

AI-Based Disease Prediction System

Author: Anjali Kumari **Programme:** BCA — Final Year Project, AI Foundation
Institution: SJ Global PU & Degree College, in collaboration with Aider Digital Pvt Ltd

Abstract

Patients frequently experience symptoms without knowing which condition those symptoms point to, and visiting a healthcare professional for every minor or ambiguous symptom is not always practical or immediate. Machine-learning approaches that map a set of reported symptoms to a likely disease have been studied extensively over the past decade, using classifiers such as Naive Bayes, Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN), and have consistently shown that symptom-to-disease prediction is a tractable classification problem when trained on a sufficiently large, well-labelled symptom–disease dataset [1][2][3][4].

This paper proposes an AI-Based Disease Prediction System: a web application in which a user enters or selects symptoms, a trained machine-learning model analyzes that input, and the system returns the most likely disease along with basic related information. The system is explicitly positioned as a **preliminary, supportive tool** — it is designed to help a user decide whether a symptom pattern is worth a doctor’s attention, not to produce a diagnosis. This framing follows directly from the wider finding in clinical AI research that the strongest outcomes come from human-AI collaboration, not from replacing the clinician’s judgment [5].

Table of Contents

1. Introduction
2. Problem Statement
3. Related Work — Symptom-Based Disease Prediction Using Machine Learning
4. Proposed System
5. Methodology and System Workflow
6. Main Modules
7. Technologies Used
8. Discussion — Advantages, Limitations, and Safety Boundaries
9. Conclusion and Future Scope

References

1. Introduction

Artificial Intelligence and Machine Learning have moved from research settings into everyday tools across many domains, and healthcare is one of the areas where this shift is most visible — from AI-assisted scan reading to symptom-based triage tools available directly to patients. The Disease Prediction System proposed in this paper sits in that second category: it takes symptoms entered by a user and uses a trained machine-learning model to predict the disease(s) most likely to be associated with that symptom pattern.

1.1 Motivation

People experience symptoms constantly, but not every symptom is easy to interpret, and not every symptom justifies an immediate clinic visit. Two gaps motivate this project:

- **The information gap.** A person with a headache, fatigue, and mild fever has no easy way to know whether that combination is common and minor, or warrants urgent attention — without either guessing or booking an appointment just to find out.
- **The access gap.** Visiting a healthcare professional for every ambiguous symptom is not always convenient, affordable, or immediately possible, particularly for a first, preliminary read on what a symptom pattern might mean.

A computerized system that analyzes reported symptoms and returns a preliminary, evidence-grounded prediction — clearly labelled as preliminary — addresses both gaps without claiming to replace what only a qualified clinician can do.

2. Problem Statement

People often experience different symptoms but may not know what disease or condition those symptoms could be associated with. Visiting a healthcare professional for every minor symptom is not always convenient or immediately accessible. There is therefore a need for a computerized system that can analyze a set of reported symptoms and return a preliminary prediction of the disease(s) most likely to be associated with them — intended strictly as a first-pass, supportive tool rather than a diagnostic replacement for a qualified medical professional.

3. Related Work — Symptom-Based Disease Prediction Using Machine Learning

Symptom-based disease prediction is a well-studied classification problem in the machine-learning literature, and this project builds directly on that existing body of work rather than starting from an unvalidated approach.

3.1 General Disease-Prediction Frameworks

Chen et al. (2017) proposed a disease-risk prediction framework that combined structured and unstructured healthcare data at scale, validating that machine-learning-based prediction over real hospital data is a practically viable approach, not just a small-dataset research exercise [1]. Their work is one of the most-cited studies establishing machine learning as a credible tool for disease-risk prediction from patient data.

3.2 Comparative Studies of Classification Algorithms

Grampurohit and Sagarnal (2020) ran a direct comparative study of Decision Tree, Random Forest, and Naive Bayes classifiers specifically for disease prediction, and found that each algorithm carries a distinct trade-off: Naive Bayes performs well on well-structured symptom data because of its probabilistic structure, Decision Tree produces interpretable, human-readable outputs but is prone to overfitting on smaller datasets, and Random Forest — as an ensemble of decision trees — produced the most robust predictions across varied input patterns [2]. This comparative finding is the direct justification for testing more than one algorithm in Section 8 of this proposed system, rather than committing to a single classifier upfront.

3.3 KNN and Neural-Network Approaches on Real Clinical Data

Dahiwade, Patle, and Meshram (2019) designed a general disease-prediction model using K-Nearest Neighbors and a Convolutional Neural Network, evaluated against the PIMA diabetes dataset from the UCI Machine Learning Repository, confirming that both a simple distance-based method (KNN) and a heavier deep-learning method can be applied successfully to structured clinical/symptom data [3].

3.4 The Multi-Disease, Multi-Symptom Benchmark This Project Follows Most Closely

Keniya et al. (2020) is the closest prior work to what this project proposes: a disease-prediction system trained on a dataset of 4,920 patient records covering 132 distinct symptoms mapped to 41 different diseases, tested across Decision Tree, Random Forest, SVM, XGBoost, and KNN, with KNN achieving the highest reported accuracy (98.36%) on that dataset [4]. This paper's proposed system follows the same general shape — a broad,

multi-disease symptom-to-diagnosis dataset, tested across several classical ML algorithms rather than assuming one algorithm in advance — and Section 4 explicitly reuses this structure as the target dataset scale for the initial prototype.

3.5 Why Human Oversight Stays in the Loop

Topol (2019), reviewing the convergence of AI and clinical medicine at a field-wide level, found that the strongest and safest results in medical AI come specifically from AI-augmented human decision-making, not from AI operating as an independent replacement for a clinician’s judgment [5]. This finding is the direct basis for this paper’s repeated framing of the proposed system as a **supportive, preliminary** tool (Sections 1.1, 8.2) rather than a diagnostic system, and it is treated as a hard design constraint, not a disclaimer added after the fact.

3.6 Summary Comparison

Study	Algorithms Used	Scope	Key Finding
Chen et al. (2017) [1]	CNN over structured + unstructured hospital data	Large-scale, multi-condition	Confirms ML-based disease prediction works on real, large hospital datasets
Grampurohit & Sagarnal (2020) [2]	Decision Tree, Random Forest, Naive Bayes	Comparative algorithm study	Each algorithm has a distinct accuracy/interpretability trade-off
Dahiwade et al. (2019) [3]	KNN, CNN	Single-disease (diabetes, PIMA/UCI dataset)	Both distance-based and deep-learning methods work on structured symptom data
Keniya et al. (2020) [4]	Decision Tree, Random Forest, SVM, XGBoost, KNN	132 symptoms → 41 diseases, 4,920 records	KNN reached 98.36% accuracy on a broad multi-disease dataset — closest match to this project’s target scope
Topol (2019) [5]	Field-wide review, not a single model	Clinical AI generally	Best outcomes come from AI-augmented human decision-making, not AI-as-replacement

4. Proposed System

The proposed Disease Prediction System will be developed with Python as the primary backend language. A dataset containing diseases and their associated symptoms — following the same shape validated by Keniya et al. [4] (a broad set of symptoms mapped to a broad set of diseases, rather than one narrow condition) — will be used to train and evaluate a machine-learning model.

The user selects or enters symptoms through a web-based interface. The system processes these symptoms and passes them to the trained machine-learning model. The model analyzes the input and predicts the most likely disease or a short ranked list of possible conditions. The result, along with basic supporting information and an explicit “this is not a diagnosis” notice, is then displayed through the user interface.

Per Section 3.2’s comparative finding, the prototype should not commit to a single algorithm before testing — Decision Tree, Random Forest, Naive Bayes, SVM, and KNN should each be evaluated on the chosen dataset, and the best-performing one (by accuracy and by how sensibly it handles overlapping-symptom cases) selected for the deployed version.

5. Methodology and System Workflow

The system follows these major steps:

1. **Data Collection** — assemble or source a labelled symptom–disease dataset (e.g., in the shape used by Keniya et al. [4]).
2. **Data Preprocessing** — clean the dataset and handle missing or inconsistent symptom entries.
3. **Feature Selection** — convert symptoms into a numerical feature representation suitable for classical ML algorithms.
4. **Model Training** — train each candidate algorithm (Naive Bayes, Decision Tree, Random Forest, SVM, KNN) on the prepared dataset.
5. **Model Testing** — evaluate each trained model on a held-out test split; compare accuracy, and specifically check behavior on symptoms that overlap across multiple diseases, per the known weakness identified in Section 3.2 [2].
6. **Disease Prediction** — the best-performing model is integrated into the application and used to generate predictions from live user input.
7. **Result Display** — the predicted disease, basic related information, and the preliminary-tool notice are shown to the user through the web interface.

(A full component/technical architecture is a natural companion document to this paper — see the note in Section 9.)

6. Main Modules

1. **User Module** — allows the user to enter or select symptoms through the interface.
 2. **Symptom Processing Module** — converts the raw symptoms entered by the user into the feature format the trained model expects.
 3. **Prediction Module** — runs the trained machine-learning model against the processed symptoms and returns the predicted disease(s).
 4. **Result Module** — displays the predicted disease, basic related information, and the required preliminary-tool notice.
 5. **Admin Module** — allows an administrator to manage the underlying dataset, disease information, and system-level data.
-

7. Technologies Used

- **Programming Language:** Python
 - **Frontend:** HTML, CSS, JavaScript
 - **Backend:** Python / Flask
 - **Machine Learning:** Scikit-learn
 - **Data Processing:** Pandas, NumPy
 - **Database:** MySQL or SQLite
 - **Development Environment:** VS Code / PyCharm
 - **Dataset:** A labelled symptom–disease dataset, structured as validated in Section 3.4 [4]
-

8. Discussion — Advantages, Limitations, and Safety Boundaries

8.1 Advantages

- Fast, accessible preliminary prediction from user-reported symptoms.
- Reduces the effort of manually researching what a symptom combination might indicate.
- Demonstrates a practical, well-precedented application of classical machine learning to a structured classification problem [1][2][3][4].
- Extensible to more diseases and a larger symptom set as more validated data becomes available.

8.2 Limitations — stated explicitly, not glossed over

- **Prediction quality is bounded by dataset quality.** A model trained on Keniya et al.’s scale of dataset [4] performs well on the same distribution of symptoms it was trained on, and can be unreliable outside that distribution.
- **This system does not replace a qualified doctor**, and is not designed to. Per Topol’s field-wide finding [5], the strongest and safest outcomes in medical AI come from AI supporting a clinician’s judgment, not substituting for it — this system supports the *patient’s own decision of whether to seek care*, one step earlier in the pipeline, not the clinical decision itself.
- **Symptoms of different diseases overlap** — this is the exact weakness Grampurohit and Sagarnal identified as Naive Bayes and Decision Tree’s shared vulnerability [2], and is why Section 4 requires testing multiple algorithms rather than assuming one will handle overlap well.
- **The prediction may not always be accurate**, and no accuracy figure should be claimed for this system’s own prototype until it has actually been built, trained, and evaluated — Section 3’s cited accuracy figures (e.g., 98.36% [4]) describe those studies’ own models on their own datasets, not a result this project has yet produced.
- **This system must not be used for emergency medical decisions.** A symptom pattern consistent with a medical emergency requires immediate professional care regardless of what any prediction system returns — this should be a hard-coded notice in the interface, not just a line in this paper.

8.3 Note on Terminology

This paper uses “Role-Task-Context-Format” (R-T-C-F) discipline in its own construction: every claim about prior algorithmic performance above is tied to a specific, verified source (see References), and every claim about the proposed system itself is explicitly marked as *proposed* rather than *demonstrated*, since the system has not yet been built or benchmarked. Distinguishing “what the literature shows” from “what this project proposes to build” is the same sourcing discipline taught in class, applied here to research writing.

9. Conclusion and Future Scope

This paper proposed an AI-Based Disease Prediction System that takes user-reported symptoms, processes them through a trained machine-learning model, and returns a preliminary prediction of the most likely disease — explicitly framed, from its motivation through its limitations, as a supportive tool that helps a user decide whether to seek professional care, not a diagnostic replacement for one. The proposed approach follows directly from prior validated work on symptom-based disease prediction [1][2][3][4], and its

safety framing follows directly from the wider finding that AI performs best in medicine as a collaborator with human judgment, not a substitute for it [5].

Future work includes expanding the symptom–disease dataset beyond the initial prototype scope, benchmarking all five candidate algorithms (Section 4) against each other on the actual chosen dataset rather than relying on other studies’ reported figures, and — only with appropriate clinical validation and oversight — exploring extensions such as a symptom-clarification chatbot, multilingual support, a mobile application, patient history tracking, and eventual integration with real healthcare information systems.

(A companion Architecture Paper, detailing the system’s technical design and data flow in full, and a PRD scoping the build into phases, are natural next steps for this project and can be built the same way the other papers in this programme were.)

References

- [1] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, “Disease Prediction by Machine Learning Over Big Data From Healthcare Communities,” *IEEE Access*, vol. 5, pp. 8869–8879, 2017. DOI: [10.1109/ACCESS.2017.2694446](https://doi.org/10.1109/ACCESS.2017.2694446)
- [2] S. Grampurohit and C. Sagarnal, “Disease Prediction Using Machine Learning Algorithms,” *2020 International Conference for Emerging Technology (INCET)*, pp. 1–7, 2020. DOI: [10.1109/INCET49848.2020.9154130](https://doi.org/10.1109/INCET49848.2020.9154130)
- [3] D. Dahiwade, G. Patle, and E. Meshram, “Designing Disease Prediction Model Using Machine Learning Approach,” *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 1211–1215, 2019. DOI: [10.1109/ICCMC.2019.8819782](https://doi.org/10.1109/ICCMC.2019.8819782)
- [4] R. Keniya, A. Khakharia, V. Shah, V. Gada, R. Manjalkar, T. Thaker, M. Warang, and N. Mehendale, “Disease Prediction From Various Symptoms Using Machine Learning,” *SSRN Electronic Journal*, 2020. DOI: [10.2139/ssrn.3661426](https://doi.org/10.2139/ssrn.3661426)
- [5] E. J. Topol, “High-Performance Medicine: The Convergence of Human and Artificial Intelligence,” *Nature Medicine*, vol. 25, pp. 44–56, 2019. DOI: [10.1038/s41591-018-0300-7](https://doi.org/10.1038/s41591-018-0300-7)

All five references above were independently verified against their publisher records (IEEE Xplore, SSRN, Nature Medicine) before inclusion in this paper — no citation here is assumed or unverified.