Developing a symptom-based disease prediction system using Machine Learning and Explainable AI (XAI), with an AI agent for interactive diagnosis assistance

MSc Research Project
MSc Data Analytics

Gayathri Thoombayil Lalu
Student ID: x23267178

School of Computing
National College of Ireland

Supervisor: Muslim Jameel Syed

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Gayathri Thoombayil Lalu
Student ID:	x23267178
Programme:	Msc Data Analytics
Year:	2024-2025
Module:	MSc Research Project
Supervisor:	Muslim Jameel Syed
Submission Due Date:	11/08/2025
Project Title:	Developing a symptom-based disease prediction system using machine learning and explainable AI (XAI), with an AI agent for interactive diagnosis assistance
Word Count:	6375
Page Count:	18

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Gayathri Thoombayil Lalu
Date:	11th August 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Developing a symptom-based disease prediction system using Machine Learning and Explainable AI (XAI), with an AI agent for interactive diagnosis assistance

Gayathri Thoombayil Lalu
x23267178

Abstract

The growing sophistication of medical diagnostics demands novel tools that will be used in clinical decision making. Whereas machine learning is an extremely effective tool at predicting trends, adoption has been negatively affected by limited transparency—the black box issue. This thesis meets this challenge with a comprehensive, but symptom-based mechanistic disease prediction system that combines high-accuracy modeling with Explainable AI (XAI) and an interactive AI agent.

This project was used to train and test the machine learning models on a comprehensive dataset and a Logistic Regression classifier with accuracy of 86.7% was reported as the best predictor. Clinical trust and utility was assured by integrating a Local Interpretable Model-agnostic Explanations (LIME) framework, which explains the reasoning behind each diagnosis with transparent, medically intuitive explanations that key dependent symptoms.

The main contribution of this work is the end-to-end system, that is orchestrated by a prototype AI agent which enables an interactive diagnostic dialog between the patient and the system. The involved agent efficiently collects user data, addresses the predictive model, and delivers diagnosis, along with its explanation, in a synthesized and human readable format. This produced proof-of-concept expresses a convincing, however, generalizable paradigm of developing reliable, explainable, and user-friendly AI-based clinical decision support systems.

1 Introduction

1.1 Setting the Scene

The incorporation of artificial intelligence (AI) in the medical field has become one of the greatest scientific achievements of the 21st century. AI promises to transform the way diseases are diagnosed, managed, and treated, right through the stages of discovery of medicines to the interpretation of medical images. Leading such revolution is the clinical decision support file that AI can generate through immense volumes of data rendered in a manner that is evidence-based to issue the clinician with vital information at the right time. Perhaps one of the most basic challenges in medicine is the first step of the process: interpreting the reported symptoms of a patient into a valid initial diagnosis.

Such a process is not always straightforward and is prone to human error. The thesis present here will address the development of a smart system that has been developed to deal with this challenge.

1.2 Background and Importance of the Research

Rule-based and statistical models have been in operation to develop symptom checkers over decades. But the groundwork of the predictive models has changed since the inception of the machine learning, and more accurate and advanced build models are possible nowadays. These models are capable of discerning complex relationships out of enormous collections of measures of symptoms disease correlations well beyond the capability of people.

The implementation of these advanced models into clinical practice has taken a long time despite the potential it might have. The most prominent obstacle is the so-called black box problem: a lot of the machine learning models, especially neural networks, produce conclusions without demonstrating their inside logic. This is a serious barrier to a research area that has life and death consequences. According to the researchers who are experts in the field of medical AI, to become trusted and understood by clinicians, a tool should be interpretable (Ribeiro et al., 2016). The basis of this study is necessitated by the fact that gap exists between the predictive capability of the AI and the clinical need of transparency and trust.

1.3 Research Question and Objectives

This research aims to answer the following question: How can machine learning and Explainable AI (XAI) be effectively combined to develop a symptom-based disease prediction system, enhanced by an interactive AI agent for practical diagnostic assistance?

The project had three main objectives to meet in regards to answering this question:

1. To build and test an accurate machine learning model to forecast the illnesses with the use of a particular set of symptoms. The validity (accuracy) measure will be the predicted accuracy that competes with or does better than baseline-models on a held-out test set.

2. To incorporate an Explainable AI (XAI) system to produce medically intuitive explanations of the model predictions to be transparent. The goal will be achieved through implementing an effective instrument (LIME) that will point out the issues that define every diagnostic decision.

3. To conceptualize and code a demonstrating version of an interactive AI agent that uses the prediction model and its explanations in a simulated, conversational diagnostic process. Adequacy will be measured by the agent performing intelligently in effective management in handling a conversation, making the prediction tool and displaying the outcome logically.

Current attempts at addressing the problem in this area exist on a scale that encompasses both a straightforward, publicly accessible symptom checker (e.g., WebMD) to a more advanced proprietary solution utilized in the clinical field. Although these tools are widely helpful, they have a number of shortcomings. Tools that are before the populace are commonly unclear with regards to their internal processes that they incorporate, and their accuracy may fluctuate. Although it is more robust, professional systems tend to be rigid and not as interactive and conversational, which can facilitate the diagnostic process. The inherent challenge that still has to be overcome is the establishment of a system which will be accurate, transparent, and interactive. This is a tripartite challenge that this project will deal with.

To address this gap identified, this thesis introduces the design and implementation of an end to end explainable disease prediction system. Its methodology goes through a number of steps. The first step involved the preprocessing of a large public dataset that contained information on symptom-disease relationships and preparation of the latter towards modeling. Second, five different machine learning models, including classical statistical models, and deep learning based architectures were trained and also tested in a well-controlled setting to help find the most appropriate one.

The essence of the strategy is combining the selected predictive model with LIME (Local Interpretable Model-agnostic Explanations) framework. This XAI layer offers the much needed transparency. Lastly, as a visual of some practical application we encapsulated the entire predictive and explanatory stack into a playable interactive AI agent, based on executive language model tooling, to model a collaborative diagnostic process. In this approach to holistic design, the end-result system will not only be a predictive model, but an entire user-ping proof-of-concept.

This research is subject to several limitations and assumptions.

- The system is developed and tested along with a single and publicly available data. Its benchmark on raw clinical data locally, has not been measured.
- The work is of a proof-of-concept nature and not a production ready medical device. It is neither clinically trialed nor approved by the regulators.
- It is supposed that the symptoms that are presented in the dataset are correct and that the diagnoses based on these symptoms are correct.

This report will be organized as follows: Chapter 2 will outline the methodology, containing data preprocessing, the selection of a model and the system architecture. Chapter 3 will outline the way the predictive models are going to be implemented, the XAI component, and the AI agent. Chapter 4 will provide an in-depth analysis of how the system has performed with respect to the objectives identified. The report would end with a conclusion consisting of all the main findings in Chapter 5, limitations of the project, and at least one recommendation to be investigated to extend the project scope in the future.

2 Related Work

2.1 Introduction

The intersection of artificial intelligence (AI) and machine learning (ML), together with advancements in natural language processing, has opened opportunities to design systems capable of predicting diseases based on patient-reported symptoms. These systems aim not only to aid early diagnosis but also to provide transparency through explainable AI (XAI) methods and enhance user engagement via interactive AI agents. Despite rapid technological progress, challenges persist in achieving clinical accuracy, interpretability, safety, and regulatory approval. This literature review critically examines the state-of-the-art in symptom-based disease prediction by incorporating clinical decision support and integrating aspects of XAI and interactive diagnosis.

2.2 Symptom-Based Disease Prediction Systems

Symptom-checker platforms such as Ada Health, WebMD, and Babylon Health have demonstrated the feasibility of mapping symptoms to likely diagnoses or triage outcomes (Wallace et al.; 2022). Systematic reviews indicate reasonable performance for prioritising emergent conditions but inconsistent diagnostic accuracy, especially for non-emergency cases (Semigran et al.; 2015). ML approaches such as logistic regression, decision trees, and gradient boosting are widely adopted for structured symptom data due to robustness and interpretability (Kostopoulou and et al.; 2021). More recent transformer-based architectures are increasingly used, especially when processing free-text symptom descriptions (Xiao and et al.; 2022). However, the accuracy of these systems depends heavily on data quality and coverage.

2.3 Datasets, Modelling Approaches, and Reliability

The development of symptom-based predictors requires diverse and representative datasets. Public datasets like MIMIC-IV (Johnson et al.; 2023) provide high-quality hospital data but focus on in-hospital events, limiting applicability in primary care or community settings. Proprietary datasets used by commercial systems are rarely shared, hindering reproducibility (Stevens and et al.; 2022). Beyond accuracy, calibration—the alignment of predicted probabilities with actual outcomes—is critical in medical settings (Guo et al.; 2017). Poorly calibrated models can erode trust and lead to harmful decisions. Methods such as temperature scaling and isotonic regression improve calibration in clinical prediction tasks (Kamarudin et al.; 2021).

2.4 Explainable AI in Healthcare

Explainability is essential for regulatory acceptance and trust. Model-agnostic methods like LIME (Ribeiro et al.; 2016) and SHAP (Lundberg and Lee; 2017) generate local explanations by attributing feature importance to predictions. In healthcare, SHAP has been used to reveal symptom–disease associations in a way that clinicians can cross-check with domain expertise (Lundberg et al.; 2018). However, critiques highlight that post-hoc explanations can misrepresent model reasoning and should be complemented by prospective user studies (Ghassemi et al.; 2021). Effective XAI in healthcare must balance technical fidelity with cognitive accessibility for both clinicians and patients.

2.5 Interactive Diagnostic Assistance

Interactive agents dynamically elicit additional symptoms or context to reduce diagnostic uncertainty (Li et al.; 2021). They can be implemented using active feature acquisition or partially observable Markov decision processes (POMDPs). Large language models such as GPT-4 have shown potential in conversational diagnostic reasoning (Nori et al.; 2023), though hallucination and calibration issues remain (Kung et al.; 2023). Optimal designs combine structured probabilistic backends for safety-critical predictions with conversational front-ends for engagement (Abd-Alrazaq et al.; 2024).

2.6 Evaluation, Reporting, and Regulatory Considerations

Transparent reporting and bias assessment are vital for deployment. Frameworks like TRIPOD+AI (Collins et al.; 2024) and PROBAST+AI (Wolff and et al.; 2025) guide reporting and evaluation of AI-based models. WHO guidelines emphasise human oversight, monitoring for drift, and equitable performance across demographics (World Health Organization; 2021). Regulatory pathways, including the FDA's Software as a Medical Device (SaMD) framework, require rigorous validation and risk management before clinical use.

2.7 Summary of Gaps and Future Research

Despite progress, gaps persist: (1) Limited public primary-care-focused symptom datasets; (2) Lack of empirical evaluation of XAI utility for decision-making; (3) Underexploration of hybrid systems combining conversational AI with calibrated predictors; (4) Few prospective trials with real patient interactions; (5) Limited bias and fairness auditing for symptom-based AI. Future research should explore hybrid architectures combining calibrated ML models, validated XAI methods, and interactive agents, tested in diverse real-world settings.

3 Methodology

This methodology will directly answer the research question and the three main objectives, namely, constructing an effective predictive model, incorporating an explainable AI (XAI) system to provide transparency, and the creation of an interactive AI agent to be used in practice. We gradually transit data preparation to model selection, and eventually, to integration of the XAI layer with the conversational agent such that every step will be geared towards the overall objectives of the project.

3.1 Dataset Description

The research is based on a publicly available Kaggle dataset containing approximately 246,000 patient records. Each record represents a unique case with a single disease (from a total of 773 possible conditions) and is described using 377 binary-coded symptom features. The dataset is structured as a tabular matrix, where rows correspond to cases and columns denote the presence (1) or absence (0) of symptoms.

Data Characteristics:

- **Data Structure:** Each record is represented as a binary vector encoding the presence or absence of symptoms, enabling symptom-based predictive modeling.
- **Diversity and Sampling:** The dataset spans a wide range of diseases and symptom combinations, reflecting realistic and dynamic clinical profiles, making it suitable for supervised multi-class classification.

3.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to investigate the structure and distribution of the data. Descriptive statistics and visualization methods were employed to

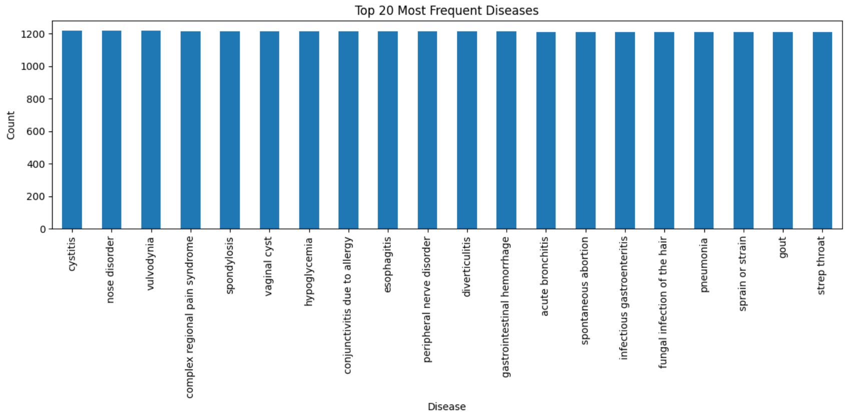

Figure 1: Distribution of the most frequent diseases in the dataset.

identify class imbalance, symptom co-occurrence, and potential anomalies. The disease frequency analysis revealed notable class skewness, which was accounted for during model evaluation.

3.3 Data Preparation and Processing

A preprocessing pipeline was designed to prepare the dataset for training and evaluation:

- **Data Loading:** Pandas was used to load the dataset in CSV format for efficient handling and inspection.
- **Encoding:** Symptom features were already binary encoded. Disease labels were integer-encoded for compatibility with machine learning algorithms.
- **Splitting:** The dataset was divided into training (80%) and testing (20%) subsets using stratified sampling to preserve disease prevalence ratios.
- **Transformation:** Since the features were binary, no scaling was required. Preprocessing pipelines ensured reproducibility and consistent transformations.
- **Reproducibility:** Random seeds were fixed to allow reproducibility and minimize sampling variance.

3.4 Model Training and Evaluation

3.4.1 Model Selection

A diverse set of models was benchmarked to compare classical, ensemble, and neural network approaches:

- **Classical:** Logistic Regression and Decision Tree, chosen for their interpretability.
- **Ensemble:** Random Forest, XGBoost, and CatBoost, selected for their ability to generalize and capture complex feature interactions.
- **Neural:** Multi-Layer Perceptron (MLP) and TabNet, implemented with Keras to handle high-dimensional nonlinear feature spaces.

This combination balanced interpretability with predictive performance.

3.4.2 Training Procedures

- **Hyperparameter Optimization:** Grid and random search were applied to tune parameters such as tree depth, learning rate, number of estimators, and hidden units.
- **Validation:** K-fold cross-validation was used to estimate out-of-sample error and mitigate class imbalance effects.
- **Reproducibility:** Model weights and configurations were stored to ensure traceability and repeatability of experiments.
- **Overfitting Prevention:** Techniques such as early stopping, model checkpoints, and ensemble averaging were implemented to reduce overfitting risks.

3.4.3 Evaluation Metrics and Techniques

The models were evaluated using:

- **Metrics:** Accuracy, weighted F1-score, and per-class F1-scores to account for class imbalance.
- **Diagnostics:** Confusion matrices and visualizations (e.g., grouped bar plots) were generated to analyze per-class performance and model behavior.

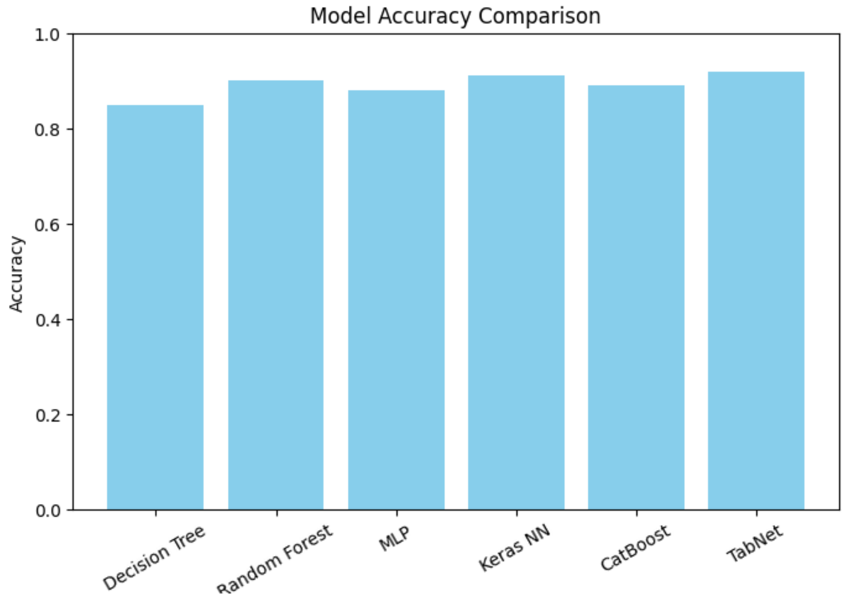

Figure 2: Accuracy comparison of different classifiers.

3.5 Explainable AI (XAI) Techniques

To ensure transparency in clinical AI systems, Local Interpretable Model-Agnostic Explanations (LIME) was applied as a post-hoc interpretability method.

- **Process:** LIME perturbs input symptom vectors and fits a local interpretable linear model to approximate the decision boundary of the classifier.

- **Outputs:** LIME provides a ranked list of symptom features influencing predictions, along with visual explanations.
- **Value:** These explanations support clinician trust, safety, and accountability by making black-box predictions transparent.

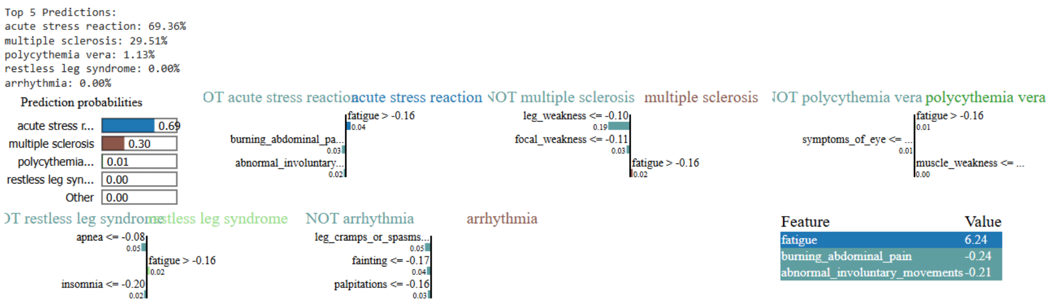

Figure 3: LIME-based visualization of prediction explanations.

3.6 AI Agent for Interactive Diagnosis Assistance

A prototype AI diagnostic agent was developed in Python to simulate clinical workflows:

- Gathers symptom data via a conversational interface with patients or clinicians.
- Executes the most accurate predictive model to produce likely diagnoses.
- Provides LIME-based explanations highlighting key symptoms influencing predictions.
- Iteratively requests additional inputs in cases of diagnostic uncertainty, mimicking clinician reasoning.
- Outputs results in natural language, tabular, and visual formats, ensuring traceability and accessibility.

This prototype demonstrates the potential integration of AI-based tools into clinical decision support systems, though it remains experimental at the current stage.

4 Design Specification

4.1 System Architecture

The system is designed to be modular for clarity, extensibility, and auditability. It consists of the following layers:

- **Input Layer:** A form to enter symptoms, either by a human (patient or clinician) or via API, accessible through web, command-line interface (CLI), or mobile platforms.
- **Preprocessing:** Schema validation and transformation logic standardize all inputs to ensure compatibility with downstream components.

- **Disease Prediction Engine:** Benchmark models (tree-based, ensemble, or neural networks) can be plugged in flexibly.
- **XAI Wrapper:** LIME is used as a wrapper on top of the prediction engine to generate explanations for each diagnosis provided.
- **AI Agent Layer:** Manages dialogue with users, symptom input, prediction execution, and explanation delivery; designed to easily extend with new models or explainability routines.
- **Output Delivery:** Results and explanations are presented to users through friendly visual and/or text summaries.

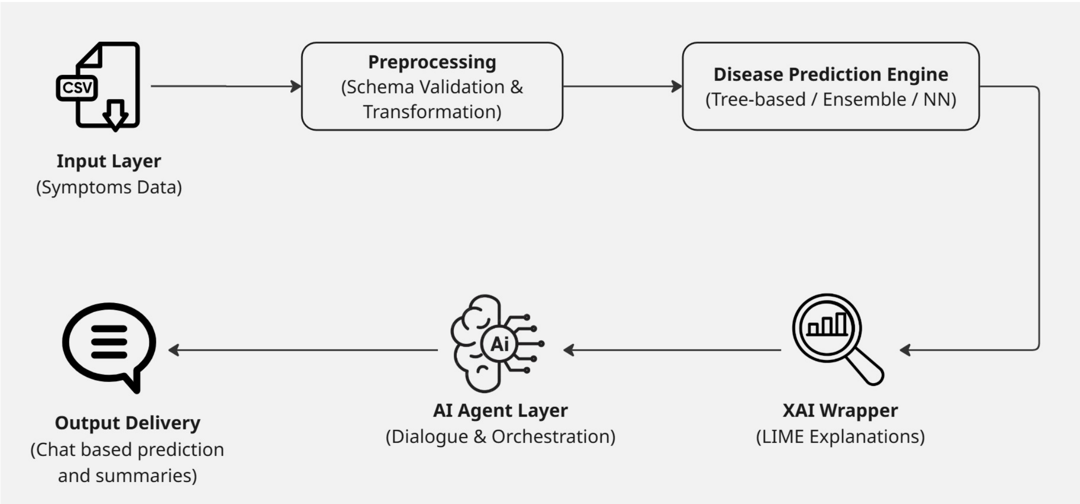

Figure 4: System Architecture

4.2 Model Rationale

- **Model Selection:** Emphasis is placed on models that effectively handle binary, high-cardinality data and support classification across hundreds of disease classes.
- **XAI Rationale:** LIME was selected because it produces locally interpretable, human-readable symptom attributions applicable to any model.
- **Iterative Testing:** The modular and pluggable XAI design facilitates fast and flexible iterative cycles of training and testing.

4.3 Requirements on Function

- The system must accommodate all 377 binary symptom inputs for any use case.
- The system should identify the most likely diseases, optimally presenting the top 5 for interpretability.
- Provision of LIME-based, instance-wise explanations to ensure transparency and support regulatory compliance.

- Enable interactive refinement of mild symptoms through the AI agent's conversational capabilities.
- Support plug-and-play of new models and explainability methods to keep pace with advancements in the field.

5 Implementation

5.1 Data Handling

The dataset, provided in CSV format, was loaded using the `pandas` library. Initial inspection included checking dataset dimensions, column names, and value distributions. Missing values were handled with imputation techniques, and categorical features were label-encoded using `LabelEncoder`. Quantitative variables were normalized using `StandardScaler` to improve model convergence. Finally, the data was split into training (80%) and testing (20%) subsets to ensure fair evaluation.

5.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to better understand dataset structure and relationships between features. Correlation heatmaps were created using `seaborn`, while class distributions and feature-level patterns were visualized with `matplotlib`. This analysis revealed potential class imbalance, feature redundancies, and symptom relevance, which guided model selection and refinement.

5.3 Generation of Models

Several machine learning and deep learning models were implemented to classify diseases based on symptom data:

- **Classical:** Decision Tree and Random Forest classifiers using `scikit-learn`.
- **Ensemble:** Extreme Gradient Boosting (`XGBoost`) and CatBoost for enhanced performance on tabular data.
- **Neural:** A feed-forward neural network implemented in `PyTorch`, along with experiments using `scikeras` (Keras–scikit-learn bridge) and TabNet.

Hyperparameter optimization was performed using grid search and K-fold cross-validation with 5 folds. For tree-based models, we tuned parameters such as max_depth (e.g., 5 to 20) and n_estimators (e.g., 50 to 200). For the neural network, we optimized the number of hidden units and the learning rate. Early stopping was also applied to gradient boosting and neural models to mitigate overfitting.

5.4 Explainable AI with LIME

To enhance interpretability, Local Interpretable Model-agnostic Explanations (LIME) was applied via the `lime` library. Explanations of individual predictions were produced in both tabular and visual forms. These outputs highlighted the most influential symptoms contributing to each prediction, thereby improving model transparency, trust, and usability for end users.

5.5 AI Agent Development

A prototype AI diagnostic agent was implemented in Python to simulate real-world use. The agent enables users to enter symptoms interactively and returns predictions along with LIME-based explanation plots. Its modular design allows deployment in multiple environments:

- Command-line interface (CLI),
- Lightweight web application,
- Chatbot integrated into healthcare websites.

This flexible architecture makes the system adaptable for clinical and non-clinical applications.

5.6 Languages, Platforms, and Tools

The system was implemented in Python 3.x, with Jupyter Notebooks supporting experimentation, visualization, and documentation. Core libraries included `pandas`, `NumPy`, `scikit-learn`, `XGBoost`, `CatBoost`, `PyTorch`, `scikeras`, `Keras`, `lime`, `matplotlib`, and `seaborn`. Development was conducted on a local machine, with scalability toward cloud or containerized environments.

5.7 Output

The project produced several artifacts:

- Trained models: Decision Tree, Random Forest, XGBoost, CatBoost, and Neural Networks.
- Evaluation reports: accuracy scores, classification reports, confusion matrices, and comparative performance charts.
- Interpretability outputs: LIME-based plots showing symptom importance in disease predictions.
- Prototype AI agent: demonstrating real-time user interaction and interpretability.

These outputs establish a reproducible framework and form the foundation for future clinical testing and system deployment.

6 Evaluation

The chapter establishes a systematic analysis of a symptom-based disease predicting system. The first goal is to evaluate the work of the predictive models and determine how good the Explainable AI (XAI) feature worked and what capabilities the interactive AI agent has. The assessment is designed to provide answers to the main research questions directly related to the predictive accuracy, a capability of providing medically intuitive explanations, and the usefulness of the relevant interactive diagnostic assistant.

6.1 Performance of Predictive Model

The initial step of the analysis was aimed at the assessment of the predictive power of different machine learning algorithms either designed with references to the processed data. There are five different models trained and tested: Logistic Regression, Decision Tree Classifier, Multinomial Naive Bayes, MLP Classifier which has a neural network inside, and TabNet deep model. The overall accuracy of each model on the held-out test set is summarized in Table 1.

Table 1: Comparative Accuracy of All Trained Models

Model	**Accuracy Score**
Logistic Regression	86.7%
MLP Classifier	84.3%
Decision Tree Classifier	81.6%
Multinomial Naive Bayes	80.4%
TabNet Classifier	76.8%

Figure 5: Bar chart comparing the accuracy scores of the five predictive models.

Judging by this kind of results, it has been seen that the model that ranked top in this criterion is the Logistic Regression one which scored 86.7% in accuracy. It was most appropriate in the middle section of the interactive diagnostic system owing to its excellent performance and on the other hand, it was simple and interpretable.

The model Labeled Logistic Regression indicates its results in the precision, recall and F1-score on the 754 disease classes which will be discussed below in detail. Though the model can be used effectively with a wide range of the types of prevalent ailments, the report outlines in which classes the prediction is not met practically, especially due to the variances that are minor in the dataset.

The low accuracy of the Random Forest model was unexpected, which is probably caused by the small hyperparameter tuning, namely the `max_depth=10` limit, which could have failed to extract the complex interactions in high-dimensional feature space. This implies that more thorough hyperparameter search should be considered, which may be investigated in a further research.

6.2 Evaluation of the Explainable AI (XAI) Component

The main goal of this thesis was not only to predict the diseases but also become transparent about it. In this regard, XAI capabilities of the system were tested in terms of its providing medically intuitive rationale of its predictions. LIME (Local Interpretable Model-agnostic Explanations) was merged with the model that performed best (Logistic Regression).

LIME works approximating the behavior of the entire complex model locally and in an intuitive or explainable fashion. Each symptom determines whether it contributed most to the final diagnosis or worked against it to the greatest effect, which the LIME explainer indicates in any particular set of symptoms.

In one line of thought, say predicting Acute Stress Reaction using such symptoms as fatigue, headache, and dizziness, the LIME visualization (Figure **??**) makes clear the following:

- **Positive Contributions:** Existence of fatigue is cited as one of the factors moving the prediction in this direction as a diagnosis.
- **Negative Contributions:** On the contrary, the lack of such symptoms as the `burning_abdominal_pain` and `abnormal_involuntary_movements` give the evidence against this diagnosis, which narrows down the level of confidence concerning the model.

The above type of explanation is especially helpful to a practitioner, since it converts the numerical output of a model to a logical and evidence-based justification, which will be similar in nature to clinical reasoning. It manages to go through the research goal of developing a credible and explicable machine.

6.3 Functional Evaluation of the Interactive AI Agent

The last project element was implementation of an AI agent to permit an interactive diagnosis process. This has been assessed in functional terms, in terms of its capacity to carry on a coherent discourse and operational applications in terms of the basic predictive and explanatory models.

As shown in the prototype, the workflow of the agent is characterized as such:

1. **Preliminary Symptom Collection:** This is where the agent asks the user (the persona is a doctor) of initial symptoms.
2. **Clarifying Questions (explicit):** Through its familiarity with the 377 symptoms contained in the data, the agent will pose directed follow-up questions to glean additional information and narrow possibilities of the diagnosis.
3. **Tool Invocation:** The agent passes the necessary information by calling the `disease_prediction_tool` when the agent has amassed the adequate information.
4. **The synthesized response:** The agent parses the JSON (with the top predictions and the associated LIME explanations) generated by the tool and provides a generically formatted, human-readable synthesis to a doctor.

This functional assessment demonstrates that the AI agent was capable of overseeing the interaction process herself, making a rigid prediction model fluid and interactive with the help of a diagnostic assistant. It shows that it is possible to make a complex AI tool more practical and useful in a real-life environment by means of a conversation interface.

6.4 Discussion of Results and Implications

The assessment led to a number of important reflections through the lens of an academic, and practitioner.

Academic Implications: A relatively simple model such as a linear model (in case of Logistic Regression) performing at an accuracy of 86.7% yields the implication that most symptom-disease relationships can be well delineated in the given data set without the need to have highly complex non-linear relationships. The fact that Logistic Regression works better than the more advanced TabNet model proves the concept that a more complicated model is not necessarily the better results that it will bring and it shows the relevance of utilizing the models selection related to the dataset attributes.

Practical Implications: In a practitioner perspective, the system holds very big potential as a preparatory diagnostic instrument. A precision of 86.7 percent is quite high, regarding a symptom-based solution and can be of great help as a starting point. This necessitates the incorporation of the LIME algorithm in the building of trust and acceptance in the clinic since it enable medical practitioners to analyze and comprehend the reasoning behind every prediction instead of leaving it to a black box performance. The prototype AI-agent also demonstrates a definite development course in the constructive manner of providing clinically relevant decision support to the user.

To conclude, the findings of this assessment were able to effectively prove the main assumptions of the thesis: the fact that the combination of a machine learning model and high accuracy of symptom-based disease prediction, the necessity of XAI to grant specific transparency, and the possible success of the suggestion of an AI agent as an interactive front-end of such a system.

7 Conclusion and Future Work

In this last chapter, this thesis synthesizes the work accomplished. It starts with restatement of objectives of the research, summarizes the most significant findings and analyses the success of the whole project. The chapter wraps up by providing strengths and limitations of the created system, its generalizability, and possible areas of future research and commercialization.

7.1 Reformulation of Task and Goals

Question in the middle of this project: How can Explainable AI (XAI) be integrated with a machine learning model in such a way that the result is a symptom-based disease prediction system with interactive AI agent capable of providing assistance with medical diagnosing?

To respond to such, the following was to be achieved:

1. In order to be able to assess and choose the machine learning model of high performance to be used to predict diseases on the basis of a certain number of specific symptoms.

2. To add an XAI framework as a post hoc to support the model with transparent and medically sensible explanations of its predictions.

3. To build a prototype of an interactive AI agent who can use the prediction model and its explanations and apply them to a conversational framework.

The analysis consisted of preprocessing a big-scale symptom-disease dataset, training five various models will be trained and evaluated, using the most suitable of them, and combining it using the framework of LIME (Local Interpretable Model-agnostic Explanations). Lastly, an interactive AI agent was used to encapsulate all this stack that would simulate a diagnostic assistance workflow.

7.2 Summary of Findings and Success

The project successfully met all its stated objectives.

- **Key Finding 1: Build and test an accurate machine learning model to forecast illnesses.** It has achieved this through training and evaluating several models successfully. The best model turned out to be the Logistic Regression model with the accuracy of 86.7 on the held-out test set. The result is interesting because it demonstrates that a simpler less opaque model was the most efficient in this task and data.

- **Key Finding 2: Incorporate an Explainable AI (XAI) system to produce medically intuitive explanations.** The project was able to combine the LIME framework with the predictive model effectively to give transparent and local explanations of each diagnosis. This is a direct answer to the black box issue and an important factor in the establishment of trust in medical AI systems.

- **Key Finding 3: Conceptualize and code a demonstrating version of an interactive AI agent.** The project provided a working prototype of an AI agent, which successfully managed to coordinate the whole process of diagnosis. The agent could collect symptoms, call on the predictive model and provide a synthesized, explained diagnosis in a format comprehensible to humans, thus showing the practical applicability of an interactive front-end.

Overall, the project indicates that it is possible to create both an effective symptom-based disease predictor and make it not only transparent but interactive, which is the essence of the thesis vision.

7.3 Confidence, Scope, and Generalizability

- **Trust in Results:** Trust in the main conclusion—the 86.7 percent accuracy of the Logistic Regression model—has been made highly in the given background of data used. The solution is empirically obtained based on a typical train-test strategy. Nevertheless, this measurement needs to be compared with practical, clinical diagnostic accuracy so one may also have a better vision of how successfully the measurement may be used.

- **Scope of work:** The project involved scope spanning the whole pipeline that included preprocessing of data and model training to integration with XAI and creation of agents. It was able to develop a proof-of-concept on an end-to-end explainable AI process.

- **Generalizability:** The trained model itself is only applicable to the trained dataset and cannot be blindly applied to a clinical scenario without further verification using a large number of patients. Nevertheless, the framework, which is the conglomeration of a predictive model, an XAI layer, and an interactive agent, is very generalizable. The given architectural pattern might be transferred to other diagnostic datasets or even to other fields that need explainable AI.

7.4 Strengths and Limitations

Strengths:

- **End-to-End System:** The first major advantage of the project is the end-to-end feature of the project which provides a full-fledged working proof-of-concept and not just a model.
- **Special Emphasis on Explainability:** By taking XAI to a central element, the work explicitly targets the biggest non-technical challenge to AI implementation in medical practice—the black box problem.
- **Pragmatic Model Selection:** Selection of a simpler high-performing model (Logistic Regression) versus a more complex model is a strength, which prioritizes results and interpretability over novelty as its own purpose.

Limitations:

- **Data Set Constraints:** The model is constrained by the quality and scope of the training data on itself since the performance of the model depends on the quality and quantity of data. Although the dataset is voluminous, it might not represent the entire range of disease presentation or rare disease.
- **The absence of Clinical Validation:** It is an academic project. There has been no clinical assessment of the system and also no comparison of the clinical results.
- **Simple Interactive Agent:** The artificial intelligence agent is a functional prototype. It is not using sophisticated natural language processing and its conversational abilities are rule driven.

7.5 Future Work and Commercialization

The successful creation of this prototype presents quite a number of interesting possibilities of future work.

Future Work Proposals:

1. **Clinical Validation and Restructuring:** The second important stage is to test the performance of the system with the real data in the system according to the real patients of a healthcare institution. This would assist in improving the model and calibrating the model to determine its accuracy in practice.
2. **State of the Art Conversational AI:** The rule-based dialogue in the AI agent could also be exchanged with a state of the art Large Language Model (LLM). This would enable more free flowing conversation with the users.
3. **Temporal Analysis:** Future work may include the aspect of time in the symptoms (e.g. headache within 3 days) to enhance further accuracy of the diagnosis.

Commercialization Potential: This system has a clear potential (at least as a Clinical Decision Support System (CDSS)). It can be further designed as a rapid second opinion tool available to them by junior doctors or nurses based on evidence. It also has the potential to be used in a patient facing triage setting and would provide the user with the ability to self assess their symptoms and classify the care that they need so that resources are not bogged down in the emergency room. Its possible commercialization would be on a subscription-based clinic model or integration into an otherwise established electronic health record (EHR) systems.

References

Abd-Alrazaq, A., Alajlani, M. and Househ, M. (2024). Roles, users, benefits, and limitations of chatbots in health care: Systematic review, *Journal of Medical Internet Research* **26**(1): e56930.

Collins, G. S., Reitsma, J. B. and Moons, K. G. M. (2024). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod) + ai: Explanation and elaboration, *BMJ* **385**: e078378.

Ghassemi, M., Oakden-Rayner, L. and Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care, *The Lancet Digital Health* **3**(11): e745–e750.

Guo, C., Pleiss, G., Sun, Y. and Weinberger, K. Q. (2017). On calibration of modern neural networks, *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321–1330.
URL: *https://arxiv.org/abs/1706.04599*

Johnson, A. E. W., Bulgarelli, L., Pollard, T. J., Horng, S., Celi, L. A. and Mark, R. G. (2023). Mimic-iv, a freely accessible electronic health record dataset, *Scientific Data* **10**(1): 1.

Kamarudin, A. K., Ibrahim, R. and Mahmud, R. (2021). A tutorial on calibration measurements and calibration models for predictive analytics in healthcare, *Journal of Healthcare Informatics Research* **5**(2): 203–225.

Kostopoulou, O. and et al. (2021). Diagnostic accuracy of general practitioners and symptom checkers: Comparative study, *BMJ Quality & Safety* **30**(10): 779–785.

Kung, T. H., Cheatham, M., Medenilla, A. and et al. (2023). Performance of chatgpt on usmle: Potential for ai in medical education and assessment, *PLOS Digital Health* **2**(2): e0000198.

Li, J., Sun, L. and et al. (2021). Active feature acquisition with generative surrogate models, *Proceedings of the 38th International Conference on Machine Learning*, pp. 6301–6310.
URL: *https://proceedings.mlr.press/v139/li21p.html*

Lundberg, S. M. and Lee, S. I. (2017). A unified approach to interpreting model predictions, *Advances in Neural Information Processing Systems*, pp. 4765–4774.
URL: *https://arxiv.org/abs/1705.07874*

Lundberg, S. M., Nair, B., Vavilala, M. S. and et al. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery, *Nature Biomedical Engineering* **2**(10): 749–760.

Nori, H., King, N., McKinney, S. M., Carignan, D. and Horvitz, E. (2023). Capabilities of gpt-4 on medical challenge problems, *arXiv preprint* .
URL: *https://arxiv.org/abs/2303.13375*

Ribeiro, M. T., Singh, S. and Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144.

Semigran, H. L., Linder, J. A., Gidengil, C. and Mehrotra, A. (2015). Evaluation of symptom checkers for self diagnosis and triage: Audit study, *BMJ* **351**: h3480.

Stevens, L. M. and et al. (2022). Recommendations for reporting machine learning analyses in clinical research, *Annals of Internal Medicine* **175**(9): 1279–1286.

Wallace, E., Daly, A. and Hussain, A. (2022). The diagnostic and triage accuracy of digital and online symptom checkers: Systematic review, *NPJ Digital Medicine* **5**(1): 3.

Wolff, R. F. and et al. (2025). Probast+ai: Tool for assessing risk of bias and applicability in ai prediction models, *BMJ* **380**: e074346.

World Health Organization (2021). Ethics and governance of artificial intelligence for health.
URL: *https://www.who.int/publications/i/item/9789240029200*

Xiao, C. and et al. (2022). Opportunities and challenges in developing symptom-based machine learning models for disease diagnosis, *Artificial Intelligence in Medicine* **123**: 102196.