

DISEASE PREDICTION THROUGH SYMPTOMS USING DEEP LEARNING

A MINI PROJECT REPORT

submitted by

S R NIKHIL KRISHNAN

TKM23MCA-2057

to

**TKM College of Engineering
(Government Aided and Autonomous)**

Affiliated to

The APJ Abdul Kalam Technological University

*in partial fulfillment of the requirements for the award of the
degree of*

MASTER OF COMPUTER APPLICATION



**Thangal Kunju Musaliar College of Engineering
Kerala**

**DEPARTMENT OF COMPUTER APPLICATIONS
TKM COLLEGE OF ENGINEERING**

NOVEMBER 2024

DECLARATION

I undersigned hereby declare that the project report **DISEASE PREDICTION THROUGH SYMPTOMS USING DEEP LEARNING**, submitted for partial fulfillment of the requirements for the award of degree of Master of Computer Application of the APJ Abdul Kalam Technological University, Kerala is a bonafide work done by me under supervision of **Dr. Nadera Beevi S.** This submission represents my ideas in my own words and where ideas or words of others have been included, I have adequately and accurately cited and referenced the original sources. I also declare that I have adhered to ethics of academic honesty and integrity and have not misrepresented or fabricated any data or idea or fact or source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the institute and/or the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been obtained. This report has not been previously formed the basis for the award of any degree, diploma or similar title of any other University.

KOLLAM

11-11-2024

S R NIKHIL KRISHNAN

DEPARTMENT OF COMPUTER APPLICATIONS

TKM COLLEGE OF ENGINEERING

(Government Aided and Autonomous)

KOLLAM - 691005



CERTIFICATE

This is to certify that, the report entitled **DISEASE PREDICTION THROUGH SYMPTOMS USING DEEP LEARNING** submitted by **S R NIKHIL KRISHNAN (TKM23MCA-2057)** to the **APJ Abdul Kalam Technological University** in partial fulfillment of the requirements for the award of the Degree of Master of Computer Application is a bonafide record of the project work carried out by him under my/our guidance and supervision. This report in any form has not been submitted to any other University or Institute for any purpose.

Internal Supervisor(s)

Mini Project Co-Ordinator

ACKNOWLEDGEMENT

It is my privilege to express profound gratitude to the God almighty and to all those who have supported the shaping of my project.

I am sincerely grateful to **Dr. Sajeeb R, Principal of TKM College of Engineering, Kollam**, for providing the facilities necessary for completing this project and for his inspiration and support. I extend my heartfelt thanks to **Professor Natheera Beevi M., Head of the Department of Computer Applications**, for her valuable suggestions, constant encouragement, and steadfast support in making this project a reality.

I owe special gratitude to my project guide, **Dr. Nadera Beevi S, Professor, Department of Computer Applications**, for her guidance and support throughout the development of this work.

I would like to extend my heartfelt gratitude to my project coordinator, **Prof. Sheera Shamsu**, Assistant Professor, **Department of Computer Applications**, for her invaluable support and guidance throughout the completion of this project

I also express my appreciation to all the faculty members of the Department of Master of Computer Application, TKM College of Engineering, Kollam, for imparting invaluable knowledge and skills that have been instrumental to my growth over the past years.

ABSTRACT

The ability to predict diseases based on symptoms is a transformative step in enhancing healthcare services. This **Symptom-Based Disease Prediction using deep learning** project utilizes machine learning and deep learning models to diagnose potential diseases from user-reported symptoms. By leveraging sophisticated predictive algorithms, the system is trained to extract critical features from natural language symptom descriptions and accurately classify a range of medical conditions

The model processes input text to identify key symptom patterns, analysing descriptions of common and complex symptoms. These extracted features are used for prediction tasks, allowing the system to identify likely conditions such as infections, chronic illnesses, and rare diseases. Additionally, the model is capable of handling diverse symptom inputs, accounting for variations in language and presentation, which enhances its prediction accuracy.

This data-driven approach provides a scalable and non-invasive solution for symptom analysis, supporting early detection of illnesses. When integrated into a clinical or self-assessment setting, the system offers valuable decision support, aiding healthcare professionals and users in making informed choices, thereby contributing to improved patient care and health outcomes.

.

TABLE OF CONTENTS

1.INTRODUCTION.....	1
1.1 Existing System.....	2
1.2 Problem Statement.....	5
1.3 Proposed System.....	6
1.3.1 Data Preprocessing.....	6
1.3.2 Deep and Machine Learning Models.....	7
1.3.3 Prediction and Treatment Information.....	7
1.3.4 Web interface.....	7
1.4 Objective.....	7
1.4.1 Disease Classification.....	8
1.4.2 Data Preprocessing.....	8
1.4.3 Model Evaluation.....	8
1.4.4 Treatment Recommendation.....	8
1.4.5 Web App.....	8
1.4.6 User Empowerment.....	8
1.5 Scope.....	8
2.LITERATURE SURVEY.....	10
2.1 Purpose of Literature Survey.....	10
2.2 Related Works.....	10
3. METHODOLOGY.....	19
3.1. Approach Taken.....	20
3.1.1. Loading and Exploring Data.....	20
3.1.2. Data Preprocessing.....	20
3.1.3. Tokenization and Padding.....	21
3.1.4. Model Architectures.....	21
3.1.5. Model Evaluation.....	24

3.1.6. Model Prediction.....	24
3.1.7. Treatment Information.....	25
3.1.8. Saving Model.....	25
3.1.9. User Interface.....	25
3.2. Software and Tools used.....	25
4. RESULTS AND DISCUSSION.....	28
4.1. Results and Performance.....	29
4.1.1. Bidirectional LSTM.....	29
4.1.2. CNN.....	31
4.1.3. KNN.....	32
4.2. Comparative Analysis.....	33
4.3. Screenshots.....	34
5. CONCLUSION.....	38
5.1. Future Enhancement.....	39
6.REFERENCE.....	42

LIST OF FIGURES

Figure 3.1 - Flowchart.....19

Figure 3.2 - CNN Architecture.....22

Figure 3.3 - LSTM Architecture.....23

Figure 4.1 - LSTM Training and Validation Curve.....30

Figure 4.2 - LSTM Validation Loss.....30

Figure 4.3 - CNN Training and Validation Curve.....31

Figure 4.4 - CNN Validation Loss.....32

Figure 4.5 - Comparative Analysis.....33

Figure 4.6 - Output 1.....34

Figure 4.7 - Output 2.....35

Figure 4.8 - Output 3.....36

Figure 4.9 - Output 4.....37

CHAPTER 1

INTRODUCTION

The rapid advancement in machine learning and artificial intelligence has paved the way for developing predictive models that can significantly enhance healthcare systems. In today's healthcare landscape, the ability to predict diseases accurately based on early symptoms holds enormous potential for improving patient outcomes. Disease prediction through symptoms is especially crucial, as it allows healthcare providers to offer timely and accurate treatment, ultimately reducing the burden on healthcare systems and preventing the escalation of diseases.

This project, "Disease Prediction through Symptoms," aims to leverage machine learning models to predict possible diseases based on patient-reported symptoms. Given that symptoms often overlap across various diseases, achieving high accuracy in such predictions presents a challenging task. For instance, symptoms like fever, fatigue, and cough can be indicators of multiple diseases, including the flu, pneumonia, or even COVID-19. Therefore, to address this challenge, the project explores advanced machine learning techniques, including Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and the K-Nearest Neighbors (KNN) algorithm, to identify the disease with higher precision by analyzing input symptom text.

The model training in this project focuses on processing symptom descriptions through natural language processing (NLP) methods. The text data undergoes cleaning, stop-word removal, tokenization, and padding to prepare it for model input. The project utilizes word embeddings to represent symptom descriptions effectively, allowing the models to discern patterns that correlate symptoms with potential diseases. The CNN and LSTM models are designed to capture semantic patterns in the text, with CNN being particularly effective for shorter, descriptive phrases, and LSTM excelling at identifying sequential dependencies. A K-Nearest Neighbors (KNN) model, combined with Term Frequency-Inverse Document Frequency (TF-IDF) vectorization, is also implemented to provide a comparative analysis of traditional

machine learning against deep learning models. Each model's accuracy and performance are evaluated to determine the best-suited approach for symptom-based disease prediction.

Moreover, this project includes a Flask-based web interface for real-time disease prediction, where users can input their symptoms to receive a list of potential diseases and their respective treatments. Treatment information is derived from an external dataset, providing a comprehensive solution that not only predicts diseases but also suggests relevant treatments. By providing a user-friendly interface, the project aims to make disease prediction accessible, contributing to patient self-care and preliminary diagnosis. Additionally, the model is saved in an accessible format, while the tokenizer and label encoder are preserved for consistent processing, ensuring that predictions remain consistent even as the model is deployed.

In summary, this project exemplifies how machine learning models can be effectively used to interpret natural language descriptions of symptoms and predict potential diseases. Through the combination of CNN, LSTM, and KNN models, this project explores multiple approaches to achieving accuracy in disease prediction. The result is an end-to-end application that holds significant potential for aiding patients and healthcare providers, demonstrating how machine learning can advance diagnostic processes.

1.1 EXISTING SYSTEMS

The number of papers dealing with disease prediction based on symptoms in literature is growing exponentially. Several researchers have played a significant role in the development of disease prediction algorithms.

D. Dahiwade et al., [1] in 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, designed a disease prediction model using machine learning approach. Due to increased amount of data growth in medical and healthcare field the accurate analysis on medical data has benefits on early patient care. Data mining finds hidden pattern information in the huge amount of medical data. In this literature, the aim is to recognize trends across various types of supervised Machine learning models in disease detection through the examination of performance metrics. The most prominently discussed ML algorithms are Convolution Neural Network(CNN), KNearest Neighbour (KNN). It is a

general disease prediction based on symptoms of the patient. The accuracy of general disease prediction by using CNN is 84.5% which is more than KNN algorithm.

S. Grampurohit et al., [2] in International Conference for Emerging Technology (INCET), Belgaum, India, 2020, proposed accurate analysis of medical database benefits in early disease prediction, patient care and community services. The techniques of machine learning are successfully employed in assorted applications including disease prediction. The aim of developing classifier system using machine learning algorithms is to immensely help to solve the health-related issues by assisting the physicians to predict and diagnose diseases at an early stage.

In this literature a sample data of 4920 patients records diagnosed with 41 diseases was selected for analysis. A dependent variable composed of 41 diseases. 95 of 132 independent variables (symptoms) closely related to diseases were selected and optimized. This research work carried out that demonstrates the disease prediction system developed using Machine learning algorithms such as Decision Tree classifier, Random forest classifier, and Naïve Bayes classifier. The paper presents the comparative study of the results of the above algorithms used.

Hong Qing Yu [3] in IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT), Dalian, China, 2019, introduced Experimental Disease Prediction on Combining Natural Language Processing and Machine Learning. He proposed a framework to evaluate the efficiency of applying both Machine Learning and Natural Language Processing technologies for disease prediction system. He used modern computational methods to develop and analyse new approaches that can efficiently predict the disease with reasonable accuracy.

S. Vijaya Shetty et al., [4] in International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2019, designed symptom based health prediction system using data mining. Taking certain prominent symptoms and their diseases to build a Machine learning model to predict common diseases based on real symptoms is the objective of this research.

Feixiang Huang et al., [5] in IEEE International Conference on Granular Computing, Hangzhou, China, 2012 proposed a model to predict a disease by using data mining based on healthcare information system. This paper applies the data mining process to predict hypertension from patient medical records with eight other diseases.

A. Gavhane et al., [6] in Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2018 has discussed about prediction of heart disease using machine learning. The problem is solved using emerging technologies like deep learning to get good results in terms of speed and Accuracy. This work investigated and showed the potential of using DNN-based data analysis for detecting heart disease based on routine clinical data. DNN data analysis techniques can yield very high accuracy.

M. Shankar et al., [7] in Second International Conference on Advances in Computing and Communication Engineering, Dehradun, India, 2015 introduced a method for disease recognition and cure time prediction based on symptoms and proposed a novel method for recognition of diseases and prediction of their cure time based on the symptoms. For predicting the cure time of a disease, reinforcement learning is used. The algorithm takes into account the similarity between the condition of the current user and other users who have suffered from the same disease.

M. Chen et al., [8] in IEEE, 2017 developed a disease prediction model based on machine learning over big data from healthcare communities using convolutional neural network (CNN)-based multimodal disease risk prediction algorithm using structured and unstructured data from hospital. To overcome the difficulty of incomplete data, we use a latent factor model to reconstruct the missing data. Regional chronic disease of cerebral infarction were experimented. The prediction accuracy of our proposed algorithm reaches 94.8%.

With the dataset of the most commonly exhibited diseases, they built a relation to predict the possible disease based on the symptoms which were given as input. The proposed model utilizes the capability of different Machine learning algorithms combined with text processing to achieve accurate prediction. In health industry, it provides several benefits such as pre-emptive detection of diseases, faster diagnosis, medical history for review of patients.

1.2 PROBLEM STATEMENT

Accurate and timely diagnosis is a fundamental pillar of healthcare, as it directly influences the course of treatment and patient outcomes. However, diagnosing diseases based on initial symptoms is challenging, especially when patients present vague or overlapping symptoms. In a healthcare setting, the accuracy of diagnosis often depends on the structured processing of patient information, which includes symptoms, medical history, and test results. Many patients, however, describe their symptoms in free-form text, making it difficult for existing systems to process and interpret this unstructured data efficiently. Additionally, the rise in healthcare data, both structured and unstructured, underscores the need for intelligent systems that can assist healthcare providers in making more informed decisions.

Currently, there is a lack of systems capable of accurately predicting diseases from natural language descriptions of symptoms alone. Traditional diagnostic tools and models primarily rely on structured clinical data and laboratory results, which may not always be accessible or timely. This gap is particularly critical in settings where diagnostic resources are limited, and healthcare providers need reliable preliminary assessments based on symptom descriptions. Moreover, the complex overlap of symptoms across various diseases poses a challenge in narrowing down diagnoses without additional testing. As a result, the need for an intelligent, flexible system that can translate symptom descriptions into probable disease predictions is increasingly vital.

If this problem remains unaddressed, several consequences are likely to persist in healthcare settings. Patients may experience delays in diagnosis and treatment due to the time-consuming nature of traditional diagnostic methods. This delay could lead to deteriorating health outcomes, particularly for conditions that require early intervention. Additionally, healthcare providers, particularly those working in high-demand or resource-constrained environments, may face heightened workloads as they struggle to interpret symptom descriptions quickly and accurately. A machine learning-based solution that can predict diseases from symptom descriptions would improve diagnosis speed and accuracy, optimize resource allocation, and ultimately enhance patient outcomes. By addressing this problem, this project aims to bridge a critical gap in healthcare diagnostics, providing a scalable tool that can assist healthcare providers in offering timely and accurate care based on initial patient-reported symptoms.

The project leverages two advanced deep learning architectures, Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks and a machine learning algorithm K-Nearest Neighbors (KNN), to analyze symptom-based data for disease prediction. The primary objective is to assess which of these models is better suited for disease classification tasks based on their accuracy and reliability. CNNs are applied to extract spatial features from symptom data, utilizing their capability to identify patterns and correlations across different symptoms. LSTMs, on the other hand, are used to capture sequential dependencies in patient histories, enabling the model to learn patterns over time that are critical for accurately predicting diseases based on symptom progression and KNN is used for correctly classifying the class of each disease.

1.3 PROPOSED SYSTEM

The proposed system aims to predict potential diseases based on user-reported symptoms using machine learning techniques. The system leverages a combination of Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and K-Nearest Neighbors (KNN) to classify diseases from textual symptom descriptions. This system is designed to aid in quick disease diagnosis, especially in cases where immediate medical consultation is not possible.

The system takes user inputs in the form of symptom descriptions, processes these texts, and classifies them into disease categories. It uses a dataset containing symptom-disease pairs to train machine learning models that can recognize patterns in the symptoms and suggest possible diseases. In addition to disease prediction, the system provides treatment suggestions for the predicted diseases.

1.3.1 Data Preprocessing

The symptom descriptions are cleaned by removing stopwords, punctuation, and unnecessary characters. This ensures that the text is in a format suitable for analysis by machine learning models. Tokenization and padding are applied to convert the textual data into numerical sequences of equal length, making it suitable for input into the models.

1.3.2 Deep and Machine Learning Models

Convolutional Neural Network (CNN), The CNN model is designed to learn spatial features from the symptom text. It uses convolutional layers to capture local patterns and a global average pooling layer to aggregate information. The model is trained to classify the symptoms into disease categories. Long Short-Term Memory (LSTM), The LSTM model is built to capture sequential dependencies in the symptom text. It is bidirectional, allowing it to understand context from both past and future words in the symptom descriptions. The model is also trained to classify the diseases based on the temporal patterns in the text. K-Nearest Neighbors (KNN), A KNN model is used as a comparison to the deep learning models. This model classifies diseases based on the most similar symptom descriptions found in the training set using the TF-IDF vectorization of the symptoms.

1.3.3 Prediction and Treatment Information

Once the models are trained, they can predict the disease for any given symptom input. The system provides the top three most likely diseases based on the models predictions. After predicting the disease, the system fetches treatment recommendations from a predefined list of diseases and their associated treatments. The treatments are stored in a CSV file and matched with the predicted disease.

1.3.4 Web Interface

The system is integrated into a web application using Flask, where users can input their symptoms through a simple form. After submitting the symptoms, the system processes the text, predicts the disease, and displays the results along with the treatment suggestions on the web page

1.4 OBJECTIVE

The objective of the "Disease Prediction through Symptoms" project is to develop a machine learning-based system capable of accurately predicting potential diseases based on a user's reported symptoms. The system aims to assist healthcare professionals and individuals in quickly identifying possible health conditions, thereby enabling timely medical interventions. Key objectives of the project include:

1.4.1 Disease Classification

To build machine learning models (CNN, LSTM, and KNN) that can classify diseases based on textual descriptions of symptoms. The system will predict diseases with high accuracy by analyzing patterns in the input data.

1.4.2 Data Preprocessing

To preprocess symptom text by removing irrelevant words, stopwords, and non-useful characters, ensuring that the data is cleaned and ready for model training.

1.4.3 Model Evaluation

To evaluate and compare the performance of different machine learning models (CNN, LSTM, and KNN) in terms of accuracy, precision, and reliability for predicting diseases.

1.4.4 Treatment Recommendations

To provide recommended treatments for the predicted diseases, using a database of known disease-treatment pairs, offering users a practical follow-up after diagnosis.

1.4.5 Web Application Development

To design and develop a user-friendly web interface using Flask that allows users to input their symptoms, receive disease predictions, and view treatment suggestions in real-time.

1.4.6 User Empowerment:

To empower users by providing them with an easy-to-use tool that can help them understand their symptoms and potential health conditions, promoting early detection and better management of health concerns.

1.5 SCOPE

The "Disease Prediction through Symptoms" project aims to develop an accessible machine learning-based system that predicts diseases from user-input symptoms and recommends treatments. By utilizing CNN, LSTM, and KNN models, the system processes symptom

descriptions, predicts potential diseases, and suggests relevant treatments through a user-friendly web interface. Key project components include data preprocessing, model evaluation, and web application development using Flask. This project seeks to empower users with preliminary health insights, promoting early detection and encouraging timely medical consultation, while also being adaptable for future scalability in medical applications. The scope of this project encompasses the development of an intelligent system capable of predicting a range of diseases based solely on user-reported symptoms. This initial implementation focuses on the most common and easily recognizable symptoms related to a predefined set of diseases. However, the project has the potential to be extended significantly in several ways like

Expansion of Disease Database:The current version is limited to predicting a set number of diseases based on a fixed dataset of symptoms. In future iterations, the system can be expanded to include a broader spectrum of diseases by integrating larger and more diverse datasets. This will enable the model to handle rare and complex diseases, providing a more comprehensive diagnostic tool.

Integration with Patient Medical History:Currently, the system only considers the symptoms reported by the user at a given time. Incorporating a patient's medical history, such as previous diagnoses, medications, and chronic conditions, could greatly enhance the model's predictive accuracy. By leveraging past health records, the system can better contextualize current

CHAPTER 2

LITERATURE REVIEW

A literature survey, also known as a literature review, involves analysing scholarly sources related to a particular subject. Examining the available literature, it provides a comprehensive overview of the state of the field, allowing you to identify relevant theories, approaches, and gaps in the existing body of knowledge. When conducting a literature review from an audit perspective, the main focus is on evaluating the relevant literature. This process covers information that has been published in a specific field of study and sometimes includes information published within a specific time frame.

2.1 PURPOSE OF LITERATURE REVIEW

1. It gives readers easy access to research on a particular topic by selecting high quality articles or studies that are relevant, meaningful, important and valid and summarising them into one complete report.
2. It provides an excellent starting point for researchers beginning to do research in a new area by forcing them to summarise, evaluate, and compare original research in that specific area.
3. It ensures that researchers do not duplicate work that has already been done.
4. It can provide clues as to where future research is heading or recommend areas on which to focus.
5. It highlights the key findings

2.2 RELATED WORKS

The study of prior research related to this project include:

[1] Human Disease Prediction using Machine Learning Techniques and Real-life Parameters

In recent years, advancements in machine learning have significantly impacted disease prediction in healthcare, with numerous models striving to enhance accuracy and efficiency. The current research landscape shows a concentration on machine learning algorithms like Support Vector Machine (SVM), K-nearest neighbors (KNN), and RUSBoost to detect diseases based on symptoms. These earlier studies focused on recognizing patterns within symptom data to predict disease probability. However, their models often lacked data

transformation and used only raw symptom data, which contributed to lower accuracy levels in predictions.

In response to these limitations, the proposed research model adopts a novel approach that addresses accuracy and reliability challenges faced by previous models. This model introduces a data preprocessing step, which assigns weights to data points based on the rarity of symptoms, thereby transforming the dataset into a more nuanced format that can better capture the relationships between symptoms and disease outcomes. The dataset used originates from a publicly available medical repository on Kaggle, making this approach replicable and transparent.

The proposed model employs a combination of three machine learning algorithms: Random Forest, Long Short-Term Memory (LSTM), and SVM, to leverage the strengths of each. Random Forest, known for handling complex datasets and reducing overfitting, enhances predictive accuracy through multiple decision trees. Meanwhile, LSTM analyzes patient history data to capture temporal patterns and dependencies that are critical for understanding symptom progression over time. Finally, SVM provides a robust classification mechanism, synthesizing insights from the other models to conclude the most probable diagnosis.

The results of this research indicate that the proposed model outperforms previously established methods in terms of accuracy and reliability. Comparative analysis illustrates the advantages and limitations of prior models, underscoring the weaknesses in Naive Bayes, KNN, and earlier SVM models, which were unable to achieve the desired level of precision. The new approach, with its data transformation techniques and combined algorithmic structure, demonstrates an improvement in predictive accuracy, achieving a 97% accuracy rate compared to the maximum of 95% in earlier models.

The Random Forest model's accuracy is further visualized in a confusion matrix, showcasing its performance in correctly identifying disease states, and Figure 5 provides a comparative breakdown of model accuracies, solidifying the proposed model's position as an optimal choice for disease prediction.

When integrated with Random Forest and SVM, the LSTM component adds a dynamic layer to the model, allowing it to make predictions based on both current symptoms and historical data. This holistic approach provides a more comprehensive view of a patient's health profile, which is particularly valuable in clinical settings where patient history and symptom onset

timing are critical factors in accurate diagnosis. SVM's final layer of classification strengthens the overall predictive framework, adding a robust decision-making stage that refines the model's output into precise disease categories.

This research offers valuable contributions to healthcare automation by facilitating early and precise disease detection. With its improved accuracy, this model presents a promising alternative for healthcare providers seeking reliable disease prediction tools, ultimately contributing to better patient outcomes and more efficient clinical decision-making.

The enhancements in the proposed model are noteworthy for addressing several critical aspects often overlooked in prior research, particularly the role of data transformation in improving model performance. By assigning weights to symptoms based on rarity, the model gives greater emphasis to unusual symptoms that may otherwise be diluted in a dataset dominated by more common indicators. This weighted approach enables the model to recognize subtler distinctions across cases, which is essential for detecting diseases with overlapping symptom profiles or rare presentations.

[2] Disease Prediction System Using Symptoms

The integration of technology in healthcare has led to innovative tools for early disease prediction, transforming patient care and improving clinical outcomes. This project leverages data mining and machine learning to develop a diagnostic model capable of predicting various diseases based on a patient's symptoms. By automating parts of the diagnostic process, this model serves as a decision-support system, potentially reducing patient wait times and enhancing accessibility to preliminary diagnoses.

The proposed system employs data mining techniques, specifically the Naive Bayes Algorithm, which is well-suited for classification tasks. Classification in this context involves grouping patients' symptoms into categories associated with specific diseases. The Naive Bayes Classifier, a probabilistic machine learning model, was chosen due to its efficiency in calculating the likelihood of a disease given a set of symptoms, making it a reliable choice for symptom-based predictions. The algorithm applies Bayes' theorem, assuming independence among symptoms, which enables it to efficiently handle large datasets by calculating posterior probabilities for each potential disease. This probabilistic approach allows the model to estimate the probability of various diseases, generating a ranked list of possible diagnoses that can guide patients toward appropriate follow-up actions.

The project's dataset, collected from a comprehensive medical database, contains a large volume of information on disease symptoms and their relationships. After preprocessing the dataset to remove inconsistencies and irrelevant features, the model was trained on this refined data to recognize symptom-disease patterns. This preprocessing phase is crucial, as it ensures that the model learns from high-quality, relevant data, thus enhancing its predictive accuracy and reliability. During testing, the model was able to reach an almost 100% accuracy on the dataset, a significant improvement over existing symptom-based diagnostic systems. This accuracy indicates the model's robustness and its potential applicability across a range of clinical settings.

The proposed disease prediction system is designed to be patient-centric, providing quick and reliable insights into potential health issues. Patients can input their symptoms into the system through an intuitive survey interface. By analyzing these symptoms, the system generates a list of potential diseases, which is a valuable resource for patients who may otherwise face long wait times or logistical barriers in scheduling doctor appointments. This predictive capability empowers patients to take early action regarding their health, facilitating proactive consultations with healthcare providers and potentially preventing the escalation of diseases.

In addition to helping patients, this system can serve as a valuable support tool for physicians. Doctors can use the predictions generated by the model to validate their own assessments, making it a complementary resource for clinical diagnosis. By offering preliminary information, the system helps physicians focus on more complex cases, ultimately improving the efficiency of healthcare delivery. Furthermore, this project embodies the ideal of "prediction is better than cure" by promoting early intervention, thus reducing the likelihood of advanced disease stages and subsequent healthcare costs.

The results achieved by this model underscore its effectiveness as a diagnostic tool, with performance metrics surpassing those of traditional symptom-based models. The almost 100% accuracy on the test dataset demonstrates the model's ability to generalize well to diverse patient cases, making it a significant improvement over comparable systems that rely on simpler algorithms or fewer parameters. This high accuracy is likely due to the Naive Bayes Algorithm's capability of handling large datasets and its efficient probability calculations, which make it suitable for analyzing the varied and often noisy data found in medical records.

To illustrate the performance difference, a comparative analysis between the proposed system and other widely-used diagnostic models. Many existing systems rely on K-nearest neighbors (KNN), decision trees, or logistic regression, which may be effective but often lack the same level of interpretability and efficiency for large-scale symptom classification. Models such as KNN, while popular, can be computationally expensive when dealing with voluminous datasets, whereas decision trees are prone to overfitting on symptom data, making them less reliable for generalization. The Naive Bayes-based model proposed here has been tested to overcome these limitations, providing high accuracy without sacrificing computational efficiency.

The model's performance is further supported by a confusion matrix analysis, showing the high rate of correct predictions. Additionally, provides a graphical comparison of accuracy scores across models, emphasizing the proposed system's lead in predictive performance.

Given the promising accuracy and utility of the proposed system, its deployment within clinical and automated healthcare platforms could greatly benefit both patients and healthcare providers. In a healthcare landscape increasingly focused on AI-driven diagnostics and patient-centered care, this system could serve as an initial screening tool for various healthcare institutions. Hospitals and clinics could implement the system to screen patients based on their symptomatology, allowing for more targeted follow-up and resource allocation.

Further development of this system could also see it integrated with electronic health record (EHR) systems, enabling real-time symptom analysis as patient data is updated. Additionally, the model's potential for expansion includes training it on more diverse datasets, allowing for increased adaptability to new diseases and symptom presentations. With future enhancements, this system could serve as a vital component of telemedicine platforms, where remote patients could receive immediate diagnostic insights before physically consulting a healthcare provider.

By enhancing accessibility to early-stage diagnosis, this project embodies the concept of technology-driven preventative healthcare. The model's high accuracy, combined with its patient-centered design, has the potential to revolutionize the way preliminary diagnostics are conducted, offering a scalable and efficient solution in the push toward automated healthcare. Through this system, the goal of "prediction is better than cure" comes closer to reality, aligning with the ongoing trend of leveraging machine learning to support healthcare and improve patient outcomes.

[3] The Prediction Of Disease Using Machine Learning

The surge in biomedical and healthcare data has made disease prediction models essential in enhancing patient care through early detection and efficient management of diseases. This paper presents a machine learning-based system to predict diseases using patient symptoms as input. By analyzing user-provided symptoms, the system calculates the likelihood of various diseases, offering an accessible diagnostic tool for both patients and healthcare providers. The primary algorithm utilized is the Naïve Bayes classifier, a supervised machine learning approach known for its effectiveness in probabilistic classification. The Naïve Bayes algorithm processes symptoms to output the probability of potential diseases, allowing users to take informed steps regarding their health.

This model includes not only the Naïve Bayes algorithm but also integrates linear regression and decision tree algorithms for broader predictive capability, particularly for diseases such as diabetes, malaria, jaundice, dengue, and tuberculosis. Linear regression helps in identifying trends and correlations between symptoms, while decision trees facilitate interpretability in predictions, making the system user-friendly and easy to understand. These algorithms, when combined, allow the model to provide comprehensive and reliable disease predictions, thus enhancing the system's accuracy and usability for real-world applications.

The project also undertakes a comparative analysis of machine learning algorithms frequently applied in disease prediction, including Decision Tree, Random Forest, Naïve Bayes, Support Vector Machine (SVM), and K-nearest neighbors (KNN). In this study, Random Forest emerged as the most accurate algorithm with an accuracy rate of 98.95%, outperforming others due to its ensemble nature that reduces overfitting and improves robustness. SVM also performed well, with an accuracy of 96.49%, followed by Naïve Bayes at 89.4%, Decision Tree at 84.5%, and KNN at 71.28%. This comparison is consistent with findings across numerous studies, where Random Forest and SVM have demonstrated high accuracy due to their resilience in handling complex datasets and minimizing classification errors.

According to the analysis, Random Forest was utilized in 40 studies and consistently achieved top accuracy scores, supporting its effectiveness in the disease prediction domain. SVM, widely implemented in 30 studies, also demonstrated reliable accuracy rates and is often favored for its precision in binary and multi-class classification tasks. Naïve Bayes, a simpler yet effective algorithm, was used in 24 studies and provides a balanced trade-off between

accuracy and computational efficiency, making it suitable for large-scale medical applications.

The use of Naïve Bayes in this system offers several key advantages. Naïve Bayes is computationally efficient, allowing it to handle large datasets with minimal processing time. This efficiency makes it well-suited for scenarios where quick and real-time predictions are essential, such as in remote health monitoring and telemedicine applications. Naïve Bayes also provides easily interpretable probabilities, which help users understand the likelihood of different diseases based on their symptoms. Additionally, this algorithm is relatively robust to noise and works effectively even with limited training data, making it accessible for initial diagnoses and cases where only partial symptom data may be available.

On the other hand, Random Forest, due to its high accuracy and ensemble structure, offers robustness and flexibility for more complex disease prediction tasks. The algorithm uses multiple decision trees to create an aggregated prediction, thereby mitigating overfitting and enhancing the reliability of its results. The combination of these two algorithms—Naïve Bayes for efficient, probability-based classification and Random Forest for high-accuracy, ensemble-based prediction—ensures that the system provides a balance between speed and precision, making it suitable for diverse healthcare scenarios.

The proposed system has significant implications for healthcare applications, particularly in areas lacking immediate access to clinical services. By allowing patients to input their symptoms and receive a preliminary diagnosis, the system can potentially serve as an initial screening tool for rural or underserved populations, reducing the need for costly and time-consuming hospital visits. This predictive capability empowers users to make informed decisions about consulting healthcare providers, saving both time and resources for patients. For diseases like diabetes and tuberculosis, where early detection is critical for effective treatment, the model's ability to identify high-risk individuals based on symptoms could facilitate timely medical intervention.

Furthermore, this system can be integrated into existing telemedicine platforms, providing remote consultations with AI-driven disease predictions. Physicians can use these predictions to prioritize patients based on disease probability, allowing them to focus on cases requiring urgent attention. For chronic disease management, the model can be extended to monitor patients over time, analyzing changes in symptom patterns to predict disease progression or recurrence.

While the current system demonstrates high accuracy and utility, future enhancements could further increase its applicability in healthcare. Expanding the model to include deep learning techniques, such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, could improve its ability to detect complex diseases and identify temporal symptom patterns. These advanced models, capable of analyzing sequential data, would be particularly valuable in chronic disease tracking, where symptom progression over time provides critical diagnostic information.

Additionally, the system's dataset could be expanded to include more diverse patient data, incorporating variables such as age, medical history, and genetic information to improve prediction specificity. By training on these additional factors, the system could provide personalized disease predictions, tailoring the output to each individual's unique health profile. Integrating the system with electronic health records (EHRs) could further enhance its functionality, allowing for seamless access to historical patient data and enabling real-time updates as new symptoms emerge.

In conclusion, this project exemplifies how machine learning can be effectively applied to healthcare for early and reliable disease prediction. With a combination of Naïve Bayes, linear regression, and decision tree algorithms, the system offers high accuracy and interpretability, making it accessible for both patients and healthcare providers. The comparative analysis highlights Random Forest as the most accurate model, though Naïve Bayes remains advantageous for its computational efficiency and ease of interpretation. The system's predictive capabilities and potential integration with telemedicine and EHRs make it a valuable asset in promoting early diagnosis and improving healthcare accessibility. Through continued enhancements and integration, this model has the potential to contribute significantly to preventive healthcare, aligning with the vision that "prediction is better than cure."

Moreover, the implementation of this system on a larger scale could support public health monitoring and epidemiological research. By aggregating anonymized data from users, healthcare organizations could detect emerging health trends or outbreaks of infectious diseases in real-time. This data could be invaluable in resource planning and allocation, especially during peak times or in response to specific disease outbreaks, thereby aiding public health initiatives.

Disease Prediction Through Symptoms Using Deep Learning

As machine learning models continue to advance, this system can also incorporate real-time symptom data from wearable health devices, which track vital signs such as heart rate, temperature, and oxygen saturation. Integrating such data would enable continuous health monitoring, alerting users to early signs of potential health risks, and facilitating timely medical consultations. With ongoing improvements, this model represents a transformative approach in healthcare, bringing disease prediction, preventive care, and early intervention closer to patients' fingertips and contributing to a healthier society.

CHAPTER 3

METHODOLOGY

The methodology section of this project outlines the various steps taken to develop and implement a disease prediction system based on patient symptoms. The goal of this project is to leverage machine learning models to predict potential diseases from a given set of symptoms, and subsequently recommend appropriate treatments. This is achieved by utilizing different types of models, each with its strengths and weaknesses, to ensure a robust solution.

The primary steps in the methodology include data preprocessing, model development, and evaluation. Initially, data containing symptoms and disease labels is explored and cleaned to ensure its suitability for training machine learning models which is shown in *Figure 3.1*. Various models, including Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and K-Nearest Neighbors (KNN), are then trained on the data. Each model's performance is assessed based on key metrics, such as accuracy, and compared to determine the best-performing model. Finally, the selected model is integrated into a user-friendly web application that predicts diseases and recommends treatments based on the user's input symptoms.



Figure 3.1-Flowchart

3.1 APPROACH TAKEN

3.1.1 Loading and Exploring Data

The primary dataset used is Symptom2Disease.csv, containing symptom descriptions in text format (text column) and corresponding diseases (label column). Initial exploration includes checking unique disease labels, text lengths, and visualizing the frequency of words in symptom descriptions using histograms and word clouds. The goal of the exploratory data analysis is to understand symptom length distributions and frequent terms to inform the model design.

In this section, you load the dataset containing symptoms and the corresponding disease labels. You can analyze the structure and the contents of the data, including checking for missing values, unique disease labels, and the distribution of symptoms. By reviewing the first few records, you get an initial understanding of the data format. Visualizations such as histograms and bar plots can help in understanding how long the symptom descriptions are and the frequency of each disease label.

You also explore word clouds to visualize the most frequent terms in the symptom descriptions, providing insights into which words or phrases are common across different symptoms.

3.1.2 Data Preprocessing

The preprocessing stage involves several tasks to clean and prepare the data for training. Key tasks include:

Label Encoding: Since machine learning models cannot process string labels, label encoding converts the disease labels into numerical form.

Text Cleaning: Symptom descriptions are typically raw, so tasks like removing punctuations, converting text to lowercase, and eliminating stopwords (common words that don't contribute to the meaning) are essential. NLTK (Natural Language Toolkit) is used to handle stopwords, with additional words added for better coverage.

Splitting the Data: The dataset is divided into training and validation sets, ensuring the model gets trained on one subset and validated on another to evaluate its generalization performance.

Tokenization and Padding: Symptom texts are tokenized into sequences (arrays of integers) and padded to a fixed length to ensure uniform input size for models.

3.1.3 Tokenization and Padding

Tokenization: A Tokenizer is created to convert text into sequences of integer tokens. Out-of-vocabulary tokens are handled with a special <OOV> token. Tokenization involves converting each word in the symptom description into a unique integer. This helps transform the text data into a format suitable for neural network processing.

Padding: Each tokenized sequence is padded to a maximum length (50) to ensure uniform input dimensions for the models. Padding is essential for CNN and LSTM models, as they require fixed input sizes. Padding ensures that all input sequences (texts) are of the same length. Padding helps prevent issues when sequences of different lengths are passed into a model, ensuring consistency.

3.1.4 Model Architectures

(i) Convolutional Neural Network (CNN)

A Convolutional Neural Network (CNN) is a type of deep learning model specifically designed to process structured grid data, such as images or time-series data. It leverages layers of convolutional filters (also called kernels) to automatically learn spatial hierarchies of features from raw input data, progressively extracting more complex patterns as the data passes through deeper layers. CNNs are particularly effective in tasks like image recognition, classification, and object detection due to their ability to detect features like edges, textures, and shapes. By utilizing pooling layers to reduce dimensionality and fully connected layers for final classification, CNNs are able to efficiently learn and make predictions from visual or sequential data. *The Figure 3.2 shows the architecture of CNN.*

A CNN is a type of deep learning model traditionally used for image data but can also be effective for sequence data, such as text. In this project:

The CNN model includes an embedding layer to convert words into dense vector representations. A convolutional layer is used to extract patterns from local regions in the sequence, helping capture the relationship between words in close proximity. Global Average Pooling is applied to reduce the dimensionality of the feature maps, aggregating the features learned by the convolutional layer. Dense layers with activation functions like ReLU are used to predict the disease class, with a softmax output layer to output a probability distribution across all possible diseases. Dropout is included in the model to prevent overfitting by randomly setting some of the activations to zero during training.

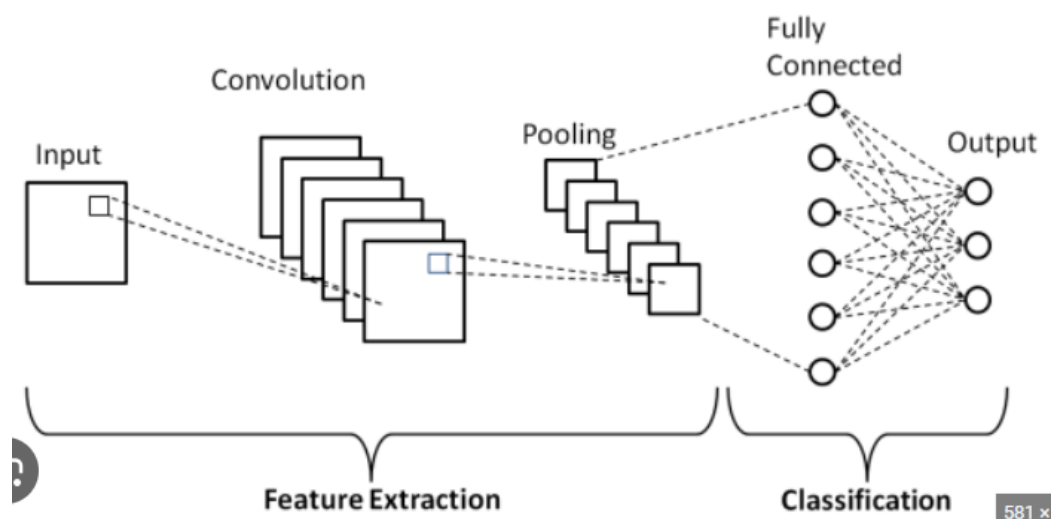


Figure 3.2 - CNN Architecture

(ii) Long Short-Term Memory (LSTM)

A Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) designed to better handle long-term dependencies in sequential data. Unlike traditional RNNs, which struggle with vanishing gradients over long sequences, LSTMs use a specialized architecture that includes memory cells and gates (input, forget, and output gates) to regulate the flow of information. These gates allow the LSTM to remember important information for long durations and forget irrelevant data, making it effective for tasks such as time-series prediction, speech recognition, and natural language processing. LSTMs are particularly

valuable in applications where capturing temporal dependencies and context across time is crucial. *Figure 3.3* shows the architecture of LSTM.

In this case, a Bidirectional LSTM is used, which processes the sequence from both directions (forward and backward) to better understand the context.

Like the CNN model, it also has an embedding layer, dropout layers, and dense layers. The LSTM layer helps capture the temporal relationships in the data, which is particularly useful for understanding sequences such as symptom descriptions.

The model is trained similarly, using sparse categorical cross-entropy as the loss function to handle multiple classes (diseases).

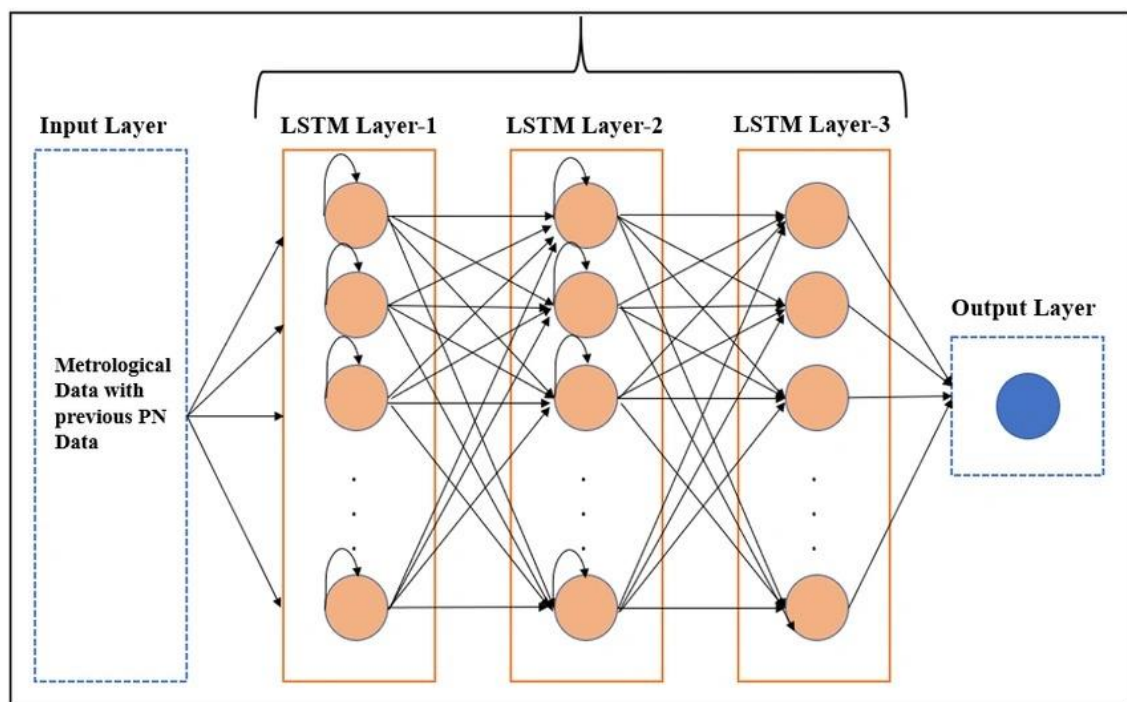


Figure 3.3 – LSTM architecture

(iii) K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a simple, instance-based machine learning algorithm used for both classification and regression tasks. The algorithm works by identifying the "k" closest data points to a given input and making predictions based on the majority class (for

classification) or the average (for regression) of these neighbors. KNN makes predictions based on the assumption that similar data points exist near each other in the feature space. The distance between data points is typically measured using metrics like Euclidean or Manhattan distance. KNN is non-parametric, meaning it makes no assumptions about the underlying data distribution, and is often used in situations where the decision boundary is complex or non-linear. However, it can be computationally expensive during inference, especially with large datasets, as it requires calculating the distance to every point in the training set.

KNN is a simple machine learning algorithm used for classification based on proximity. In this case: Symptom descriptions are first transformed into TF-IDF (Term Frequency-Inverse Document Frequency) features, which represent the importance of each word in the context of the entire corpus. The KNN model is trained on these TF-IDF features, and predictions are made by looking at the nearest neighbors (most similar symptom descriptions) and classifying based on the majority class. The accuracy score is calculated to evaluate the model's performance on the validation dataset.

3.1.5 Model Evaluation & Comparison

The performance of each model (CNN, LSTM, and KNN) is evaluated using the validation dataset. Key metrics include accuracy, which measures how often the model predicts the correct disease label. The models are compared side-by-side through bar plots to visualize which one performs best.

This section helps in choosing the most effective model for predicting diseases from symptoms based on their performance metrics (accuracy).

3.1.6 Model Prediction for Symptoms

After training the models, they are used for making predictions on new, unseen symptom descriptions. Symptom descriptions provided by users (e.g., "chills, fatigue, a cough, a high fever, and difficulty breathing") are preprocessed (tokenized and padded) before being fed into the trained models.

The models predict the most likely disease labels, which are then mapped back to their original names using the label encoder. This process helps in identifying the top diseases that could match the symptoms described.

3.1.7 Treatment Information Retrieval

Once the disease is predicted, the model's output is mapped to a treatment suggestion using a treatment dataset. This dataset contains disease-treatment pairs. The treatment recommendation is displayed to the user, providing information on the necessary medical steps to take for the predicted disease.

3.1.8 Saving the Models

Once the models (CNN, LSTM) are trained and evaluated, they are saved for future use. The Keras save function is used to save the CNN model in HDF5 format. The tokenizer and label encoder are saved using Pickle to ensure that they can be loaded later during the prediction process. This allows the Flask web application to load the models and necessary preprocessing components without needing to retrain them.

3.1.9 Flask Application: User Interface

The Flask web framework is used to build a simple web interface for users to input their symptom descriptions and receive disease predictions and treatment suggestions. The application allows users to input symptoms through an HTML form. When the form is submitted, the backend performs the necessary preprocessing steps, uses the trained model to predict the disease, and then retrieves the corresponding treatment. The results are displayed on the web page, offering users a seamless experience in predicting diseases based on symptoms.

3.2 SOFTWARES AND TOOLS USED

Python: Python is the primary programming language used for this project due to its simplicity, readability, and the extensive ecosystem of libraries and frameworks available for machine learning and data science. Python is used for data preprocessing, model building, training, evaluation, and web application development.

Disease Prediction Through Symptoms Using Deep Learning

TensorFlow/Keras is an open-source framework for machine learning, and Keras is a high-level API for building deep learning models. These libraries are used for building and training deep learning models like CNNs and LSTMs.

Scikit-learn: A library for traditional machine learning algorithms.: Scikit-learn is used for simpler machine learning models like KNN and for preprocessing tasks like feature extraction and scaling.

Pandas: A Python library for data manipulation and analysis. Pandas is used for handling structured data, cleaning datasets, and analyzing the data before feeding it into machine learning models.

NumPy A library for numerical computations and matrix operations. NumPy is essential for handling large arrays and matrices of data during the preprocessing phase and while training the models.

Matplotlib/Seaborn Visualization libraries in Python for creating static, animated, and interactive plots. These libraries are used to plot graphs for data exploration, model performance evaluation, and visualizing metrics such as accuracy and loss over epochs.

NLTK (Natural Language Toolkit): A suite of libraries and programs for natural language processing. NLTK is used for text-based symptom data preprocessing, including tokenization, stemming, and stop-word removal.

WordCloud: A library for generating word clouds. WordCloud is used for visualizing frequent symptoms in the dataset, helping to understand the most common symptoms for disease prediction.

Google Colab: A cloud-based Jupyter notebook environment with access to GPUs and TPUs for faster computations. Google Colab is used to train deep learning models, especially when using large datasets or complex models requiring significant computational resources.

Flask: A lightweight web framework for Python that helps to create web applications. Flask is used for building a RESTful web service that interacts with the trained machine learning model to provide real-time predictions via a web interface.

Disease Prediction Through Symptoms Using Deep Learning

HTML/CSS/JavaScript: The fundamental technologies for web development. These are used to build the front-end of the application where users can input symptoms and view the predicted results

Visual Studio Code (VS Code): Visual Studio Code (VS Code) is a free, open-source code editor developed by Microsoft. It is lightweight, highly customizable, and comes with built-in support for many programming languages, including Python, JavaScript, and HTML. It also provides integrated Git version control, debugging tools, and various extensions for enhanced productivity. VS Code is used for writing and editing the project code. It supports Python, HTML, CSS, JavaScript, and other languages used in the project.

CHAPTER 4

RESULTS AND DISCUSSION

The disease prediction system was evaluated using three different machine learning models: Convolutional Neural Network (CNN), Bidirectional LSTM, and K-Nearest Neighbors (KNN), each leveraging distinct methodologies for analyzing symptom descriptions.

The dataset contained pairs of textual descriptions of symptoms and their corresponding disease labels. After preprocessing, the data was split into training (80%) and validation (20%) sets to evaluate model performance.

The Models Evaluated were:

CNN Model: This model utilizes convolutional layers to extract key patterns from text input, making it effective at recognizing common n-gram features within symptom descriptions.

LSTM Model: The Bidirectional LSTM model captures sequential patterns and dependencies, helping to better understand the context and flow of symptom descriptions.

KNN Model: The KNN classifier was trained using TF-IDF features, a method that transforms text data into vectorized form by emphasizing important terms, allowing the KNN model to classify diseases based on symptom similarity.

The primary metric used for model evaluation was validation accuracy, which measures the proportion of correct predictions on unseen data.

Additionally, the models were assessed based on their ability to provide the top 3 most probable diseases, improving the usability of predictions in cases where symptoms are ambiguous or indicative of multiple conditions.

Accuracy Comparison: Among the models tested, the CNN achieved the highest validation accuracy, followed closely by the KNN and LSTM models.

Model Interpretability: While deep learning models like CNN and LSTM offered higher accuracy, the KNN model provided easier interpretability based on symptom similarity.

Real-world Applicability: The system was tested on several unseen symptom inputs, demonstrating consistent and reliable predictions, suggesting its potential utility for aiding in preliminary disease diagnosis.

4.1 RESULTS AND PERFORMANCE

The implemented project on disease prediction using symptom descriptions utilizes three models: Convolutional Neural Network (CNN), Bidirectional LSTM, and K-Nearest Neighbors (KNN). In this section, we delve into the performance of three predictive models used for symptom-based disease classification: the Convolutional Neural Network (CNN), the Bidirectional Long Short-Term Memory (LSTM), and the K-Nearest Neighbors (KNN). By examining both training and validation metrics, we aim to provide a comprehensive understanding of each model's strengths, weaknesses, and overall effectiveness. The performance of these models was evaluated based on their accuracy on the validation dataset, and the results are summarized as follows:

4.1.1 Bidirectional LSTM

The Bidirectional Long Short-Term Memory (LSTM) model exhibited high training accuracy, reaching 98.78%, with a low training loss of 0.0529. However, the validation accuracy was significantly lower at 85.42%, accompanied by a higher validation loss of 0.7758. This discrepancy highlights a considerable level of overfitting, where the model performed well on the training data but struggled to generalize to new, unseen data. The LSTM architecture, designed to capture sequential dependencies in text, was advantageous for understanding the order of symptoms described by patients. Its bidirectional nature allowed the model to learn from both past and future contexts within the text, enhancing its ability to interpret nuanced symptom descriptions. Despite these advantages, the pronounced overfitting indicates that the model's complexity, combined with the limited size of the training dataset, hindered its ability to generalize. Future improvements could involve techniques such as reducing the model complexity, using more extensive datasets, or incorporating attention mechanisms to better focus on key symptoms. The *Figure 4.1* shows the training and validation accuracy of the bidirectional LSTM model and the *Figure 4.2* shows the validation loss of the bidirectional LSTM model during evaluation.

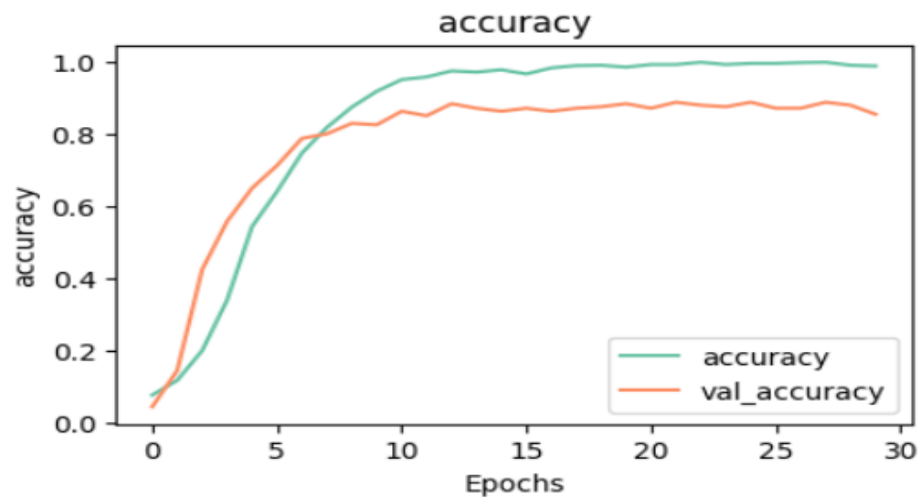


Figure 4.1-LSTM Training and Validation

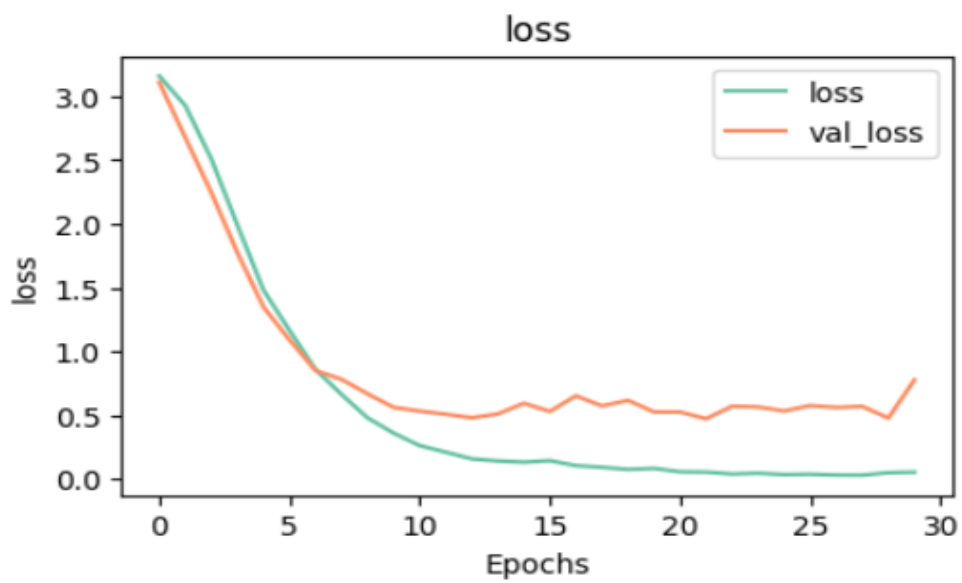


Figure 4.2-LSTM Validation Loss

4.1.2 Convolutional Neural Network (CNN)

The Convolutional Neural Network (CNN) demonstrated strong predictive capabilities, achieving a training accuracy of 98.28% with a relatively low training loss of 0.0786. During validation, the CNN reached an accuracy of 88.33%, with a validation loss of 0.4322. The close alignment between training and validation accuracy suggests that the model effectively generalized to unseen data, although the slight decrease in validation accuracy indicates some degree of overfitting. The CNN's architecture, which leverages convolutional layers to capture local features in text, proved effective for identifying patterns in medical symptom descriptions. However, the small gap between training and validation performance indicates room for improvement, potentially through additional regularization techniques or data augmentation. Overall, the CNN model's strong performance in capturing relevant text features made it a suitable choice for this task, particularly in handling common patterns found in symptom data. The *Figure 4.3* shows the training and validation accuracy during evaluation and the *Figure 4.4* shows the validation loss during the evaluation.

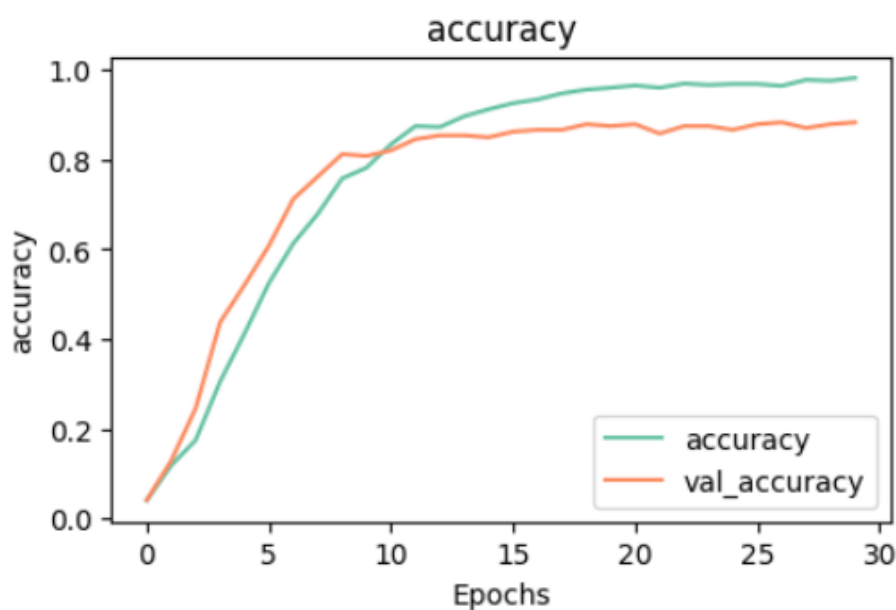


Figure 4.3 -CNN Training and Validation

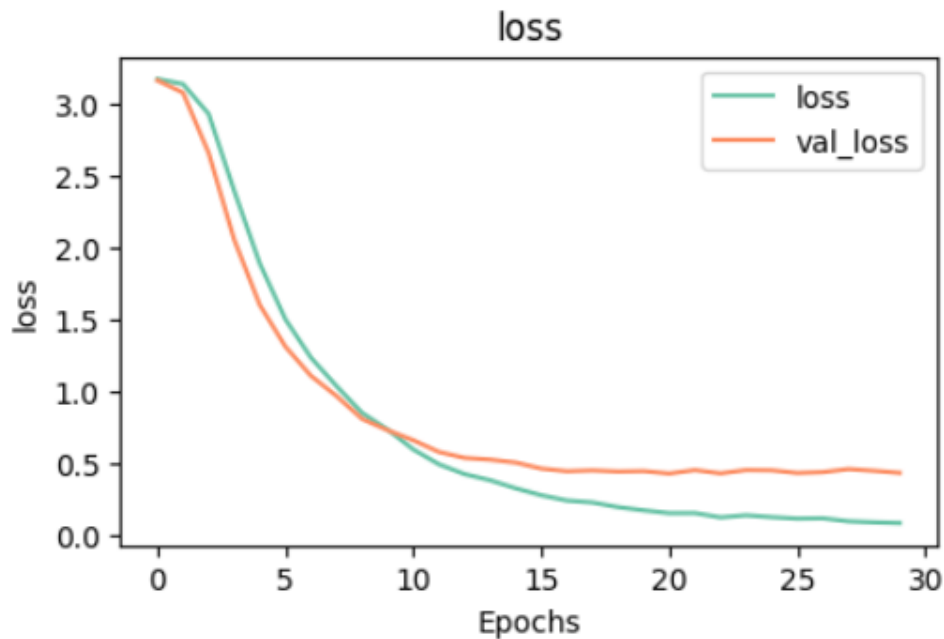


Figure 4.4 – CNN Validation loss

4.1.3 K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) model, using TF-IDF vectorization for feature extraction, achieved a consistent validation accuracy of 87%. Unlike deep learning models, KNN does not undergo a traditional training process; instead, it relies on the proximity of data points in the feature space during prediction. This characteristic makes it less prone to overfitting, as evidenced by the stability in its performance across different data splits. The effectiveness of KNN in this context can be attributed to the high-quality feature representation provided by TF-IDF, which successfully transformed symptom descriptions into numerical vectors capturing their semantic meaning. Despite its simplicity, the KNN model performed robustly, particularly with well-represented classes in the dataset. However, its performance might degrade with more complex, unseen data or when faced with symptoms that are significantly different from those in the training set.

4.2 COMPARATIVE ANALYSIS

Comparatively, the CNN model emerged as the top performer with the highest validation accuracy (88.33%), indicating a good balance between learning the training data and generalizing to new cases. While the LSTM model had the highest training accuracy (98.78%), it suffered from significant overfitting, suggesting the need for additional techniques to enhance its generalization capabilities. The KNN model, although simpler, showcased reliable performance without extensive training, highlighting the importance of effective feature extraction. The Figure 4.5 shows the comparison of each models CNN,LSTM and KNN.

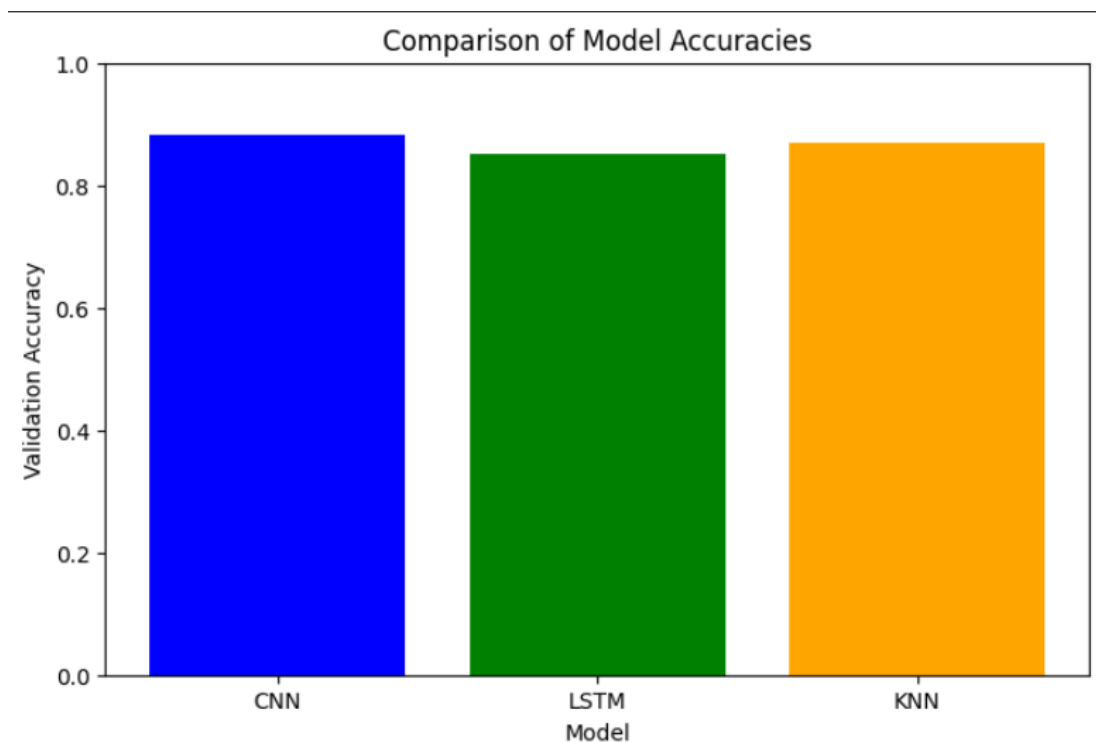


Figure 4.5-Comparative Analysis

4.3 SCREENSHOTS

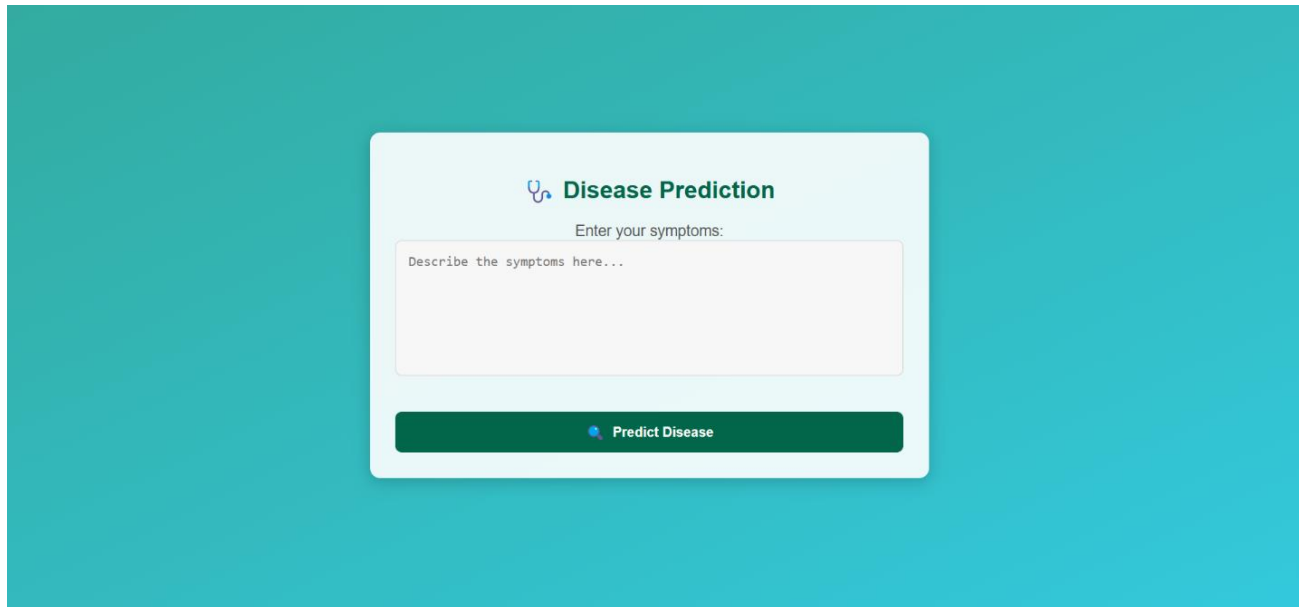




Figure 4.6-Output 1

 **Disease Prediction**

Enter your symptoms:


Iam having rashes and itching on my skin and my skin also have discolorations

 **Predict Disease**

Predicted Disease: Acne

Treatment: Antibiotics,Light therapy,Steroid injection


Figure 4.7-Output 2



Disease Prediction

Enter your symptoms:


Iam having rashes and itching on my skin and my skin also have discolorations

 **Predict Disease**

Predicted Disease: Acne


Treatment: Antibiotics,Light therapy,Steroid injection

Figure 4.8-Output 3

 **Disease Prediction**

Enter your symptoms:

slow healing of wounds and dry throat

 **Predict Disease**

Predicted Disease: diabetes

Treatment: Insulin injections, Lifestyle changes, Oral medications

Figure 4.9-Output 4

CHAPTER 5

CONCLUSION

This project aimed to develop a machine learning-based system for the early prediction of diseases based on user-provided symptoms, leveraging three different models: Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory (LSTM), and K-Nearest Neighbors (KNN). The integration of these models within a Flask web application offered an accessible interface for users, enhancing the practical utility of the system in a real-world scenario. By comparing different machine learning techniques, this project provided valuable insights into the effectiveness of various algorithms in handling symptom-based text data for disease prediction.

The CNN model emerged as the best performer, demonstrating strong generalization capabilities and effectively learning the underlying features of symptom descriptions. Its convolutional layers excelled in capturing local patterns and distinguishing key symptoms, making it highly accurate and reliable. Despite the slight overfitting observed, the CNN maintained high validation accuracy, indicating its suitability for the task at hand. The LSTM model, on the other hand, was proficient in understanding the sequential nature of the symptom descriptions but faced significant overfitting issues. This highlights the LSTM's sensitivity to the training data and its potential requirement for larger datasets or enhanced regularization techniques to perform optimally. The KNN model, although simpler, provided consistent and stable results, showing the effectiveness of using TF-IDF vectorization for feature extraction and underscoring the potential of distance-based algorithms as a baseline approach.

Overall, the project successfully demonstrated the feasibility of using machine learning models for early disease prediction based on symptoms. It emphasized the importance of selecting appropriate algorithms, addressing overfitting, and fine-tuning models to improve predictive performance. The findings suggest that while advanced deep learning models like CNN and LSTM offer significant advantages in feature extraction and sequence learning, simpler models like KNN can still be valuable, especially when combined with robust feature extraction techniques.

Future work could focus on expanding the dataset to include a broader range of symptom descriptions and diseases, exploring ensemble methods that integrate multiple models to boost accuracy, and incorporating transfer learning with state-of-the-art language models. Such enhancements would likely improve the system's predictive capabilities and broaden its applicability in healthcare settings, paving the way for more accurate and timely disease diagnosis.

In conclusion, this project lays a solid foundation for symptom-based disease prediction using machine learning, showcasing the potential of AI-driven solutions in improving early diagnosis and supporting medical decision-making. The insights gained through model comparison and performance analysis provide a clear direction for future research and development, with promising implications for advancing patient care through intelligent, symptom-aware prediction systems.

5.1 FUTURE ENHANCEMENTS

1. **Expansion of the Dataset:** The current model was trained on a relatively small dataset. Expanding the dataset to include more symptoms, diseases, and diverse patient profiles would improve the model's ability to generalize and make more accurate predictions for a wider range of cases. Additionally, incorporating multi-lingual data could make the system more accessible globally.
2. **Model Optimization and Regularization:** The overfitting observed in the LSTM model could be mitigated with more advanced techniques like dropout regularization, batch normalization, or using smaller model architectures. Hyperparameter tuning using methods such as grid search or Bayesian optimization could further enhance model performance.
3. **Integration of Multi-modal Data:** Incorporating additional types of medical data such as test results, patient demographics, or medical history could provide more context and improve prediction accuracy. Multi-modal models that combine symptom text data with other data sources can be more robust and reliable in predicting diseases.
4. **Real-time and Continuous Learning:** Developing a system where the model can update and retrain itself with new data over time, either through user input or integration with healthcare

data sources, would allow the model to continuously improve and adapt to new medical knowledge and trends.

5. Explainable AI (XAI): To increase trust in the predictions, implementing explainable AI techniques could help users and healthcare professionals understand how the model arrived at its decision. This would be particularly useful for medical applications where interpretability is crucial for clinical decision-making.

6. Deployment and Integration into Healthcare Systems: The model could be integrated into existing healthcare platforms to provide decision support for doctors and healthcare professionals. Incorporating the model into electronic health record (EHR) systems could help in real-time diagnosis and treatment recommendations.

7. Mobile and Voice Assistant Integration: Making the disease prediction model available on mobile devices or integrating it with voice assistants like Siri or Alexa would make it more user-friendly and accessible, especially for people who may not be familiar with using web applications.

8. Evaluation on Different Performance Metrics: Further evaluation using additional metrics, such as precision, recall, and F1-score, would give a better understanding of the model's performance, especially in handling rare diseases where accuracy alone may not be sufficient.

9. Integration with Medical Databases: Integrating the model with large medical knowledge bases like ICD-10 codes, PubMed articles, or clinical trial data could provide a more comprehensive context for symptom-disease relationships, making the system more accurate and robust. This could also allow the system to suggest more diverse disease possibilities based on real-time updates from medical research.

10. Personalized Health Profiling: Creating personalized profiles for patients using their medical history, lifestyle factors, and genetic predispositions could help refine predictions and tailor the model's recommendations. This would allow for more accurate risk assessments and disease predictions based on individual factors.

11. Predictive Model for Disease Progression: Instead of simply predicting the presence of a disease, the model could be enhanced to predict the progression of symptoms or the likelihood

of certain diseases developing over time. This would be particularly useful for chronic diseases, allowing for early interventions or preventative measures.

12.Integration with Telemedicine Platforms: Incorporating the disease prediction tool into telemedicine platforms could help healthcare providers with initial consultations, giving them a starting point for further questions or testing. This could speed up the process of diagnosis and improve the efficiency of telehealth services.

By focusing on these enhancements, the system can become more accurate, accessible, and effective, ultimately improving its potential to assist in early disease detection and improving patient outcomes.

REFERENCE

- [1] **K. Gaurav, A. Kumar, P. Singh, A. Kumari, M. Kasar, T. Suryawansh** (2023), Human Disease Prediction using Machine Learning Techniques and Real-life Parameters, International Journal of Engineering, IJE TRANSACTIONS C: Aspects Vol. 36 No. 06, (June 2023) 1092-1098
- [2] **Er. Harjeet Singh, Mr. Ankit Mehta** (2023), Disease Prediction System Using Symptoms, International Research Journal of Engineering and Technology, Volume: 10 Issue: 06 | Jun 2023
- [3] **Dr. C K Gomathy, Mr. A. Rohith Naidu** (2021), The Prediction Of Disease Using Machine Learning, International Journal of Scientific Research in Engineering and Management (IJSREM), Volume: 05 Issue: 10 | Oct - 2021