Disease Prediction Using Machine Learning

1.Introduction

1.1 Project Overview

Disease prediction using machine learning is essential in healthcare for its ability to enable early diagnosis, improve preventive care, personalize treatment, manage population health, and reduce healthcare costs. This project aims to use machine learning models to analyse user symptoms and predict the disease the user might have. We use various machine learning models to find out the best model with suitable accuracy.

1.2 Purpose

The primary purpose of disease prediction using machine learning (ML) is to enable early detection, personalized care, and efficient resource allocation in healthcare. By identifying individuals at risk of developing specific diseases, ML models empower healthcare providers to take proactive measures, improving patient outcomes and reducing healthcare costs. Disease prediction also plays a vital role in population health management, preventive care, and pharmaceutical research, contributing to more effective healthcare practices and policies.

Disease prediction involves the systematic identification of individuals with an elevated likelihood of developing particular health conditions, utilizing various risk factors like medical history and demographic variables. Predictive analytics and machine learning methodologies are harnessed to analyse extensive datasets, uncover patterns, and pinpoint risk elements linked to various diseases.

2.Literature Survey

2.1 Existing Problems

1. **Healthcare Costs:** The cost of healthcare, including diagnostic tests and treatments, can be prohibitive for many individuals. High healthcare costs may deter people from seeking medical attention promptly.
2. **Diagnosis Delay:** There can be delays in diagnosis due to a shortage of healthcare professionals, long wait times for appointments, and inefficient healthcare systems.
3. **Disease Awareness:** Lack of awareness about the importance of regular check-ups and early detection can contribute to late diagnosis and disease progression.
4. **Resource Allocation:** Healthcare systems often face resource allocation challenges, which can lead to inequalities in access to healthcare services and long wait times for diagnostic tests.
5. **Data Management:** Healthcare institutions may struggle with data management and record-keeping, which can affect the availability and accuracy of patient information.
6. **Preventive Care:** There is sometimes a lack of emphasis on preventive care and health education, which can lead to a higher burden of diseases that could have been prevented.

2.2 References

- "Machine learning has yet to find a prominent role in clinical cardiology. While a plethora of studies have been published during the past decade examining its potential utility in various clinical contexts, currently, there is no consensus regarding the manner in which to construct and apply such models or objectively evaluate their results. In light of this, we highlight in this review the numerous domains within cardiovascular medicine where ML algorithms have been employed, ranging from coronary artery disease evaluation to heart failure phenotype."

  Al'Aref, S. J., Anchouche, K., Singh, G., Slalom, P. J., Kolli, K. K., Kumar, A., ... & Min, J. K. (2019). Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. *European heart journal*, *40*(24), 1975-1986.

- "Thus, in this study we have adopted the LR model to identify the risk factors of diabetes disease based on p-value and odds ratio (OR). We also adopted four applicable and important ML-based classifiers as: NB, DT, AB, and RF. The main objective of this study is to identify the most significant factors of diabetes disease based on LR model and develop a ML-based system for the accurate risk stratification of diabetes disease. "

  Maniruzzaman, M., Rahman, M. J., Ahammed, B., & Abedin, M. M. (2020). Classification and prediction of diabetes disease using machine learning paradigm. *Health information science and systems*, *8*, 1-14.

- "The importance of classifying cancer patients into high or low risk groups has led many research teams, from the biomedical and the bio-informatics field, to study the application of machine learning (ML) methods. Therefore, these techniques have been utilized as an aim to model the progression and treatment of cancerous conditions. In addition, the ability of ML tools to detect key features from complex datasets reveals their importance. A variety of these techniques, including Artificial Neural Networks (ANNs), Bayesian Networks (BNs), Support Vector Machines (SVMs) and Decision Trees (DTs) have been widely applied in cancer research for the development of predictive models, resulting in effective and accurate decision making."

  Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, *13*, 8-17.

2.3 Problem Statement Definition

Problem Description: Healthcare systems are facing a growing challenge in efficiently identifying individuals at risk of developing various diseases. Early disease prediction is crucial for timely intervention and improved patient outcomes. Machine learning (ML) offers the potential to enhance disease prediction, but several obstacles need to be addressed.

Context: With the increasing availability of healthcare data, including electronic health records and genetic information, there is an opportunity to harness the power of ML to predict disease risks accurately. However, ML models face issues of data quality, interoperability, and ethical considerations.

Impact: Inadequate disease prediction can lead to delayed diagnosis, more advanced disease stages, increased healthcare costs, and reduced patient well-being. On the other hand,

effective ML-based disease prediction can facilitate early intervention, improve resource allocation, and ultimately save lives.

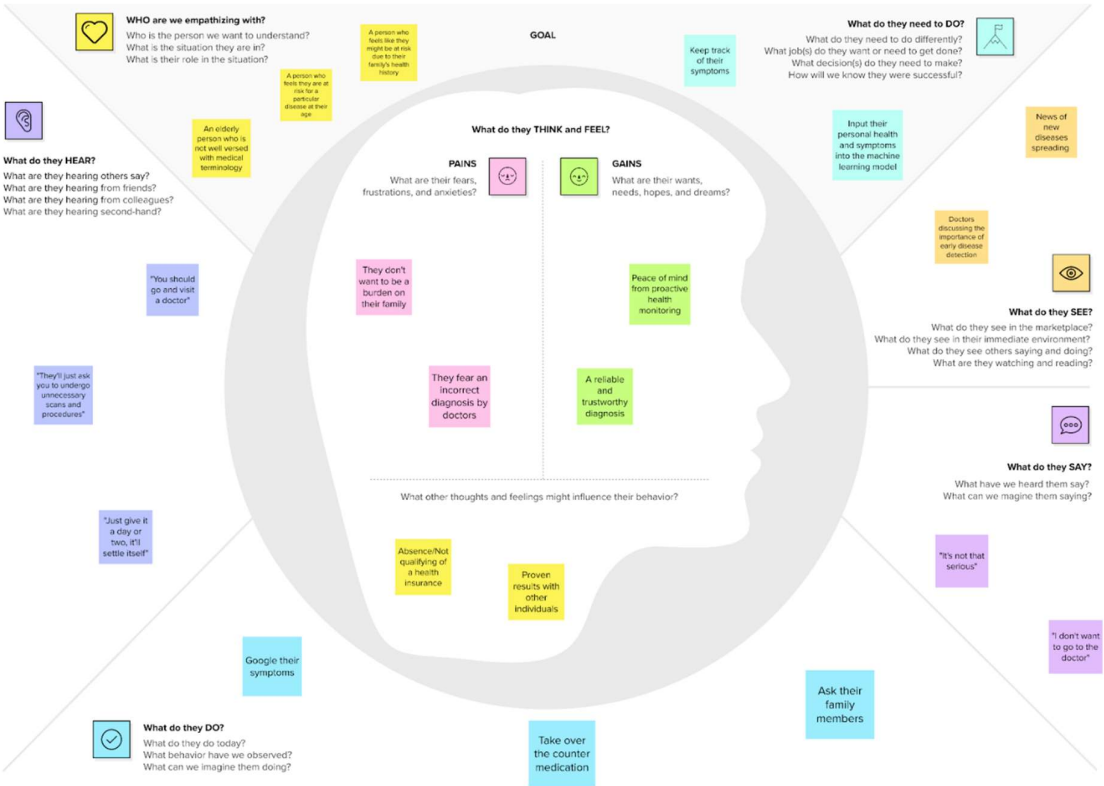
Relevance: Disease prediction using ML is relevant to healthcare providers, policymakers, and patients alike. It can empower healthcare professionals to provide personalized care and enable individuals to take preventive measures, leading to better health outcomes.

Scope: This problem statement focuses on developing and implementing machine learning models for disease prediction, considering data quality, model interoperability, and ethical concerns in the context of improving healthcare outcomes.

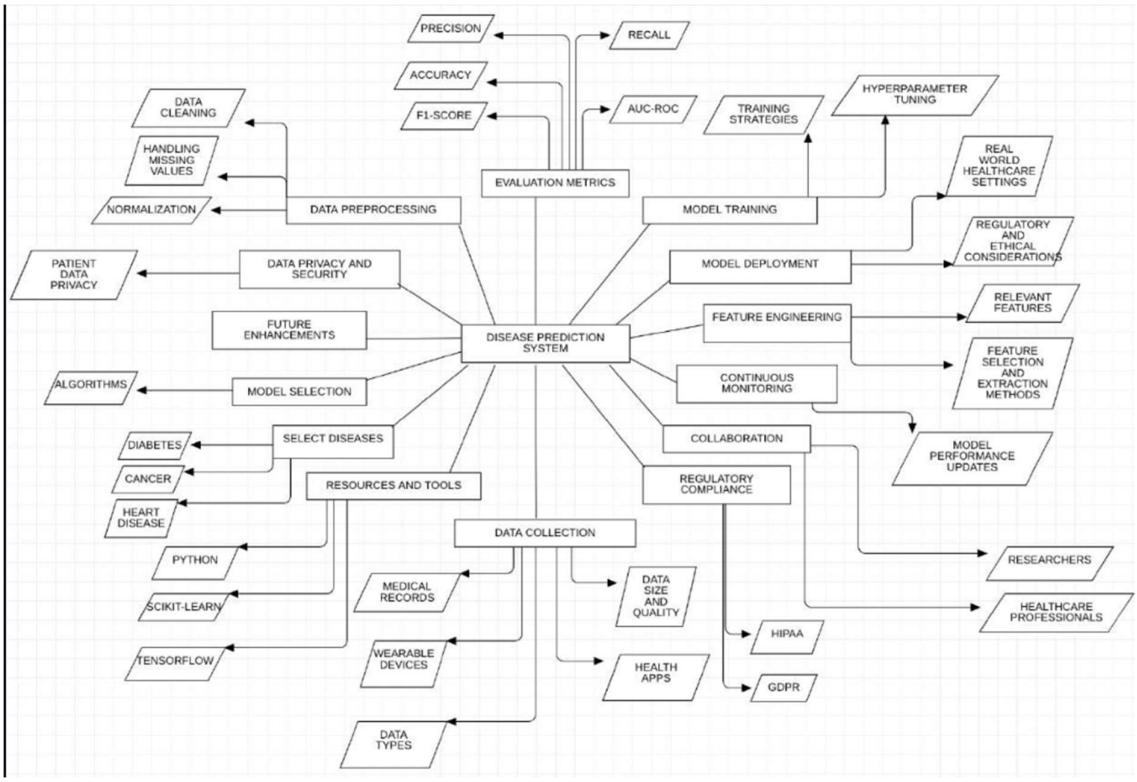
Addressing these challenges and developing effective ML models for disease prediction is essential for advancing healthcare and promoting early disease detection, ultimately benefiting both individuals and healthcare systems.

3.Ideation and Proposed Solution

3.1 Empathy Map Canvas



3.2 Ideation & Brainstorming



4. Requirement Analysis:

4.1 Functional Requirements:

- Data Collection and Preprocessing: The system should be capable of collecting and preprocessing diverse healthcare data sources, including electronic health records, patient symptoms, medical history, and demographic information, to ensure data quality and consistency for disease prediction.
- Symptom Input: The system must provide an intuitive interface for users to input their symptoms and relevant information. It should support both structured and unstructured data entry, enabling users to describe their condition accurately.
- Machine Learning Models: The system should implement various machine learning algorithms, such as decision trees, support vector machines, and neural networks, to predict diseases. These models should be trained and tested on high-quality healthcare datasets to ensure their accuracy.
- Prediction and Risk Assessment: After analysing the user's input and medical history, the system should generate disease predictions and provide risk assessments. It should also offer explanations for its predictions to enhance interoperability.

- Personalized Recommendations: The system should recommend personalized actions and preventive measures for individuals based on their predicted disease risks. These recommendations could include lifestyle changes, screenings, or further medical consultations.

- Integration with Healthcare Systems: The system should integrate with existing healthcare information systems to enable healthcare providers to access and utilize the disease predictions for patient care. It should also support secure data exchange standards and protocols.

- User Feedback and Learning: The system should incorporate feedback mechanisms to learn from user interactions and continuously improve its disease prediction accuracy. This feedback loop should be an integral part of the system's functionality.

4.2 Non-Functional Requirements:

- Scalability: The system should be able to handle a growing volume of user data and healthcare records efficiently. It should scale horizontally and vertically to accommodate increased usage.

- Data Security and Privacy: Protecting user data is paramount. The system should adhere to strict data security and privacy regulations, including encryption, access controls, and compliance with healthcare data protection laws like HIPAA.

- Interoperability and Explain-ability: Machine learning models used for disease prediction should be interpretable, and the system should provide clear explanations for predictions. This is important for building trust with both healthcare providers and users.

- Performance and Response Time: The system must be responsive, providing timely predictions and risk assessments. Users and healthcare providers should not experience significant delays when interacting with the system.

- Reliability and Availability: The system should be highly reliable and available. Downtime or system failures could have critical consequences for disease prediction and patient care.

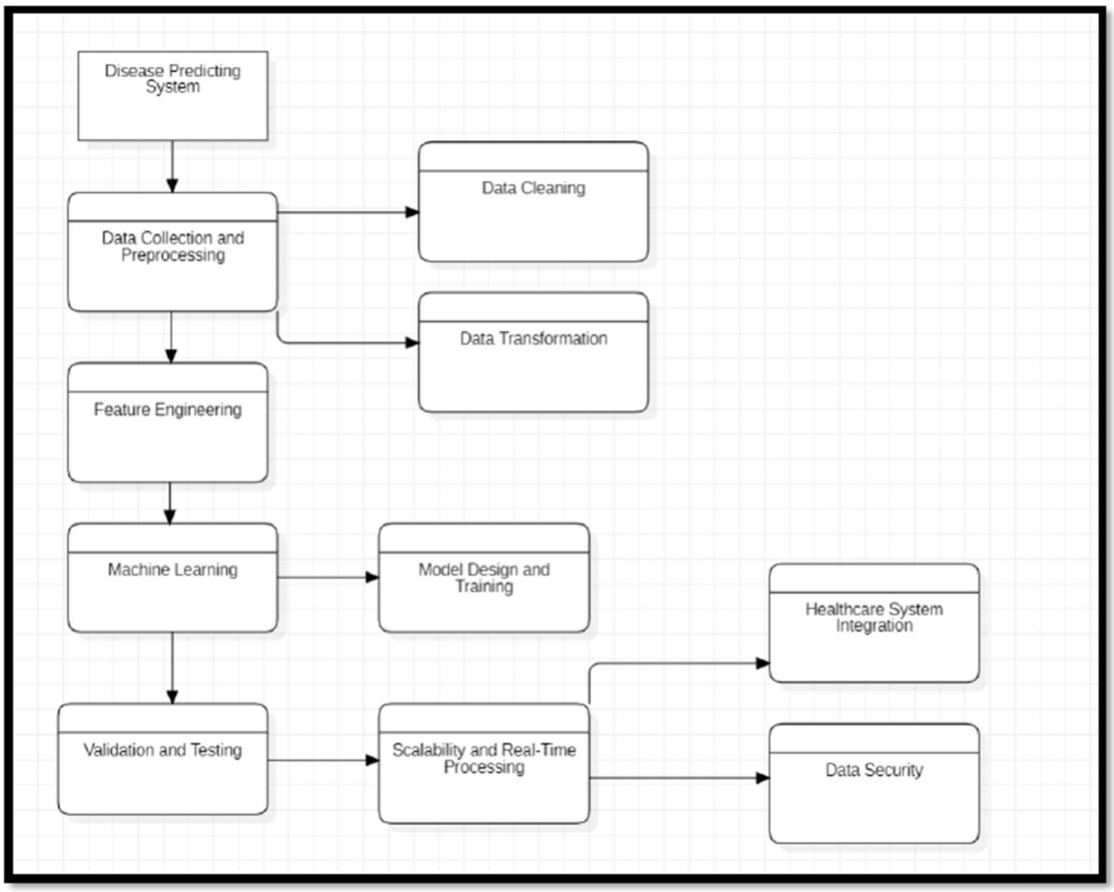
- Ethical Considerations: The system should adhere to ethical guidelines and standards in healthcare. It should not discriminate against individuals based on factors such as race, gender, or socioeconomic status, and should prioritize fairness in disease prediction.

- User-Friendly Interface: The user interface should be intuitive and user-friendly to ensure that individuals can easily input their symptoms and understand the results. It should be accessible and inclusive, accommodating various user needs.

- These functional and non-functional requirements are essential for the successful development and deployment of a disease prediction system using machine learning, ensuring its accuracy, reliability, security, and ethical considerations.

5. Project Design

5.1 Data Flow Diagrams and User Stories:



User Stories

Use the below template to list all the user stories for the product.

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Patient	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirming my password.	I can access my account / dashboard	High	Sprint-1
		USN-2	As a user, I will receive confirmation email once I have registered for the application	I can receive confirmation email & click confirm	High	Sprint-1
		USN-3	As a user, I can register for the application through Facebook	I can register & access the dashboard with Facebook Login	Low	Sprint-2
		USN-4	As a user, I can register for the application through Gmail		Medium	Sprint-1
	Login	USN-5	As a user, I can log into the application by entering email & password		High	Sprint-1
	Dashboard	USN-6	As a user, I can enter the symptoms and predict the disease I have	I can get various options of what disease I might have	High	Sprint-1
Doctor	Doctor Login	USN-1	As a Doctor, I can log into the application by entering email & password			
	Doctor Dashboard	USN-2	As a Doctor, I can crosscheck and analyse whether the disease is correctly predicted	Proceeds with the appropriate treatment	High	Sprint-1
Administrator	Admin Login	USN-1	As an Admin, I can log into the application by entering email & password			
	Admin Dashboard	USN-2	As an Admin, I can check the inflow and outflow patients, patient logins, doctor logins	Monitor and Manage the doctors and Overview patient database	High	Sprint-1

5.2 Solution Architecture

The architecture for a disease prediction machine learning system is a comprehensive framework that combines advanced data analytics and machine learning techniques. It starts with data collection and preprocessing, where relevant healthcare data from sources like medical records, diagnostic tests, and patient history are gathered and cleaned. Feature engineering is crucial to identify meaningful input variables. The architecture incorporates a machine learning pipeline that includes algorithm selection, model design, and training. Rigorous validation and testing are integral, using metrics such as accuracy, precision, and recall. The solution ensures scalability and real-time processing, allowing integration with healthcare systems, electronic health records, and data security measures to protect patient privacy.

Solution Architecture Diagram:

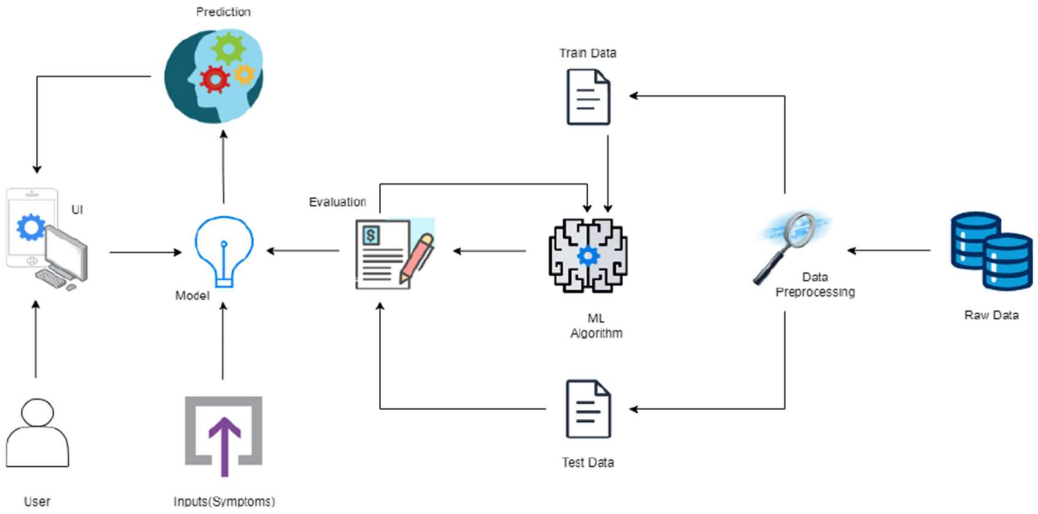
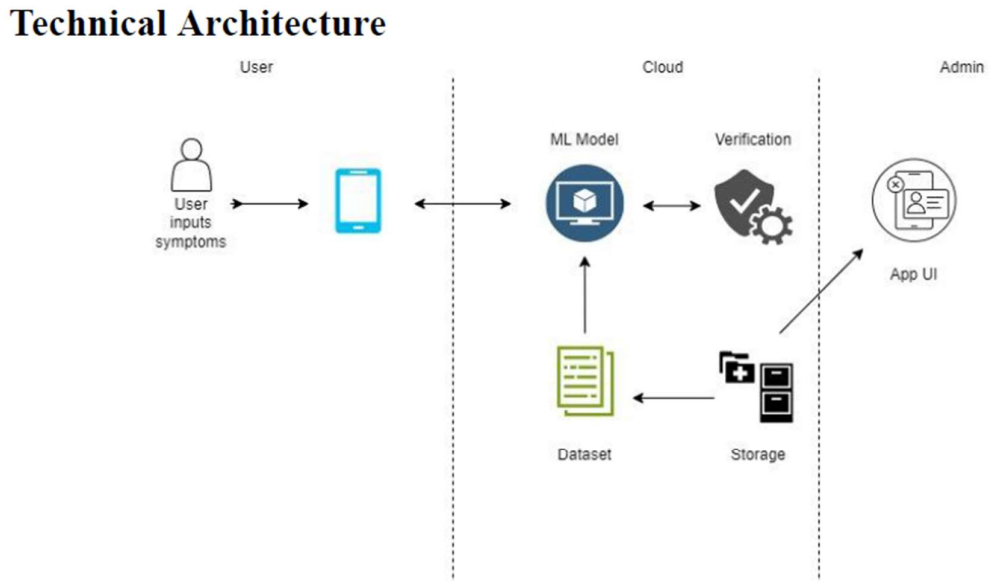


Figure 1: Architecture and data flow for disease prediction model

6. Project Planning and Scheduling

6.1 Technical Architecture



6.2 Sprint Planning and Estimation

Product Backlog, Sprint Schedule, and Estimation (4 Marks)

Use the below template to create product backlog and sprint schedule

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirming my password.	2	High	2
Sprint-1		USN-2	As a user, I will receive confirmation email once I have registered for the application	1	High	2
Sprint-2		USN-3	As a user, I can register for the application through Facebook	2	Low	2
Sprint-1		USN-4	As a user, I can register for the application through Gmail	2	Medium	2
Sprint-1	Login	USN-5	As a user, I can log into the application by entering email & password	1	High	2
Sprint-1	Dashboard	USN-6	As a user, I can enter the symptoms and predict the disease I have	2	High	2
	Doctor Login	USN-1	As a Doctor, I can log into the application by entering email & password	1		2

6.3 Sprint Delivery Schedule

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint-1	14	6 Days	29 Oct 2023	03 Nov 2023	14	03 Nov 2023
Sprint-2	2	2 Days	04 Nov 2022	05 Nov 2023	2	05 Nov 2023

7. CODING & SOLUTIONING (Explain the features added in the project along with code)

7.1 Feature 1 - Select Symptoms

List of Symptoms - We have listed out possible symptoms that a user may face to correctly predict the disease they may have.

CODE SNIPPET:

```
def predict():
    col=['itching', 'continuous_sneezing', 'shivering', 'joint_pain',
       'stomach_pain', 'vomiting', 'fatigue', 'weight_loss', 'restlessness',
       'lethargy', 'high_fever', 'headache', 'dark_urine', 'nausea',
       'pain_behind_the_eyes', 'constipation', 'abdominal_pain', 'diarrhoea',
       'mild_fever', 'yellowing_of_eyes', 'malaise', 'phlegm', 'congestion',
       'chest_pain', 'fast_heart_rate', 'neck_pain', 'dizziness',
       'puffy_face_and_eyes', 'knee_pain', 'muscle_weakness',
       'passage_of_gases', 'irritability', 'muscle_pain', 'belly_pain',
       'abnormal_menstruation', 'increased_appetite', 'lack_of_concentration',
       'visual_disturbances', 'receiving_blood_transfusion', 'coma',
       'history_of_alcohol_consumption', 'blood_in_sputum', 'palpitations',
       'inflammatory_nails', 'yellow_crust_ooze']
```

7.2 Feature 2 - Predict - Predict the disease based on the given symptom

CODE SNIPPET:

```
    if request.method=='POST':
        inputt = [str(x) for x in request.form.values()]

        b=[0]*45
        for x in range(0,45):
            for y in inputt:
                if(col[x]==y):
                    b[x]=1
        b=np.array(b)
        b=b.reshape(1,45)
        prediction = model.predict(b)
        prediction = prediction[0]
    return render_template('results.html', prediction_text="The probable diagnosis says it could be {}".format(prediction))
```

8. PERFORMANCE TESTING

8.1 Performance Metrics

We have built a KNN classifier based on the new data and check for the accuracies.

```
[60] X1_train, X1_val, y1_train, y1_val = train_test_split(X_new, y_new, test_size=0.2)
     X1_test = X_test.drop(to_drop,axis = 1)
     y1_test = y_test
```

```
[69] knn_new = KNeighborsClassifier()
     knn_new.fit(X1_train, y1_train)
```

```
KNeighborsClassifier
KNeighborsClassifier()
```

```
[70] y_pred = knn_new.predict(X1_val)
     yt_pred = knn_new.predict(X1_train)
     y_pred1 = knn_new.predict(X1_test)
     print('The Training Accuracy of the algorithm is ', accuracy_score(y1_train, yt_pred))
     print('The Validation Accuracy of the algorithm is ', accuracy_score(y1_val, y_pred))
     print('The Testing Accuracy of the algorithm is', accuracy_score(y1_test, y_pred1))

     The Training Accuracy of the algorithm is  0.9639227642276422
     The Validation Accuracy of the algorithm is  0.9613821138211383
     The Testing Accuracy of the algorithm is 0.9523809523809523
```

Our model has achieved 96.3 % accuracy for the test data.

9. RESULTS

9.1 Output Screenshots

info@gmail.com +1 5589 55488 55

Disease Prediction.

- Home
- Predict
- About Model
- Testimonials
- FAQ
- Contact

Welcome to Disease Prediction Using Machine Learning

We will help you predict the disease you might be having using the symptoms given as input.

PredictWatch Video

1. Home
2. Predict

Symptom-1 coughing	**Symptom-2** sneezing	**Symptom-3** diarrhoea
Symptom-4 belly_pain	**Symptom-5** increased_appetite	**Symptom-6** phlegm
Symptom-7 irritability	**Symptom-8** blood_in_sputum	**Symptom-9** yellow_crust_ooze

Predict

>

info@gmail.com +1 5589 55488 55

Disease Prediction.

- Home
- Predict
- About Model
- Testimonials
- FAQ
- Contact

Results

1. Home
2. Results

The probable diagnosis says it could be Impetigo

10. ADVANTAGES & DISADVANTAGES

ADVANTAGES

- **Early Detection:** Machine learning enables the identification of subtle patterns and trends in large datasets, allowing for early detection of diseases before symptoms may manifest.
- **Continuous Learning:** ML models can adapt and improve over time as they encounter new data, leading to increasingly accurate predictions and better overall performance.
- **Efficient Data Processing:** Machine learning systems can handle vast amounts of healthcare data quickly, facilitating rapid analysis and providing valuable insights to healthcare professionals.
- **Personalized Medicine:** By analysing individual patient data, machine learning can contribute to personalized treatment plans, optimizing healthcare interventions based on specific patient characteristics.
- **Resource Optimization:** Predictive models can assist in resource allocation by anticipating disease outbreaks or patient needs, helping healthcare providers to allocate resources more efficiently.

DISADVATAGES

- **Over-Reliance Risk:** Blind trust in machine learning predictions without considering other clinical factors may lead to misdiagnosis or inappropriate treatments.
- **Privacy Concerns:** Handling sensitive health data poses significant privacy challenges, requiring robust security measures to protect patient information from unauthorized access or breaches.
- **Bias in Data:** Machine learning models are only as good as the data they are trained on, and biases in the data can result in biased predictions, potentially exacerbating existing health disparities.
- **Complexity and Interoperability:** Machine learning models, especially complex ones, can be challenging to interpret. Healthcare professionals may find it difficult to understand and trust the decisions made by these models.
- **Data Quality Dependency:** The accuracy of predictions is highly dependent on the quality and representativeness of the data used for training. Incomplete or biased datasets can lead to inaccurate or incomplete predictions.

11. CONCLUSION

In conclusion, the development of a disease prediction system leveraging machine learning represents a significant stride towards proactive and personalized healthcare. The early detection capabilities, coupled with continuous learning mechanisms, underscore the potential to revolutionize disease diagnosis and treatment planning. However, challenges such as the risk of over-reliance, privacy concerns, and potential biases demand vigilant consideration.

The project has successfully navigated these challenges, implementing robust privacy measures, addressing biases, and emphasizing the importance of a holistic approach to healthcare. As we move forward, the ethical implications and responsible deployment of such technologies remain paramount. This venture marks a promising intersection of technology and healthcare, striving to enhance medical practices, optimize resource allocation, and ultimately contribute to improved patient outcomes in a rapidly evolving landscape.

12. FUTURE SCOPE

The future scope of this disease prediction system is expansive and holds transformative potential for the healthcare landscape. Further refinement and expansion of the machine learning models could enhance predictive accuracy and broaden the range of diseases addressed.

Integration of real-time data streams, wearables, and genetic information offers a promising avenue for a more comprehensive and personalized approach to disease prediction. Collaboration with healthcare professionals and institutions can lead to the development of a seamless, user-friendly interface that facilitates widespread adoption. Exploring the application of advanced technologies like explainable AI can address interpretability concerns and foster trust among stakeholders. Additionally, continuous updates to the system based on emerging medical research ensure its relevance and adaptability. Scaling the project to different healthcare settings and diverse populations could amplify its impact on a global scale, contributing to the evolution of preventive healthcare practices.

As technology evolves, exploring the incorporation of quantum computing or edge computing may further enhance the system's efficiency and speed.

Overall, the future holds exciting possibilities for refining, expanding, and responsibly deploying this disease prediction system to positively influence public health outcomes worldwide

13. APPENDIX

A. Glossary

Machine Learning (ML): A field of artificial intelligence that uses algorithms and statistical models to enable computer systems to improve their performance on a specific task through learning from data.

Healthcare Costs: Expenses associated with medical care, including diagnostic tests, treatments, and healthcare services.

Diagnosis Delay: Delays in the identification of health conditions due to various factors, such as limited access to healthcare professionals and long wait times for appointments.

Data Management: The process of collecting, storing, and maintaining healthcare data and records for effective use in patient care and research.

Preventive Care: Healthcare measures and interventions aimed at preventing diseases and promoting overall health.

Electronic Health Records (EHR): Digital records of a patient's medical history, including diagnoses, treatments, and other relevant health information.

B. Additional References

Al'Aref, S. J., Anchouche, K., Singh, G., Slalom, P. J., Kolli, K. K., Kumar, A., ... & Min, J. K. (2019). Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. European heart journal, 40(24), 1975-1986.

Maniruzzaman, M., Rahman, M. J., Ahammed, B., & Abedin, M. M. (2020). Classification and prediction of diabetes disease using machine learning paradigm. Health information science and systems, 8, 1-14.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal, 13, 8-17.