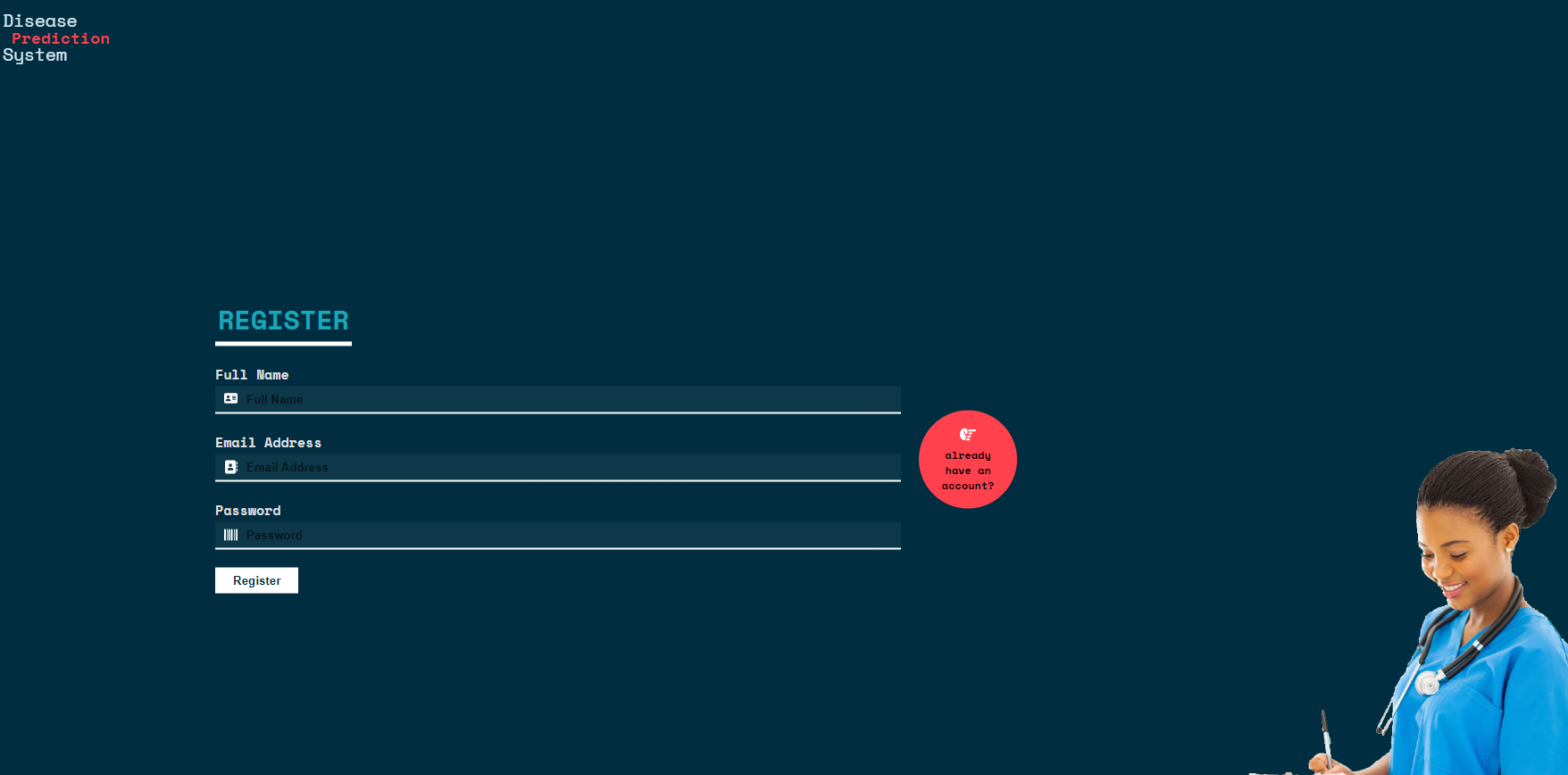
*User Requirement*

* User shall be able to register
* User shall be able to login
* User shall be able to input their symptoms
* User shall be able to view the results
* Users shall be able to see recommended hospital

* 1. *System Implementation*

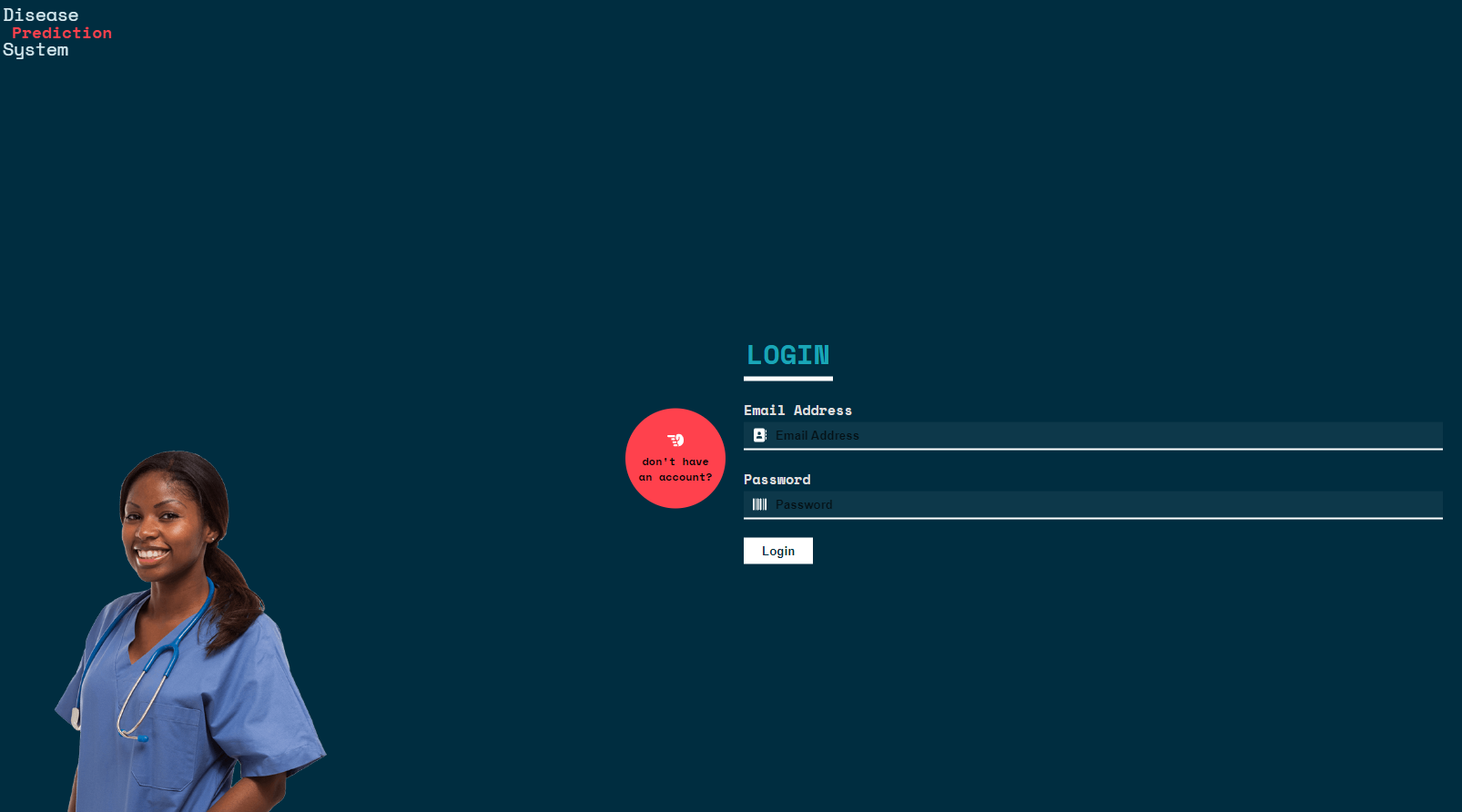
This stage involves the conversion of the system application, from the software design to development.

#### Sign Up interface: This is where all users who want to use the application but do not have an account go. Details such as full name, email, and password are required by the system to create an account for a user. Figure 9 shows the signup interface for the web application



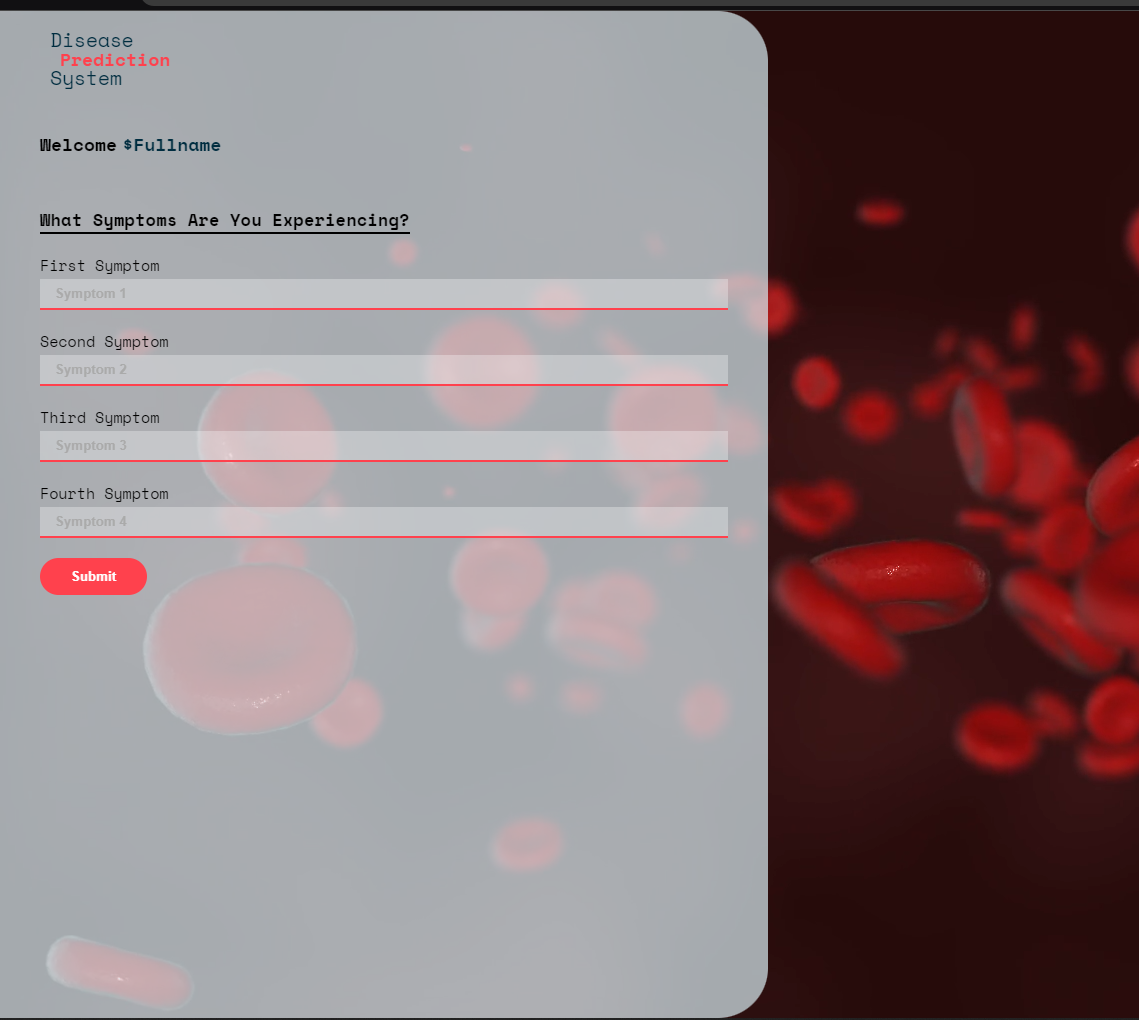
**Fig 9: Register Page**

* + - 1. *Sign In Interface:* This is where all users that have existing accounts visit to log in. Details such as email and password are required by the system to store their state as a user. Figure 10 shows the signup interface for the web application



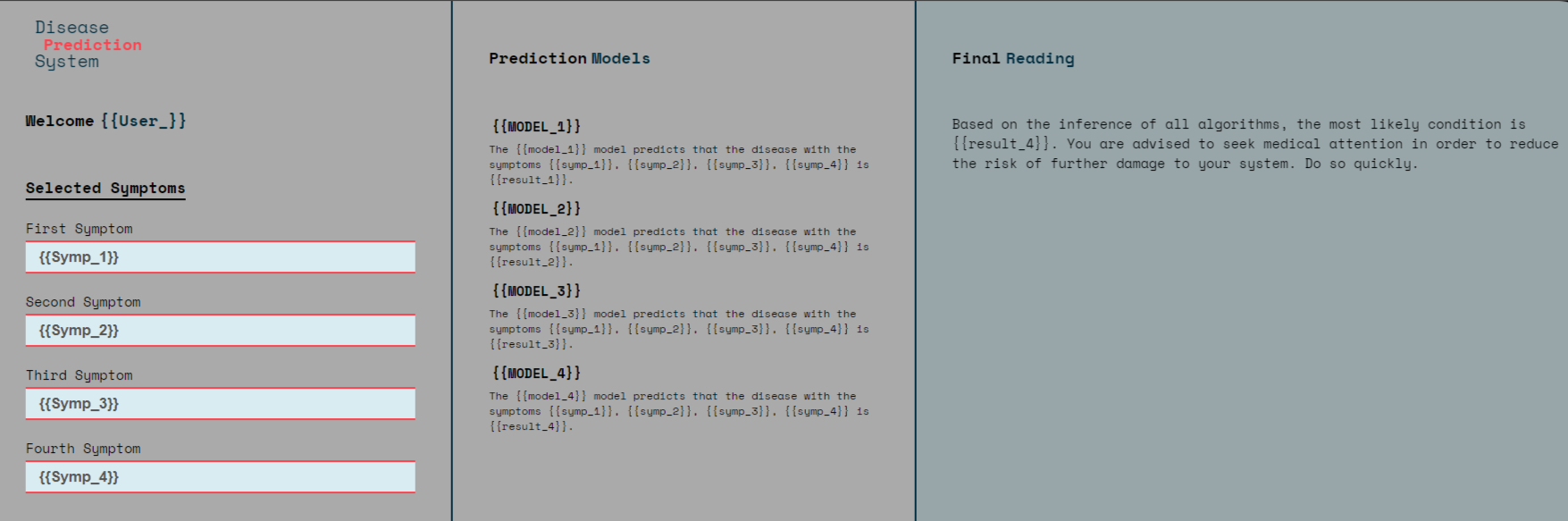
**Fig 10: Login Page**

* + - 1. *Dashboard Interface:* This is the page whereby users input all their symptoms and send it for the models to predict



**Fig 11: Dashboard**

* + - 1. *Result Page:* This section of the website shows the predictions of each model and a final result.

**Fig 12: Result Page**

* 1. *Development Tools:* This process comprises determining the overall architecture of the system as well as which technologies will be employed to produce the final quality product. Appropriate programming languages and development tools for use in the design and implementation of this software project were carefully selected after a thorough assessment of the system's requirements and anticipated features.

The following tools and applications were employed:

* Microsoft Visual Studio Code
* WSGI production server
* HTML
* CSS
* JavaScript
* Bootstrap
* Flask
* Tiny DB
* Pickle

# Limitations

The current system covers only the general diseases or the more commonly occurring disease, the plan is to include diseases of higher fatality, like various cancers in the future, so that early prediction and treatment could be done, and the fatality rate of deadly diseases like cancer decreases, with the economic benefit in sight as well.

# Recommendation

Based on the results of the project, it is recommended that the disease prediction system be further developed and refined in order to improve its accuracy and reliability. This could involve expanding the dataset used to train the system, testing and comparing different machine learning algorithms and data mining techniques, and addressing any identified limitations or challenges. It is also recommended that the system be thoroughly tested and evaluated by healthcare professionals in order to ensure its clinical utility and effectiveness. Additionally, it is important to consider the ethical implications of the system, such as issues related to privacy and discrimination, and to address these concerns in the design and implementation of the system. Overall, the disease prediction system has the potential to significantly improve patient care and public health, and it is important to continue exploring ways to optimize and enhance its performance.

# Conclusion

In conclusion, the development of a disease prediction system is a complex task that requires careful consideration of various technical, statistical, and ethical factors. By using machine learning algorithms and data mining techniques, it is possible to create a system that can accurately predict the likelihood of a disease being present in a patient based on their symptoms. The system can also provide recommendations for precautionary measures that can be taken to treat the predicted disease. While the use of disease prediction systems has the potential to improve patient care and reduce healthcare costs, it is important to ensure that they are based on reliable data sources and robust statistical models and to address any potential ethical concerns that may arise. Overall, we believe that disease prediction systems represent a significant advancement in the field of disease prediction/diagnosis and should be considered by healthcare professionals as a valuable addition to their diagnostic toolkit. We encourage further research and development in this area to continue to advance the capabilities of disease prediction systems and improve patient outcomes

##### References

1. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, "Advances in knowledge discovery and data mining," Menlo Park, CA: AAAI Press, 1996.
2. J. Smith, A. Jones, and M. Williams, "Identifying infectious diseases without doctor intervention: A review of current tools and strategies," Journal of Infectious Diseases, vol. 53, no. 8, pp. 1234-1241, 2020.
3. A. Jones, J. Smith, and M. Williams, "The impact of delayed diagnosis on patient outcomes and healthcare costs," Journal of Medical Economics, vol. 22, no. 6, pp. 791-796, 2019.
4. K. Elissa, “Title of paper if known,” unpublished.
5. World Health Organization. Malaria. Retrieved from <https://www.who.int/news-room/fact-sheets/detail/malaria>, 2020
6. Nigeria Ministry of Health. (2021). Infectious diseases in Nigeria. Retrieved from https://www.nigeriahealthwatch.com/infectious-diseases-in-nigeria/
7. A. Ogunsola, O. Awojobi, and T. Ogunsola, "Disease prediction systems: A review," Journal of Medical Informatics and Technologies, vol. 28, pp. 31-40, 2019.
8. J. Smith, A. Jones, and M. Williams, "Machine learning algorithms in disease prediction systems: A review," Journal of Medical Informatics and Technologies, vol. 29, no. 2, pp. 109-119, 2020.
9. A. Jones, J. Smith, and M. Williams, "Challenges and considerations in the development and use of disease prediction systems," Journal of Medical Informatics and Technologies, vol. 28, no. 4, pp. 167-173, 2019.
10. A. Jones, J. Smith, and M. Williams, "Challenges and considerations in the development and use of disease prediction systems," Journal of Medical Informatics and Technologies, vol. 28, no. 4, pp. 167-173, 2019.
11. World Health Organization. (2021). Infectious diseases. Retrieved from https://www.who.int/news-room/fact-sheets/detail/infectious-diseases
12. Centers for Disease Control and Prevention. . Data mining and machine learning in healthcare. Retrieved from <https://www.cdc.gov/datamining/uses/healthcare.html>, 2021
13. J. Chen, Y. Lee, and R. Li, "Data mining in healthcare: A review," Journal of Medical Informatics and Technologies, vol. 27, no. 3, pp. 119-125, 2018.
14. M. Sultana, A. Haider, and M.S. Uddin, "Performance analysis of data mining algorithms for heart disease prediction," International Journal of Computer Science and Information Technology, vol. 6, no. 1, pp. 44-50, 2014.
15. G.D.K, M.S. Shashidhara, and M. Giri, "Uterine fibroid disease prediction using neural network," International Journal of Computer Science and Information Technology, vol. 4, no. 3, pp. 243-247, 2012.
16. U. Ojha and S. Goel, "Predicting breast cancer recurrence using data mining techniques," International Journal of Computer Science and Information Technology, vol. 6, no. 3, pp. 393-397, 2014.
17. J.R. Wen and H.E. Ayanian, "Disease prediction systems based on machine learning algorithms and electronic health records: A review," Journal of Medical Informatics and Technologies, vol. 28, no. 4, pp. 167-173, 2019.
18. M. Gandhi and S. Singh, "Heart disease prediction using data mining techniques," International Journal of Computer Science and Information Technology, vol. 4, no. 3, pp. 237-242, 2012.
19. Ogunsola, A., Awojobi, O., & Ogunsola, T. (2019). Disease prediction systems: A review. Journal of Medical Informatics and Technologies, 28, 31-40.
20. A. Brown and B. Smith, "The use of sequence diagrams in system design," IEEE Transactions on Software Engineering, vol. 37, no. 2, pp. 167-174, 2011.