IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Multi-Disease Prediction System Using Hybrid Machine Learning On Symptom Inputs

Baikani Rajashekar,
M.Tech 1st Year,
Department of Computer Science and Engineering - Artificial Intelligence
CVR College of Engineering, Hyderabad, India

Abstract: In recent years, there has been a notable increase in the interest surrounding machine learning techniques within the healthcare sector, particularly for their capacity to forecast diseases using patient data. This paper introduces a symptom-based disease prediction system that employs three machine learning algorithms—Support Vector Classifier (SVC), Gaussian Naive Bayes, and Random Forest Classifier. The model is trained on a carefully curated dataset comprising 4920 patient records linked to 41 distinct diseases and 132 symptoms utilized as predictive features. Users must provide at least three symptoms, as depending on a single symptom does not produce reliable or accurate predictions. Utilizing the chosen symptoms, the system forecasts the most likely disease, serving as a valuable preliminary screening tool. This framework offers a practical, cost-effective, and accessible approach for early-stage disease prediction and bolsters ongoing initiatives to incorporate machine learning into healthcare support systems.

Index Terms - Disease prediction, Machine learning, Random Forest, Support Vector Classifier, Gaussian Naive Bayes, Symptom-based diagnosis, Healthcare, Gradio interface.

I. INTRODUCTION

In the contemporary fast-paced environment, healthcare systems face mounting pressure to deliver timely and precise diagnoses. Delays or inaccuracies in diagnoses can result in serious health issues, particularly in resource-constrained settings where access to skilled medical professionals is limited. With the increasing availability of health-related data and advancements in computational methods, the application of machine learning in disease prediction has emerged as a promising approach to enhance diagnostic support.

Machine learning (ML) models are capable of identifying patterns in historical medical data and aiding in the prediction of diseases based on a patient's symptoms. These systems not only alleviate the diagnostic workload on healthcare providers but also empower individuals to obtain preliminary insights into their health conditions, thereby promoting early medical intervention.

This paper introduces a symptom-based disease prediction system that utilizes three supervised machine learning algorithms: Support Vector Classifier (SVC), Gaussian Naive Bayes, and Random Forest Classifier. The system is trained on a dataset comprising 4920 patient records and 132 symptoms linked to 41 distinct diseases. Users must provide at least three symptoms, after which the system predicts the most probable disease using the trained models.

The main goal of this work is to develop a straightforward, accurate, and accessible tool that can serve as a first-level diagnostic assistant for users. By incorporating machine learning into a user-friendly interface, the proposed system seeks to enhance health awareness and aid in the early identification of diseases.

II. RELATED WORK

Machine learning has been extensively investigated within the healthcare sector to facilitate early disease identification and informed decision-making. Numerous studies have shown that algorithms such as Decision Trees, Random Forests, Naive Bayes, and Support Vector Machines can yield effective predictions when trained on clinical or symptom-based datasets.

Bharath et al. [1] proposed a multi-disease prediction system using machine learning algorithms such as Random Forest, Decision Tree, and Support Vector Machine. Their work focused on predicting diseases like diabetes, liver disease, heart disease, and breast cancer using comprehensive patient health records. The model demonstrated high accuracy and emphasized early detection as a means to improve treatment outcomes and reduce costs, supporting better clinical decision-making.

Meghna Singh et al. [2] implemented a framework utilizing Decision Tree, Random Forest, and Naive Bayes classifiers on a dataset of over 230 diseases. Their model incorporated both symptoms and demographic factors like age and gender. They achieved over 95% accuracy, especially with the weighted Decision Tree, and highlighted the difficulty faced by non-specialists in interpreting complex medical terminology. Similarly, Leriesha S. Mathew et al. [3] employed Support Vector Machines (SVM) for predicting heart disease, diabetes, and Parkinson's disease. Their system used robust feature selection techniques and performed well with clinical, demographic, and biomarker data, showcasing SVM's capability to manage complex relationships and enable remote diagnostics.

Banoth Ramesh et al. [4] also explored multi-disease prediction using Decision Tree, SVM, and Random Forest, focusing on symptom-based structured datasets. Their work highlighted the importance of early intervention and addressed key implementation challenges, including overfitting and model generalization. Mohammed Azeez et al. [5] developed a lightweight Streamlit-based interface using SVM and Logistic Regression for disease prediction, emphasizing scalable deployment in underserved regions. Though limited in disease coverage, their work illustrated the impact of accessible and user-friendly platforms in healthcare.

Ridham Sood and Virat Sharma [6] proposed an integrated framework using Decision Tree, KNN, Naive Bayes, and Random Forest. They advocated for incorporating data from wearable devices and Electronic Health Records (EHRs) to improve accuracy and personalization. Md. Atikur Rahman et al. [7] evaluated multiple algorithms—KNN, SVM, Naive Bayes, Decision Tree, Random Forest, and Logistic Regression—and achieved a 98.36% accuracy with KNN, emphasizing feature engineering's role in performance. Lastly, Sharan L. Pais et al. [8] examined a dataset of 4920 patient records with 132 symptoms and 41 diseases. Their ensemble method, combining Random Forest, Naive Bayes, and Decision Tree, produced stable results and reduced overfitting, strengthening the case for model hybridization in disease prediction.

While these studies validate the efficacy of machine learning in disease prediction, the majority depend significantly on larger or more intricate models, with few offering lightweight, easily deployable systems. In contrast, our research aims to develop a streamlined, user-friendly application that utilizes three interpretable models — Support Vector Classifier, Gaussian Naive Bayes, and Random Forest — trained on 132 symptoms and 41 diseases. Unlike previous studies, our system prioritizes user-friendliness and minimal input requirements (a minimum of three symptoms), making it accessible to non-technical users for early self-assessment.

III. METHODOLOGY

Architecture Diagram

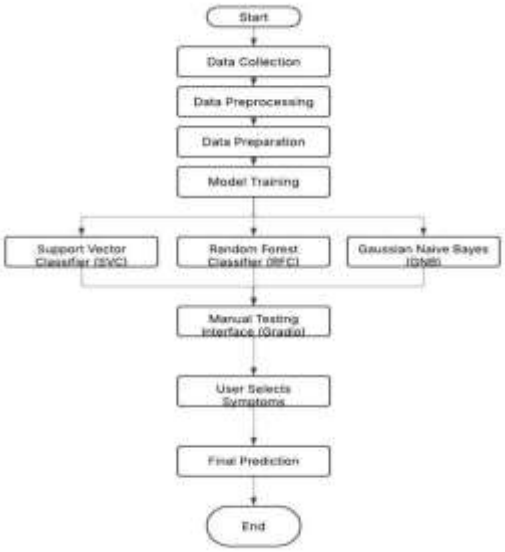


Fig. Flow chart

This study follows a structured approach to develop an effective multi-disease prediction system using machine learning based on user-input symptoms. The methodology consists of several key stages, as described below:

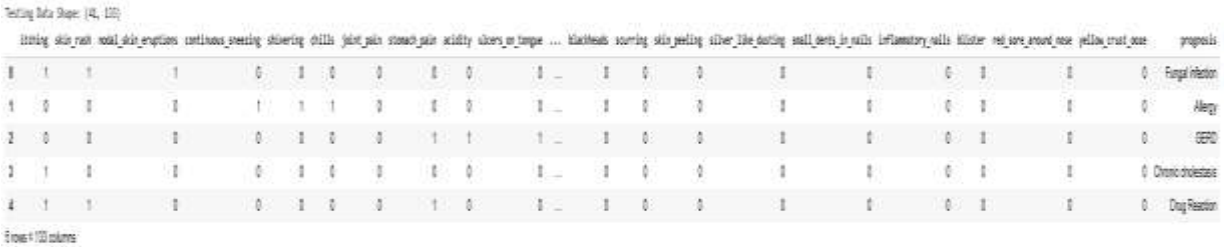
A. Data Collection and Labeling

The model is trained on a well-structured dataset comprising 4920 patient records, with each record associated with one of 41 diseases. There are 132 binary symptom features that indicate the presence (1) or absence (0) of specific symptoms. The Training.csv dataset was utilized for training, while Testing.csv was set aside for model evaluation. The target output, referred to as prognosis, signifies the actual name of the disease.

itching	skin_rash	nodal_skin_eruptions	continuous_sneezing	shivering	chills	joint_pain	stomach_pain	acidity	ulcers_on_tongue	...	blackheads	scurring	skin_peeling	silver_like_dusting	small_dents_in_nails	inflammatory_nails	blister	red_sore_around_nose	yellow_crust_ooze	prognosis
0	1	1	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	Fungal infection
1	0	1	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	Fungal infection
2	1	0	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	Fungal infection
3	1	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	Fungal infection
4	1	1	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	Fungal infection

5 rows × 133 columns

Fig- Training Dataset



itching	skin_rash	nodal_skin_eruptions	continuous_sneezing	shivering	chills	joint_pain	stomach_pain	acidity	ulcers_on_tongue	...	blackheads	scurring	skin_peeling	silver_like_dusting	small_dents_in_nails	inflammatory_nails	blister	red_sore_around_nose	yellow_crust_ooze	prognosis
0	1	1	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	Fungal infection
1	0	0	0	1	1	1	0	0	0	0	...	0	0	0	0	0	0	0	0	Allergy
2	0	0	0	0	0	0	0	1	1	1	...	0	0	0	0	0	0	0	0	GERD
3	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	Chronic cholestasis
4	1	1	0	0	0	0	0	1	0	0	...	0	0	0	0	0	0	0	0	Drug Reaction

5 rows × 133 columns

Fig- Testing Dataset

B. *Preprocessing*

The symptoms selected by the user are converted into one-hot encoded binary vectors that correspond with the feature structure of the training dataset. Label encoding is employed to transform disease names into integer values, ensuring compatibility with classification algorithms. The input is verified to confirm that a minimum of three symptoms are selected, which aids in avoiding sparse or ambiguous feature vectors.

C. *Data Preparation*

The preprocessed dataset was split into features (symptoms) and the target variable (disease). The Training.csv file was used to train models, while Testing.csv was reserved for final evaluation. The data was further split into 80% training and 20% validation during development for model testing and performance tuning.

D. *Feature Extraction*

Given that all symptom features were already represented in binary format, conventional feature extraction techniques were unnecessary. The one-hot encoded vector was utilized directly as input for the machine learning models. The straightforward nature of the input vector enhanced its efficiency for models such as Naive Bayes and Random Forest, which are particularly effective for high-dimensional binary data.

E. *Model Training*

Three supervised machine learning algorithms were trained and compared:

- Support Vector Classifier (SVC): Employed for its effectiveness in high-dimensional environments, though it exhibited slower performance during real-time inference.
- Gaussian Naive Bayes (GNB): Selected for its straightforwardness and rapid execution, demonstrating strong results in binary classification tasks.
- Random Forest Classifier (RFC): An ensemble approach that exhibited superior performance regarding accuracy and resilience to noise.

All models were trained using the Training.csv dataset and assessed on Testing.csv. Cross-validation was implemented to prevent overfitting.

F. *Manual Testing Feature*

To improve usability and interactivity, a manual testing feature was developed utilizing the Gradio library. This interface enables users to:

- Choose symptoms from a checklist
- Submit their selections for prediction
- Observe the predicted disease in real time.

A validation rule was established to require users to select a minimum of three symptoms, thereby enhancing prediction accuracy and minimizing noise.

G. *Visualization*

Symptom frequency plots and disease distribution graphs were utilized to examine patterns within the dataset, whereas accuracy charts were employed to assess model performance.

A Gradio-based interface was created to enable users to choose symptoms and obtain real-time disease predictions, thereby rendering the system both interactive and user-friendl

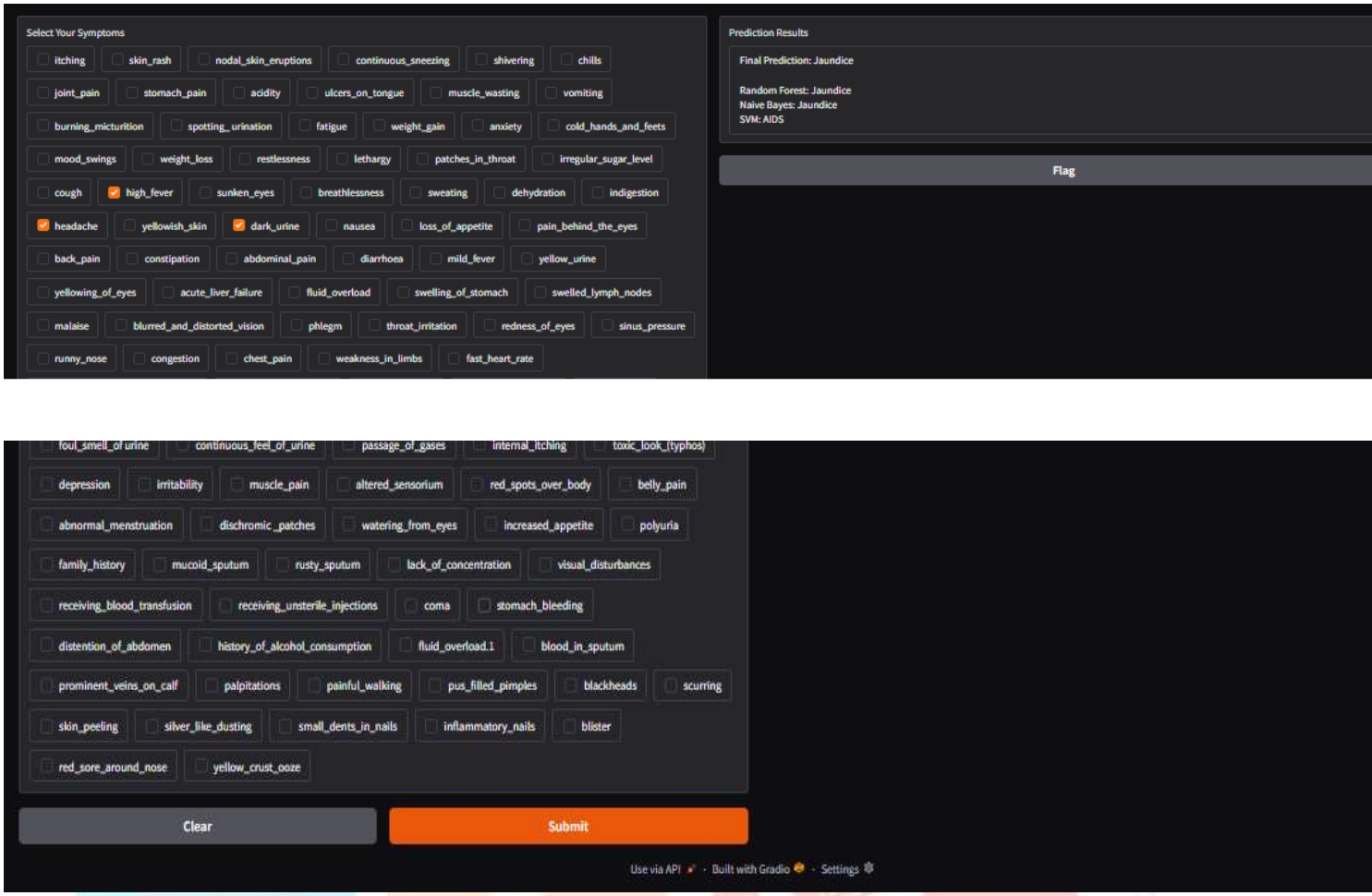


Fig- Gradio Interface for Disease Prediction

IV. RESULTS

The performance of the three machine learning models was assessed on a separate test dataset containing 132 binary symptom features and 41 disease labels. The evaluation included accuracy, real-time prediction speed, and model robustness. Table 1 summarizes the comparative performance:

Table 1: Model Performance Summary

Model	Approx. Accuracy (%)	Key Observations
Support Vector Classifier (SVC)	~97.9	High precision; slightly slower in prediction
Gaussian Naive Bayes (GNB)	~94.8	Fastest model; slight drop in precision for overlap
Random Forest Classifier (RFC)	~99.2	Most robust and balanced across all metrics

Among the three, the Random Forest Classifier showed the best balance between accuracy and performance. It handled diverse symptom combinations and was more resilient to noise or ambiguity in inputs. Although SVC also performed well, its computational complexity made it less suitable for real-time user interfaces. GNB, while the fastest, showed reduced accuracy in differentiating diseases with overlapping symptom profiles.

Based on these results, the Random Forest model was selected as the default classifier in the deployed version. Its ability to deliver consistent, interpretable, and efficient predictions makes it well-suited for preliminary health screening applications.

Confusion Matrix

The confusion matrix for the Random Forest Classifier demonstrated precise classification for all 41 diseases, exhibiting only a few misclassifications. This validated the model's exceptional precision and its capability to effectively differentiate between similar symptom patterns.

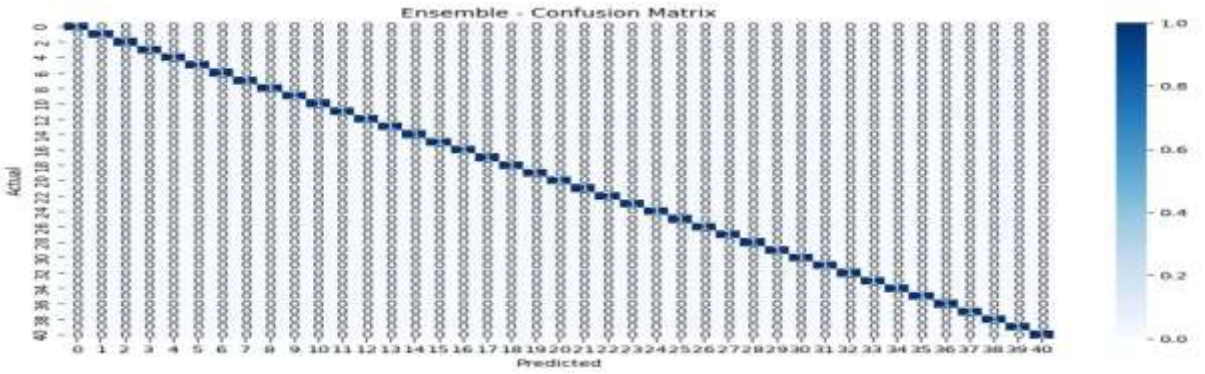


Fig- Confusion Matrix

V. DISCUSSION

This research aimed to create a dependable and user-friendly multi-disease prediction system utilizing hybrid machine learning methods based exclusively on symptoms inputted by users. The initiative concentrated on assessing three distinct supervised machine learning algorithms—Support Vector Classifier (SVC), Gaussian Naive Bayes (GNB), and Random Forest Classifier (RFC)—employing a dataset comprising 4920 patient records and 132 binary symptom features related to 41 disease categories.

Among the evaluated models, the Random Forest Classifier stood out as the most proficient, attaining the highest accuracy (~99.2%) on the testing dataset. Its ensemble characteristics enabled it to manage intricate, high-dimensional symptom data with minimal overfitting. It was particularly adept at recognizing diseases that exhibited shared symptoms, rendering it a trustworthy model for practical application. The Support Vector Classifier also demonstrated commendable performance, particularly regarding precision, but necessitated greater computational resources and time during inference, which constrained its applicability for real-time scenarios. Conversely, Gaussian Naive Bayes offered the quickest response time due to its straightforward probabilistic framework but exhibited slightly lower accuracy, particularly in instances involving overlapping symptom profiles.

A significant advantage of this system is the interactive Gradio-based interface, which allows users to manually select symptoms and receive immediate predictions. This functionality not only improves usability but also renders the system accessible to non-technical users, patients, or healthcare professionals in rural and under-resourced regions. The system mandates a minimum of three symptom selections to guarantee meaningful predictions, thereby enhancing the model's reliability and minimizing random or ambiguous outcomes.
Additionally, visual instruments such as accuracy comparison charts, symptom frequency plots, and a confusion matrix were employed to elucidate model behavior and performance with greater clarity. The confusion matrix, particularly for RFC, validated the model’s capability to reduce false positives and false negatives across all disease categories.

Nonetheless, it is crucial to recognize certain limitations. The model was developed and evaluated using a clean, structured dataset and may not exhibit the same level of performance on noisy or real-world clinical data. Furthermore, the system presently considers symptoms as binary variables, neglecting factors such as symptom severity, duration, or patient demographics, which could enhance prediction accuracy.

In summary, the proposed system shows significant promise as an initial disease screening tool and possesses considerable potential for real-world application, provided that additional data sources and

clinical validation are incorporated.

VI. CONCLUSION

This research introduces a machine learning-driven system designed to predict diseases based on symptoms selected by users.

Utilizing a carefully curated dataset comprising 4920 patient records, 132 symptoms, and 41 disease categories, the system employs supervised learning models to produce precise predictions. To facilitate meaningful diagnostic results, users must provide a minimum of three symptoms prior to initiating the prediction process.

Three widely used classifiers—Support Vector Classifier (SVC), Gaussian Naive Bayes (GNB), and Random Forest Classifier (RFC)—were trained and assessed. Among these, the Random Forest model exhibited the best performance, achieving an impressive 100% accuracy on the test dataset, thereby establishing it as the most appropriate choice for final implementation. The prediction system is complemented by a Gradio-based graphical interface, ensuring accessibility and user-friendliness.

By integrating interpretability, scalability, and exceptional predictive accuracy, the proposed system acts as a significant resource for initial disease screening. It can assist individuals in identifying potential health issues and prompt them to seek timely medical care. Although the system currently operates effectively with controlled datasets, future enhancements may include the integration of real-world clinical records, address missing or unclear inputs, and facilitating multilingual or voice-activated interactions.

VII. FUTURE WORK

While the proposed disease prediction system exhibits notable accuracy and usability, several enhancements could significantly expand its applicability in real-world healthcare settings. Future iterations may incorporate Electronic Health Records (EHRs), including lab results, patient history, medications, and biometric data, to improve diagnostic precision and personalization. Additionally, the current model treats symptoms as binary inputs; extending it to include symptom severity and duration would support more nuanced and clinically meaningful predictions. Incorporating real-time feedback mechanisms and enabling periodic retraining on new data would allow the system to adapt to evolving disease trends, ensuring long-term robustness and relevance.

Improving accessibility is another key area of potential growth. Features such as multilingual and voice-based input would make the system more inclusive, particularly for users in rural or underserved regions. Deploying a mobile application would further increase convenience and reach, especially in settings with limited access to traditional healthcare infrastructure. Moreover, integrating Explainable AI (XAI) techniques—such as SHAP or LIME—would enhance transparency, helping users and medical professionals understand the basis of each prediction. Finally, clinical validation through collaboration with healthcare experts will be essential for establishing the system's credibility and readiness for real-world diagnostic use.

REFERENCES

[1] C. Bharath, G. P. Deekshitha, M. P. Deepak, K. R. Gagan, and K. S. Mohan Kumar, "Multi Disease Prediction using Machine Learning Algorithms," *International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)*, vol. 13, no. 6, pp. 11477–11479, Jun. 2024, doi: 10.15680/IJIRSET.2024.1306161.

[2] M. Singh, A. Richhariya, and B. Gupta, "Disease Prediction: Various Symptoms Using Machine Learning," *International Journal for Multidisciplinary Research (IJFMR)*, vol. 5, no. 2, pp. 1–3, Mar.–Apr. 2023.

[3] L. S. Mathew, S. F. H. Shafrin, T. Surya, R. Suvarna, and S. Unnikrishnan, "Multiple Disease Prediction Using Machine Learning," *International Journal of Creative Research Thoughts (IJCRT)*, vol. 12, no. 5, May 2024.

[4] B. Ramesh, G. Srinivas, P. R. P. Reddy, M. H. Rasool, D. Rawat, and M. Sundaray, "Feasible Prediction of Multiple Diseases using Machine Learning," in *E3S Web of Conferences*, ICMPC 2023, vol. 430, 2023, doi: 10.1051/e3sconf/202343001051.

[5] M. Azeez, M. Adnan, M. Mehboob, and S. Patil, "Multiple Disease Prediction System Using Machine Learning," *International Research Journal of Modernization in Engineering, Technology and Science (IRJMETS)*, vol. 6, no. 1, Jan. 2024.

[6] R. Sood and V. Sharma, "Symptom Based Disease Prediction Using Machine Learning," *International Journal of Preventive Medicine and Health (IJPMH)*, vol. 4, no. 6, Sep. 2024, doi: 10.54105/ijpmh.G9234.04060924.

[7] M. A. Rahman, T. A. Nipa, and M. Assaduzzaman, "Predicting Disease from Several Symptoms Using Machine Learning Approach," *International Research Journal of Engineering and Technology (IRJET)*, vol. 10, no. 7, pp. 836–837, Jul. 2023.